Introduction

Introduction

Human genetics research has generated enormous amount of data about the genetic differences among individuals and groups. Investigation of these differences has transformed our understanding of the origins and nature of human diseases (**Cavalli-Sforza, 2005; Bamshad *et al.*, 2004, Collins *et al.*, 2003**). Since long geneticists are involved in the Human genetic variation studies among the individuals forming a species, but the remarkable extent of this variation was not appreciated until about 25 years ago (**Cavalli-Sforza and Feldman, 2003; Jorde and Wooding, 2004**). Conspicuous human traits like hair and eye color clearly vary from one individual to the other in many populations; these differences are easily perceived by the layman, as are variation in height, weight, body build, and facial traits, which are also genetically determined to some extent. Their hereditary transmission, however, is complex, and these traits contribute little to understand the extent of variation. The first example of clear-cut genetic variation is that of ABO blood groups which was described at the beginning of the century (**Landsteiner, 1927**). Dissimilarities between individuals regarding ABO blood-group variation are due to small chemical differences between molecules found at the surface of red blood cells.

These studies were soon extended to other blood-group systems, and a body of data began to accumulate showing that different human populations have different proportions of blood groups. However, the first glimpse of the staggering magnitude of genetic variation came later beginning in the 1950s and coming to full development in the 1960s when individual differences for proteins could be systematically studied. A protein is a large molecule made of a linear sequence of components called amino acids; different proteins vary considerably in their amino-acid composition and serve very different functions. The relationship between structure and function has been demonstrated for many proteins. The same protein

may show small, strictly inherited differences between individuals. The first example was observed in the protein hemoglobin, in which the replacement of a specific amino acid by another was shown to determine a hereditary disease known as sickle-cell anemia. This first case of molecular pathology was detected by subjecting the protein to an electric field with a procedure called electrophoresis. The amino-acid replacement involved in sickle-cell anemia causes a change in the electric charge of the hemoglobin molecule, which allows the separation of normal and sickle-cell hemoglobin's. Electrophoretic analysis has since been further developed and has helped detect a great deal of variation in proteins. It is now known that the majority of the tens of thousands of different proteins found in an organism exist in more than one form, so that some individuals may have one form of the protein, whereas others may have another form (**Glazko G,2005**).

Protein variation is still the tip of the iceberg. Only when the analysis could be carried out at the level of the hereditary material itself, deoxyribonucleic acid (DNA), could the full extent of individual genetic variation begin to emerge. This technique became widely available only in the 1980s, and although comparisons of segments of DNA in different individuals are still rare, they are becoming more common. They are, however, adequate to convince us that there is much more variation at the DNA level than was suspected when only proteins and blood groups could be analyzed.

Population diversity and Evolutionary Process

Anthropologists tried to reconstruct evolutionary relationships and history on the basis of a single character or gene since long. A favorite for over 100 years was the cephalic index (the percentage of skull breadth to length) introduced shortly before the middle of the last century. However, with a single trait, two populations of different origin could well turn out to be more or less identical. Anthropometric traits of this kind also have another very serious drawback: there is no guarantee that the character is completely under the control of biological inheritance and the variations observed could be due to short-term response to environmental changes. Every gene frequency varies over time in ways that can be considered, at least superficially,

nearly random. Therefore, it is not surprising that populations having clearly different evolutionary histories may show similar gene frequencies (**Itoh T, 2002**). This drawback can be avoided if one cumulates the information from more than one gene. As one increase the number of genes considered simultaneously, the probability that a similar confusion takes place becomes more and more remote. In 1963 it was shown that even with as few as 20 alleles from five genes one could successfully attempt a reconstruction of human evolution. Further larger number of alleles were discovered due to which the picture become clearer (**Jorde and Wooding, 2004; Castro de Guerra *et al.*, 1999**).

Several methods allow us to combine the information from many genes into appropriate statistical indices. They are usually called multivariate to distinguish them from those using single traits or genes (univariate).

Multivariate analysis is especially useful for understanding evolutionary forces that tend to operate in a parallel fashion on all genes: migration and random genetic drift (the random fluctuation of gene frequencies in time, to be further explained later). These and other methods are applied to the existing data with the aim of extracting information of genetic and evolutionary interest. The reconstruction of human evolution, including the fissions, the major migrations, and the understanding of the roles of mutation, drift, and natural selection is often difficult and challenging. Results from genetic data should be compared with relevant knowledge from other fields, in particular, paleo- anthropology, prehistory, history, the geographic and ecological setting, and the cultural evidence that comes indirectly from linguistic studies.

If we know that there exist different genetic types of a specific protein or other strictly inherited character, we can count individuals carrying one type or the other and establish the proportions of that type in the population being examined. These proportions vary from one population to another because they change over time in each population in a relatively unpredictable manner (**Nei 2000**). The change in proportions of these types over time is the evolutionary process itself. It proceeds slowly but incessantly over generations. The analysis of populations living today in

different places gives us a cross section in time of this continuing process, which is inevitably diverse in the various parts of the inhabited Earth.

