Chapter 1:

Introduction

1. Introduction

Researches on human genetics over decades have played a major role in exploring intra and inter population genetic differences, which have not only strengthened our base for understanding human migration patterns on global scale but also enhanced our knowledge regarding the etiology of different diseases. (Bamshad, *et al.*, 2001;Cavalli-Sforza and Piazza, 1975).

Although anthropometric traits such as cephalic index were once considered a favorite tool for reconstructing evolutionary relationships, but later on, it was realized that such traits might not be under complete control of biological inheritance. Furthermore, variations in such traits may have arisen due to short-term responses to environmental changes. On the other hand, certain human traits such as height, weight, skin colorations, facial features and many more, which can be easily perceived by a non-professional, are also genetically determined to some extent, but have little contributions to the understanding of the patterns of genetic variations.

Genetic studies of ABO blood group system defined by Landsteiner (Landsteiner, 1961), provided a precise definition of genetic variation for the first time. However, the full extent of individual genetic variation began to emerge only when genetic analysis could be carried out at the level of hereditary material itself. This technique became widely available only in the 1980s and are gaining importance in present day researches since greater variations exist at the DNA level compared to proteins and blood groups.

Populations with clearly different evolutionary histories demonstrated similar frequencies of a single gene owing to random genetic variation. Therefore, cumulative information on variations of more than one gene is necessary to overcome such confusion. Informations from many loci may be combined through multivariate analyses, which are useful for extracting informations of genetic and evolutionary interest.

It is indeed a very difficult and challenging task to interpret the role of evolutionary forces in the reconstruction of major human migration events. This is executed by comparing genetic information with relevant knowledge from other fields, which include ecogeographical, historical, paleo-anthropological, cultural, and linguistic evidences.

It is a well-established fact that the number of genes vary among organisms. Therefore, molecular mechanisms must exist that generate new gene structures and govern their evolution and sustenance in the population. A brief description of the sources of new genes and their evolution has been given below:

1.1. New genes: Sources and Evolution

The creation of new gene structures involves several molecular mechanisms, which are as follows:

1.1.1. Exon shuffling

A new exon–intron structure can be created by bringing two or more exons from different genes together, or by duplicating the same exon (Gilbert, 1978). Genomic evidences suggest that exon shuffling, also known as domain shuffling, often recombines sequences that encode various protein domains to create mosaic proteins (de Souza, *et al.*, 1996;Kaessmann, *et al.*, 2002). Numerous genes created by exon shuffling were identified using direct sequence comparison (Patthy, 1996).

1.1.2. Gene duplication

Duplicate genes created by this process may show different function(s) compared to that of its ancestral copy (Kimura, 1983;Ohno, 1970). In fact, gene duplication not only contributed to the evolution of new genes with different functions but also contributed substantially to the developmental evolution of an organism (Prince and Pickett, 2002).

1.1.3. Retroposition of genes.

Through this mechanism, duplicate genes are created by reverse transcription of expressed parental genes (Long, *et al.*, 2003). A functional retroposed gene has a chimeric structure consisting of a retroposed coding region with a new regulatory sequence from another gene, thereby functioning differently from that of the parent gene. In mammals, the L1 retro-element is responsible for retroposing nuclear genes (Esnault, *et al.*, 2000;Long, *et al.*, 2003;Moran, *et al.*, 1999).

1.1.4. Mobile elements

Analyses of human genome sequences (Nekrutenko and Li, 2001) and vertebrate genes (Lorenc and Makalowski, 2003) have shown that the integration of mobile elements into nuclear genes generate new functions as was seen in case of the human decay-accelerating factor (DAF) gene (Makalowski, *et al.*, 1994).

1.1.5. Lateral gene transfer

This molecular mechanism may result in the exchange of genes between prokaryotic organisms. Additionally, this process may also cause the recruitment of new genes in an organism resulting in new phenotypes (Ochman, 2001). It is also evident that lateral gene transfer might be important in the evolution of eukaryotic genes (Bergthorsson, *et al.*, 2003).

1.1.6. Gene fusion/fission

A single gene can be produced as a result of fusion of two adjacent genes by read through transcription. This may occur due to the deletion or mutation of the translation stop codon and the transcription termination signal in the upstream gene (Long, *et al.*, 2003). Several cases of gene fusion events were reported in higher eukaryotes by Thomson and colleagues (Thomson, *et al.*, 2000). In contrast to the above mentioned phenomenon, a single gene may get split to form two new genes by a still unclear mechanism (Long, *et al.*, 2003).

