Chapter 1

Introduction
Introduction

Origin and evolution has always been a curiosity of human beings. Traditionally, historian, archeologist and paleontologist have investigated the past, while indirect evidence from modern human population has been sought by linguistics and increasingly today by molecular biologist. The most useful information comes from ‘genetic polymorphism’- variations transmitted in strict Mendelian fashion. DNA that is inherited to us through generations under goes mutations on its way, which generate polymorphism. The arsenal of available markers has let the path to analyse such polymorphism, and it is here that the molecular revolution, especially the more recent development in the direct study of DNA, are generating great progress in the field of population genetics.

Major demographic event like migration bottlenecks and population expansion; eave genetic imprint by which gene frequencies are altered in general, human being group themselves units in such away that members within a unit freely interbreed and thus freely exchange genes, but members between the units rarely exchange there genes due to cultural and geographical barriers resulting in genetic divergence of the populations. From an estimate of genetic distance between two
populations their time of divergence can be estimated. Thus, genetic distance between populations provides useful data for reconstructing and dating evolutionary history.

Recent advances in the molecular genetic technology and the bioinformatics tools have given a new era for the science of human evolution (Cavalli-Sforza and Feldman 2003; Jorde and Wooding, 2004). Emergence of lineage based analysis of human non-recombining chromosome (NRY) and mt-DNA molecules and detection of numerous unique event polymorphisms-UEPs (both single nucleotide polymorphisms-SNPs and insertion deletion events-indels) have paved the way for reconstructing the genealogies of both extinct and extant haplotypes in the form of bifurcating trees (Underhill et al, 2001; Jobling and Tyler Smith, 2003; Ingman et al, 2000). Genotyping of length-repeat polymorphisms in the form of highly mutating short tandem repeat (STR) markers became a successful indicator of the extent of diversity and branching of both population and mutation splits (Jorde et al, 1997; Bowcock et al, 1994; Rosenberg et al, 2002). More recently, haplotype blocks of strong Linkage Disequilibrium (LD) depicted by genome-wide SNPs are projected to be an ideal candidate to infer not only the pattern of genetic variation but also the genetic basis of complex multi-factorial diseases (Reich et al, 2001; Wall and Pritchard, 2003; McVean et al, 2004). Furthermore, availability of high throughput genotyping techniques and user-friendly software packages based on more logistic and reliable algorithm has made the collection and interpretation of the genetic data more feasible.

More than 80 years of research of genetic variation have revealed that a substantial degree of diversity lies both within and between extant human populations. Despite the huge amount of data gathered during this relatively long time span of classical ("pre-DNA") genetics, recently reviewed in a fundamental way by Cavalli-Sforza et al. (1994), the "classical era" raised many basic problems, leaving them largely unsolved. Increasingly more resourceful "DNA era", rapidly expanding during the last 20 years, took up the same list of problems and is formulating new ones (reviews see Foley 1998; Jorde et al. 1998; Paabo 1996; Relethford 1999)
Largely irrespective of what kind of general questions related to demographic history are being asked, the present-day genetics investigates variation in three different systems: autosomal chromosomes, paternally inherited Y chromosome and maternally inherited mitochondrial DNA. While the size of autosomal genes exceeds manifold that of Y chromosome and vastly the content of mtDNA, the latter two have powerful advantages, making them the favorite tools in the hands of population geneticists. Besides uniparental mode of inheritance, the list of advantages includes also the lack of recombination. Taken together, they allow, at least in theory, to reconstruct genetic lineages back to the most recent common ancestors, Y chromosomal Adam and mitochondrial Eve. And in combination with information about the variation in autosomal genes, a promising synthesis is hopefully possible in the future.

1.1 Human Genetic Variation

Much of the genetic variations in human population may well be selectively nearly neutral; indeed, neutrally has been assumed in some of the analysis we have discussed. However, a great deal of human genetic variations, especially in genes that can cause disease or deficiencies, in certainly or probably subject to natural selection. Much of the huge field of human genetics concerns such traits.

Human genetics is inspirable from evolutionary genetics. Genes underline hereditary disorders can be located by tracing the inheritance of a trait in family pedigree by its association with genetic markers (linkage mapping). Because humans can not be experimentally crossed like fruit flies, such analysis depends on estimates of allele frequencies in population and the theory of linkage disequilibrium among loci, both of which are the subject matter of population genetics. Many of the statistical methods for determining inheritance, as well as relative importance of genes and environment, have been developed by evolutionary genetics.

Our knowledge of human genetics is growing explosively as genetics use molecular methods to identify and sequence genes, identify gene products and their functions, and identify the mutation that may impair their those functions. To facilitate the growth of such knowledge, genetics throughout the world contributing
to the Human Genome Project, an effort initiated in 1989 to sequence all 3 billion base pairs, including perhaps 100,000 functional genes in the human genome (Cooper 1994).

1.2 Tools to study human genetic variation

Generally identification of different alleles at genetic level for the study of population through observation of the expressed traits or outward, physical manifestation of a gene, called the phenotype. Mendelian genetics allowed population geneticists to identify the heritable form of a gene (genotype) including individual variants (alleles). Advances in molecular genetics facilitated identification of single genes at the molecular or biochemical level. Regardless of the method used to identify genes and their alleles, allelic data obtained on different populations is analyzed by statistical analyses of allele frequencies to understand and make prediction about gene flow in populations. Estimated gene frequencies of alleles at different loci are the estimates of the relative frequencies of alleles and are of widest application in the studies of structural dynamics and evolution of natural (particular in human) populations. Information about the relative frequencies of different alleles is of paramount importance in the study of population structure. However, the estimation of gene frequencies is plagued by the phenomenon of sampling fluctuations and misclassification of genotypes however this can be solved by using right markers and more robotic techniques. The markers that hold high importance in studying the pattern of genetic diversity are distributed throughout the human nuclear and cytoplasmic genome.