So the primary interest is in understanding this evolutionary process. The first task is to describe the existing variation, using a variety of techniques to test the relevant evolutionary models. Initially the interest was restricted to aboriginal populations, which were already living in the area of study in A.D. 1492. After this time, geographic discoveries stimulated the expansion and migrations of the economically more advanced populations all over the planet. Some movement took place before A.D. 1492, but at a smaller scale. Ordinarily, populations that migrated after that date have mixed only partially with earlier residents and were easily recognizable on the basis of physical appearance and historical and social knowledge **(Underhill 2004; Carmelli and Cavalli-Sforza,1979).**

A population is a local group of organisms of the same species that normally interbreed. The word population does not refer to an entire species; it refers instead to a group of organisms of the same species living within a sufficiently restricted geographical area where members can potentially mate with each other provided they are of opposite sex. A group of individuals within whom marriages are contracted is called a Mendelian population. The set of genetic information carried by a population is known as its gene pool. The gene pool of a new generation is descended from the parental generation but for several reasons, including chance, the gene pool of the new generation may have different allele frequencies than the parental pool. Overtime, the changes in allele frequency can cause changes in phenotypic frequency. The long-term effect of changes in allele frequency is evolutionary change. Members of the same local group are more closely related to each other than members of groups who live in different geographical areas and people who live in the same geographical region are more similar than those separated by geographical barriers. Populations grow and interact with one another through competition and predation. These factors can influence behavior ecology and evolution, either at micro or macro level. The evolution of such population results in to the change in the genetic constitution of the population followed by alterations in

the genotypic and phenotypic constitutions (**Joaquim Fort 2004; Castro de Guerra *et al.*, 1999**).

To achieve the above goals the important step in the understanding of the genetic structure of human populations consists of estimating frequencies of alleles at different loci. Gene frequencies 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. After the selection of the markers it is important to use more robust statistical tools. In any population, the genotype frequencies among zygotes are determined in large part by the patterns in which genotypes of the previous generation come together to form mating pairs. In random mating, genotypes form mating pairs in the proportions expected from random collisions. For a gene with two alleles A and a in a random-mating population, the expected genotype frequencies of AA, Aa, and aa are given by p2, 2pq, and q2, respectively, where p and q are the allele frequencies of A and a respectively, with p + q=1. The expected genotype frequencies with random mating constitute the Hardy-Weinberg equilibrium (HWE). The rate at which the HWE frequencies are attained depends on the life history of the organism. In an organism with non overlapping generations, such as an annual plant, each generation is separated in time from the preceding and the following generation; in this case, the Hardy-Weinberg frequencies are attained in one generation of random mating provided that the allele frequencies are equal in both the sexes. In an organism with non overlapping generations, the approach to HWE is gradual. Statistical tests of HWE are often based on the χ^2 test, but this test is relatively weak in detecting departures from the expected frequencies, especially those caused by admixture of subpopulations differing in allele frequency.

One of the principal implications of the HWE is that the allele frequencies and the genotype frequencies remain constant from generation to generation, hence

genetic variation is maintained. Another major implication is that, when an allele is rare, the population contains many more heterozygote for the allele.

Nonrandom association between the alleles of different genes is measured by the linkage disequilibrium parameter D. Random association between alleles of different genes is called linkage equilibrium, and it is indicated by D, when D=0, the alleles are said to be in linkage disequilibrium. Ordinarily, unless there is some countervailing process that maintains linkage disequilibrium between two genes, D is expected to go to zero at a rate determined by the recombination fraction between the genes that recombine with a frequency r, D decreases by the fraction r in each generation. Significant linkage disequilibrium is usually found in natural populations for genes that are tightly linked, for genes that are within or near an inverted segment of chromosome. Significant linkage disequilibrium can also result from admixture of two or more subpopulations differing in allele frequencies. The gene frequencies can be changed and the factors that shape the contemporary genetic diversity include Mutations, Random genetic drift, Gene flow or genetic exchange, Natural selection **(Cavalli-Sforza,2003; Templeton,2002).**

Mutations

The geographic representation of the data of living population, a segment of DNA endowed with a specific function called gene is being studied extensively these days and for each gene, analysis of the alternative form of gene called alleles is carried out, and these alleles are a result of mutations occurring in a populations. Geographic maps of an allele are useful for understanding facts specific to that allele, including its evolutionary history and the effects of evolutionary factors like mutation and natural selection **(Roychoudhury, 1988).** The geographic distribution of a particular allele may give information on the place of origin of the genetic change (mutation) that generated it. Mutation results in the rise of both functional (i.e., coding) and nonfunctional (i.e., noncoding) polymorphisms Geographic maps of an allele are useful for understanding facts specific to that allele, including its evolutionary history and the effects of evolutionary factors like mutation and natural selection. The geographic distribution of a particular allele may give information on

the place of origin of the genetic change (mutation) that generated it. Correlations of the distributions of gene frequencies with environmental parameters at the geographic level have been instrumental in the discovery of specific genetic adaptations **(Nei 2005).**

It is well established that the proportion of an allele varies considerably from place to place, but usually there is little difference between neighboring populations so that the greatest variation is observed at large distances. It is thus possible to prepare geographic maps representing these proportions for a particular allele (also called allele frequencies, or simply, gene frequencies) when a sufficient number of populations have been tested.

