Unity in diversity: an overview of the genomic anthropology of India

by

Sarabjit S. Mastana

Human Genomics Lab., Centre for Global Health and Human Development, School of Sport, Exercise and Health Sciences, Loughborough University, Loughborough, LE11 3TU. UK

Keywords:

India, Castes, Tribes, SNPs, STR, Alu MtDNA, Y-chromosome, DNA collections, genomic anthropology, Genetic variation and diversity.

Address for correspondence

Dr Sarabjit Mastana

Human Genomics Lab.

Centre for Global Health and Human Development,

School of Sport, Exercise and Health Sciences, Loughborough University,

Loughborough LE11 3TU

Tel:+44-1509-223041

Email: S.S.Mastana@LBORO.AC.UK

Abstract

Context: India is considered a treasure for geneticists and evolutionary biologists due to its vast human diversity, consisting of more than 4500 anthropologically well-defined populations (castes, tribes and religious groups). Each population differs in terms of endogamy, language, culture, physical features, geographic and climatic position and genetic architecture. These factors contributed to India-specific genetic variations which may be responsible for various common diseases in India and its migratory populations. As a result, interpretations of the origins and affinities of Indian populations as well as health and disease conditions require complex and sophisticated genetic analysis. Evidence of ancient human dispersals and settlements is preserved in the genome of Indian inhabitants and this has been extensively analysed in conventional and genomic analyses.

Objective and Methods: Using genomic analyses of STRs and *Alu* on a set of populations, we estimate the level and extent of genetic variation and its implications.

Results: The results show that Indian populations have higher level of unique genetic diversity which is structured by many social processes and geographical attributes of the country.

Conclusion: This overview highlights the need to study the anthropological structure and evolutionary history of Indian populations while designing genomic and epigenomic investigations.

Genomic Anthropology of India, 'Unity in diversity': an overview

by

Sarabjit S. Mastana

Genomic anthropology is a fast growing branch of anthropology that holds great promise for future genomic and epigenomic studies. Recent developments in the field have highlighted the role that genomic studies can play in understanding past as well as present diversity: DNA is an unbroken link to our ancestors, populations and relatives. Human genetic diversity among populations is an excellent source of information about historical events such as migrations, expansions, colonisation and selection. By examining patterns of genetic polymorphisms one can infer how past demographic events and selection have shaped variation in the genome. Genomic anthropology is useful in estimating the contribution of different gene pools to the make-up of the present-day populations and test hypotheses about origin of linguistic and historical population movements. In addition, it increases our understanding of gene-environment interactions and the contribution of populations to the detection of genes in common and complex diseases. Thus, modern human biology, in the form of DNA analysis, can be used to explore past population history, and reconstructing this human biology of the past can also provide vital information that aids our understanding of modern disease and enables better experimental design.

India is the second most populous country in the world with more than 1.21 billion people and is considered to be a major southern coastal route of human migration (Kivisild et al. 1999). The anthropological, historical, linguistic and genetic evidence for early peopling is found imprinted all over the country. The evolutionary antiquity of Indian ethnic groups and subsequent migrations from different parts of the world has contributed to the rich tapestry of socio-cultural, linguistic and biological diversity. In this overview, the genomic history of India is reviewed with reference to the impact of historical and social events on contemporary human genetic diversity and their implications for disease and health from the perspective of personalised medicine. This topic is vast so this overview cannot discuss every aspect of the genomic history of India, but will focus on selected major aspects that provide the best examples of how knowledge of current human biology can help us understand the past.

1. Genesis of genomic diversity in India

The contemporary populations of the present day India are a panorama of social, cultural geographical and ethnic diversities. The early history of its populations is like a jigsaw puzzle with many missing pieces. However there is enough anthropological and archaeological evidence to show that from time immemorial peoples of many different ethnic stocks, cultures and languages have entered India and contributed to the present day gene pool of the subcontinent (Bhasin et al. 1994; Papiha, 1996; Kivisild et al. 1999; 2003; Bhasin and Walter 2001; Kashyap et al. 2004; Metspalu et al. 2004;2011; Reich et al 2009; Tamang and Thangaraj 2012; Moorjani et al. 2013). There are traces of human activity in India around 200,000 B.C. (Misra 2001), and by the middle Palaeolithic period, humans had spread to many parts of India (Bhasin et al. 1994; Papiha, 1996; Kivisild et al. 1999).

Mirroring other types of diversity, languages spoken in India belong to several different language families, reflecting the complex history of the subcontinent. The peoples of India can be classified into speakers of four major linguistic families: Indo-European, Dravidian, Austroasiatic and Tibeto-Burman. The majority of the people speak Indo-European languages, followed by Dravidian (mostly in southern parts of the country). The Austroasiatic speakers are dispersed mostly in the central and eastern parts, while the Tibeto-Burman speakers are concentrated in the northeast states and the Himalayan foothills. In addition to these major linguistic families, a few isolated and specialist languages (like Andamanese) are also spoken by isolated groups. Neolithic settlements are numerous in India, and it appears that Austric languages came from these Neolithic peoples. The Dravidian speaking tribes of South and Central India are considered to be the descendants of the original inhabitants of the Indian subcontinent who adopted the Dravidian in preference to their own original language. However there is still a considerable debate over whether Dravidian languages developed with Neolithic peoples of India or were brought into India (Bhasin et al. 1994; Papiha, 1996; Kivisild et al. 1999; 2003; Bhasin and Walter 2001). The Indus valley civilization, which began around 3000 B.C. and lasted for about 1500 years, saw flourishing trade contacts with Persian Gulf area and Mesopotamia. The circumstances that led to the collapse of the Harappa and Mohenjodaro civilisations is not known, but this period is synchronous with the arrival of Indo-Aryans (2000-1400 B.C.) who migrated/invaded from central Asia via the Iranian plateau.