(i) Restriction Fragment Length polymorphism: The nuclear genome contains the great majority of human polymorphisms. These include the “classical” systems (blood groups and protein polymorphisms) as well as single nucleotide variants, usually ascertained as restriction fragment length polymorphisms (RFLPs). It is estimated that the mutation rate for single nucleotides is approximately $10^{-7}$ to $10^{-8}$ per generation (Jorde and Barnshad, 1998). The slow mutation rate of these polymorphisms provide means of visualizing the ancient history of our species, but these polymorphisms are likely to be relatively uninformative about recent history.
(ii) **Y-chromosome markers:** Y chromosome is a haploid marker which is transferred from father to son and reveals paternal lineages. Most of the paternally inherited Y chromosome does not undergo recombination (Non-Recombining Y, NRY); although recombination occurs at the tips of the Y chromosome in the pseudo-autosomal region -PAR (Jobling and Tyler Smith 2000) and thus genetic information is passed intact from father to son. Due to this property of Y chromosome, it has been widely studied to trace the origin of human populations. Analysis of Y chromosome variation in studying human evolution got off to a slow start due to a lack of identification of informative polymorphic sites. However, an abundant number of polymorphisms were recently identified and analyzed for variation across ethnically diverse human populations (Underhill et al, 1997, 2000, 2001). These include studies of SNPs (Hammer et al, 2000; Underhill et al, 2001; YCC, 2002; Jobling and Tyler smith, 2003), Alu polymorphisms (Rowold and Herrera, 2003), microsatellite (Kayser et al, 2000, 2003; Heyer et al, 1997; Karafet et al, 2001) and combination of these markers.

To estimate the migration event which might have occurred several thousand years ago, Y chromosomal markers have been used to investigate the genetic contribution. The biggest advantage of working with Y-chromosomal lineages is the presence of UEPs i.e. unique event polymorphisms because of absence of recombination and rarity of back or recurrent mutation found on Y-chromosome. A classical combination of Y-UEPs and Y-STR assigns different haplogroups A-Q mentioned in YCC binary tree (YCC, 2002) based on unique occurrence of UEPs while the highly mutating YSTR helps in assigning the age of each such haplogroup clusters (Jobling and Tyler smith, 2003).

(iii) **mt-DNA:** The mitochondrial DNA (mtDNA) plays a very important role in human evolutionary genetics, because mtDNA is inherited only through the maternal cytoplasm, variation in mtDNA provides a record of the maternal lineage of our species. There is no recombination in mtDNA as in autosomal DNA. Hence, the difference between any two-mitochondrial sequences represents only the mutations that have taken place since each sequence was derived from a common ancestor (Ingman et al, 2000). The non-coding mitochondrial control region is especially rich
in polymorphisms, and the mutation rate is estimated to be approximately $2 \times 3 \times 10^{-7}$ per nucleotide per generation (Horai et al., 1995). The mtDNA accumulates mutations about ten times faster than nuclear DNA. As a result, mtDNA is altered so quickly that it is easy to measure the differences between one human population and another, since separate groups have accumulated different sets of mutations. Two closely related populations should have only minor differences in their mtDNA. Conversely, two very distantly related populations should have large differences in their mtDNA (Ingman et al., 2000; Torroni et al., 1998; 2000). The mt-DNA also offers a unique combination of highly mutable Hyper Variable Region-I and II (HVR-I and II) in the non-coding control region and coding region RFLPs to assign number of haplogroups that can define the phylogeographic distribution of genetic diversity.

(iv) **Microsatellites:** Nuclear polymorphism consists of 2-5 bp tandemly repeated DNA sequences in (Jaffery et al., 1994). These are termed microsatellite systems, and the number of repeats in each system varies from one individual to the next. Thousands of microsatellite polymorphisms have now been identified in the human (Dib 1996). Their mutation rate is much higher than that of single nucleotides, approaching $10^3$ per generation (Weber and Wong, 1993). Because of this high mutation rate, microsatellite has the potential to provide information about recent evolutionary events. In particular, they can assign dates to events that have occurred since the origin of humans with greater accuracy.

Due to high mutation rate microsatellite also has a heterozygosity (>0.5) and a large number of alleles (Tishkoff and Kidd 2004; Weber and Wong 1993). They are suitable marker for population genetic studies because they are found highly polymorphic and show co-dominant expression of alleles. These features, along with automated DNA typing procedure (Khan and Agrawal 2005), make them most desirable tool for scoring genetic variability in the human population, inferring evolutionary relationships and demographic histories (Bowcock et al., 1994; Calafell et al., 1996), and in human identification for determining parentage and relatedness of individuals (Budowle et al., 1998).
1.3 Evolutionary event affect the genetic variation

Migration is an important factor in human evolution that can affect the genetic variation within a population. Most populations are relatively isolated, however, although the rare exchange of marriage patterns between groups does occur. An average of one immigration/generation in a population is sufficient to keep drift partially in check and to avoid complete fixation of alleles. Sometimes a whole population (or fraction of it) migrate and settle elsewhere.

Figure:-1.1 Genetic structuring of Populations in relevance to geographic distances.

Summary tree of world populations and relationship between genetic and geographic distance. (a) Phylogenetic tree based on polymorphism of 120 protein genes in 1,915 population grouped by continental sub-areas and Fst genetic distances (b) Genetic distance of population pairs measured by Fst as a function of geographic distance between number of pairs.