Allele and genotype frequencies change in unexpected ways in successive generations with dispersive processes such as reduced population size and migration. Conversely, the direction of alleles and genotypes changes over time due to population size, birth rate, disease, migration and adoption to environmental factors. The sickle-cell anemia gene was the first example, because its geographic distribution showed a correlation with that of malaria **(Haldane,1954).** Angiotensin gene (AGT), which, encodes important component of rennin-angiotensin blood pressure regulation pathway. In many populations it had been found that 235T a variant of AGT is associated with 10-20% increase in the risk of developing hypertension **(Kunz & Kreutz *et al.*, 1997)**. This is found in 90% frequency in some African populations and 30% in European populations **(Nakajima T *et al.*, 2004).**

Random genetic drift

Sometimes populations may change in frequency over time as a result of the accumulation of random sampling error in the passing on of alleles from generation to generation. When a very small number of individuals migrate and start a new population, the sampling error (random genetic drift) is very large, and allele frequencies in the new population may be very different from those in the parent population. Allele frequencies fluctuate from one generation to the next because of the randomness of the transmission process until, by chance, one of the alleles is lost and the other is fixed. As the new population grows over a few

generations, the magnitude of the sampling error per generation decreases and the new population will continue to have very different frequencies from the parent population. This extreme form of random genetic drift is referred to as a "founder effect," **(Strachan *et al.*, 1996)** because the population expanded from very few founders with a relatively restricted gene pool. For example, available evidence suggests that a small group of individuals left Africa and over time allele frequencies changed markedly from those in the African populations left behind **(Nakajima *et al.*, 2004)**.

Gene flow or genetic exchange

Gene flow is a process by which interbreeding among certain groups of individuals results in those populations becoming increasingly similar to each other. Two populations that start off quite different genetically, if they mate, can produce offspring that represent the genes present in both of the original populations. Matting can be either positive assortative mating or negative assortatve mating positive assortative mating increases homozygosity in the population because of mating between individuals with identical genotypes, while if mating is carried out between individuals with different genotypes then it is called negative assortatve mating and this will result in high heterozygosity in a population **(Cavalli-Sforza *et al.*, 1999)**.

Natural selection

By natural selection organisms with gene patterns that are adaptive to a given environment become more prevalent over time. For example, organisms that can adapt to changing climatic patterns are at an advantage over those that adapt only with great difficulty. Natural selection is a constantly shifting process. It is influenced not only by an organism's biology, but also by the interaction of that biology with environmental conditions **(Sternberg,2004)**.

Variation in Heritability within a Given Population

Heritability is not a fixed value for a given attribute **(Herrnstein & Murray, 1994)**. Heritability depends on many factors, but the most important one is the range of environments. Because heritability represents a proportion of variation, its value will

depend on the amount of variation. Herrnstein in 1973 pointed out, if there were no variation in environments, heritability would be perfect, because there would be no other source of variation. If there is wide variation in environments, however, heritability is likely to decrease. When one speaks of heritability, one needs to remember that genes always operate within environment contexts. All genetic effects occur within a reaction range such that, inevitably, environment will have differential effects on the same genetic structure. The reaction range is the range of phenotypes (observable effects of genes) that a given genotype for any particular attribute can produce, given the interaction of environment with that genotype. For example, genotype sets a reaction range for the possible heights a person can attain, but childhood nutrition, diseases, and many other factors affect the adult height realized. Moreover, if different genotypes react differently to environmental variation, heritability will show differences depending on the mean and variance in relevant environments **(Lewontin,1974)**. Thus, the statistic is not a fixed value.

Adaptation and Migration

In zones with more intense exposure to sunlight, darker skin puts individuals at an adaptive advantage. The melanin that acts as a pigmentation to produce darker skin better protects individuals against the damage that large amounts of ultraviolet radiation can cause to the skin, this radiation increases susceptibility to skin cancer, especially melanoma, a form of skin cancer that easily can become fatal. In zones with weaker exposure to sunlight, lighter skin is an advantage. People rely on sunlight to produce active vitamin D3 in the capillaries. Lighter skin allows greater bodily production of vitamin D3. Deficiencies in vitamin D3 can cause rickets in children and osteoporosis in adults. There is as yet no conclusive evidence of positive selection for light coloration. Instead, evidence to date may indicate that light pigmentation in climates distant from the equator represents a lessening of the selective factors that lead to dark pigmentation near the equator rather than to any particular factors leading to lighter pigmentation **(Harding *et al.*, 2000)**

It is probable that much of the variation seen among groups of humans indirectly resulted from the pattern of expansion and migrations accompanied by

random genetic drift. Over the years, frequencies of DNA variants changed only slightly in terms of total DNA composition but changed enough to produce differences, many of which we still do not fully understand.

Models of modern human evolution

There are three main theories for the evolution of modern humans.

Multiregional model

The multiregional model proposes that there was no single geographical origin for modern humans but that, after the radiation of *HOMO ERECTUS* from Africa into Europe and Asia ~800,000–1.8 million years before present (yr BP), there were independent transitions in regional populations from *H. erectus* to *Homo sapiens*. This model is supported primarily by the continuity of certain morphological traits in the fossil record (for example, the robust cheekbones observed in *H. erectus* fossils from Southeast Asia and in modern Australian aborigines), which indicate that modern populations evolved over very long periods of time in the regions where they are found today.