During the period 1500 B.C. to about 1100 A.D., north-west and northern India turned into a melting pot. The year 1500 B.C. saw the entry of Indo-European speakers from Iran and witnessed the beginning of a long period of conflicts with and conquest of indigenous peoples (Papiha, 1996; Bhasin and Walter 2001). The caste system was formed soon after the entry of the Indo-European speakers. During the period 800-500 B.C., iron was introduced which provided the means for large-scale expansion of the Indo-Aryan speakers into the Ganges valley (Mishra 2001; Papiha, 1996; Bhasin and Walter 2001). Several important historical migrations took place after the Indo-Aryan arrival. Greeks (400-200 B.C.), Sakas (200 B.C.), Kushanas (100 A.D.), Huns (200-500 A.D.) and Arabs (800 A.D.) are some of the important groups who came to India and slowly merged with the local populations. During the medieval age, the northern region of the Indian subcontinent experienced massive invasions from the Turks and Afghans. Mohammed Ghazni and Mohammed Gori (998-1030 A.D.) brought Islamic rule to India, although it was limited to the lowland districts of northern India. The greatest impact of Islam came during the period of the Mughal Empire (1526-1608 A.D.). The kingdom of the last Mughal King, Aurngzeb, extended over the greater part of the Indian subcontinent. The Europeans started colonisation of the Indian subcontinent in the 16th century. The Portuguese captured Goa in 1510 A.D. Due to the rivalry between the Portuguese and the Dutch, the British gained power and supremacy in India and at one time the British Empire ruled the whole of India except for a few small enclaves and extended upto Burma. A small proportions of the European genes mixed with the indigenous populations, and their descendants today form a large Anglo-Indian and Indo-Portuguese communities in Bombay and Goa respectively (Papiha, 1996; Bhasin and Walter 2001).

The contemporary Indian population is socially organised in distinct groups that are largely endogamous and reproductively isolated (Figure 1). The population is sub-divided by caste, tribe, religion, region and language. Tribal populations are considered to be the indigenous populations of India and constitute around 8.2% of the total population. Most non-tribal populations are Hindus, who are hierarchically organized in social classes know as castes. Cultural norms act as barriers to inter-caste marriages which are strictly adhered to even in the present day. The caste system may have been established for social and economic organisation; traditionally, each caste pursued a hereditarily prescribed occupation and were linked to each other through a determined pattern of barter of services and produce (Karve 1961). The four caste system from low to high comprises: Shudra (menial labour class), Vysya (business class), Kshatriya (warrior class) and Brahmin (priestly class). These

regulated marriage patterns which led to creation of a large number of isolated endogamous groups. Many other communities, often determined by religion (such as Muslim, Christian and Sikh) do not belong to the Hindu or tribal groups.

Figure 1 around here.

Within each linguistic and religious group, socio-cultural and biological characteristics delineate numerous endogenous ethnic groups. This scheme of social organisation may be further complicated due to territorial affiliation of various tribes and caste groups. Overall, a range of migrations into India contributed to an extensive genetic diversity and stringent socio-cultural barriers structured the genetic variation into different endogamous groups. A collection of genotypes, in a group of individuals belonging to a particular socio-economic stratum, represents the genetic constitution of that group. 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, populations should be defined and selected on the basis of regional, linguistic, religious, tribal or ethnic/caste affiliations. The genetic constitution of a population consists of the genes of its constituent individuals, who received them from their parents who themselves strictly belonged to the same population. The genetic constitution or gene pool of a population can be expressed in terms of an array of allele frequencies. The genetic constitution of a large random mating population remains constant over generations provided that there is no natural selection, mutation, genetic drift or gene flow. Among these four primary genetic mechanisms that provide change in the genetic constitution of a population, gene flow may be the most rapid and very relevant in the context of the population of the Indian subcontinent (Papiha 1996). In this article, I briefly review gene and genotypic variation and show some examples from my own research highlighting the dynamics of the maintenance of genetic variation in Indian populations.

2. Conventional/Classical Genetic Variation

India has long history of human population genetic studies using several single gene polymorphisms like blood groups, serum proteins, red cell enzymes, human leukocyte antigen and immunoglobulin allotypes from various geographical regions. Three blood groups (ABO, MN, D), three serum proteins (Haptoglobin, Transferrins and Group Specific

Components) and several red cell enzyme systems (ADA, AK, PGM1, AP, 6PGD, ESD, PGI, GLO, GPT, MDH and LDH) have been extensively analysed in many thousands of populations to document genetic variation and anthropological relationships of populations (Gill et al. 1991; Bhasin et al. 1994; Mastana and Papiha 1994; Papiha 1996; Bhasin and Walter 2001). Each system shows distinct geographic and ethnic distributions. In India more than 2000 studies of ABO blood groups have been carried out, highlighting that the Indian populations are characterised by high ABO*B allele frequency (0.233) compared to allele ABO*A (0.186). There is considerable variation from sample to sample; however overall there is a general trend of diminishing frequencies of the B gene and increase of the O gene from north to south (Bhasin and Walter 2001). Studies of conventional genetic markers (using many alleles and loci in different parts of the country) have demonstrated that there are small but distinct differences in tribes, castes and non-tribal groups (Bhasin et al. 1994; Mastana and Papiha 1994; Papiha 1996; Papiha and Mastana 1999; Bhasin and Walter 2001). A number of classical polymorphisms such as ABO blood group, Haptoglobin, Group Specific component, Transferrin, and APOE serum proteins show clinal and geographic variation in India (Mastana and Papiha 1998; Bhasin and Walter 2001).

Descriptive analysis of a single marker or locus-by-locus descriptive study of several markers can give information about diversity but do not provide any meaningful information about the genetic structure of the population or genetic affinities between populations. Using multivariate analyses on allele frequency data for 48 alleles from 32 Indian populations, Mastana and Papiha (2006) demonstrated significant differences between populations and their geographic, linguistic and cultural affiliations. The tribal groups showed early differentiation and were isolated from the urban and caste groups (Figure 2). It is interesting to note that Brahmin groups from different parts of the country do not form a close cluster but instead join with geographic counterparts, suggesting the caste groups may have originated from the local populations.

