Indian Journal of History of Science, 51.2.2 (2016) 406-416 DOI: 10.16943/ijhs/2016/v51i2.2/48454

The Dazzling Diversity and the Fundamental Unity: Peopling and the Genomic Structure of Ethnic India

Analabha Basu*

(Received 12 February 2016; revised 14 June 2016)

Abstract

India with a census size of 1.25 billion, harbours more than one-sixth of the world population. Contemporary India has a rich tapestry of cultures and ecologies. However, when we refer to the heritage and the glorious diversity of the land, we are generally referring to the immense diversity of its people. There are about 400 tribal groups and more than 4000 groups of castes and sub-castes, speaking 22 'officially' recognized languages and 461 'tongues' belonging to four major language families.

Recent advances in molecular and statistical genetics have enabled the reconstruction of human history by studying living humans. The ability to sequence and study DNA by calibrating the rate of accumulation of changes with evolutionary time has enabled robust inferences about how humans have evolved and retrace their early paths of migration.

Modern humans after their origin in Africa about 150,000 years before present, has migrated out-of-Africa to populate the entire globe. It is postulated that one of the earliest of these migration waves passed through India. Subsequently, the Indian sub-continent has occupied the center-stage of many Paleolithic and Neolithic migrations. Waves of migration at different points of time in history have resulted in India being a genetic melting pot. The contemporary social structure of Indian populations is characterized by endogamy with different degrees of porosity. The social structure, possibly coupled with large ecological heterogeneity, has resulted in largescale genetic diversity and local genetic differences within India. In this essay, I provide genetic evidence of how India may have been peopled, the nature and extent of its genetic diversity, and genetic structure among the extant populations of India.

Studies of uniparental markers like the mtDNA and Y-chromosome has established the antiquity of the Indian populations. It has also outlined the possible earliest migration routes from out-of-Africa into India. However, for fine-graining the intricacies of migrations and admixture, uniparental markers are not extremely useful. An early effort with genomewide markers modeled the peopling of India as an admixture with only two distinct ancestries. However, recent studies have shown that the peopling in India has a much more complicated history, which has also been reshaped by cultural and demographic events. Evidence shows that populations of mainland India have atleast four distinct genetic ancestries. A distinct ancestry of the populations of Andaman archipelago is also identified and found to be co-ancestral to Oceania populations.

Key words: Admixture, Ancestry, Ethnicity, Genomic Diversity, Migration

1. Importance of Studying Genetic Diversity of Indian Populations

South Asia, including India, constitutes about one-sixth of the world population. Understanding the genomic diversity of ethnic populations of India holds a major key to the understanding of global genetic diversity and human population structure. Though the cultural, linguistic and social diversity in Indian populations is widely appreciated, its genomic

* National Institute of Biomedical Genomics, Kalyani, West Bengal, Email: ab1@nibmg.ac.in

diversity is grossly underpresented to the global scientific community (Reich, Thangaraj et al., 2009). The contemporary population of this region is structured as an array of sub-populations that are largely isolated in terms of mating, with minimal admixture between some sub-populations. This social custom governed mating patterns not only has shaped the Indian population diversity in an unique fashion for genomewide studies of diseases, but also requires deep understanding in interpretation and inferences relating to population genetic studies. This understanding is also of critical importance to the understanding of the nature and extent of human migration from out-of-Africa to Oceania including the early migration to the Australia (Rasmussen, Guo et al., 2011, Basu, Sarkar-Roy et al., 2016). Recent genetic data also show that India may have served as the primary access route to the peopling of East Asia(Abdulla, Ahmed et al., 2009) and the milieu of populations in India have a more complex genomic history than originally postulated(Basu, Sarkar-Roy et al., 2016).