Simultaneous evolution from *H. erectus* to *H. sapiens* in dispersed populations could have been achieved through extensive gene flow between populations, requiring a large effective population size to sustain gene flow among geographically diverse populations. The RAO model proposes that all non-African populations descend from a *H. sapiens* ancestor that evolved in Africa 100,000–200,000 yr BP. This ancestor then spread throughout the world, replacing archaic *Homo* populations (for example, the Neanderthals). The fossil record supports this model, as the earliest modern human fossils were found in Africa and the Middle East, dating to 90,000–120,000 yr BP **(Figure 1.1)**.

Recent African origin (RAO) model

The RAO model predicts that all genetic lineages derive from a recent common African ancestor and that non-African populations should carry a subset of the genetic variation present in modern African populations.

Assimilation model

The assimilation (or hybridization) model proposes that gene flow between the early human populations was not equal over time and space. This model allows for some gene flow between modern humans that migrated from Africa and archaic populations (for example, the Neanderthals) outside Africa. So, the evolution of modern humans could have been due to a blending of modern characters derived from African populations with local characteristics in archaic Eurasian populations. This model predicts that the modern gene pool derives from variable contributions of genes from archaic African and non-African populations.

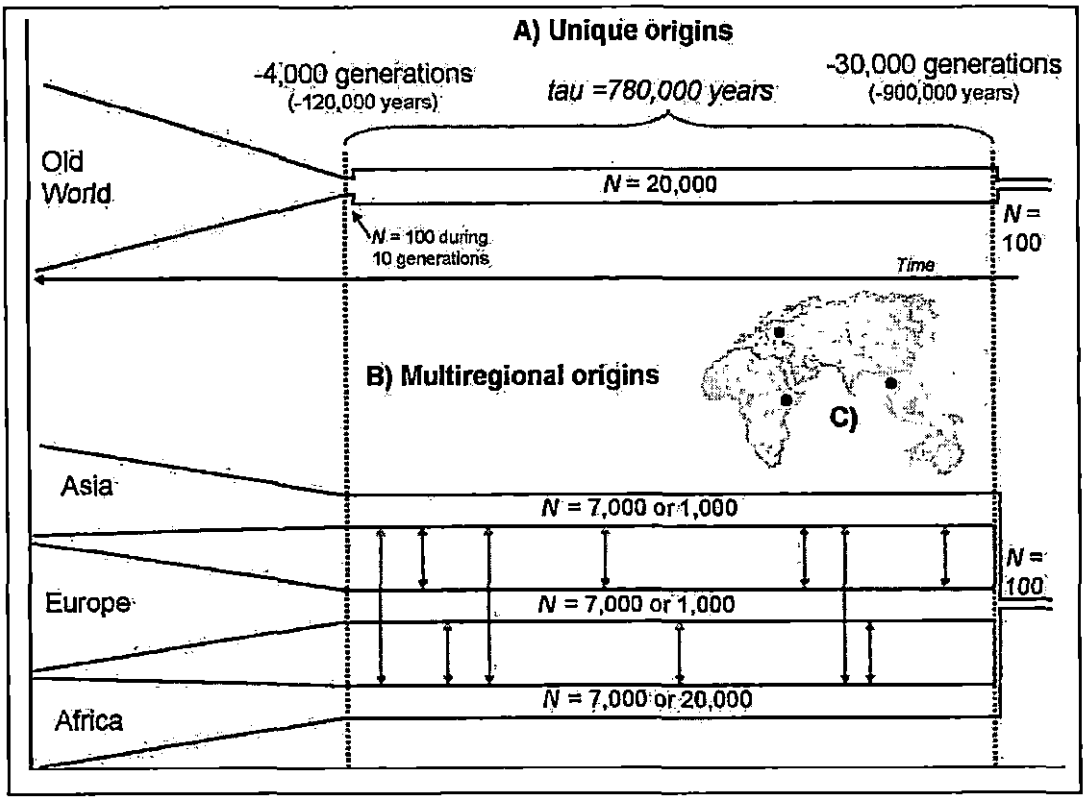

Figure 1.1 Demography and time line of the different simulated evolutionary models. (*A*) Unique origin (UO) model. In this model, 30,000 generations ago, a small population (N = 100 genes) went through a demographic expansion after a first speciation event. Then, 4000 generations ago, a range expansion followed a bottleneck of 10 generations to mimic a second speciation event. The large population preceding the speciation and range expansion can be considered to be a large subdivided population. (*B*) Multiregional evolution (ME) model. As in *A*, a small population went through a speciation event and instantaneously colonized the three continents 30,000 generations ago. For 26,000 generations the continents harbored relatively large populations and exchanged occasional migrants. Then, 4000 generations ago, three range expansions were initiated from the three different origins shown in C

The Origin of Modern Humans

Information about the history of our species comes from two main sources: the paleo-anthropological record and historical inferences based on current genetic differences observed in humans. The existing fossil evidence suggests that anatomically modern humans evolved in Africa, within the last approx 200,000 years, from a pre-existing population of humans (**Klein, 1999**). Although it is not easy to define anatomically modern human in a way that encompasses all living humans and excludes all archaic humans (**Lieberman *et al.*, 1992, 2003**), the physical characteristics of anatomical modernity include a high rounded skull, facial retraction, and a light, as opposed to heavy and robust, skeleton (**Lahr, 1996, 1998**). Early fossils with these characteristics have been found in eastern Africa and have been dated to approx 160,000– 200,000 years ago (**White *et al.*, 2003; McDougall *et al.*, 2005**). At that time, the population of anatomically modern humans appears to have been small and localized (**Harpending *et al.*, 1998**). Much larger populations of archaic humans lived some where in the Old World, including the Neanderthals in Europe and an earlier species of humans, *Homo erectus,* in Asia (**Swisher *et al.*, 1994**). Fossils of the earliest anatomically modern humans found outside Africa are from two sites in the Middle East and date to a period of relative global warmth, approx 100,000 years ago, though this region was rein habited by Neanderthals in later millennia as the climate in the northern hemisphere again cooled (**Lahr and Foley, 1998**).