Figure 2 around here

3. ***Molecular Genetic Variation***

The human genome consists of 3 billion base pairs of nuclear DNA and around 16.6 Kb of extra- nuclear mitochondrial DNA. 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, forensic and evolutionary studies. In the late-1990s, a number of molecular genetic studies on Indian populations tried to untangle or reconstruct the complex relationship of Indian population groups (Bhasin et al. 1994; Papiha, 1996; Kivisild et al. 1999; 2003; Bhasin and Walter 2001; Kashyap et al. 2004; Metspalu et al. 2004;2011; Reich et al 2009; Tamang and Thangaraj 2012; Moorjani et al. 2013). Some of the early studies dealt with populations which were neither anthropologically well-defined nor were really representative Indian populations. However, recent systematic molecular genetic studies have provided important evidence about the evolutionary history of peoples in India. Most accept that the peopling of India is very ancient along with recent gene flow from west and east Eurasia (Kivisild et al. 1999; 2003; Bamshad et al. 2001; Misra 2001; Basu et al. 2003; Kashyap et al. 2004; Metspalu et al. 2004; 2011; Thangaraj et al. 2006a, b; 2009;2010; Sengupta et al. 2006; Eaaswarkhanth et al. 2010; Chaubey et al. 2011). It seems likely that a major demographic expansion of modern humans took place within India, and although the period of this expansion remains uncertain, it appears to have taken place 60,000-85,000 years before present (ybp) (Kivisild et al. 1999; 2003). The traces of socio-cultural, linguistic physiographical boundaries and evolutionary forces leading to diversity are well documented in the recent studies (Kivisild et al. 1999; 2003; Metspalu et al. 2004;2011; Reich et al 2009; Tamang and Thangaraj 2012; Moorjani et al. 2013). It has been shown that, with the exception of Africa, India harbours more genetic diversity than other comparable global regions (Papiha 1996). In this section, a historical perspective on molecular genetics studies is used to document the level of genetic variation among Indian populations.

3.1 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. Both minisatellites and microsatellites are highly variable, polymorphic, co-dominant and heterozygous and thus are excellent tools for genetic individualization and population genetic studies (Papiha et al. 1996a; Das et al. 2002; Mastana and Singh, 2002; Das and Mastana, 2003; Ranjan et al. 2003; Agrawal et al. 2003; Sachdeva et al. 2004; Kashyap et al. 2004, 2006; Mastana et al. 2007; Khan et al. 2008). 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. Minisatellite loci have high heterozygosity and their high mutation rate makes them most useful in exploring

recent population history. However, their analyses are laborious and require high quality DNA and complex statistical methods. Minisatellites were a popular choice in forensic and population genetic analyses in the late 1980s and early 1990s but were replaced by other genetic markers such as Microsatellites or Short Tandem repeats (STRs), which were analysed by Polymerase Chain Reaction (PCR) and showed discrete allelic variation. STRs are used extensively to analyse intra and inter-population affinities and evolutionary history.

Studies using microsatellites highlight that genetic variation is geographically and socially structured among Indian populations. With my colleagues, I have analysed a number of geographically and socially structured populations from different regions of India using a battery of STR loci. I illustrate this using one example in which 20 STR loci were analysed in three tribal populations from Andhra Pradesh (Chenchu, Koya, and Lambadi), five populations from North West (Brahmin-W, Gujarati Patel, Jat Sikh, Lobana Sikh, and Kanet tribe from Himachal Pradesh) and two populations from Eastern India, Santal tribe and Brahmin-West Bengal. In the majority, common alleles are shared but their frequency varies greatly. Correspondence analysis of STR allele frequencies (Figure 3) shows that tribal populations are isolated and are scattered on the periphery of variation in caste populations. This study along with other studies (Kashyap et al. 2004; Khan et al. 2008; Krithika et al. 2009) confirmed that microsatellite genetic variation in Indian populations is structured on geographical, linguistic and caste hierarchy basis. Overall, these studies strongly show that microsatellite loci are polymorphic in Indian populations and there is no significant deficiency in heterozygosity levels as one would expect if the population was endogamous. The F_{ST} values for the minisatellites and microsatellites are of low to moderate range (1.2-6.8%) in various studies indicating the genetic differentiation among Indian populations is of moderate level.

Figure 3 around here.

***3.2 Alu Insertion Polymorphisms*:**

Alu insertion 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 and are a common

source of mutations in humans. *Alu* insertions are approximately 300 bp in length, dimeric in structure, and composed of two nearly identical monomers joined by a middle A-rich region. *Alu* elements increase in number by retrotransposition – a process that involves reverse transcription of an *Alu*-derived RNA polymerase III transcript (Batzer and Deininger 2002). A variety of *Alu* subfamilies, defined by variations of the base DNA sequence, have been discovered. Approximately 5000 *Alu* elements of the youngest sub-family have been integrated into the human genome after the divergence of humans and African apes in the past 4-5 million years (Batzer et al. 1996; Stoneking et al. 1997; Roy-Engel et al. 2001). About 25% of this small subset are not fixed in human populations and therefore provide polymorphic markers in the form of presence/absence variants.

Alu insertions/repeats are convenient genetic markers. First, the insertion of an *Alu* element at a certain chromosomal site is a 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 mechanisms for removing them. Even when a rare deletion occurs, a significant remnant is left behind-molecular fossils. 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 (Batzer et al. 1996, ; Stoneking et al. 1997). Studies have shown that the root of population tree is located near the African Sub-Saharan populations, presenting evidence for an African origin of modern human populations (Batzer et al. 1996; Stoneking et al. 1997; Watkins et al. 2003). 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 (Bamshad et al. 2003). Indian samples, however, failed to form or group with any discrete cluster, indicating significant genetic heterogeneity among populations studied (Bamshad et al. 2003).