2. Social Structure of Indian Ethnic Populations and its Implication on Genetic Structure

India occupies the center-stage of human evolution (Cann, 2001). Contemporary India is a rich tapestry of largely intra-marrying ethnic populations. These populations belong to a diverse set of culture and language groups. There are four distinct language families in India, namely, Austro-Asiatic, Dravidian, Tibeto-Burman and Indo-European. The geographical distribution of the language groups within India is largely non-overlapping. The Dravidian speaking groups inhabit southern India, Indo-European speakers inhabit northern India and Tibeto-Burman speakers are confined to northeastern India. By contrast, the numerically small group of Austro-Asiatic speakers, who are exclusively tribal, inhabit fragmented geographical areas of eastern and central India. Interesting population isolates, like the Brahuis of Pakistan, speak the Dravidian language, whereas, numerous isolated populations in South-East Asia speak variants of the Austro-Asiatic language; thus giving rise to the hypothesis that these languages were probably far more widespread that they are now and could have been relic evidences of complex ancient migrations. Culturally, the vast majority of the people of India belong to either tribal or caste societies. The tribal populations are characterized by their traditional modes of subsistence: hunting and gathering, unorganized agriculture, slash and burn agriculture and nomadism (practiced by a limited number of groups). They also have no written form of language and speak a variety of 'tongues' or dialects. On the other hand, the 'mainstream' of Indian populations, within the Hindu fold, comprises castes, characterized by a wide range of occupations and have written forms of language. There is a long-standing debate about the genesis of the caste and tribal populations of India. One model suggests that the tribes and castes share considerable Pleistocene heritage, with limited recent gene flow between them (Kivisild, Bamshad et al. 1999; Kivisild, Rootsi et al., 2003), whereas an exact opposite view concludes that caste and tribes have independent origins (Cordaux, Aunger et al., 2004). Another under-appreciated factor in some studies is the innate complexity of caste origins, including the recognition that some castes have tribal origins (Basu, Mukherjee et al., 2003).

There are over 400 tribal groups and over 4,000 caste groups in India. Though caste and tribe have been administrative social categories in British India; their existence as a social construct probably predates any other similar social categories found elsewhere in the world. Populations belonging to the caste fold have a ranked social order. There are four broad caste groups; however, commonly the caste populations are now ranked as high, middle and low. While

the usage of the ranks as 'high', 'middle' and 'low' can appear to have a value judgment attached to it, but it is scientifically imperative to take cognizance of this hierarchy, since, as D. D. Kosambi has pointed out ''stratification of Indian society reflects and explains a great deal of Indian history, if studied in the field without prejudice''. What is important from a scientific point of view is that there is virtually no exchange of genes among tribal populations or between tribal and caste populations. There is also little exchange of genes between castes, primarily due to strict social rules of marriage within the caste system. Social stratification and norms governing mate-exchange between social strata impact on genetic structures of populations. Therefore, human geneticists have studied the genetic structures, similarities and dissimilarities of ranked caste groups in India aiming to shed light on human history, without prejudice. Historical and anthropological studies suggest that in the establishment of the caste system in India there have been varying levels of admixture between the tribal people of India and the later immigrants bringing along with them knowledge of agriculture, artisanship and metallurgy from central and west Asia. The migrants from central and west Asia who likely entered India through the north-western corridor, spread to most areas of northern India, but not to southern India. There is a distinct gradient of decreasing genetic similarity (representing a cline) of Indian populations with the West- and Central-Asian gene pools as we move eastward or southward from the north-western corridor (Basu, Mukherjee et al., 2003) (Reich, Thangaraj et al. 2009) (Sengupta, Zhivotovsky et al. 2006) (Consortium 2008) (Basu, Sarkar-Roy et al. 2016). In other words, southern and northern India have had differential inputs of genes from central and west Asia. This differential admixture is expected to have differential impacts on the genetic structures of castes of different ranks (Basu, Sarkar-Roy et al., 2016).

3. Molecular Genetics in the Study of Human Evolution and Dispersal

Paleontology and archeology have traditionally provided inferences on early human footfalls, particularly about periods of pre-history. However, such disciplines often rely heavily on scanty, ill-preserved evidence, such as extremely small fragments of skeletal and artifactual remains buried under the soil for centuries, or even exposed to the ravages of nature. Recently, molecular genetics has come to make increasing contributions to our understanding of human evolution, migration and population structure (Cavalli-Sforza 1966) (Cavalli-Sforza, Barrai et al. 1964) (Cavalli-Sforza 1994) (Li, Absher et al. 2008). Since we pass on our genes from one generation to the other, and since random mutations occur and accumulate in our genomes over time, the extant variation we carry in our genomes comprises footprints of the history of our species. The pattern in which these mutations and polymorphisms (mutations that increase to high frequencies in populations) accumulate in our genomes is indicative of the dynamics of the population history (Kingman 2000) (Nordborg 1997) (Hudson 1990). Besides the random accumulation of mutations, the gene pool of a population is also systematically and randomly shaped by genetic drift, demographic changes and differential selective effects (Hudson 1990)(Barreiro, Laval et al. 2008) (Cavalli-Sforza, 1994) (Cavalli-Sforza, Barrai et al. 1964).