A classical model describing the fate of a new gene was put forwarded by Haldane (Haldane, 1933) and Fisher (Fisher, 1935), and was later extended by Kimura (Kimura, 1983). This model described the mechanism through which duplicated gene could acquire new functions and ultimately be preserved in a lineage. However, the impact of the evolutionary forces in the sustenance or loss of a newly generated gene from a population cannot be denied.

1.2. The four forces of evolution

A gene is a DNA segment encoding a functional RNA or protein product and is considered as a molecular unit of heredity. The extensive diversity of genes on earth is based on four fundamental forces of evolution: mutation, gene flow, natural selection and random genetic drift (Long, *et al.*, 2003).

A **mutation** is considered as a permanent change in the nucleotide sequences of genetic elements. Mutations of a single gene produce its alternative forms called alleles, which are the main reasons behind variations in protein products.

Random sampling of alleles when passing from generation to generation alters the frequency of an allele in a population. Such a change is referred to as the **Genetic drift** (Masel, 2011). Fluctuations in the allele frequencies occur due to the randomness of the transmission process until, by chance, one of the alleles is lost and the other is fixed. After few generations, the newly emerged population may show extreme variations in gene frequencies from those of parental population. Such a phenomenon of extremely random genetic drift is referred to as "**founder effect**," (Templeton, 1980) as the population started with few founding members with restricted gene pool.

The transfer of genes from one population to another through interbreeding is referred to as **Gene flow,** which results in the increase of similarity between those populations. If two genetically different populations mate, offspring having genes from both the populations may be produced. Further, additions of new genetic variants to the established gene pool of a particular species or population may also result due to immigration (Cavalli-Sforza, *et al.*, 1993).

Natural selection is a constantly shifting process by which organisms with specific gene patterns that are adaptive to a particular environment become more prevalent over time. It is influenced not only by an organism's biology, but also by the interaction of that biology with environmental conditions (Sternberg, 2004).

Allelic distribution in context to geographical locations may give information on the evolutionary history of the allele, on the place of origin of the genetic change (mutation) that generated it and the influence of selection pressure on it (Roychoudhury and Nei, 1988). Furthermore, frequencies of alleles in a particular region may also be correlated with the environmental parameters to interpret any specific genetic adaptations (Nei, 2005). In general, degree of allelic variations between two regions is directly proportional to their geographic distances. Thus, geographic maps corresponding to allele or gene frequencies of a number of populations may be constructed. Furthermore, geographic distribution of a gene may also be correlated with disease prevalence and adaptation to environmental factor. For example, a variant of Angiotensin gene (AGT) which has been found to be 90% frequent in some African populations and 30% in European populations (Nakajima, *et al.*, 2004))**,** has been found to increase the risk of developing hypertension by 10-20% (Kunz, *et al.*, 1997).

Much of the genetic variations seen among the humans resulted indirectly from the pattern of expansion and migrations accompanied by random genetic drift. However, the surrounding environment is the most important factor, which influences the heredity of a given attribute. In fact, environment affects the same genetic structure differentially, thereby generating a range of phenotypes for a particular genotype. Moreover, heritability will also differ based on interactions of different genotypes with varying environments (Lewontin, 1974).

1.3. Evolution of Modern Human

According to fossil evidences, modern humans evolved in Africa about 200 KYA (Klein, 1999)**.** The physical characteristics of anatomical modern man consist of a high rounded skull, facial retraction, and a light skeleton (Lahr and Foley, 1998;Lahr, 1996). Fossils

with such characteristics of anatomically modern man have been found in eastern Africa dated to approximately 160–200 KYA (McDougall, *et al.*, 2005;White, *et al.*, 2003).