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; Vishwanathan et al. 2003; Tripathi et al. 2008, Kanthimathi et al 2008; Yadav and

Arora 2011; Meitei et al 2010, Kshatriya et al. 2012). In order to extend and document genetic variation of Alu polymorphisms, we have analysed systematically 40 *Alu* insertion polymorphisms in 18 endogamous populations (n=2200). These were sampled from 5 North India (Punjab) populations, 5 from Western India, 4 from Central India, 2 from South India (Andhra Pradesh) and 2 Sri Lankan populations (Sinhalese and Moor). Samples were collected as part of ongoing population and anthropological studies (Papiha et al. 1996; Mastana and Papiha 1994; Mastana 1999). *Alu* dimorphisms were selected in which both presence and absence alleles are common and informative. The details of analysed *Alus* were collected from published literature including primer sets, expected amplicon sizes, and comparative population allele frequencies (Carroll et al. 2001; Roy-Engel et al. 2001; Watkins et al. 2003). Insertion allele frequencies with associated standard errors and heterozygosity were calculated for each locus. Unbiased DA distance was computed from allele frequencies and a unrooted UPGMA dendrogram was constructed from DA distance matrix. Correspondence analysis of allele frequencies was conducted as an independent method to assess the level of genetic affinity among the different populations. Relative amount of gene flow was assessed using Harpending and Ward's method of regression analysis (Harpending and Ward 1982).

It was found that 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 (Carroll et al. 2001; Roy-Engel et al. 2001; Watkins et al. 2003; Majumder et al. 1999; Viswanathan et al. 2003; Tripathi et al 2008, Kanthimathi et al 2008, Meitei et al 2010; Yadav and Arora 2011; Kshatriya et al 2012). Single locus-by-locus comparisons do not provide useful information about overall level of variation, so are not given here, but a database of allele frequencies is available from the author on request. Significant departures from Hardy-Weinberg equilibrium expectations were observed for only 22 of 720 comparisons. Average population heterozygosity ranged from 0.381 (Parsee) to 0.478 (Koya). North Indian populations showed lower of genetic diversity (G_{ST}: 0.043) compared to Western Indian (0.059). Overall, the level of G_{ST} at all loci in the present set of populations is moderate (6.4%). This level of differentiation is slightly lower than that observed in populations from East and North (6.8%), and South India (8.3%) (Majumder 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), and provides evidence that an appreciable amount of inter-population differentiation is reflected in polymorphic *Alu* insertions.

Alu elements provide useful amounts of variation in evolutionary heritage, which is shown by the use of phylogenetic tree analysis. Figure 4 shows the un-rooted dendrogram produced from a 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 (from different geographical areas) do not show close genomic proximity. Instead, they are close to geographically proximal populations. Brahmins, Parsee and Muslim from Western India also cluster together showing lower levels of differences. The Chenchu tribe joins with caste populations from North and West India, while other tribal groups (Koya, Borodeshi, Baiga and Gond) are quite distinct. Sinhalese from Sri Lanka are also well isolated from the main cluster of Indian populations and join with tribal groups. Similar conclusions have been observed in other studies on conventional and molecular genetic markers (Mastana and Papiha, 1994, Papiha et al. 1996b). The plot of the correspondence analysis (figure 5) shows that caste populations form a loose central cluster and tribal populations are spread sporadically but maintaining a geographical proximity.

Figures 4 and 5 around here.

Harpending and Ward's method of regression of heterozygosity on genetic distance enables evaluation of whether, in a group of incompletely isolated populations distributed over a geographical space, observed patterns of genetic diversity are the outcome of the process of drift and migration among the populations or if the patterns are generated by interactions with populations outside those under consideration (Figure 6). Approximately half the studied populations experienced lesser gene flow than predicted. These populations, such as Brahmin-N, Lobana, Brahmin-West, Patel, Parsee, Borodeshi, are below the regression line therefore exhibit lower levels of gene flow. Interestingly, two tribal groups (Gond and Koya) showed higher levels of gene flow. Brahmin groups from North and West India along with other endogamous populations showed lower levels of gene flow.

Figure 6 around here.

Overall, *Alu* insertion results confirm that Indian subcontinent populations are geographically and socially structured. Similar conclusions were observed for other studies from the Indian subcontinent (Majumder et al.1999; Viswanathan et al. 2003; Tripathi et al 2008, Kanthimathi et al 2008, Yadav and Arora 2011). In order to evaluate the relationship of Indian populations with other world populations using *Alu* polymorphisms, we collected allele frequency data on representative European, Chinese, African populations to evaluate genetic affinities of Indian populations. The results are presented in the principal component analysis (PCA) plot (Figure 7). In this plot, it is clear that Indian subcontinent populations (caste, tribal and Sri Lankan groups), though differentiated from each other, still cluster together in the centre of the plot, with other major groups distant from Indian populations, indicating genomic unity.

Figure 7 around here.

3.3 *Mitochondrial and Y chromosome DNA markers*:

Mitochondrial DNA (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 (Jobling and Tyler-Smith 2003; Kivisild et al. 1999, 2003; Corduex et al. 2003). mtDNA is inherited through the maternal cytoplasm therefore variation in mtDNA provides a record of maternal lineage. Y chromosome DNA documents the paternal lineage (Jobling and Tyler-Smith 2003). Essential data and information can be obtained using Y chromosome 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 documented mtDNA and Y-chromosome variation among Indian populations (Bamshad et al. 2001; Kivisild et al. 1999, 2003; Corduex et al. 2003; Basu et al. 2003; Palanichamy et al. 2004; Sengupta et al. 2006; Sahoo et al. 2006; Gutala et al. 2006; Thanseem et al. 2006; Zerjal et al. 2007 ; Thangaraj et al. 2006a, b; 2008; 2009;2010; Sengupta et al. 2006; Chaubey et al 2008; Mittal el al. 2008; Eaaswarkhanth et al. 2010; Chaubey et al. 2011; Debnath et al. 2011;Wang et al. 2011; Rai et al. 2012; Chaubey et al 2014). Barnabas et al. (1996) was one of the earliest