Traditionally the maternally-inherited (transmitted by a mother to all her children) mitochondrial DNA (mtDNA) and the paternally-inherited Y-chromosomal DNA (transmitted by a father to all his sons), particularly the non-recombining part of the Y chromosome (NRY), were popular genetic materials to draw inferences about our past (Cann, Stoneking et al., 1987) (Cavalli-Sforza, 1966) (Ke, Su et al., 2001) (Paabo, 1995). Both mtDNA and NRY have their own merits of usage. The uniparental lineages

provide a tool to separately investigate the history of males and females; these provide easy and direct ways of studying gender-specific migrations, which are often different, and contrast them. It is inherently simple to analyze because it is non-recombining and the only variation that we see is because of mutations that have accumulated over time. Both mtDNA and NRY harbor regions which are 'hypervariable' or have high mutation rates and also regions which rarely vary and highly conserved allowing very few mutations to accumulate. The more stable regions are used to draw inferences about distant history, while the fast-mutating regions are used to draw more recent historical inferences.

However, mtDNA and NRY are strongly impacted upon by a non-genetic phenomenon, known as genetic drift, leading to chance fluctuations in frequencies of mutants over generations. By contrast, the impact of genetic drift on autosomal chromosomes is less pronounced, indicating that frequencies of mutations over generations are dictated to a smaller extent by chance. Thus, dates of population divergence and migration events that are usually estimated on the basis of frequencies of genetic markers can potentially show wide differences depending on whether mtDNA/Y-chromosomal or autosomal markers are used. A more profound problem with the uniparental markers is that it represents a very small portion of individual's ancestry, as the mtDNA is inherited from ancestral great grandmother and the NRY from an ancestral great grandfather, whereas, the autosomal data for an individual is a genomic representative of all ancestors the individual has had.

Molecular genetic diversity studies on Indian ethnic populations have primarily used mtDNA and NRY markers (Basu, Mukherjee et al., 2003) (Sengupta, Zhivotovsky et al., 2006; Bamshad, Kivisild et al., 2001; Kivisild, Bamshad et al., 1999; Kumar, Padmanabham et al., 2008; Underhill, Myres et al., 2010); the number of studies that have used autosomal markers have been comparatively fewer. Most past studies that have used autosomal markers, have either been based on a restricted set of loci (Consortium 2008; Rosenberg, Mahajan et al., 2006), or have been based on a large set of markers but on a small number of individuals or populations, although there are some exceptions (Abdulla, Ahmed et al., 2009; Reich, Thangaraj et al., 2009; Basu, Sarkar-Roy et al., 2016).

4. Which Exit Route?

Africa has been the hotbed of human evolution, both modern and archaic. By looking at the confinement of big primates in the African continent, none other than Charles Darwin put forth the hypothesis that humans have a single origin (monogenesis) in Africa. The concept was speculative until the 1980s, when it was corroborated by a study of present-day mtDNA, combined with evidence based on physical anthropology of modern and archaic human remains. According to genetic and fossil evidence, archaic humans evolved to anatomically modern humans (*Homo sapiens sapiens*) solely in Africa, between 200,000 and 100,000 years ago. It is now established beyond reasonable doubt that anatomically modern humans evolved in Africa and expanded and spread throughout much of Africa about 100,000 years ago (Cann, Stoneking et al., 1987; Ramachandran, Deshpande et al., 2005; Li, Absher et al., 2008). Genetic data indicate further migration into Asia, Europe and later to other parts of the world (Ramachandran, Deshpande et al., 2005; Li, Absher et al., 2008).

Which route did the first humans take when they moved out of Africa? The intuitive would have been a walking trail along the river Nile, across the Sinai Peninsula ('northern exit route'). Then, modern humans could have gone towards Europe or Central East Asia via the

Levant. This evidence is also supported by ancient archeological finds in the Levant, postulated to be remnants of modern human settlements out-of-Africa. However, the devil lies in the data, because most studies of human dispersal 'out of Africa' postulates that there was a 'southern exit route' from the Horn of East Africa across the mouth of the Red Sea along the coastline of India to southeastern Asia and Australia (Macaulay, Hill et al., 2005; Oppenheimer, 2012; Lahr and Foley, 1994; Forster, 2004; Forster and Matsumura, 2005) which pre-dates the 'northern exit route'. Fossil records and evidence from genetic data shows that islands in South-East Asia and Southern Pacific including Australia were populated at least as early as the oldest record suggestive of the presence of modern humans outside Africa via the Northern exit route. It is also noted that, the archeological evidence of the northern exit route is associated with the Upper Paleolithic blade-dominated technology, whereas the remnants of the Southern-exit route are associated with much simpler Middle Paleolithic technology. Whether there have been atleast two waves (Lahr and Foley, 1994; Lahr and Foley, 1998) with a time-lag between them, or most of the "out-of-Africa" peopling is a resultant of a single rapid wave of migration (Macaulay, Hill et al., 2005; Oppenheimer, 2012), are unclenched debates of modern population genetics and paleo-archeology.

