1.1. Development of the Field of Population Genetics | 3
1.2. Tools of Population Genetics | 20
  1.2.1. Population Genetics and the Classical Approach | 21
  1.2.2. Population Genetics and the Molecular Approach | 23
    1.2.2.1. Molecular Markers | 24
      1.2.2.1.1. Repetitive DNA Sequence Variants | 24
      1.2.2.1.2. Non-Repetitive DNA Sequence Variants | 27
      1.2.2.1.3. Lineage Markers | 28
    1.2.2.2. Laboratory Techniques | 29
1.3. Statistical Measures | 32
1.4. Evolutionary and Genome Diversity Studies | 33
1.5. A Look Inside the Genome | 37
  1.5.1. Human Genome Project (HGP) | 37
  1.5.2. Human Genome Diversity Project (HGDP) | 38
  1.5.3. HapMap Project | 38
  1.5.4. Indian Genome Variation (IGV) Consortium | 39
1.6. Rationale of the Study | 40
1.7. Objectives of the Study | 42
Introduction

“In nature’s infinite book of secrecy a little I can read”
William Shakespeare (1623), Antony and Cleopatra

Anthropology is a multidisciplinary and multifaceted approach to study the biological and socio-cultural attributes of human societies/populations; both extinct and extant. Combining information from a wide range of interrelated branches of behavioural and biological sciences is perhaps one of the most interesting aspects of Anthropology. It focuses on studying population groups as opposed to individuals. In its earliest stages of development the subject was mainly concerned with documenting socio-cultural aspects of tribal communities but now it has come to encompass various aspects of any and every human ethnic groups. It understands the importance of studying the similarities and differences within and between the population groups in context of their cultural and genetic boundaries.

Biological or Physical Anthropology is the branch of Anthropology that is concerned with documenting and understanding biological attributes of human populations; and their interaction with culture. Although in the beginning Physical Anthropology started off mainly as a descriptive and classificatory science of morphological variation of human population groups, it has now come to include other aspects in its ambit. For Biological anthropologists one of the foremost areas of interest has been the origin of man within the framework of human evolution. The subject also deals with the process of microevolution that has been responsible for shaping the present world population scenario whereby similarities and differences exist within and between population groups. Implications of the resulting population structures and the contributing factors are a matter of research inquisitiveness as well. To address questions about human origins, migration patterns, evolution and diversification at regional and global levels, anthropologists had mainly relied on tools and inputs from palaeontology and archaeology in the past (Smith and Spencer, 1984; Nitecki and Nitecki, 1994; Stringer and McKie, 1996; Stringer, 2000; Mellars, 2006). Findings from such studies have helped immensely in developing a framework for understanding origin of man.
Evolution refers to changes, small and large, adaptive and non-adaptive, that are passed on from one generation to next and consequently the concept of genetic variation comes into the picture. So, understanding of origin and maintenance of genetic variation is crucial to understanding evolution. All the issues pertaining to evolution are dealt with in the branch of science known as **Population Genetics**. Population genetics is the study of gene and genotype frequencies in populations of interbreeding organisms (small or large, natural or artificial) and predicting the way these frequencies are maintained or changed under the combined influence of various factors. It is concerned with applying models of gene frequency change involving different factors in the context of Mendelian genetics to examine evolution in a quantitative manner. In order to understand the pattern of allele frequencies we need to have a defined population, in this case a Mendelian population. Dobzhansky (1951) defined it as the reproductive community of individuals which share a common gene pool. Evolutionary studies involve reconstructing past demographic events that have led to the present day diversity patterns. Use of various models allows one to examine interplay of various factors and make inferences about the past based on present day data. But one has to be careful about interpreting results obtained from any model considering that all models have some assumptions inherent to them.

The field of population genetics as we know today is a result of the synthesis of evolutionary theory of Darwin and Mendel’s principles of genetics. This was not so from the starting when the two areas were considered incompatible. It took a long time for this synthesis to take place and was made possible by the remarkable contributions of three great scholars; namely Ronal Aylmer Fisher, John Burdon Sanderson Haldane and Sewall Wright who showed mathematically, the consequences of selection and drift acting on populations obeying the Mendelian rules of inheritance. Due to the contribution of several scholars and their ideas about evolution and genetics, the field has come a long way and has firmly established its importance. In order to understand the development of the field of population genetics, an overview of developments in fields of evolution and genetics is essential.
1.1. Development of the Field of Population Genetics