If the migrant group is initially small but subsequently expand by chance alone the frequencies of alleles among the founder of the new population will differ from those of the original population and even more so far those among which it settles. In this situation, group migration has an effect that is in some respects is opposite to that of extent of selection, the degree to which pairs of sites interact in response to selection (epistasis), with population scale forces, such as drift, migration and non random mating, genomic pattern of LD be expected to be fairly complex. Recent studies of relatively long (200-500 kb) stretch of DNA however produce a
picture of blocks of high LD inter-spread by short intervals of low LD. Within the blocks of high LD there is evidence of lack of recombination, where as the regions between the blocks seem to be hot spots in which recombination occurs frequently. It has been therefore suggested that the next phase of research in to human variation should focus on these blocks of high LD, for which haplotype, rather than single markers, will become unit of variation. All though it has been known for many years that the extent of LD among specific sets of genes shows great variation around the world for example, it is usually much weaker in African than in European populations.

1.4 Molecular approaches to study human genetic variation

The basic mile stone of all the human genetic diversity carried out till date, is DNA. Still, the choice of population, polymorphic marker, genotyping procedures and statistical methods designs the main strategies for human genetic diversity analysis.

Selection of population so far has been concentrated on non-Admixed populations of widely separated geographical zones like sub-Saharan Africans, Han Chinese and Europeans or that of rare ethnic groups like tribes of native American (Karafet et al, 2002) or north eastern Indian (Cardaux et al, 2004) or Andaman island origin (Thangaraj, et al, 2005). However, the admixed populations of geographically intermediate zones like Central Asia, Middle East or Southwest Asia are generally unexplored or rarely studied (Kayser et al, 2000 and 2003). Furthermore the populations with heavy admixture of various gene pools like Indian caste and tribal populations are usually a neglected lot. Such population will provide the missing links in phylogeography of lineage based NRY and mt-DNA genealogies and present a continuous gradient or cline of genetic variation between geographic extremes.

Normally markers for genetic studies starts from blood grouping to recent lineage based NRY and mt-DNA and LD based SNP haplotype blocks. The existence of genetic variation among human population was first demonstrated by Hirschfeld and Herszfeld, 1919, in the first human gene to be described-ABO that determines the ABO blood groups. The subsequent identification of blood group
protein markers like MNS and Rh expanded the repertoire of polymorphic markers that were analyzed using immunological techniques. Pauling et al, 1949, has introduced the electrophoresis technique and that saw a rapid expansion I protein variant analysis among human groups. R.A. Fisher showed that that the evolution could be reconstructed by analyzing multilocus genotype on a chromosome. Cepellini et al, 1967 describe the term ‘haplotype’ for multilocus combination of alleles on a chromosome based on their study on HLA class II loci. However, in last 3 decades, with the emergence of PCR, restriction enzyme technology and sequencing methods has revolutionized the studies of genetic variation. Now DNA based markers like RFLPs, Alu insertions, STRs, SNPs and indels are preferred for their neutral aspect of information. If the purpose of study is to trace the evolutionary pattern, then the UEPs of NRY and control region (HVRI and II) of mt-DNA are preferred, while YSTR are used to determine the TMRCA of each of the haplogroups clusters. Autosomal STR loci and Alu insertions are preferred for genetic diversity and sub-structuring studies. Highly polymorphic and rarely recombining HLA class I and II loci are other commonly studied markers. Next phase of research of the human variation is focused on high LD blocks of genome-wide SNP haplotypes as a marker of variation instead of one marker. By the recent discovery of denaturing HPLC (Underhill et al, 1997, 2000, 2001), mass spectrophotometry, array based re-sequencing automated DNA sequencing, automated fragment size analysis and SnapShot PCR has not only increased our repertoire of mutation but have also provided a high throughput techniques for quick and reliable genotyping. However, still the choice of technique varies from RFLP, ARMS, gene-scanning and re-sequencing to dHPLC depending upon the type of marker studies.

The selection of statistical approach has become well-targeted aspect of genetic variation studies. Various statistical parameters and software packages are available for different requirements. For Phylogenetic reconstruction, two preferred approaches are ‘cluster’/’clade’ based and ‘cline’/’gradient’ based. Cluster based approach involve population trees constructed from the genetic distances between pair of populations and then averaging the genetic distance for numerous of genes.
Such trees include NJ and UPGMA tree (Feistlein 1993). If instead of genetic distances, allele frequency variation is only considered then the radial phylogram is called as ML-tree. Alternative approach for cluster-based analysis is Principal component (PC) analysis that summarizes independent pattern present in the data matrix and carry information of allele frequency variation for hundreds of gene frequencies (Cavalli-Sforza 1994). PC-analysis involves plotting of first two PCs as co-ordinates, as a result populations with similar genetic structure clusters together. On the contrary if only single PC or the frequency pattern of an allele or a haplotype is plotted on the geographical map, it produces a cline or gradient of variation (Menozzi et al, 1978).

Construction of trees based on genologies of haplotype is the most defined statistical strategies for lineage-based marker (NRY and mt-DNA) and these trees are constructed from the genetic information of individual and not the populations. Further highly resolving YSTR and their mutation rates are used to carry out TMRCA analysis for calculating the dates of branching points or the coalescence periods. As far as analysis of genetic variance is considered, then Wrights F-statistic (Fst) is the universally acclaimed and most superior method. Fst can be calculated for both within and between population variance. Furthermore, various other algorithms are also used to analyze the level and pattern of sub-structuring (STRUCTURE), creating a network of mutation and its order for individual population (NETWORK) and detecting level of admixture from two different sources (ADMIX) etc. Pair wise Mismatch analysis is carried out to determine the level of population expansion, while Tajima’s‘d’ and Wu’s “Fs” are used to determine the neutrality of a locus.