What is agreed upon is that the probable date of dispersal through the southern exit route was substantially earlier (about 70–80 kya) (Oppenheimer, 2012; Lahr and Foley, 1994). However, as the estimated time of this migration coincides with the last-glacial age, sea-levels were depleted and the coastline extended. Hence the archaeological evidence in favour of the 'beach-combing' first 'out-of-Africa' migrations at best would be scanty, primarily because the coastlines of that period have become deeply submerged because of the rapid rise of sea levels (Stringer, 2000; Oppenheimer, 2012).

Some recent archaeological finds from India have indicated that the major route of dispersal to India from out-of-Africa was through the southern route (Mellars, 2006). The strongest genetic evidence in favour of an early southern exit into India comes from the mtDNA signatures of Indian and other Asian populations. A vast majority of the populations of this region harbor derivatives of the mitochondrial DNA lineages (M and N), which are closely related to the Africa specific L3 lineage (Quintana-Murci, Semino et al., 1999). The southern exit hypothesis is also supported by analyses of mtDNA data from Andaman Islands (Kivisild, Rootsi et al., 2003; Endicott, Gilbert et al., 2003; Thangaraj, Chaubey et al., 2005) and New Guinea (Tommaseo-Ponzetta, Attimonelli et al., 2002; Forster, Torroni et al., 2001). Further, the Y-chromosomal lineages (C and D) are found only in Asian continent and Oceania (Underhill, Passarino et al., 2001; Kivisild, Rootsi et al., 2003), but not in Eurasia or north Africa. The absence of the Y and mtDNA haplogroups (groups of chromosomes bearing similar sets of mutations) in Eurasia and northern Africa pose a strong argument in favour of an early southern exit route of migration and settlement into India. The virtual absence of the mtDNA M haplogroup in the Eurasian as well as north African populations and its extreme high frequency in India and Oceania supports the southern exit route hypothesis (Forster, Torroni et al., 2001; Oppenheimer, 2012). A major limitation of more direct genetic evidence however is the absence of reliable and comparable data from populations which might lie on the geographic coastline connecting Africa and India via the southern exit route (Stringer, 2000).

5. The Early Settlers

Who are the earliest inhabitants of India? The Austro-Asiatic speaking populations of India, who are exclusively tribal, show the highest frequencies of the ancient mtDNA lineage M.

They also show the highest frequency (about 20%) of sub-lineage M2, which has the highest nucleotide diversity within a fast evolving segment (HVS1) of the mtDNA compared to other sub-lineages (Kumar, Padmanabham et al., 2008). Based on these patterns, it has been suggested that the Austro-Asiatic speakers are autochthones and possibly the earliest settlers of India (Basu, Mukherjee et al., 2003; Kumar, Padmanabham et al., 2008). Recent results on Y-chromosomal markers provide further support for this inference. The Y lineage OM95, found in high frequency in India, had originated in the Indian Austro-Asiatic populations around 65 kya. These findings are consistent with linguist Colin Renfrew's observation that the present distribution of the Austric language group is due to the initial dispersal process out of Africa, whereas later agricultural dispersal can account for distribution of the Tibeto-Burman languages. However, recent studies have found that many Dravidian tribal populations also have M2 frequencies comparable to those of the Austro-Asiatic tribal people (Kumar, Padmanabham et al., 2008; Chandrasekar, Kumar et al., 2009). The antiquity of Dravidian speakers in India, who are not all tribal but also belong to the organized caste system, has been extremely enigmatic. There is evidence that they were widespread over nearly the entire landmass of India, which may have overlapped with regions inhabited by the modern Austro-Asiatic speakers, and were possibly pushed to their predominant current habitat in the southern region of India probably after the advent of the Indo-European speakers (Basu, Mukherjee et al., 2003). A hypothesis putting forward 'some' ancestral Dravidian speaking population as an anchor ancestral population of India has been suggested invoking information from autosomal data(Reich, Thangaraj et al., 2009); while more recent studies show further complex relationship between Dravidian and Austro-Asiatic speakers of peninsular and South India (Basu, Sarkar-Roy et al., 2016).