Since time immemorial people have been interested in knowing how the different organisms, especially humans, came into existence and reached their present form; and also the reasons behind the observed diversity. For thousands of years explanations were sought in the realm of religion and it was considered taboo to think and talk otherwise. So, when people started seeking answers on evolution outside of religious beliefs they were shrouded in controversy. Till 18th century species were considered to be unchanging entities, God’s special creations as opposed to being products of natural phenomena. It changed when French naturalist, Georges-Louis Buffon (1707-1788) proposed a theory of reproduction that was in opposition to the prevailing theory of pre-existence in his work *Histoire naturelle* (1749) which was published in 36 volumes during his lifetime. He favoured transformation of species but rejected the possibility of descent from a common ancestor for humans and apes. He thought that climate change may have driven the worldwide spread of species from their centres of origin. According to Mayr (1981), Buffon was the father of all thought in natural history in the second half of the 18th century. He brought to the notice of scientific community many evolutionary problems that had not been talked about by anybody before him. But his opinions varied at different periods. His ideas were furthered by his student Jean Baptiste de Lamarck (1744-1829).

In an effort to understand evolution, French naturalist Jean-Baptiste de Lamarck propounded the theory of inheritance of acquired characteristics in his work *Philosophie Zoologique* (1809). According to the theory, an organism can change its morphological characters in response to environmental changes and can pass on these characteristics that it acquired during its lifetime to its offspring. This leads to transformation/evolution from simpler life forms to complex ones. It was thought that organisms would acquire only those characteristics during their lifetimes that were beneficial to them and would pass on these adaptive features to their offspring. Popularly known as Lamarckism, this notion of evolution received little acceptability. Lamarck was one of the earlier influential proponents of concept of biological evolution. In past few years there has been a fresh interest in Lamarckism. Several epigenetic studies, both in animals and humans, have shown that there is a possibility of transmission of behavioral traits acquired by the previous generation to the next
generations. Transmission of information across generations which does not involve traditional inheritance of DNA-sequence alleles has been observed in these studies. Most such studies have focused on food availability as the environmental stress (Lumey et al., 1995; Richards, 2006; Natt et al., 2009).

Then came Charles Darwin (1809-1882) with his theory of evolution by Natural Selection and changed the course of evolutionary studies forever. He was mainly interested in understanding the evolution of species over geological time scales. In his seminal work *On the Origin of Species by Means of Natural Selection* (1859), he discussed at lengths his work on the diversity that he observed during his 5 year voyage around the world on Naval ship H.M.S. Beagle. During this voyage he accumulated evidence for the theory that he had started formulating in the late 1830s. Based on his observations of different organisms he proposed that the diversity of present day living organisms could be attributed to ‘descent with modification’ from one or perhaps a few archaic forms, and suggested natural selection to be the driving force behind this. Darwin was not the first one to realize that species have evolved from a common ancestor. Geologists and paleontologists had found a lot of evidence indicating that life had existed on Earth for a long time, that it had changed over that time, and that many species had become extinct. But it was Darwin who showed how all this evidence favored the evolution of species from a common ancestor and offered a credible mechanism by which life might evolve. Darwin was not alone in realizing the importance of selection as the main evolutionary mechanism. British biologist Alfred Russel Wallace (1823-1913) had independently conceived of this natural way for life to evolve although it was Darwin who had called the process as natural selection. Both of them were inspired by English economist Thomas Malthus’ work *An Essay on the Principle of Population* (1798) in which he had discussed about the pressures of population growth in humans on available resources and their implications. It occurred to both Darwin and Wallace that same population pressure would also apply to all other organisms, both plants and animals. They realized that since the Earth is not overburdened by any species, it implies that the organisms do not reproduce to their full potential. Also, not many survive to reach adulthood as they are vulnerable to environmental stresses and have limited food supply. So, they must compete, although unconsciously, for the little available resources. They both
appreciated that in this struggle for existence, survival and reproduction cannot depend entirely on chance factor. They thought that if an organism has some trait that helps it to withstand the adverse conditions or reproduce more successfully, it will leave behind more offspring than others and the favorable trait would thus become more common in the next generations. They contemplated that nature "selects" individuals better suited to surviving their local conditions. In 1858, Wallace communicated his paper *On the Tendency of Varieties to Depart Indefinitely From the Original Type* to Darwin in which he had discussed about natural selection. According to some Wallace seemed to have envisioned natural selection as a kind of feedback mechanism keeping species and varieties adapted to their environment (Smith, 2004). Both their theories, which were remarkably similar, were presented in a meeting of the Linnaean Society in 1858 and published in its proceedings. Darwin was working on his book on evolution at that time which was published in 1859. In 1889 Wallace published the book *Darwinism* in response to the criticisms to natural selection. In this book he also discussed the idea that if two populations of a species are adapted to their local environments and have diverged enough such that the hybrid offspring is less well adapted than either parent, natural selection will tend to eliminate the hybrids. This will lead to reproductive isolation of the two populations by encouraging development of barriers against hybridization. This has come to be known as Wallace effect.