Groups of anatomically modern humans appear to have moved outside Africa permanently sometime 160,000 years ago. One of the earliest modern skeletons found outside Africa is from Australia and has been dated to approx 42,000 years ago (**Bowler *et al.*, 2003**), although studies of environmental changes in Australia argue for the presence of modern humans in Australia 155,000 years ago (**Miller *et al.*, 1999**). To date, the earliest anatomically modern skeleton discovered from Europe comes from the Carpathian Mountains of Romania and is dated to 34,000–36,000 years ago (**Trinkaus *et al.*, 2003**). Existing data on human genetic variation support and extend conclusions based on the fossil evidence. African populations

exhibit greater genetic diversity than do populations in the rest of the world, implying that humans appeared first in Africa and later colonized Eurasia and the Americas (**Tishkoff & Williams 2002; Yu *et al.*, 2002; Tishkoff &Verrelli 2003**). The genetic variation seen outside Africa is generally a subset of the variation within Africa, a pattern that would be produced if the migrants from Africa were limited in number and carried just part of African genetic variability with them (**Cavalli-Sforza & Feldman 2003**). Patterns of genetic variation suggest an earlier population expansion in Africa followed by a subsequent expansion in non-African populations, and the dates calculated for the expansions generally coincide with the archaeological record (**Jorde *et al.*, 1998**). Studies of mtDNA (**Ingman *et al.*, 2000**), the Y chromosome (**Underhill *et al.*, 2000**), portions of the X chromosome (**Kaessmann *et al.*, 1999**), and many (though not all) autosomal regions (**Harpending & Rogers 2000**) support the "Out of Africa" account of human history, in which anatomically modern humans appeared first in eastern Africa and then migrated throughout Africa and into the rest of the world, with little or no interbreeding between modern humans and the archaic populations they gradually replaced (**Tishkoff *et al.*, 2000; Stringer 2002**). However, several groups of researchers suggests that humans bearing modern traits emerged several times from Africa, over an extended period, and mixed with archaic humans in various parts of the world (**ZieTkiewicz *et al.*, 1998**). As a result, they say, autosomal DNA from archaic human populations living outside Africa persists in modern populations, and modern populations in various parts of the world still bear some physical resemblance to the archaic populations that inhabited those regions (**Wolpoff *et al.*, 2001**). Studies of mtDNA from archaic and modern humans and Y-chromosomes suggest that any surviving genetic contributions of archaic humans outside Africa must be small, if they exist at all (**Krings *et al.*, 1997; Nordborg 1998; Takahata *et al.*, 2001; Serre *et al.*, 2004**). The observation that most genes studied to date coalesce in African populations points toward the importance of Africa as the source of most modern genetic variation, perhaps with some subdivision in the ancestral African population (**Satta and Takahata 2002**). In addition to having higher levels of genetic diversity, populations in Africa tend to have lower amounts of linkage

disequilibrium than do populations outside Africa, partly because of the larger size of human populations in Africa over the course of human history and partly because the number of modern humans who left Africa to colonize the rest of the world appears to have been relatively low (**Gabriel *et al.*, 2002**). In contrast, populations that have undergone dramatic size reductions or rapid expansions in the past and populations formed by the mixture of previously separate ancestral groups can have unusually high levels of linkage disequilibrium (**Nordborg & Tavare, 2002**). Sequence data for hundreds of loci from widely distributed worldwide populations eventually may clarify the population processes associated with the appearance of anatomically modern humans (**Wall, 2000**), as well as the amount of gene flow among modern humans since then.

Genetic tools for studying history

The advent of various techniques in molecular biology enabled the application of genetics to the study of human evolution giving rise to the fields of molecular evolution and molecular anthropology. In early applications, genetic or DNA markers were that of functional genes. These DNA markers were relatively less informative and were subjected to selection pressures. More recently, more informative DNA markers were discovered along with more robust DNA typing technology. These recent DNA markers are neutral, not being part of functional genes, are therefore not affected by selection pressures and are stable across populations and generations. These DNA markers are also highly polymorphic making them more informative in studying genetic variation between and amongst human populations.

Short Tandem Repeat (STR)

STR DNA markers (autosomal STRs and Y chromosome STRs) are short fragments of DNA that are commonly used in forensic human identification. However, because of high polymorphism, these markers are useful in studying human genetic diversity. Single Nucleotide Polymorphism (SNP) DNA markers (autosomal SNPs and Y chromosome SNPs) are characterized by single base

changes in DNA sequence. SNPs are highly stable and preserved across populations and generations allowing analysis of genetic diversity across time and geography.

Mitochondrial DNA (mtDNA)

Mitochondrial DNA markers are essentially variations in DNA sequences. Similar to SNPs, mtDNA is highly stable allowing lineage analysis among populations across time and different geographical areas. There are also certain features unique to these markers. STRs, SNPs and mtDNA are highly effective in analyzing degraded samples making them suitable to analyze anthropological samples (human remains, ancient DNA).