6. Social, Geographical and Linguistic Structures as Evident from mtDNA and NRY Markers

Analysis of genetic structure has shown that Indian ethnic populations when grouped as tribal versus non-tribal, or by geographical region of habitat, or by linguistic affiliation, have a dazzling diversity much larger than the diversity of Europe and is only comparable to Africa (Abdulla, Ahmed et al., 2009; Consortium, 2008; Reich, Thangaraj et al., 2009; Basu, Sarkar-Roy et al., 2016). The Tibeto-Burman speaking tribals separate from the non Tibeto-Burman speakers as the most distinct cluster (Abdulla, Ahmed et al., 2009; Consortium, 2008; Basu, Mukherjee et al., 2003). As the ancestral source of the Indian gene pool, in addition to the original African source population, west Asia (by demic diffusion of agriculture) and central Asia have been the major contributors, although migrations from Asia may have taken place only in historical times, perhaps not earlier than 8 kya. The extent of variation of female lineages (mtDNA) in India is rather restricted (Basu, Mukherjee et al., 2003; Roychoudhury, Roy et al., 2001), indicating a small founding group of females. By contrast, the variation of male lineages (Y-chromosomal) is very high (Basu, Mukherjee et al., 2003; Sengupta, Zhivotovsky et al., 2006). This pattern may be indicative of sex-biased ancient gene flow into India with more male immigrants than female (Bamshad, Watkins et al., 1998), possibly occurring within the last 8000 years.

This phenomenon of large-scale sex-biased post-agriculture migration and displacement, coupled with rapid population growth, obscures ancient genetic signatures and results in the quick introduction of high genetic variability, often mimicking extreme natural selection (Zerjal, Xue et al., 2003). The success of some of the Y-chromosomal haplotypes that arose in Central Asia to spread across vast regions of Eurasia (Zerjal, Xue et al., 2003), as well as

South and Southeast Asia, is indicative of the "success" of the cultural and technological dominance of west Eurasia and Central Asia (Zerjal, Xue et al., 2003; Underhill, Myres et al., 2010).

The mtDNA lineage U, which is likely to have arisen in central Asia, has a high frequency in India implying that large-scale migration brought a large number of copies of this lineage into India. However, this lineage was shown to comprise two deep sub-lineages, U2i and U2e, with an estimated split around 50 kya. The sub-lineage U2i is found in high frequencies in India (particularly among tribes~77%) but not in Europe, whereas U2e is found in high frequencies in Europe (>10%) but not in India (except at very low frequencies among castes, but not among tribals). Thus, a substantial fraction of the U lineage – specifically, the U2i sublineage – may be indigenous to India (Basu, Mukherjee et al., 2003; Kivisild, Rootsi et al., 2003).

Analysis of the complete mtDNA genome sequence has revealed a large number of sequence variants within major haplogroups in Indian populations, many of which, however, occur sporadically. This indicated a common spread of the root haplotypes of haplogroups M, N and R around 70–60 kya along the southern exit route. The analysis has further revealed that entry of the haplogroup U2 post-dates the earliest settlement along the southern route.

Central and Central South Asian populations are supposed to have been the major contributors to the Indian gene pool, particularly to the northern Indian gene pool, and the migrants had supposedly moved into India through what is now Afghanistan and Pakistan. Using mitochondrial DNA variation data collated from various studies, it has been shown that populations of Central South Asia and Pakistan show the lowest genetic distance with the north Indian populations (F_{ST} = 0.017), higher (F_{ST} = 0.042) with the south Indian populations, and the highest (F_{ST} = 0.047) with the northeast Indian populations. Northern Indian populations are genetically closer to Central Asians than populations of other geographical regions of India (Bamshad, Kivisild et al., 2001; Basu, Mukherjee et al., 2003).

Although considerable cultural impact on social hierarchy and language in south Asia is attributable to the arrival of nomadic central Asian pastoralists, studies using NRY polymorphisms reveal that the influence of central Asia on the pre-existing gene pool was minor. The ages of accumulated microsatellite variation in the majority of Indian haplogroups exceed 10-15 kya, which attests to the antiquity of regional differentiation. Therefore, NRY data do not support models that invoke a pronounced recent genetic input from central Asia to explain the observed genetic variation in south Asia. Haplogroups R1a1 and R2 indicate demographic complexity that is inconsistent with a recent single history. Associated microsatellite analyses of the high-frequency R1a1 haplogroup chromosomes indicate independent recent histories of the Indus Valley and the peninsular Indian region. These data are also more consistent with a peninsular origin of Dravidian speakers than a source with proximity to the Indus and with significant genetic input resulting from demic diffusion associated with agriculture; rather they indicates that pre-Holocene and Holocene-era – not Indo-European – expansions have shaped the distinctive South Asian Y-chromosome landscape (Sengupta, Zhivotovsky et al., 2006).