Informations gathered from human genetic variations express a similar view on human variation as that of the fossil records. Greater genetic diversity in populations from Africa imply the first appearance of human in Africa and subsequent colonization in Eurasia and the Americas (Tishkoff and Williams, 2002;Yu, *et al.*, 2002) (Tishkoff and Verrelli, 2003). Interestingly the dates calculated for human expansion based on genetic variations and archaeological record generally coincide (Jorde, *et al.*, 1998). There are three main theories for the evolution of modern humans:

1.3.1. Multiregional model

This theory does not support a single geographic origin for the modern *Homo sapiens* (Wolpoff, *et al.*, 2002). This theory suggests the spread of *Homo erectus* out of Africa into the different regions of the World, who later gradually evolved into archaic *Homo sapiens*. Later these archaic populations are modified simultaneously to modern *H. sapiens* in various parts of Europe, Asia and Africa **(**Figure 1**).** This model has received considerable support from the fossil records. According to archaeological reports, there is a continuum of certain morphological traits from the first H. erectus to those of modern populations from a particular geographic area. For example, skeletal remains of early H. sapiens from various location of China, North Africa, and Europe resemble modern populations in those areas in some aspects (Thorne and Wolpoff, 1992).

1.3.2. Model of Recent African origin (RAO)

This theory suggest the evolution of archaic humans to anatomically modern humans solely in Africa (Hetherington and Reid, 2010). It further extends that individuals belonging to one such branch of *Homo sapiens* left Africa and populated other regions of the World by replacing other populations of the genus Homo such as Neanderthals and *Homo erectus* (Figure 1). Genetic and fossil evidences strongly supported this concept and made this theory the most accepted one within the scientific community (Stringer, 2003). Evidences from researches on mtDNA (Ingman, *et al.*,

2000), the Y chromosome (Underhill, *et al.*, 2000), the X chromosome (Kaessmann, *et al.*, 1999), and many autosomal regions (Harpending and Rogers, 2000) extended support to this RAO model of human origin.

1.3.3. Assimilation model

According to this model, gene flow occurred between modern and earlier human populations in varying degrees (Smith, *et al.*, 1989) (Figure 1). This model suggested that modern human emerged in Africa and migrated into other regions of the World where they hybridized with earlier settled archaic populations and replaced them gradually. Therefore, modern human evolution in different parts of the World may result due to the blending of modern characters from the recent African populations with the characters of the local archaic populations. As evident from fossil evidences, this model may more correctly represent the complex and gradual nature of the processes represented.

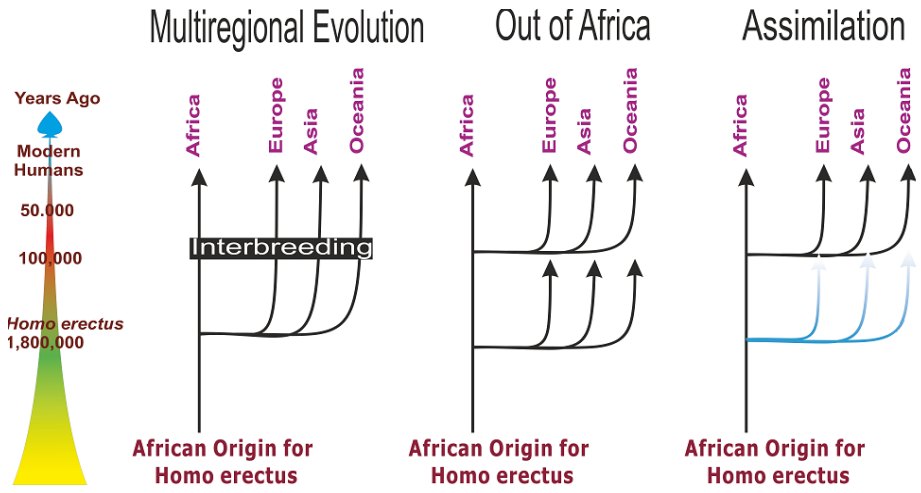

Figure 1: The three models of Human Origin and Evolution. Adapted from Gibbons, 2011 (Gibbons, 2011).

1.4. Genetic tools for studying history

The word ‘population’ refers to a group of organisms of the same species inhabitating a restricted geographical area where members of opposite sex can potentially interbreed. The set of genetic information carried by a population is known as its gene pool and a population sharing a common gene pool is referred to as the Mendelian population. Although evolutionary forces cause minor variations in the gene pool of an organism through a passage from generation to generation, quite interestingly, it has been observed that, the individuals belonging to a local group are related more closely to each other than with those occupying other geographical areas. Such evolutionary characteristics cause change in the genetic constitution of a population followed by genotypic and phenotypic alterations (Guerra, *et al.*, 1999). Thus, it can be said that gene frequencies have gained enough importance in evolutionary studies especially in human. However, selection of the right marker should be accompanied by the application of robust statistical tools in order to avoid sampling fluctuations and erroneous clarification of the data. Researches on human evolution and on anthropogenetics have been greatly benefitted by the advent of various molecular markers along with more robust DNA typing technology. Apart from the functional genes, many other markers have gained wide applications in the recent times. These DNA markers are highly polymorphic making them more informative in studying genetic variations between and amongst human populations.