1.5 Improvement of human genetic variation studies

To find out evolution and characterization of variations among different population groups, world wide scientists have worked using numerous of coding and neutral markers and various populations. The most surprising among them is the near validation of the ‘A flag model of human evolution’ based on lineage based molecular approaches and determination of ‘genetic variation’ (amount, distribution
and pattern), ‘genetic sub-structuring’, ‘genetic ancestry’ (of populations and individuals) and ‘implication of genetic variation in biomedical research’.

(i) A flag model of human evolution

Scientists have given many models of human evolution, one of them is Polygenism in which human race originated independently with every little or some times no gene flow between them. Then comes an other model which is based on fossil records called multi regional origin model starting ~8-1.8 billion years ago when homo erectus migrated from Africa there has been parallel evolution from Homo erectus to Homo sapiens in geographically dispersed population with little gene flow (Wolpoff 1996).

Then comes a newer model ‘Recent African Origin (RAO)’ or ‘Out of Africa’ as the standard model of human evolution. According to this model, modern humans ancestors were evolved in Africa ~ 200 Kya and contemporary world populations are largely the descendants of people who migrated out of Africa ~ 50-100 Kya and expanded into Asia and from there to other continents (Cavalli-Sforza et al, 2003) supplanting any archaic population like Neanderthals. This model is based on the data collected from the synthesis of NRY and mt DNA genologies and numerous studies on autosomal loci and other genetic data as this model is most accepted both the above cited models were almost discarded. Spread of humans in Asia occurred through two routes, southern and central-northern routes (Stringer 2000). The first was the southern route perhaps along the coast of south and south East Asia (through Indian sub-continent), from where it bifurcated to south till Sahul and Oceania ~ 60 Kya (Quintana Merci 1999; Underhill 2001) and north towards China, Japan and second wave towards America (Greenberg 1987). Second was the central-northern route through, Middle East, Arabia or Persia (~45 Kya), from where the migration occurred in all direction reaching Europe ~ 40 Kya, central Asia, east and northeast Asia and the first wave to America ~30 Kya (Fagan 1987). Out of Africa model is depicted in Figure 1.2
Figure 1.2 Out of Africa model of human evolution

Possible route of migration of modern human out of Africa deduced from NRY phylogeography

The chronology of mutation signifying evolution, migration, contraction and expansion population groups runs in parallel for mt-DNA and NRY genealogies but TMRCA varies as ~160-225 Kya for mt DNA (Ingman et al, 2000) and ~60-130 Kya for NRY (Tang et al, 2002). This discrepancy could be to high evolving rate of mt-DNA or gender based movement events like practice of patrilocality in most of the human societies.

(ii) Enumeration of ‘genetic variation’

Thousands of studies from past three decades ranging from blood groups, protein polymorphism, RFLPs, SNPs and STR loci have shown that most of the genetic variation (~85-90%) is enumerated among individuals rather than among populations that constitutes only 5-14% of the total genetic variation (Lewontin 1972; Jorde et al, 2000; Cavalli-Sforza 2003; Tishkoff et al, 1996; Tishkoff and Kidd 2004; Tishkoff and Virelli et al, 2003; Kidd et al, 2004). Recently, Rosenberg et al, 2002 in his study of 377 STR in 52 populations has shown that only 5-7% of
variation is accounted between 52 populations of 5 principal continental groups, while residual 93-95% variation is confined to between individuals. The Fst value of between populations genetic variation varies for autosomal loci (Fst-0.10-0.14), mt-DNA (Fst-0.24-0.27) and Y-chromosome (Fst-0.23-0.55) (Tishkoff and Kidd 2004). This is probably due to smaller population size and more susceptibility to genetic drift in case of lineage based markers. Furthermore, Fst value also differs with the type of polymorphism studied, from 3–5% base on STR loci (Rosenberg et al, 2002) to 14% on SNPs (Kidd et al, 2004). Over all variance calculated for within and between populations for protein and DNA loci is shown in

<table>
<thead>
<tr>
<th>Polymorphism</th>
<th>Loci (n)</th>
<th>Group (n)</th>
<th>Average variance component (%)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Within sample</td>
<td>Among samples</td>
<td>Among groups</td>
</tr>
<tr>
<td>Protein</td>
<td>17-25</td>
<td>3-7</td>
<td></td>
<td>83.8-87</td>
<td>2.8-8.3</td>
<td>6.3-11.2</td>
</tr>
<tr>
<td>DNA and protein</td>
<td>109</td>
<td>4-5</td>
<td></td>
<td>84.4</td>
<td>4.7</td>
<td>10.8</td>
</tr>
<tr>
<td>DNA</td>
<td>377</td>
<td>5-7</td>
<td></td>
<td>93.2-94.1</td>
<td>2.4-2.5</td>
<td>3.6-4.3</td>
</tr>
</tbody>
</table>

(iii) A model of ‘genetic variation’

The genomic diversity is observed in autosomal and X chromosomal loci (80-90% observed heterozygosity on autosomal STR loci) among African populations (Calafell et al, 1996; Kidd et al, 2001; Tishkoff et al, 2000). The haplotype genealogy of NRY and mt-DNA also supports accumulation of large number of mutation in the branches of African haplogroups like “A”, “B” and “C” in NRY binary tree (Underhill et al, 2000 and 2001) and “L1”, “L2” and “L3” of mt DNA haplogroup genealogy (Quintana Merci et al, 1999; Cann et al, 1987). Furthermore, all these molecular studies have revealed that Africans have largest numbers of
found in Africans (Calafell et al., 1996; Underhill et al., 2000). Only exception to this pattern is the Indian populations, which are reported to harbor more genetic diversity than any contemporary population other than Africans (Roychudhary, 2000; Khan et al., 2003; 2004 and Agrawal and Khan 2005). Diversity analysis based on 94 dinucleotide STR loci in 14 global populations revealed a clinal pattern of heterozygosity and number of private alleles (Tishkoff and Kidd 2004) as shown in Figure 1.3a.