7. Analyses Using Large Sets of Genomic Markers Provides Better Estimates and New Paradigms

The uniparental markers, each represent a single locus in the entire genome and is representative of a very small portion of individual's ancestry. To tease out intricate details

of migration, particularly to infer about historic and pre-historic admixture events, the autosomal markers are hugely more informative.

Analyses of data on 405 SNPs from a 5.2 Mb region on chromosome 22 in 1871 individuals from diverse 55 Indian populations have revealed that Indian populations form a genetic bridge between West Asian and East Asian populations (Consortium, 2008). The HUGO Pan Asian SNP Consortium's study (Abdulla, Ahmed et al., 2009) also showed that most of the Indian populations shared ancestry with West Asian populations, which is consistent with the recent observations and our understanding of the expansion of Indo-European speaking populations. The study also provided evidence that the peopling of India (and also of southeast Asia) was via a single primary wave of migration out-of-Africa (Abdulla, Ahmed et al., 2009).

Using over 500,000 biallelic autosomal SNPs, Reich et al. (2009) have also found a north to south gradient of genetic proximity of Indian populations to western Eurasians/central Asians. In general, the central Asian populations were found to be genetically closer to the higher-ranking caste populations than to the middle- or lower-ranking caste populations. Among the higher-ranking caste populations, those of northern India are, however, genetically much closer (F_{ST} = 0.016) than those of southern India (F_{ST} = 0.031). Phylogenetic analysis of Y-chromosomal data collated from various sources yielded a similar picture.

Reich et al. (2009) have also proposed an elegant model where that extant populations of India were 'founded' by two hypothetical ancestral populations, one ancestral north Indian (ANI) and another ancestral south Indian (ASI). Presumably, these ancestral populations were derived from ancient humans who entered India via the southern and the northern exit routes from out of Africa. All extant Indian populations are derived from admixture between the two ancestral populations, with the ANI contribution being higher among extant north Indian populations and that of ASI being higher among extant south Indian populations. The relative lack of importance of Tibeto-Burman and Austro-Asiatic speakers in shaping of the contemporary genomes of the Indian population was conspicuously detectable in the Reich et al. paper (2009). In a more recent study (Moorjani, Thangaraj et al. 2013), these investigators have shown that between 1900 – 4200 ybp, there was extensive admixture among Indian population groups, followed by a shift to endogamy. This model is simplistic, but intuitive and consistent with findings of earlier studies. It is simplistic because the origins of populations in northeastern region of India cannot be explained by this model since many past studies (cited earlier in this essay) have indicated genetic inputs into these populations from populations of southeast Asia. A more recent study analyzing data on more than 80k SNPs on 367 individuals from 18 mainland and 2 island populations, has identified 2 more components in the mainland of the Indian sub-continent besides the ANI and ASI (Basu, Sarkar-Roy et al., 2016). These 2 newly identified ancestral components are primarily seen among Tibeto-Burman(TB) speaking tribal populations of northeast India and Austro-Asiatic(AA) speaking tribal populations of central and east India, and the authors have identified them as (AAA for Ancestral Austro-Asiatic and ATB as Ancestral Tibeto-Burman) (Basu, Sarkar-Roy et al., 2016). Of these the authors have shown that the ANI is co-ancestral to Central South Asians and the ATB to be co-ancestral to the East Asians, particularly the Southeast Asians of the HGDP. The absence of significant resemblance with any of the neighboring populations is indicative of the ASI and the AAA being early settlers in India, possibly arriving on the 'southern exit' wave out-of-Africa. Differentiation between the ASI and the AAA may have taken place after their arrival in India. The ANI and the ATB, can clearly be rooted

to the CS-Asians and E-Asians, respectively; they likely entered India through the NW and NE corridors, respectively. Ancestral populations seem to have occupied geographically separated habitats.

8. Discussion

Reconstructing past events has always been a challenge to science, because it almost axiomatically rules out the possibility of a replication experiment for validation. Automatically it warrants caution when we try to reconstruct the past events based on genomic data. Though compared to disciplines like paleontology and archeology, molecular genetics has an advantage, yet inferences from genomic data should also be considered in conjunction with inferences from other disciplines like paleontology, archeology, linguistics, anthropology etc. A major caveat in the reconstruction procedure is the error variance that comes with the statistical estimation procedure. Though the broad-brush sequence of events can be reconstructed with confidence, any direct estimation of timing or dating of past events generally comes with a large variance. The difficulty of the above problem is that it would not improve substantially by increasing the sample size of individuals. Instead, it has be shown, both theoretically and empirically that including larger portion of individual's genome would substantially improve the inference procedure.