Other than natural selection, other key element of Darwin’s theory was existence of variation on which selection acted. He was unaware of the process by which variation arises but visualized how selection would operate on genetic variation once it is produced. Scientific community accepted the notion that evolution has taken place but was doubtful of the role of natural selection. Darwin realized that for his theory of natural selection to be plausible and for the adaptive modifications, observed as phenotypes, to be passed on from one generation to the next, a mechanism of hereditary transmission was required otherwise the offspring will not resemble their parents and fitness-enhancing traits will not spread through the population but he was not able to shed light on this. The fact that he could not explain the mechanism of inheritance was the main objection to his theory of evolution. He accepted a weak form of Lamarck’s use and disuse theory to explain origin of variation and attempted
to work out a way for explaining mechanism of inheritance by assuming that in response to environmental stimuli, somatic cells of body would release entities called ‘gemmales’. According to him, these ‘gemmales’ carried information on the traits of the organism and would accumulate in germ cells and get passed on to the next generation. But he could not prove this. He knew that for selection to operate in a population and gradually alter it, continuous supply of variation was required. There were scholars who believed in importance of selection for creating new adaptive changes. Others doubted that selection acting on continuous variations was sufficient to transform one species into another. Still others advocated Lamarckism. It took the discovery of genes and mutations in the 20th century to render natural selection feasible and unavoidable as an explanation for evolution.

The concept of hereditary transmission of traits or in other words mechanism of inheritance was made comprehensible by a monk who made experimental observations on *Pisum sativum* in monastery gardens of Austria for eight years (1856-1863). He inferred from his breeding experiments that each trait is governed by ‘factors’ that come in pairs and an organism gets one ‘factor’ from each of his parent. He was Gregor Mendel (1822-1884) and is now known as the father of Genetics. During his experiments he grew several thousand pea plants and cross bred individuals that differed in some kind of observable trait such as colour and form of seeds and pods, colour of seed coat, position of flower and length of stem to produce hybrids. He then kept track of the progeny type and number. In 1866, based on his work he postulated that individual traits are determined by discrete inherited ‘factors’ governed by three laws and made unprecedented contribution to the field of genetics by discovering the fundamental laws of inheritance. Mendel’s first law, the *law of dominance* states that some inherited ‘factors’ are dominant and can mask expression of other recessive ‘factors’. So, if an organism has alternate forms of ‘factors’ for a trait, the dominant form will be expressed. Second law or *law of segregation* states that the ‘factors’ keep their identities separate even when they occur together i.e. Mendel suggested that inheritance is of Particulate nature as opposed to ‘blending’ kind suggested by Fleeming Jenkin (1867) in his opposition to Darwin’s theory. In other words, during sexual reproduction parental ‘factors’ for a trait get distributed into the sex cells such that each sex cell has one ‘factor’. When a sex cell each from
both the parents unite to form the progeny, two discrete ‘factors’ get transmitted to the next generation. Third law or principle of independent assortment states that a diploid organism heterozygous for two traits produces gametes in which all the four possible combinations of the ‘factors’ for the two traits are equally represented. Factors for both the traits are sorted independent of each other and inheritance of one trait is not dependent on the other. The term ‘gene’ was later coined by botanist Wilhelm Johannsen in 1909 for the unit of heredity and the alternate forms of a gene were called ‘alleles’ by William Bateson. Mendel recognised the mathematical patterns of inheritance from one generation to the next by working out the ratios in which different types of progeny will be produced in a particular type of cross. The principles were described by him in a two-part paper, Experiments on Plant Hybridization that he presented to the Natural History Society of Brunn in 1865, published in 1866. Although we have come a long way in our understanding of inheritance and associated factors since the work of Mendel, the path was undoubtedly laid by him. His laws of inheritance have shaped the present day understanding of genetics. Mendel’s work lay largely neglected till 1900 when several years after his death, it was rediscovered almost simultaneously by Hugo de Vries in Holland, Carl Correns in Germany and Erich Von Tschermak in Austria, who had reached at the same results as Mendel in his own breeding experiments and later came across Mendel’s forgotten work. Immediately afterwards, his conclusions began to be confirmed and extended by experiments carried on in various parts of the world on many kinds of plants and animals.