studies, using low resolution mtDNA analysis among linguistically different Indian populations to document a high level of nucleotide diversity among Indians. Similar results have also been reported from studies using high resolution RFLP and sequencing analysis (Bamshad et al, 1996, Roychoudhury et al. 2001, Kivisild et al. 1999, 2003; Basu et al. 2003; Corduex et al. 2003; Thangaraj et al. 2006a, b; 2009;2010; Chaubey et al. 2011). mtDNA studies have shown that among Indians the basic clustering of lineages is not language or caste specific (Kivisild et al. 1999; Bamshad et al. 2001) although a low number of shared haplotypes indicates that recent gene flow across linguistic and caste borders has been limited (Bamshad et al. 1998; Bhattacharyya et al. 1999; Roychoudhury et al. 2001, Metspalu et al. 2004). 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 macrohaplogroup M. (Kivisild et al. 1999, 2003; Corduex et al. 2003; Thangaraj et al. 2006a, b; 2009;2010; Chaubey et al. 2011). More than 60% of Indians find their maternal roots in Indian-specific branches of macrohaplogroup M (Kivisild et al. 1999, 2003; Tamang and Thangaraj 2012) . 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 by the earliest migration wave of anatomically modern humans approximately 60 thousand years ago (Kivisild et al. 1999, Kivisild et al 2003, Basu et al. 2003). In a recent studies, M haplogroup frequency varied from 19% (Uttar Pradesh Brahmins) to 97% (Kota tribe)(Kivisild et al. 1999, 2003). Within the macrohaplogroup M, there is an extensive variation with many M sub-haplogroups (M2, M3, M4, M5, M6, M18, M25, M31, M32, M34, M35-M4, M48, M49 and M50) which are the India specific lineages. An analysis of 170 studied populations belonging to distinct language families and geographical regions revealed a wide diversity in major haplogroups of M (Krithika et al 2009).

Mitochondrial DNA profiles from a larger set of populations all over the sub-continent have bolstered the view that fundamentally there is genomic unity within Indians (Roychoudhury et al. 2001, Basu et al. 2003, Kivisild et al. 2003). Compared to Indian caste populations, tribal groups are characterized by the rarity of haplogroup U and by the lack of West Eurasian lineage clusters HV, TJ, N1, X (Kivisild et al. 2003, Basu et al. 2003). These studies also confirm that both caste and tribal populations share a common mitochondrial heritage (Roychoudhury et al. 2001, Basu et al. 2003, Kivisild et al. 2003, Corduex et al. 2003). Thanseem et al. (2006) found no 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. A greater level of geographic variation has been documented in many studies in respect of both tribals and castes. 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.

Y-chromosome analyses showed that the Indian caste populations are 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 hints 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; Tamang and Thangraj 2012). Indian Y chromosome lineages revealed distinct distribution patterns among caste and tribal populations. The paternal lineages of Indian lower castes showed significantly closer affinities 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; Krithika et al 2007). Overall these studies suggest that the origin of the caste system is mainly rooted in male mediated Indo-Aryan migration that pushed indigenous Dravidian populations towards southern India and Sri Lanka and establishing themselves as upper castes.

3.4 Single Nucleotide Polymorphisms (SNPs)

Single Nucleotide Polymorphisms (SNPs, commonly pronounced as “SNIPs”) are the most abundant types of polymorphisms in the human genome. Due to the improvement and automation of sequencing methodologies, and the development of DNA microarrays, these markers are now extensively studied in the human genome for their association with different complex diseases, and for understanding various aspects of population differentiation and human evolution. 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, drug efficacy, forensics and analysis of complex diseases. Until recently, India was unrepresented in the comprehensive genome-wide studies of human genetic diversity despite the evidence of extensive genetic variation among its populations. Recently, there have been successful concerted efforts to study anthropologically interesting populations from India (The Indian Genome Variation database, IGVdb, 2005, HUGO Pan-Asian SNP consortium, Mapping Human Genetic Diversity in Asia).

Rosenberg et al. (2006) analysed 1,200 genome-wide polymorphisms (microsatellites and insertion deletion polymorphisms) in 432 individuals from 15 Indian populations sampled in the United States and demonstrated a low level of genetic difference and differentiation among these populations. The selection of these populations was based on spoken language and as such may not be representative of true endogamous and heterogeneous nature of population of India. In addition, the number of individuals studied from each group/region/language group was small. Their results suggested that the frequencies of many genetic variants are distinctive in India compared to other parts of the world.

Reich et al. (2009), Metspalu et al. (2011), Chakrabarti et al. (2012) and Moorjani et al. (2013) carried out high density genome-wide SNP microarray analyses using thousands of SNPs on selected Indian populations. Some of the sample sizes are small so caution is warranted in some interpretations and predictions. Reich et al. (2009) postulated that present day Indian populations originated from differential admixture of two ancestral groups in pre-historic India, ancestral North Indian (ANI) and ancestral South Indian (ASI) (Figure 9). These studies showed that while ANI shares genetic affinity with Western Eurasian populations, ASI has no relation with any population outside India. Onges from Andaman Islands are an example of a population comprising a purely ASI component. The range of ANI admixture in the extant Indian populations was suggested to be between 30-70%, with North Indian populations having a higher ANI component than did South Indian populations. Reich et al. (2009) also reported higher than expected levels of homozygosity within Indian groups due to caste endogamy. Metspalu et al. (2011) extended genome wide SNP studies among Indian populations and showed that Indian populations form their own distinct sub group that is clearly separated from Pakistani populations on the principal component cline stretching from Europe to South India (figure 8). Tibeto-Burman groups show a closer relationship with East Asian populations, as expected based on their ethnohistory. Within the Indian population cluster there is significant overlap between Indo-European and Dravidian speaking populations. Similar conclusions were drawn by Chakrabarti et al. (2012) who showed Indian populations clustered separately to major global populations in multidimensional analysis plots. Metspalu et al. (2011) also showed that the shared genetic affinity between the ANI component of Northern India and West Eurasia is older than the Aryan invasion, thus rejecting the Aryan invasion hypothesis (Mallory, 1989; Kivisild et al. 1999; 2003; Reich et al. 2009; Metspalu et al. 2011) and lending support to ancient

demographic and genetic history, which is consistent with mtDNA analyses. Admixture between ANI and ASI has been estimated to have taken place ~4.2-1.9 thousand years ago (Moorjani et al. 2013; figure 9). They correlated this to profound changes in India, namely the de-urbanisation of the Indus civilisation and increasing population density.

Figures 8 and 9 around here.