The timecourse of molecular genomic studies of human evolution has naturally progressed from studies of handful of markers or variable sites on the genome to the contemporary state-of-the-art study designs where the entire human genome with all its variation is studied. Though, this timecourse is a natural resultant of the advancement of technology, enabling us to investigate large proportion of the human genome in a time and cost efficient manner; yet it has naturally enabled a much more insightful and comprehensive understanding of human evolution, because of the phenomenon that including larger portion of individual's genome substantially improves the information content and the inference procedure.

As a postscript, studies in population genetics overlaps with disciplines of social science and are often interpreted to make strong social statements. Caution should be applied when findings in one study are stretched to infer conclusions in some other related discipline. One sensitive such broad area is what surrounds the concept of 'caste' and 'tribe' in India. As population genomics and studies in evolutionary genomics are intertwined with studies of genetic epidemiology and the evolution of diseases and traits, it is mandated by looking for clustering in populations. Though it has been clearly established that more than 90% of the total variation in human populations is because of variations between individuals irrespective of any population label, there has been a tendency to over interpret the ~10% of variation that is systematic and would be enough to find clusters within the global and local populations. This homogeneity in human population has extremely strong underpinnings in designing studies of genetic epidemiology and hence there is huge scientific merit in D D Kosambi's observation that "stratification of Indian society reflects and explains a great deal of Indian history, if studied in the field without prejudice". Mechanically equating genetic homogeneity with labels of 'caste' and 'tribe', which are administrative categories of British rule, can be a huge pitfall in our understanding. Though 'endogamy', the primary cause of the resultant homogeneity is often one of the cultural practices within an ethnic group, we should minimally appreciate that the 'caste' and 'tribe' are complex bolus of long socio-cultural history and a dynamic interaction over a complex multi-dimensional landscape.

Acknowledgement

This article has drawn heavily upon previous publications co-authored by me with Partha Majumder.

Bibliography

Abdulla, M. A.; I. Ahmed, et al. "Mapping human genetic diversity in Asia." *Science* 326.5959 (2009): 1541-5.

Bamshad, M.; T. Kivisild, et al. "Genetic evidence on the origins of Indian caste populations." *Genome Res* 11.6 (2001): 994-1004.

Bamshad, M. J.; W. S. Watkins, et al. "Female gene flow stratifies Hindu castes." *Nature* 395.6703 (1998): 651-2.

Barreiro, L. B.; G. Laval, et al. "Natural selection has driven population differentiation in modern humans." *Nat Genet* 40.3 (2008): 340-5.

Basu, A.; N. Mukherjee, et al. "Ethnic India: a genomic view, with special reference to peopling and structure." *Genome Res* 13.10 (2003): 2277-90.

Basu, A.; N. Sarkar-Roy, et al. "Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure." *Proc Natl Acad Sci U S A* (2016).

Cann, R. L. "Genetic clues to dispersal in human populations: retracing the past from the present." *Science* 291.5509 (2001): 1742-8.

Cann, R. L., M. Stoneking, et al. "Mitochondrial DNA and human evolution." *Nature* 325.6099 (1987): 31-6.

Cavalli-Sforza, L. L. "Population structure and human evolution." *Proc R Soc Lond B Biol Sci* 164.995 (1966): 362-79.

Cavalli-Sforza, L. L., I. Barrai, et al. "Analysis of Human Evolution under Random Genetic Drift." *Cold Spring Harb Symp Quant Biol* 29 (1964): 9-20.

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

Chandrasekar, A., S. Kumar, et al. "Updating phylogeny of mitochondrial DNA macrohaplogroup m in India: dispersal of modern human in South Asian corridor." *PLoS One* 4.10 (2009): e7447.

Consortium, I. G. V. "Genetic landscape of the people of India: a canvas for disease gene exploration." *J Genet* 87.1 (2008): 3-20.

Cordaux, R., R. Aunger, et al. "Independent origins of Indian caste and tribal paternal lineages." *Curr Biol* 14.3 (2004): 231-5.

Endicott, P., M. T. Gilbert, et al. "The genetic origins of the Andaman Islanders." *Am J Hum Genet* 72.1 (2003): 178-84.

Forster, P. "Ice Ages and the mitochondrial DNA chronology of human dispersals: a review." *Philos Trans R Soc Lond B Biol Sci* 359.1442 (2004): 255-64; discussion 264.

Forster, P. and S. Matsumura. "Evolution. Did early humans go north or south?" *Science* 308.5724 (2005): 965-6.