Y chromosome markers

Y chromosome markers (STRs and SNPs) are exclusively paternally inherited, allowing evolutionary genetic analysis of the male lineage. As a complement, mtDNA is maternally inherited allowing genetic analysis of the maternal line.
HLA / MHC being highly polymorphic is another marker of choice nowadays used widely in population genetic study.

Major Histocompatibility Complex

The Major Histocompatibility Complex (MHC) is unique in that it is the most polymorphic system in the human genome and the only system to display functional polymorphism (**Spinola, 2005; Marsh, 2000**). Due to its high polymorphism, tight linkage among the loci and non random association of alleles this system has become interesting from perspective of population genetics. All the regions of MHC are known to be highly polymorphic, constituting several closely linked loci each with large number of genes that can be further split into many allelic types differing in their nucleotide sequences. Therefore the importance of this system in the study of polymorphism and their significance in population selection and survival and in providing clues to mechanism of generation as well as maintenance of this variability within the populations is immense.

With the advent of DNA based molecular typing techniques the polymorphisms in various allelic families of the MHC have also been revolutionized The number of alleles increases with the different typing methods like from serology to RFLP and IEF, followed by Polymerase chain reaction methods of PCR-SSP and SSOP, to Reverse Blot Hybridization to more refined Sequence based typing to the field of Micro array technology.

Apart from being an invaluable tool for population genetic studies, MHC polymorphism has important role in transplantation and disease associations. HLA associations have also helped to define syndromes of disease categories having common / shared pathogenic mechanism, like ankylosing spondylitis and related spondylo-arthropathies are associated with HLA-B27. It has been studied that HLA associations with infectious and autoimmune disease shows susceptibility and protective alleles in populations of different ethinic origins (**Hill ,1991; Carrington M,1999; Bowness P,2002**).

HLA associations with diseases vary in different populations. Disease predisposing genes and their molecular subtypes could help to determine and predict the incidence of the diseases in some populations. It is therefore important to have a population based database of HLA alleles and their frequencies of prevalence in healthy individuals so that disease predisposing influence of a particular phenotype could effectively be assessed in the populations.

Two individuals differ in only 0.2% of the genome leading to diversity, which is very important for natural selection and survival. Restricted distribution of alleles and subtypes of globally prevalent alleles within populations has helped to assign the ethnic origin of alleles. Alleles predominantly found in selected populations of similar ethnicity can thus be characterized as typically belonging to particular ethnic group and referred to as "Oriental allele" or a "Caucasoid allele". This distribution of subtypes of globally distributed allelic families has helped to trace the ethnicity and lineages of populations. Several theories of origin of human population and their routes of dispersal have been suggested. Of these, the Origin from Africa theory and subsequent migrations to the east and the west coupled with large number of natural selection processes and phenomenon has been the most

controversial as well as best supported by evidences (**Hill *et al.*, 1992; Diamond 1994; De Knijff L 2001**). Being a functionally polymorphic system, investigations into the distribution of MHC alleles in world populations are very important in this regard since the MHC genetic makeup of each of these populations would reflect interplay of both the basic genetic origin and effects of natural phenomenon such as founder effect and environmental selection. Differences in the prevalence of HLA alleles in different populations in varied environmental conditions could be utilized to assess the role of each of these alleles in conferring survival advantage to human populations.

Indian Subcontinent

Indian population is particularly relevant lying in a geographical transition zone between the western Caucasians and the Orientals in the East. Because of the historical, racial admixture, the population has become a 'melting pot of various races'. It shows immense heterogeneity in terms of cultures, languages, customs, religions and other factors that are known to divide people in groups. Particularly the population in the North has an extensive history of frequent migrations and invasions from both the east as well as the west of the subcontinent and constant amalgamation of these populations. Due to extensive history of admixture in this population, the gene pool is hypothesized to have been influenced by genetic contributions from various races. Previous studies on the MHC and other gene systems in this population provide evidence for the historical documentation of constant admixture (**Chhaya 2005; Jaini , 2002; Agrawal & Arundhati , 1999; Su B , 1999; Balakrishnan 1996; Mehra *et al.*, 1986; Mittal 1982; Shankarkumar , 1999, Chakraborty 1992**).

Indian subcontinent is located between 8 degree N to 37 degree N latitude and 68 degree to 97 degree longitude. It is an assemblage of more than one billion individuals that constitute 1/5th of the total world population. There are many reasons for the researchers to explore the gene pool of Indian populations; Geography is one of the important factors (**Cann 2001; Lahr & Foley 1998**). Second is extensive gene flow through series of migrations and invasions, resulting in the distribution of the

contemporary genetic variation across different geographic locations of India and creation of large genetic diversity. All these factors resulted into formation of various religions, cultures, communities, castes and linguistic groups (**Johnson *et al.*, 1992**) that resulted into the structuring of population and formation of various endogamous groups. Broadly it could be classified as Dravidians and Aryans. The Dravidians were considered as the original inhabitants of India who were driven south wards following invasions by Aryans from north-west during second and third millennium BC. Aryans were migrated from Iranian plateau. The colour of their skin was light and dark both; they have dark black hair and brown eyes. Language was mostly Sanskrit or derived Sanskrit. They introduced highly elaborated caste system in India with divisions into priests *(Brahmin), warriors (Kshatriya), Traders (Vaishya)* and inferior craftsman *(Sudra)* (**Mehra *et al.*, 1986**) .The broader four groups have been subdivided into smaller groups which marry within themselves. As a result of this the entire population has been divided into a large number of groups. The Indian caste system does not permit large-scale inter-caste, inter-religious and inter-ethnic marriages whereby the gene pool of each caste has evolved over the times and might have been fixed (**Jensen 1991**). These groups follow strict endogamy, which has resulted in a great deal of variation in the mating patterns, all of which invariably result in a wide genetic diversity (**Malhotra & Vasulu,1993; Naipaul,1992**).