Short Tandem Repeats (STR) are short fragments of highly polymorphic DNA which are very useful in studying human genetic diversity (Butler, 2007). Single Nucleotide Polymorphism (SNP) DNA markers (autosomal SNPs and Y chromosome SNPs) are characterized by single nucleotide changes in DNA. SNPs are preserved across populations and generations thereby allowing phylogenetic analyses across time and geography (Butler, 2007).

Another essential phylogenetic marker based on variations in mitochondrial DNA sequences is the Mitochondrial DNA (mtDNA) markers (Wallace, 1994). Similar to STRs and SNPs, mtDNA are highly effective in analyzing degraded samples making them suitable for analyses of anthropological samples (human remains, ancient DNA). Y

chromosome markers (STRs and SNPs) are paternally inherited allowing genetic analysis of the paternal line (Jobling and Tyler-Smith, 2003) while, mtDNA show maternal inheritance allowing genetic analysis of the female line.

Human Leukocyte Antigen / Major Histocompatibity Complex is another group of highly polymorphic markers (Shiina, *et al.*, 2009) which are widely used in population genetic study nowadays. This system not only exhibit high degree of polymorphism but also show tight linkage among the loci and non-random association of alleles (Carrington, 1999). These characters have made the HLA system interesting from perspective of population genetics. The number of HLA alleles increased with the advent of different DNA based molecular typing techniques. HLA polymorphism not only has an inestimable role for population genetic studies but also play an important role in transplantation and disease associations. Earlier studies have reported associations of both susceptible and protective HLA alleles with infectious and autoimmune diseases in populations of different ethnicity (Bowness, 2002;Carrington, *et al.*, 1999;Hill, *et al.*, 1991).

Killer Cell Immunoglobulin-like Receptors (KIRs) are the members of a receptor family that are found on the surface of NK cells and cytotoxic T cells (Lanier, 1997). They interact with HLA class I molecules, regulate NK cell activity, and protect healthy cells from spontaneous destruction or killing by NK cell mediated cytolysis. These receptor molecules exhibit either inhibitory or activating functions or both. The structure of the KIR molecules determine their activating or inhibitory function (Middleton and Gonzelez, 2010). The KIR gene family exhibit great diversity of genes and alleles, all of which are immunoglobulin superfamily members. This gene cluster is present on chromosome 19 within the Leukocyte Receptor Complex (LRC) (Uhrberg, *et al.*, 1997;Vilches and Parham, 2002;Wilson, *et al.*, 2000). Fourteen genes are present in the KIR family along with two pseudogenes. KIR haplotypes are of two types depending on gene content, namely A and B, (Wilson, *et al.*, 1997). The A haplotypes have seven KIR loci of which only one encode NK cell-activating receptor - *KIR2DS4*. The B haplotypes contain KIR loci most of which encode activating receptors with only two loci encoding

inhibitory receptors. Inspite of their differences in gene content (Uhrberg, *et al.*, 1997;Vilches and Parham, 2002), all KIR haplotypes contain the 'framework genes, namely *KIR3DP1, KIR3DL2, KIR3DL3 and KIR2DL4*. These framework genes are well conserved in virtually all individuals (Martin, *et al.*, 2000;Vilches and Parham, 2002;Wilson, *et al.*, 2000). Based on the analysis of KIR genotypes, variations have been observed in the frequencies of both the haplotype groups between populations. Recent studies on KIR interaction with HLA ligands have suggested its role in immunopathology. Thus, it has become important to characterize the genotypes based on KIR genes in different endemic populations. The study may also help to decipher the relatedness of ethnic populations among themselves and with the other Indian and world populations. Furthermore, the study may illuminate the possible origin of these populations and probable human migratory patterns in this part of the world. Consequently, the study may help in analyzing gene-disease co-relatedness and disease susceptibility within the local/endemic populations.