Forster, P., A. Torroni, et al. "Phylogenetic star contraction applied to Asian and Papuan mtDNA evolution." *Mol Biol Evol* 18.10 (2001): 1864-81.

Hudson, R. R., Ed. Gene genealogies and the coalescent process. Oxford surveys in evolutionary biology. New York, Oxford University Press (1990).

Ke, Y., B. Su, et al. "African origin of modern humans in East Asia: a tale of 12,000 Y chromosomes." *Science* 292.5519 (2001): 1151-3.

Kingman, J. F. "Origins of the coalescent. 1974-1982." *Genetics* 156.4 (2000): 1461-3.

Kivisild, T., M. J. Bamshad, et al. "Deep common ancestry of indian and western-Eurasian mitochondrial DNA lineages." *Curr Biol* 9.22 (1999): 1331-4.

Kivisild, T., S. Rootsi, et al. "The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations." *Am J Hum Genet* 72.2 (2003): 313-32.

Kumar, S., P. B. Padmanabham, et al. "The earliest settlers' antiquity and evolutionary history of Indian populations: evidence from M2 mtDNA lineage." *BMC Evol Biol* 8 (2008): 230.

Lahr, M. M. and R. Foley. "Multiple dispersals and modern human origins." *Evolutionary Anthropology: Issues, News, and Reviews* 3.2 (1994): 48-60.

Lahr, M. M. and R. A. Foley. "Towards a theory of modern human origins: geography, demography, and diversity in recent human evolution." *Am J Phys Anthropol Suppl* 27 (1998): 137-76.

Li, J. Z., D. M. Absher, et al. "Worldwide human relationships inferred from genome-wide patterns of variation." *Science* 319.5866 (2008): 1100-4.

Macaulay, V., C. Hill, et al. "Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes." *Science* 308.5724 (2005): 1034-6.

Mellars, P. "Going east: new genetic and archaeological perspectives on the modern human colonization of Eurasia." *Science* 313.5788 (2006): 796-800.

Moorjani, P., K. Thangaraj, et al. "Genetic evidence for recent population mixture in India." *Am J Hum Genet* 93.3 (2013): 422-38.

Nordborg, M. "Structured coalescent processes on different time scales." *Genetics* 146.4 (1997): 1501-14.

Oppenheimer, S. "Out-of-Africa, the peopling of continents and islands: tracing uniparental gene trees across the map." *Philos Trans R Soc Lond B Biol Sci* 367.1590 (2012): 770-84.

Paabo, S. "The Y chromosome and the origin of all of us (men)." *Science* 268.5214 (1995): 1141-2.

Quintana-Murci, L., O. Semino, et al. "Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa." *Nat Genet* 23.4 (1999): 437-41.

Ramachandran, S., O. Deshpande, et al. "Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa." *Proc Natl Acad Sci U S A* 102(44) (2005): 15942-7.

Rasmussen, M., X. Guo, et al. "An Aboriginal Australian genome reveals separate human dispersals into Asia." *Science* 334.6052 (2011): 94-8.

Reich, D., K. Thangaraj, et al. "Reconstructing Indian population history." *Nature* 461.7263 (2009): 489-94.

Rosenberg, N. A., S. Mahajan, et al. "Low levels of genetic divergence across geographically and linguistically diverse populations from India." *PLoS Genet* 2.12 (2006): e215.

Roychoudhury, S., S. Roy, et al. "Genomic structures and population histories of linguistically distinct tribal groups of India." *Hum Genet* 109.3 (2001): 339-50.

Sengupta, S., L. A. Zhivotovsky, et al. "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* 78.2 (2006): 202-21.

Stringer, C. "Palaeoanthropology. Coasting out of Africa." *Nature* 405.6782 (2000): 24-5, 27.

Thangaraj, K., G. Chaubey, et al. "Reconstructing the origin of Andaman Islanders." *Science* 308.5724 (2005): 996.

Tommaseo-Ponzetta, M., M. Attimonelli, et al. "Mitochondrial DNA variability of West New Guinea populations." *Am J Phys Anthropol* 117.1 (2002): 49-67.

Underhill, P. A., N. M. Myres, et al. "Separating the post-Glacial coancestry of European and Asian Y chromosomes within haplogroup R1a." *Eur J Hum Genet* 18.4 (2010): 479-84.

Underhill, P. A., G. Passarino, et al. "The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations." *Ann Hum Genet* 65.Pt 1 (2001): 43-62.

Zerjal, T., Y. Xue, et al. "The genetic legacy of the Mongols." *Am J Hum Genet* 72.3 (2003): 717-21.