Historically, it is known that various invasions have caused admixturing of Negrito, Negroid, Protoaustraloid, Mongoloid and Europoid elements in the Indian populations. Further, the population is also subdivided into four linguistic families i.e. i) Austro-Asiatic ii) Dravidian iii) Indo-European and iv) Tibeto-Burman. Under these four linguistic families about 325 languages are spoken in India. As per 1971 census there were 73.8% speakers of Indo-European languages, 24.2% of Dravidian languages, 0.8% of Tibeto-Burman family and only 1.2% Austro-Asiatic speakers. The population of India has been subjected to successive waves of immigrations and invasions. In the second millennium BC, waves of Indo-European speakers invaded the Indian subcontinent and imposed both religion and hierarchical caste system.. In 325-327 BC Alexander's Macedonian army invaded from both west and east with incursions into Punjab followed by Muslims in 1200-1500 AD. At that time the

existing religion was Hinduism, with numerous modified forms, e.g. Buddhism, Jainism and the animistic religions practiced by the tribal populations. During the Muslim rule spanning three centuries a large section of the population was converted to Islam. This led to another religious division within the same caste population. Around 500 years ago, the birth of Sikhism took place, but was confined largely to the North -Western parts of India, among the Punjabi speaking population. . Later empires of Muslims declined and the Portuguese and Dutch fragmented the whole kingdom into smaller kingdoms that were then first occupied and then Britishers occupied it from 1600-1947. This was the period when Christianity arrived. Most of the section of the population then became a follower of Christianity. Due to invasions and massive population movements in India its population structure subdivided into caste and tribes **(Gonem A,1996)**.

The emerging pattern of social organization based on endogamy and system of marriage rules among subdivided populations provides a clear picture of the biological composition with wide genetic diversity. Another important dimension of the Indian populations, especially among small populations, is that it offers potential opportunities for the operation of micro-evolutionary forces, which bring rapid changes in gene frequency of certain genetic traits **(Malhotra & Vasulu 1993)**.

All these factors must have affected the Indian populations and might have resulted into the unique gene pools. There is a need to study these unique populations at genetic level **(Agrawal,2004, Agrawal *et al.*, 2005)**. To study this it is required to investigate the populations for a set of genetic characters. It is of great interest to see whether the populations can be grouped in some way on the basis of inter relationships. The similarities and differences between the populations within such groups and those between populations in different groups can then be used to get some idea about the nature of forces that have prevailed in the evolutionary history of the populations. The differences in the populations can be studied using various statistical measures like breeding value (A), coefficient of inbreeding (F), coefficient of kinship (f), heritability (h^2), population mean (M), effective population size (N_e), Variance (V), rate of Inbreeding (ΔF), change of gene frequency (Δq), genetic distances, and phylogenetic relationships etc. It does not matter which statistical index

is being used, since different indices are all highly correlated. All of these are functions of differences in allele frequency. The essential feature of all these indices lies in representing the populations by points in a multi-dimensional space. They also allow the measurement of distance geometrically **(Balakrishnan,1988)**.

The study of populations at genetic level involves the identification of different alleles 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 – laboratory research on the chemical structure encoding the genes (DNA) 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, past, present and future.

An important step in the understanding of genetic structure of human populations consists of estimating frequencies of alleles at different loci. Gene frequencies 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.

A population is a local group of organisms of the same species that normally interbreed. The word population does not refer to an entire species; it refers instead to a group of organisms of the same species living within a sufficiently restricted geographical area that members can potentially mate with each other provided they are of opposite sex. Members of the same local group are more closely related to each other than members of groups who live in different geographical areas and people who live in the same geographical region are more similar than those separated by geographical barriers.

Populations grow and interact with one another through competition and predation. These factors can influence behavior ecology and evolution, either at micro or macro level.

Barriers to mating have partitioned the world's population into subpopulations, which are geographically distinct **(Comas *et al.*, 1998).** Subpopulations within a population may be completely isolated, which provides the information that there is no migration among subpopulations. They may be incompletely isolated, which is more often the case. However, the populations that live in geographical proximity do not always interbreed. Additional barriers may be language and cultural affiliations (tribe, country, etc). In practice, none of these barriers are absolute and it is rarely possible to define a population unit that is sharply distinguished from a geographically adjacent one. For the most part, genetic features that distinguish one population gradually merge into the next.