1.5. Indian Subcontinent

India has a unique geographical location between 8° N to 37° N latitude and 68°E to 97° E longitude, which has immense effects on its resident populations. The Indian population 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 populations. The Dravidians are generally considered as the original inhabitants of the country who were driven southwards following Aryan invasions from north-west around 1500- 100 BC. They introduced highly elaborated caste system in India with divisions into priests (Brahmin),warriors (Kshatriya), Traders (Vaishya) and the inferior craftsman (Sudra) (Mehra, *et al.*, 1986). These broader four groups have been subdivided into smaller groups who marry within themselves. As a result, of this the entire population has been divided into a large number of groups.

Further historical invasions have caused admixture of Negrito, Negroid, Proto-australoid, Mongoloid and European elements in the Indian populations (Rao, 1986), followed by subdivision into four linguistic families i.e. i) Austro-Asiatic ii) Dravidian iii) Indo-

European and iv) Tibeto-Burman respectively. Several waves of human invasions and immigrations occurred in the Indian sub-continent. In the late 16th century BC, Indo-European speakers invaded the Indian subcontinent and started imposing religious beliefs and hierarchical caste systems (Kurien, 2002). 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 (McCrindle, *et al.*, 1896). 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, that spanned 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 whole kingdom got fragmented into smaller kingdoms which were then first occupied by the Portuguese and Dutch and then it was occupied by the British from 1600-1947. This was the period when Christianity arrived. (Mayhew, 1929). Thus, invasions and massive population movements in India subdivided its population structure into caste and tribes. Such historical and racial admixture has made the Indian population a 'melting pot of various races' exhibiting extensive cultural, religious and linguistic diversity. 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 resulted in a wide genetic diversity (Naipaul, 2010). 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

Thus, it can be said that geographical location and extensive genetic diversity are the two primary reasons behind the growing interests of researchers worldwide to explore the genetic diversity among Indian populations. The primary requirement for such

exploration is the screening of some essential genetic markers among different population groups of India and then analyzes their relationships.

.

1.6. Utility of population studies

Genetic knowledge of population sub-structuring and stratification is a prerequisite for proper selection of controls and for identification of disease pre-disposing genes among different ethnic population groups. Furthermore, genetic profile also facilitates molecular sub-classification of the diseases. The genetic ancestry of an individual helps in better determination of the presence of a disease marker gene. In addition, differences in the genetic structure of different ethnic groups help to explain differences in drug responses among the groups. Therefore, information about the genetic ancestry of an individual may also prove beneficial to improve medical diagnosis and treatment.

Overall knowledge of the genetic ancestry of population sub-group, and information on population diversity, sub-structuring, stratification and phylogenetic relationship are the essential requisites in the biomedical research arena. However, reliability of the genetic markers stand as major hurdle in the path of correct inference of the genetic origin of a sub population and the resolving power of a genetic marker by which it can differentiate between two close population subgroups.

However, in India, the population structure is very much complex and therefore better planning and approach are required to conduct such studies in India. More defined genetic studies on the Tribal and Dravidian populations are required to reveal the exact composition of Indian gene pool. Populations with well-defined geographical and cultural identities should be chosen. Similarly, studies on more endogamous groups can reveal the structure of castes in genetic context and analyze the effect of endogamy on genetic composition of the Indian population.

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 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 aimed to analyze frequency and distribution of KIR genes in five ethnic population groups of India namely Rajbanshi, Rabha, Bengali, Gurkha and Muslims. This study also attempts to explore the association of KIR genes with Rheumatoid arthritis if any. Thus, this study may be able to explain numerous unanswered questions raised on the genetic structure of these populations compared to other Indian and World populations. Furthermore, studies of KIR genes in Rheumatoid Arthritis patients may reveal associations, if any. Thus, researches on differential distribution of KIR genes in different populations may be used to assess the role of each KIR gene in granting advantage to the survival of human populations in conditions of varying environment.

1.7. Objectives of the Study

1) To analyse the frequency and distribution of KIR genes within the human population of Sub-Himalayan region of India.

2) To study the heterogeneity among the local population(s).

3) To trace the phylogenetic relationships of the studied populations with that of the different World populations based on KIR gene profiles.

4) To correlate the disease association (if any) with KIR genes.