4. Indian genetic diversity and its implications in health and disease

Overall, Indian populations have larger genetic differences than observed in European populations in terms of allele frequencies and F_{ST} values, suggesting that social–cultural practices, religion, and geographical and linguistic differences have contributed to these significant differences (Reich et al. 2009; Metspalu et al. 2011; Tamang and Thangaraj 2012; Moorjani et al. 2013). This also leads to the possibility that there are larger numbers of population and region specific diseases in India. Indeed, a number of population specific diseases (eg Handigodu disease, Madras motor neuron disease and pseudocholinestrase deficiency among Vysyas) are found, which could be due in part to an accumulation of various mutations because of endogamy. Another Indian population-specific gene deletion of 25bp in myosin binding protein C3 (MYBPC3) is responsible for 45% of sudden heart attack cardiac deaths (Dhandapany et al. 2009, Mastana et al. unpublished data). This deletion is very common in Indian subcontinent populations, and its origin is dated to around 33000 years ago (Tamang and Thangaraj 2012). Recent studies on differential selection signals on four DOK5, MSTN, CLOCK and PPARA genes, which are implicated in lipid metabolism and type 2 diabetes, suggest these genes may influence the diabetes epidemic among Indian populations (Metspalu et al. 2011). Further studies are needed to confirm these observations. The study of complex genomic architecture and its implications in health and disease requires an inter-disciplinary approach to explain the origin of Indian populations and disease specific variants. This is particularly important as a number of studies have documented that Indian populations have a large arsenal of unique SNPs which have not been found in other HapMap populations. Similarly genetic associations observed in other populations are not readily replicated in Indian populations due to population structure and distribution of allele frequencies.

Conclusions:

Since data of any particular kind are often too fragmentary to enable reconstruction of a composite picture of the peopling of any large geographical area, a multi-disciplinary genomic approach may provide some resolution. Studies on genetic diversity and affinities among contemporary human populations are useful for reconstruction of the peopling of an area, which includes tracing of past population movements and identifying ancestral populations. Examining the human biology of the past via DNA analyses, as reviewed above, shows that geographic proximity, ethno-history, biosocial and cultural affiliation all seem to be important determinants of the genetic diversity among the populations of India. For genetic association and epidemiological studies, understanding population history helps to elucidate patterns of underlying genetic variation. The genetic differentiation in regional and Caste/Tribal populations is low to moderate but is large enough to warrant care in selecting patients and the controls for genomic/epigenomic investigations, so that they should not suffer from any stochastic error and led to erroneous results. Use of endogamous groups also has the potential to provide a unique medical database, with DNA studies aiding the understanding of genetic diseases and the development of biotechnology designed for the development of diagnostics and therapeutics. Indian populations with their unique genetic diversity, different cultural and social practices, food habits, and geographic locations are an excellent model for epigenomic studies. Hopefully, the decreasing cost of next generation sequencing combined with new bioinformatics and modelling tools will pave the way for better understanding of origins, disease load and planning for personalised medicine among Indian populations.

REFERENCES

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

Bamshad M, Fraley AE, Crawford MH, Cann RL, Busi BR, Naidu JM, Jorde LB.1996. mtDNA variation in caste populations of Andhra Pradesh, India. Hum Biol 68: 1-28.

Bamshad M, Kivisild T, Watkins WS, Dixon ME, Ricker CE, Rao BB, Naidu JM, Prasad BV, Reddy PG, Rasanayagam A, Papiha SS, Villems R, Redd AJ, Hammer MF, Nguyen SV, Carroll ML, Batzer MA, Jorde LB. 2001. Genetic evidence on the origins of Indian caste populations. Genome Res 11:994-1004.

Bamshad, MJ, Wooding S, Watkins WS, Ostler CT, Batzer MA, Jorde LB.2003. Human population genetic structure and inference of group membership. Am J Hum Genet 72:578-589.

Barnabas S, Apte RV, Suresh CG. 1996. Ancestry and interrelationships of the Indians and their relationship with other world populations: A study based on mitochondrial DNA polymorphisms. Ann. Hum Genet 60:409-422.

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 H.1994. People of India: An investigation of biological variability in ecological, ethno-economic and linguistic groups. Delhi, Kamla Raj Enterprises.

Bhasin MK, Walter H. 2001. Genetics of castes and tribes of India. Delhi, Kamla Raj Enterprises.

Carroll ML, Roy-Engel AM, Nguyen SV, Salem AH, Vogel E, Vincent B, Myers J, Ahmad Z, Nguyen L, Sammarco M, Watkins WS, Henke J, Makalowski W, Jorde LB, Deininger PL, Batzer MA. 2001. Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. J Mol Biol. 311:17-40.

Chakrabarti B, Kumar S, Singh R, Dimitrova N. 2012. Genetic diversity and admixture patterns in Indian populations. Gene. 508:250-255

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

Chaubey G, Karmin M, Metspalu E, Metspalu M, Selvi-Rani D, Singh VK, Parik J, Solnik A, Naidu BP, Kumar A, Adarsh N, Mallick CB, Trivedi B, Prakash S, Reddy R, Shukla P, Bhagat S, Verma S, Vasnik S, Khan I, Barwa A, Sahoo D, Sharma A, Rashid M, Chandra V, Reddy AG, Torroni A, Foley RA, Thangaraj K, Singh L, Kivisild T, Villems R.2008. Phylogeography of mtDNA haplogroup R7 in the Indian peninsula. BMC Evol Biol. 2008.8:227. doi: 10.1186/1471-2148-8-227

Chaubey G, Metspalu M, Choi Y, Mägi R, Romero IG, Soares P, van Oven M, Behar DM, Rootsi S, Hudjashov G, Mallick CB, Karmin M, Nelis M, Parik J, Reddy AG, Metspalu E, van Driem G, Xue Y, Tyler-Smith C, Thangaraj K, Singh L, Remm M, Richards MB, Lahr MM, Kayser M, Villems R, Kivisild T. 2011. Population genetic structure in Indian Austroasiatic speakers: the role of landscape barriers and sex-specific admixture. Mol Biol Evol 28:1013-1024.