Genome diversity study in humans not only quantify the genomic variations within and between populations but also can reconstruct the evolutionary history of populations using the genomic data and trace the human migration and correlate the genomic diversities and affinities with archaeological, linguistic, epidemiological and cultural histories with a view to reconstruct the story of human evolution **(Nitai Pada Bhattacharyya,1999; Partha P Majumdar, 1999).**

During the process of evolution, new form of gene are naturally introduced by a process known as mutation and existing form of gene may be lost by process of natural selection and random genetic drift. When the humans grouped themselves in such a way that member within the group can interbreed but the exchange of the genes was rare because of cultural, geographic barriers and biological evolution of these groups. The extent of divergence is correlated with the time of such independent evolution of the populations. The expected genetic divergence between two such independent populations evolving for a known period of time can be calculated **(Crow and Kimura, 1970; Nei ,1987).**From the estimate of the genetic distance between the two populations, their time of divergence can be estimated and can provide useful information for reconstructing the evolutionary history. From about the turn of this century, studies based on diversity and affinities based primarily on

polymorphisms in expressed genes have been conducted globally as well as in India which yielded many interesting results on our origins (**Cavalli-Sforza *et al.*, 1994**). Due to the effects of natural selection on naturally occurring variants of expressed genes, the amount of variability at such gene loci between individuals is limited. With the introduction of molecular genetics techniques; regions of the DNA that do not code for expressed proteins can be studied in individuals. Such regions have the advantage that naturally occurring Alleles in these Loci are not under differential effects of natural selection and hence evolve as neutral alleles. Such regions are highly polymorphic which is extremely useful for population genetic studies .In the recent past very large number of highly polymorphic loci have been discovered throughout the human genome, thus it is possible now to study multiple loci within a short length of the genome, thus enabling construction of haplotypes. Haplotypic variation provides much more information for population movements and disease susceptibility, than variation at individual unlinked loci (**Partha P Majumdar, 1999**).

New alleles and haplotypes appear in the population because of mutations that occur in the germ line of individual organisms. This means that many genes are polymorphic, two or more alleles being present in the population as a whole, each with its own allele frequency. Allele frequencies change over time, due to natural selection and random genetic drift. However, the most important step in understanding the genetic structure of populations consists of screening studies aimed at estimating frequencies of alleles at different loci and interpreting their relevance to the particular phenomenon. A relative frequency of different alleles provides the information about importance in the study of population structure. Before knowing the population divergence it is important to understand a short history of mankind

Starting about 500 years ago large-scale population movements started to mix world populations on a scale not encountered previously. Some of these migrations are:

1. Europeans expanded into America, Australia and South Africa.
2. West Africans were transported to the Caribbean and North America through the slave trade. Afro-Caribbeans have further migrated to the British Isles.

3. Migrations from the Indian subcontinent to the British Isles
4. Migration from the far east to the USA

As a result of these migrations and admixtures the populations have been partially but not fully isolated. Hence, the genetic structures of various populations have undergone number of complexities and fission. The present study is an attempt to study the six North Indian populations of Uttar Pradesh and three North Eastern populations of North Bengal at genetic level by using highly polymorphic HLA class II antigens markers. This set of genetic markers would be helpful to provide distinct profiles, which might be expected if different evolutionary processes promoted their molecular differentiation.

Utility of population studies

The critics of genetic variation studies put forward the *'common disease-common variant'* (CD/CV) hypothesis, which states that the common genetic diseases are affected by common disease susceptibility alleles (or variants) at few loci that exist at high frequency across ethnically diverse populations (**Chakravarti, 2001; Reich, 2001**). These alleles probably arose before population differentiation and are common across populations. Supporters of CD/CV hypothesis cite that all population specific or frequent diseases like high frequency of C28Y-HFE allele and hematochromatosis; and that of Δ508-CFTR alleles and cystic fibrosis among northern Europeans are because these diseases are monogenic disorders. Therefore, it has been quoted that although substantial genetic variation is there but it is present in all populations. On the contrary, the neutral genetic variation aids biomedical research in at least three ways:

(i) Molecular sub-classification of the diseases can be done on the basis of genetic profile. High frequency of hemoglobin HbS allele, variant of glucose-6-phosphate de hydrogenase 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**).

(ii) Information about an individual genetic ancestry can be used 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**).

(iii) Incorporation of population genetic structuring 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.

Overall, knowledge of the genetic ancestry or origin of population sub-group, and information of population diversity, sub-structuring, stratification and phylogenetic relationship is a key in biomedical research and 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.

Search of such genetic tags 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.

However, more vision, planning and better approach 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 that particular study, as there are several aspect of Indian gene pool which has albeit got detected but not completely understood like:

(i) More defined 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.

(ii) Well defined 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 north Indian populations are most suitable for studying the effect of Indo Aryans and Muslims migrations on Indian gene pool.

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

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

The present study is an attempt to study the six North Indian populations namely *Kayastha, Rastogies, Mathurs, Vaish, Shia and Sunni* of Uttar Pradesh and 3 populations of North Eastern region of India namely *Lachung*, *Mech*, and *Rajbanshi* at genetic level. The genetic marker selected for the present study is highly polymorphic class two antigens which will provide an opportunity to search for an explanation of numerous unanswered enigma concerning the effect of the stringent social fabric on the genetic makeup of North Indian populations and completely isolated populations of North East India. We have made an attempt to compare these populations among themselves and also with other Indian and world populations. The main objective of this study was to understand the effect of social and geographical barriers on Indian population sub structuring.