Figure 1.3: Examples of highest diversity among African populations

(a) Number of private alleles and average observed heterozygosity at 94 dinucleotide STR loci in African and non-African populations; (b) The average LD for 83 SNPs across 21 haplotypes for 32 populations

A haplotype is the variation along a chromosome, is also responsible for the differences in the population. Some regions show strong LD while some are hotspots for recombination. Level and pattern of LD also depends on numerous factors like selection, recombination, and rate of mutation as well as genome wide demographic effects like population size, structure, founder effect and admixture. Various studies based on LD between SNPs have shown that Eurasian have higher LD levels than Africans (Gabriel et al., 2002, Kidd et al., 2001) as shown in Figure 1.3b. This is because Africans being the most ancestral population have maintained a larger effective population size and have had more time for mutation and recombination to
effective population size and have had more time for mutation and recombination to spoil the LD. The bottleneck associated with the expansion of modern humans out of Africa in loss of many haplotypes causing an increased LD in non-African populations.

(iv) Genetic lineage

Different genetic markers which is based on lineage or to some extent of autosomal loci (~60-150 STR) is reported to successfully assign the genetic ancestry to a sub-population (Rosenberg et al, 2002; Jorde et al, 2000; Underhill et al, 2001). NRY binary tree offer best method to assign a genetic ancestry to a sub-population or even to an individual because of the availability of an exquisite and well defined order of mutational events that have determined 18 haplogroup clusters (A-Q) owing to ~ 400 UEPs (Jobling and Tyler Smith 2003). When a sub-population group or an individual fall in one of these 18 clusters, it becomes easier to assign his ancestry. However, the tasks compounded in a population like that of Indian population which on one hand experienced heavy gene flow from numerous parts of the world (Balakrishnan 1978).

All prior information about ethnicity, geographical location and language etc. and analyzed solely on the basis of their genetic profile. If they still cluster with the groups of their demographic and linguistic profile, then the power of the marker is reliable. Such instances are possible in NRY or mtDNA markers to a large extent, but only maternal or paternal lineages do not represent the entire genetic makeup of an individual. For example, we might have 75% of western European genetic ancestry in an individual if 3 of his grand parents were western European still, he can have an African NRY lineage if the 4th grand parent (paternal grandfather) was an African. Therefore, sufficient number of autosomal loci is also equally important to assess the genetic ancestry of a sub-population or an individual along with the lineage based markers. This also describes the ardent need of identifying more number of ancestry markers and defining the effect of different evolutionary forces on the genetic structure of a population.
(v) Genetic sub-structuring

Autosomal lineage based DNA loci and X chromosomal studies have revealed the fact that genetic variations are parallel to the geographic separation. Members of same local groups are more closely related to each other than to members who live in distant geographical area. When Fst based distances are used to cluster the populations in trees (Figure 1.1), it was clearly evident that (i) Africans are most divergent populations; (ii) there is large genetic distance between Africans and non-African groups suggesting a population bottleneck and considerable genetic drift associated with initial expansion out of Africa (Jorde and Wooding 2004). Still, however, these genetic clusters mainly refer to geographically far locations and non-admixed populations. Therefore, labeling them only with the ethnicity will be a faux classification. The genetic sub-structuring is most visible in the NRY and mtDNA genealogies, where even considerable structuring within a geographical region can be done like exclusive presence of NRY haplogroups “A” and “B” in Africans, “O” in Chinese and “Q” in Americans (YCC, 2002, Underhill et al, 2000). As mentioned earlier that even the studies carried out on larger number of autosomal loci like Bamshad et al, 2003 or Rosenberg et al, 2002 are successful in inferring the genetic structuring in varied population groups.

(vi) Biomedical Implications of Genetic Variation.

Differential distribution of normal genetic variation and also that of genetic variation affecting diseases are the result of mutation, migration, random genetic drift and selection (Tishkoff and Kidd 2004). Therefore, understanding the structure of neutral human genetic variation provides an insight about the allelic structure of health related genetic variation. ‘Common Disease- Common Variant’ (CD/CV) hypothesis, which states that the common genetic diseases are affected by common disease susceptibility alleles (or variants) at a few loci that exist at high frequency across ethnically diverse populations was cited by the critics of variations studies (Chakravarti 2001; Reich 2001). These alleles probably arose before population differentiation and are common across populations. Supporters of CD/CV hypothesis cites the reason of being monogenic disorder to discard the distinctive examples of parallel presence of high frequency of HbS allele and sickle-cell anemia among sub-
Saharan Africans and Mediterranean populations or that of C28Y-HFE allele and hematochromatosis; and that of V508-CFTR alleles and cystic fibrosis among northern Europeans.

Examples of 235T variant of angiotensin (AGT) gene that codes a key component of rennin-angiotensin, blood pressure regulatory pathway supports the CD-CV hypothesis. This variation is present across all human groups, as high as 90% among Africans and as low as 30% among Europeans (Nakagima et al, 2004). The allele is associated with 20-30% increase risk of developing hypertension. Similarly, frequency of the null allele of CYP2D6 gene varies from 6% in Asians to 7% in Africans and upto 30% in Europeans (Bradford et al, 2003). CYP2D6 encodes a member of cytochrome P450 family involved in metabolism of important drugs (Weinshilboum 2003), and its null allele render the gene product inactive to an extent that homozygous null allele individuals experience little or no analgesic effect. Therefore it has been quoted that although substantial genetic variation is there but it is present in all populations.