Chaubey G, Singh M, Crivellaro F, Tamang R, Nandan A, Singh K, Sharma VK, Pathak AK, Shah AM, Sharma V, Singh VK, Selvi Rani D, Rai N, Kushniarevich A, Ilumäe AM, Karmin M, Phillip A, Verma A, Prank E, Singh VK, Li B, Govindaraj P, Chaubey AK, Dubey PK, Reddy AG, Premkumar K, Vishnupriya S, Pande V, Parik J, Rootsi S, Endicott P, Metspalu M, Lahr MM, van Driem G, Villems R, Kivisild T, Singh L, Thangaraj K. 2014. Unravelling the distinct strains of Tharu ancestry. Eur J Hum Genet. Mar 26. doi: 10.1038/ejhg.2014.36.

Cordaux R, Saha N, Bentley GR, Aunger R, Sirajuddin SM, Stoneking M. 2003. Mitochondrial DNA analysis reveals diverse histories of tribal populations from India. Eur J Hum Genet. 11:253-264.

Das B, Ghosh A, Chauhan PS, Seshadri M. 2002. Genetic polymorphism study at four minisatellite loci (D1S80, D17S5, D19S20, and APOB) among five Indian population groups. Hum Biol 74: 345-361.

Das K, Malhotra KC, Mukherjee BN, Walter H, Majumder PP, Papiha SS (1996) Population structure and genetic differentiation among 16 tribal populations of central India. Hum Biol. 68(5), 679-705.

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

Debnath M, Palanichamy MG, Mitra B, Jin JQ, Chaudhuri TK, Zhang YP. 2011. Y-chromosome haplogroup diversity in the sub-Himalayan Terai and Duars populations of East India. J Hum Genet. 56:765-771.

Dhandapany PS, Sadayappan S, Vanniarajan A, Karthikeyan B, Nagaraj C, Gowrishankar K, Selvam GS, Singh L, Thangaraj K. 2007.Novel mitochondrial DNA mutations implicated in Noonan syndrome. Int J Cardiol. 120:284-285.

Eaaswarkhanth M, Dubey B, Meganathan PR, Ravesh Z, Khan FA, Singh L, Thangaraj K, Haque I. 2009. Diverse genetic origin of Indian Muslims: evidence from autosomal STR loci. J Hum Genet. 54:340-348.

Gill PS, Chahal SM, Blangero J, Corruccini RS, Bansal IJ, Kaul SS, Bhalla V. 1991, Genetic epidemiology of non-insulin-dependent diabetes mellitus in north India: preliminary analyses of some genetic markers in Punjabis. Hum Biol. Aug;63(4):549-53.

Harpending H, Ward RH.1982. Chemical systematics and human populations, pp 213-256, In: Biochemical aspects of evolutionary biology. Nitechi, M.H, (ed.). Chicago University of Chicago Press.

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

Jobling MA, Tyler-Smith C. 2003. The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet.4:598-612.

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

Kanthimathi S, Vijaya M, Ramesh A. 2008. Genetic study of Dravidian castes of Tamil Nadu J Genet. 87:175-179.

Kashyap VK, Guha S, Sitalaximi T, Bindu GH, Hasnain SE, Trivedi R.2006. Genetic structure of Indian populations based on fifteen autosomal microsatellite loci. BMC Genet 7:28.

Kashyap VK, Ashma R, Gaikwad S, Sarkar BN, Trivedi R. 2004: 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.

Khan F, Pandey AK, Borkar M, Tripathi M, Talwar S, Bisen PS, Agrawal S. 2008. Effect of sociocultural cleavage on genetic differentiation: a study from North India. Hum Biol 80:271-286.

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

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

Krithika S, Maji S, Vasulu TS. 2007. Molecular genetic perspective of Indian populations: A Y chromosome scenario. Anthropologist special volume no 3:385:392

Krithika S, Maji S, Vasulu TS. 2009. A microsatellite study to disentangle the ambiguity of linguistic, geographic, ethnic and genetic influences on tribes of India to get a better clarity of the antiquity and peopling of South Asia. Am J Phys Anthropol. 39:533-546.

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.

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.

Mastana SS, Murry B, Sachdeva MP, Das K, Young D, Das MK, Kalla AK. 2007 Genetic variation of 13 STR loci in the four endogamous tribal populations of Eastern India. Forensic Sci Int 169:266-273.

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

Mastana SS, Papiha SS.1998. Genetic variability of Transferrin subtypes in the populations of India. Hum Biol 70:729-744.

Mastana S, Singh PP. 2002 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.

Mallory JP. 1989. In Search of the Indo-Europeans: Language, Archaeology, and Myth, London: Thames & Hudson,

Meitei KS, Meitei SY, Asghar M, Achoubi N, Murry B, Mondal PR, Sachdeva MP, Saraswathy KN. 2010. A genomic insight into the peopling of Manipur, India. Genet Test Mol Biomarkers 14:765-773

Metspalu M, Romero IG, Yunusbayev B, Chaubey G, Mallick CB, Hudjashov G, Nelis M, Mägi R, Metspalu E, Remm M, Pitchappan R, Singh L, Thangaraj K, Villems R, Kivisild T. 2011. Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. Am J Hum Genet 89:731-744.

Misra VN.2001. Prehistoric human colonization of India. J Biosci. 26: 491-531.

Mittal B, Tripathy V, Aruna M, Reddy AG, Thanseem I, Thangaraj K, Singh L, Reddy BM. 2008. Mitochondrial DNA variation and substructure among the tribal populations of Andhra Pradesh, India. Am J Hum Biol. 20:683-692.

Moorjani P, Thangaraj K, Patterson N, Lipson M, Loh PR, Govindaraj P, Berger B, Reich D, Singh L. 2013. Genetic evidence for recent population mixture in India. Am J Hum Genet 93:422-438.

Palanichamy MG, Sun C, Agrawal S, Bandelt HJ, Kong QP, Khan F, Wang CY, Chaudhuri TK, Palla V, Zhang YP. 2004. Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia. Am J Hum Genet. 75:966-978.

Papiha SS.1996.Genetic variation in India. Hum Biol 68:607-628.

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

Papiha SS, Mastana SS, Purandare CA, Jayasekara R, Chakraborty R.1996b. 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.