A couple of years back, Bamshad et al, 2004 has shown that a sequencing based mega-analysis of 63, 724 SNPs in the coding and regulatory regions of 3931 human genes reveals that large number of private alleles are present in different population groups. Total 50,736 SNPs were found polymorphic in that study and when 10% frequency was considered as ‘common’, then, total 48% were found common in African-Americans and Europeans, out of which 41% differ significantly, between the two populations. Interestingly, 23.1% (~ 4,704SNPs) were found private in Africans and 2.9% (~ 585) were private in Europeans. This clearly indicates that CD/CV hypothesis was supported because only commonly occurring SNPs are studies but all the complex or multifactorial diseases involves a combined effect of numerous genes, most of which have not been identified so far. Further, even if the variants are shared between the groups, their frequencies differ substantially.

The main goal of biomedical research related to genetic variations are incorporation of population genetic structuring in study designs of association studies and information about individual genetic ancestry to improve medical
diagnosis and treatment on individualized basis. However, at present, both of these aspects are resolve by 'proxy' of self reported ancestry, geographical location, ethnic identity, language, community or religion or caste. However, for most biomedical research applications, there are little empirical data that compares the reliability self-reported ancestry or bio-geographical ancestry with ancestry inferences from explicit genetic data.

Undetected population stratification in case-control studies could lead to false positive associations (Pritchard et al, 1999, Tishkoff and Verrelli 2003). Genetic knowledge of population sub-structuring and stratification is an essential requirement for proper selection of controls and for identifying disease pre-disposing alleles that may differ across ethnic groups like high occurrence of Tay Sachs disease, Torsion Dystonia, Breast cancer and Gaucher disease is high among Jewish populations and not among Arabs, probably owing a result of founder effect or inbreeding despite of belonging to same geographical area. Similarly, absence of factor V leiden mutation (A2086G) and prothrombin gene mutation (C10965T) among north Indians despite of frequent occurrence of thrombotic event and various reports of possible Caucasian genetic ancestry (Agrawal et al, 2004).

Like wise information for medical diagnosis and treatment can be obtained by the information collected from individual ancestry. Such an approach holds importance because even if CD/CV hypothesis is correct then also differential effects of risk allele in people with different genetic ancestry have been reported. An allele of APOE- APOE4, frequent in Africans, Asians and Europeans is associated with Alzheimer's disease. However, homozygous APOE4 Asian individuals have -5 fold higher risk of developing this disease than homozygous Africans (Farrer et al, 1997). Several polymorphisms in the 5' cis-regulatory region of CCR5 influence the progression of AIDS and even death in HIV patients (Martin et al, 1998, Gonzalez et al, 1999). However, one CCR5 haplotype (HHE) is associated with delayed progression of AIDS in European-Americans but with faster progression African-Americans (Bamshad et al, 2003). Similarly, three important variants of CARD15 or NOD2- R702W, G908R and 1007fs have been associated with an inflammatory bowel disorder- Crohn's disease in European-Americans but not in Europeans or
Asians. Therefore even if the same risk allele for a complex trait is present in different group, it might be associated with different outcomes. The knowledge of the genetic ancestry of an individual or a sub-group, and information of population sub-structuring and stratification is key in biomedical research and it is high time to replace the proxy tags of geography, ethnicity, race or caste by an accurate genetic profile. However, the question likes in the reliability of the genetic markers to infer the correct ancestries and the level of resolution required to infer genetic tags i.e. differentiation between European and African genetic ancestry is sufficient or it should be resolved till west African and east African or north Indian and south Indian or more resolutely an upper caste or a lower caste Indian population.

To overcome these aspects, more information about the genetic ancestry and genetic structure of contemporary populations is required, more knowledge about of various evolutionary and socio-cultural factors that has shaped the present day genetic diversity is required, more data on admixed population is desired and more identification of population specific private marker, private alleles or haplotypes is required and that is what the genetic variation studies are doing worldwide.

1.6 Genetic structure of Indian populations

1/5th of the total world population lives in Indian sub continent and that is more than 1 billion individuals. The genetic structure, affinity and diversity of these 1 billion Indians is often contested and postulated of holding an important key about numerous unanswered questions concerning the evolution of modern human and a range of factors that shape the contemporary pattern of genetic variation (Bamshad et al, 2004).

Genetic researchers still have many areas to work on to explore the gene pool of Indian populations; one of them is its crucial geographical location that lies on the postulated southern coastal route followed by the anatomically modern H. sapiens out of Africa (Cann 2001; Lahr and Foley 1998). Secondly, the impending role of extensive gene flow through a series of migrations and invasions, that have shaped and distributed the contemporary genetic variation across different geographical locations of India and have created enormous amount of genetic diversity. Finally, a
unique social framework fabricated with the threads of numerous religions, communities, castes and linguistic groups (Renfrew, 1989) that have resulted into extensive population structuring and created numerous endogamous groups. These features have immensely affected the genetic composition of the contemporary Indian population. Analysis and interpretation of the genetic ancestry or origin and the current pattern of genetic variation among different endogamous groups will not only help in gaining an insight into past human movements but also provide an ideal experimental chunk to get an answer of the five imperative questions of the human genetic variation studies. However, despite of such a rich gene pool, Indian populations are relatively unexplored or under represented in the studies of human genetic variation.

1.7 Important geographical position

There are several pathways proposed for the migration of modern human out of Africa (Harpending et al, 1998; Jorde et al, 1998; Cavalli-Sforza and Feldman, 2003; Jobling and Tyler-Smith, 2003; Underhill et al, 2001; Hammer et al, 1997, 1998; Kivisild et al, 1999). In the history of human evolution India was the central place for the dispersal of modern human (Cann 2001). The early waves of migration of modern humans from out of Africa, through West Asia, were into India. According to the ‘Weak garden of Eden hypotheses, India acted as an incubator of early human populations. Some geneticists have also suggested that another route of migration from India has populated western Asia and Europe (Palanichamy et al, 2004; Kivisild et al, 1999, 2003).

Modern human have dispersed in the various parts of Indian sub continent, but conform data is still a controversy it is supposed Paleolithic period (50-20 Kya) (Misra 2001). However, the signs of upper and middle Paleolithic period are scanty, but the recent studies based on the tribal and caste population groups of India have provided various clues about the persistence of genetic signatures of these erstwhile human settlers among existing population (Kivisild et al, 1999, 2003; Quintana-Murci et al, 2001). Haplogroup genealogies based on NRY and mt-DNA have revealed that population and mutation splits have its deep routes in Indian genealogy.
Mitochondrial studies have revealed the fact that less than 10% of the maternal lineage of India has migratory ancestor in past 12000 years (Kivisild et al, 1999). More than 60% of Indians have their maternal roots in Indian-specific branches of haplogroup M, which is virtually absent in western Eurasians. Furthermore, the M haplogroup found in East Asia is suggested to have originated from Indian “M” haplogroup. This haplogroup might have reached India from East Africa, along the southern route, during earliest migration wave of anatomically modern humans ~60 Kya (Kivisild et al, 1999, 2003; Quintana-Murci et al, 1999). However, such strong signs are not visible in NRY genealogy, barring the presence of high frequency of exclusively Indian “H” haplogroup among both caste and tribes, presence of “D” among 27 Andaman tribes (Thangaraj et al, 2003) and recent reports of “C2” in south Indian tribes and castes (Kivisild et al, 2003). However, the subsequent migrations and gene flow from Indo European speakers of central Asia and Western Eurasia have almost diminished the paternal signals of these earliest settlers. The accurate information of genetic structure and ancestry of Indian population groups can provide the ‘missing links’ of standard model of evolution.

1.8 Area of Gene flow

Neolithic migration has occurred at least twice in Indian subcontinent which is the main source of massive gene flow in the subcontinent. A part of an eastward wave of human migration about 10-15 KYA, when agriculture developed in the Fertile Crescent region, (Cavalli-Sforza et al, 1994; Renfrew 1989) entered India. This wave brought Dravidian languages into India (Renfrew, 1989) mainly, Elamo-Dravidian languages (Ruhlen 1991), which may have originated in the Elam province (Zagros Mountains, southwestern Iran) and are now confined to southeastern India and to some isolated groups in Pakistan and northern India. A later episode, the arrival of pastoral nomads from the central Asian steppes to the Iranian plateau, ~4,000 YBP, brought with it the Indo-European language family, which eventually replaced Dravidian languages from most of Pakistan and northern India, perhaps by an elite dominance process (Renfrew 1989; Quintana-Murci et al, 2001). These nomadic migrants may have consolidated their power by admixing
Introduction

with native Dravidic-speaking (e.g., Telugu) proto-Asian populations (Cavalli-Sforza et al, 1994), and subsequently established the Hindu caste hierarchy to legitimize and maintain this power (Poliakov 1974; Cavalli-Sforza et al, 1994). These Indo-Aryan speakers constitute nearly 75% of present day Indian populations and have been responsible for forcible or freely southern retreat of Dravidians.

Indian population is an admixture of western Eurasian (Bamshad et al, 2001, 2003; Khan et al, 2004) and central Asian populations is also affected by the Indo-Aryan speakers (Majumder, 2001; Basu et al, 2003). Apart from these two major events, there are numerous large movements of people with diverse genetic backgrounds into India. The period from 520 B.C. to 300 A.D., was marked by successive invasions of Persians, Greeks, and Scythians, and was followed until 1100 A.D. by invasions of central Asian nomadic tribals (Huns), Arabs, and Turks (Majumder, 2001). The effect of such an extensive and diverse gene flow was evident in the study of Bamshad et al, 2004 based on 60 STR loci and 100 alu insertions. The study has shown that when geographically far located sub-Saharan Africa, Europe and East Asia were structured in a triangular plot, they occupy three corners of triangle, but when same analysis was done along with 8 south Indian populations, they got spread around the edges between Europeans and East Asians. The analysis of Indian genetic structure will eventually lead us to understand that different demographic movements have created ‘how much’ effect on the genetic diversity of human populations and will surely provide the clue about the genetic sources of Indian population groups.

1.9 Place of huge diversity

The world has created immense ethnic, geographical, morphological, linguistic and religious diversity among contemporary human groups of India. Four different morphological groups, namely Negrito (Andaman Island, Nilgiri hills of Tamil Nadu), Atrwaloid (Central and southern region), mongoloids (sub-Himalayan and north eastern region) and Caucasian (spreaded all over India especially north and north-western regions) are found in India (Majumder 1998). Linguistically, there are three most frequent language families, Elamo-Dravidian is mostly spoken in
southern India, Indo-European language is spoken almost all over India and Austro-Asiatic mainly spoken by tribes of India (Misra 2001). The first two families are structured into 18 official languages and ~750 dialects and Sino-Tibetan speakers found exclusively in northeastern part of the country (Cordaux et al, 2004a).