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

Rai N, Chaubey G, Tamang R, Pathak AK, Singh VK, Karmin M, Singh M, Rani DS, Anugula S, Yadav BK, Singh A, Srinivasagan R, Yadav A, Kashyap M, Narvariya S, Reddy AG, van Driem G, Underhill PA, Villems R, Kivisild T, Singh L, Thangaraj K.2012. The phylogeography of Y-chromosome haplogroup h1a1a-m82 reveals the likely Indian origin of the European Romani populations. PLoS One. 7(11):e48477. doi: 10.1371/journal.pone.0048477.

Ranjan D, Trivedi R, Vasulu TS, Kashyap VK.2002. Geographic contiguity and genetic affinity among five ethnic populations of Manipur, India: further molecular studies based on VNTR and STR loci. Ann Hum Biol 30:117-131.

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

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. 2002. Genetic structure of human populations. Science 298(5602):2381-2385.

Rosenberg NA, Mahajan S, Gonzalez-Quevedo C, Blum MG, Nino-Rosales L, Ninis V, Das P, Hegde M, Molinari L, Zapata G, Weber JL, Belmont JW, Patel PI.2006. Low Levels of Genetic Divergence across Geographically and linguistically Diverse Populations from India. PLoS Genet 2(12).

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

Roy-Engel, AM, Carroll ML, Vogel E, Garber RK, Nguyen SV, Salem AH, Batzer MA, Deininger PL.2001. Alu insertion polymorphisms for the study of human genomic diversity. Genetics 159:279-290.

Sachdeva MP, Mastana SS, Saraswathy KN, Elizabeth AM, Chaudhary R, Kalla AK.2004. Genetic variation at three VNTR loci (D1S80, APOB, and D17S5) in two tribal populations of Andhra Pradesh, India. Ann Hum Biol. 31:95-102.

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.

Sengupta S, Zhivotovsky LA, King R, Mehdi SQ, Edmonds CA, Chow CE, Lin AA, Mitra M, Sil SK, Ramesh A, Usha Rani MV, Thakur CM, Cavalli-Sforza LL, Majumder PP, Underhill PA.2006. 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.

Stoneking M, Fontius JJ, Clifford SL, Soodyall H, Arcot SS, Saha N, Jenkins T, Tahir MA, Deininger PL, Batzer MA.1997. Alu insertion polymorphisms and human evolution: evidence for a larger population size in Africa. Genome Res 7:1061-1071.

Tamang R, Thangaraj K. 2012. Genomic view on the peopling of India. Investig Genet. 3(1):20.

Thangaraj K, Chaubey G, Kivisild T, Selvi Rani D, Singh VK, Ismail T, Carvalho-Silva D, Metspalu M, Bhaskar LV, Reddy AG, Chandra S, Pande V, Prathap Naidu B, Adarsh N, Verma A, Jyothi IA, Mallick CB, Shrivastava N, Devasena R, Kumari B, Singh AK, Dwivedi SK, Singh S, Rao G, Gupta P, Sonvane V, Kumari K, Basha A, Bhargavi KR, Lalremruata A, Gupta AK, Kaur G, Reddy KK, Rao AP, Villems R, Tyler-Smith C, Singh L. 2008. Maternal footprints of Southeast Asians in North India. Hum Hered 66:1-9. doi: 10.1159/000114160. Epub 2008 Jan 28.

Thangaraj K, Chaubey G, Reddy AG, Singh VK, Singh L. 2006a. Unique origin of Andaman Islanders: insight from autosomal loci. J Hum Genet 51:800-804.

Thangaraj K, Chaubey G, Singh VK, Vanniarajan A, Thanseem I, Reddy AG, Singh L. 2006b. In situ origin of deep rooting lineages of mitochondrial Macrohaplogroup 'M' in India. BMC Genomics 7:151.

Thangaraj K, Chaubey G, Reddy AG, Singh VK, Singh L. 2007. Autosomal STR data on the enigmatic Andaman Islanders. Forensic Sci Int 169:247-251.

Thangaraj K, Naidu BP, Crivellaro F, Tamang R, Upadhyay S, Sharma VK, Reddy AG, Walimbe SR, Chaubey G, Kivisild T, Singh L. 2010. The influence of natural barriers in shaping the genetic structure of Maharashtra populations. PLoS One. 5(12):e15283.

Thangaraj K, Nandan A, Sharma V, Sharma VK, Eaaswarkhanth M, Patra PK, Singh S, Rekha S, Dua M, Verma N, Reddy AG, Singh L. 2009. Deep rooting in-situ expansion of mtDNA Haplogroup R8 in South Asia. PLoS One 4(8):e6545

Thanseem I, Thangaraj K, Chaubey G, Singh VK, Bhaskar LV, Reddy BM, Reddy AG, Singh L.2006. Genetic affinities among the lower castes and tribal groups of India: inference from Y chromosome and mitochondrial DNA. BMC Genet 7:42.

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.

Vishwanathan H, Edwin D, Usharani MV, Majumder PP.2003. Insertion/deletion polymorphisms in tribal populations of southern India and their possible evolutionary implications. Hum Biol 75:873-887.

Wang HW, Mitra B, Chaudhuri TK, Palanichamy MG, Kong QP, Zhang YP. 2011. Mitochondrial DNA evidence supports northeast Indian origin of the aboriginal Andamanese in the Late Paleolithic. J Genet Genomics. 38:117-122.

Watkins WS, Rogers AR, Ostler CT, Wooding S, Bamshad MJ, Brassington AM, Carroll ML, Nguyen SV, Walker JA, Prasad BV, Reddy PG, Das PK, Batzer MA, Jorde LB.2003. Genetic variation among world populations: inferences from 100 Alu insertion polymorphisms. Genome Res 13:1607-1618.

Yadav AS, Arora P. 2011.Genomic diversities and affinities among eight endogamous groups of Haryana (India): a study on insertion/deletion polymorphisms. Ann Hum Biol 38:114-118.

Zerjal T, Pandya A, Thangaraj K, Ling EY, Kearley J, Bertoneri S, Paracchini S, Singh L, Tyler-Smith C. 2007. Y-chromosomal insights into the genetic impact of the caste system in India. Hum Genet. 121:137-44