1.10 Advantages of population studies

This has been noted that although substantial genetic variation is present in all the populations. On the contrary, the neutral genetic variation aids biomedical research in at least three ways:

(a) Genetic profiling can give a molecular sub classification of any genetic disease. High frequency of hemoglobin HbS allele, variant of glucose-6-phosphate dehydrogenase and sickle-cell anemia are found among sub-Saharan Africans (Tishkoff et al, 2001; Luzzato and Mehta, 1995). However, the same disease with the underlying mutations is also found in Hispanics and inhabitants of northwestern India (Braun, 2002) and central Greece (Braun 2002; Kevles 1995). Therefore, labeling the disease only on the basis of ethnic affiliation or phenotypic occurrence can be wrong interpretation and could possess serious health consequences. The concept of genetic ancestry is a much better indicator than race or ethnicity to determine that whether one carries the marker of a genetic disease. It has also been reported that differential effects of risk allele is seen in people with different genetic ancestry like homozygous APOE4 Asian individuals have ~5 fold higher risk of developing Alzheimer’s disease than homozygous Africans (Farrer et al, 1997).

(b) Genetic ancestry of an individual can be very useful to improve medical diagnosis and treatment. The genetic differences among ethnic groups often cause differences in drug responses. The null allele of CYP2D6, a drug-metabolizing enzyme (DME) that encodes a member of cytochrome P450 family (Weinshilboum 2003) renders the gene product inactive to an extent that homozygous null allele individuals experience little or no analgesic effect. This null allele occurs in a frequency of 10% among north European ancestry and therefore they do not experience an analgesic effect from the prodrug codeine
(Bradford 2002). On the contrary, about 98% Arabs are able to transform codeine into the active form morphine (Mclellan 1997).

(c) In study designs of association studies, undetected population stratification in case-control studies could lead to false positive associations (Pritchard et al, 1999) therefore, genetic knowledge of population sub-structuring and stratification is an essential requirement for proper selection of controls and for identifying disease pre-disposing alleles that may differ across ethnic groups.

The key of biomedical research is the study of genetic ancestry or origin of population sub-group, and information of population diversity, sub-structuring, stratification and phylogenetic relationship. Worldwide efforts are going on to replace the proxy tags of geography, ethnicity, race or caste by an accurate genetic profile. However, the major hurdle is in the reliability of genetic markers to infer the correct genetic origin of a sub population and the level till which a marker can resolve the genetic-sub-structuring i.e. it can differentiate between an Indian and an African or it can even resolve a north Indian from a south Indian. Such genetic milestones require information about the genetic ancestry and presence of genetic structuring in contemporary populations. More knowledge about of various evolutionary and socio-cultural factors that have shaped the present day genetic diversity is required, more data on admixed population is desired and more identification of population specific private marker, private alleles or haplotypes are required.

More vision, planning and better approaches are required to conduct such studies in Indian sub-continent due to its complex structure as illustrated above. Each study should be able to raise a specific question and selection of populations and markers should comply with the need of a particular study, as there are several aspect of Indian gene pool which has albeit got detected but not completely understood like: (a) The studies on the Tribal and Dravidian populations based on markers additional to the one used in Caucasian populations are required to reveal the exact composition of Indian gene pool specially to know about the pre-Indo Aryan admixture.
(b) The geographical and cultural groups should be chosen to infer their role on Indian gene pool like north east population are best suited for studying East Asian gene flow while east Indian populations are most suitable for studying the effect of Austro-Asiatic and Dravidian migrations on Indian gene pool.

(c) Similarly, studies on more endogamous groups can reveal the structuring of tribes in genetic context. Moreover as described tribal groups of India observe strict endogamy hence it is important to analyze the effect of endogamy on genetic structuring of Indian populations.

As a result studies on pattern and distribution of genetic variation in Indian population will aid in assessing the level of genetic sub-structuring and correct genetic ancestry in different endogamous groups. Furthermore, such studies along with offer an unsullied elucidation on human genetic diversity of Indian population which will help in tracing the missing block of ancestral human settlers that will form the connecting link of standard model of human evolution.

1.11 Objective of the present study

Since the studies of human genetic variation have shifted from “pre DNA era” to the “DNA era”, we have also made an attempt to determine the genetic variation among East Indian tribal populations and to investigate the possible sources of this genetic variation.

Present study has been exploring the genetic composition of four East Indian tribal populations and there by offering a fresh interpretation on the pattern, distribution and structuring of genetic variation existing in East Indian tribal populations.

All the four populations selected for the present study are ethnically Austroloid and linguistically belong to Austro-Asiatic, Dravidian and Indo-European clad of languages.

The study was designed in such a manner that the information obtained can be useful in some imperative facts about the origin and diversity of east Indians. For this fact a total of 32 unique event polymorphisms (UEPs) and 4 STR loci located on
non-recombinating region of Y-chromosome (NRY) were studied to delineate the paternal lineages.

(a) Determination of the paternal genetic lineages of these populations allowed us to detect their possible ancestries and to trace their origin.

(b) We further tried to quantify and apportioned the genetic variation existing in these populations into within and between population categories.

(c) Furthermore, we have made an attempt to assess that whether these populations are genetically differentiated or not, if not then does the reason lies in their common origin or in recent admixing.

(d) Lastly, the genetic data generated from the study have been systematically compared with that of other Indian tribes and global populations to determine

   (i) The genetic legacy of Austro-Asiatic on Jharkhand tribal gene pool

   (ii) To determine whether genetic variation in Jharkhand runs parallel with geography or ethnicity or languages.

The major objectives that we have tried to achieve in this study include:

1. To analyze the amount, pattern and distribution of genetic variation among four endogamous tribal groups of Jharkhand state using unique event polymorphism (UEPs) of non-recombinating region of Y chromosome.

2. To study the genetic diversity and paternal lineages of four endogamous tribal groups of Jharkhand state using Y-chromosome specific STR’s markers.

3. To analyze the polymorphic markers which can be used as signature of specific geographical region or ethnicity like African specific YAP insertion at Y-chromosome.