Carl Erich Correns (1864-1933) was a tutor at the University of Tübingen when he began to experiment with trait inheritance in plants in 1892. At that time he was somewhat aware of Mendel’s work on hawkweed plant from Carl Nageli whom Mendel had earlier written to. But Correns was unaware of Mendel’s work on pea plants. But when he submitted his work in 1900 for publication, it was titled as G. Mendel’s Law Concerning the Behavior of the Progeny of Racial Hybrids. He also discovered cytoplasmic inheritance while working with Mirabilis jalapa.

Erich Von Tschermak-Seysenegg (1871-1962) studied agriculture at the University of Vienna. His botanist grandfather, Eduard Fenzl, had taught Mendel at one point in time. In 1898, he started doing plant breeding experiments using peas, and by 1900,
he had written up his results, published as *Concerning artificial crossing in Pisum sativum*, and contributed to the rediscovery of Mendel’s laws. He was a plant breeder, and his hybridization experiments were done with the idea of improving crops using the laws of heredity. He subsequently produced high-yielding food crops such as wheat, barley, and oats.

Hugo de Vries (1848-1935) was a professor of Botany at the University of Amsterdam. He began his genetic experiments on multiple plants species in 1880 and completed his work without knowing anything about Mendel’s work. He published his work, *Concerning the law of segregation of hybrids*, first in French and then German. Only the German version had mention of Mendel’s work. He was also a strong proponent of the idea of discontinuous variations and believed that speciation took place through sudden, large changes of character traits. During the time of controversies over importance of selection as prime mechanism of evolution, de Vries proposed the "theory of mutation" or theory of formation of new species by single mutation in 1903. He also coined the term ‘mutation’. He based the theory on his work on evening primrose, *Oenothera lamarckiana*. He observed that the plant had several aberrant forms even in the wild populations. He experimentally bred the plant for many generations and observed that *O. lamarckiana* continually produced small proportions of offspring with significant phenotypic differences such as leaf shape and plant sizes. The aberrant forms would then either bred true or segregated into *O. lamarckiana* and the aberrant forms; these de Vries designated as new species. Considering that genetics has just been born, the theory was not examined. It is now known that most of the variants that de Vries isolated from *O. lamarckiana* were due to aberrant chromosomal segregations, and not to mutations associated with specific genes as shown by Renner (1914) and Cleland (1923).

de Vries’ idea of role of mutation in causing evolution was taken forward by Thomas Hunt Morgan (1866-1945). He worked on fruit fly *Drosophila melanogaster* and observed several mutations, most of them governed by Mendelian genes (Morgan et al., 1915). Morgan and colleagues showed that these mutations occurred with very low but measurable frequencies. He favoured the role of mutations in evolution based on this work. While he agreed that most new mutations are harmful and thereby
quickly eliminated from the population, he believed that a small proportion is advantageous and will spread through the population. He viewed that replacement of existing genes by such mutations eventually leads to evolution. In his opinion role of natural selection is limited to preservation of beneficial genotypes and elimination of unfit ones. His theory is known as mutationism or mutation-selection theory. Ahead of his times, in his 1932 work *The Scientific Basis of Evolution* he also discussed about the ‘neutral’ mutations and how their fate will depend on the chance factor. He also showed that genes are carried on chromosomes and are the mechanical basis of heredity for which he received a Nobel Prize in Physiology or Medicine in 1933. He was initially against the idea that behavior of chromosomes can explain inheritance but then along with his students (Alfred Sturtevant, Calvin Bridges, Hermann Muller and others) provided the proof for the chromosomal theory of heredity, genetic linkage, chromosomal crossing over and non-disjunction. After an initial phase of acceptance till 1930s, popularity of the theory of mutationism declined as neo-Darwinism gained importance.

In the beginning of the 20th century, scientific community was broadly divided into two groups regarding understanding of mechanism of evolution. One group was of biometricians, headed by Karl Pearson, who were interested in continuously varying traits and believed in Darwinian gradualism. Other group was of Mendelians, headed by William Bateson, who emphasised importance of discontinuous variations and believed that major adaptive change could be produced by single mutational steps. Differences on such issues led to formulation of principle of Hardy-Weinberg Equilibrium which forms the backbone of population genetics. This principle, demonstrated in 1908 independently by Godfrey Hardy, an English mathematician and Wilhelm Weinberg, a German physician, is concerned with maintenance and transmission of gene and genotype frequencies over generations in Mendelian populations and the contributing factors (Hardy, 1908; Weinberg, 1908). Hardy-Weinberg principle states that after one generation of random mating, single locus genotype frequencies can be represented by a binomial or multinomial function of allele frequencies and that the gene and genotype frequencies would remain constant from one generation to the next, provided the population meets certain assumptions such that:
Introduction

- It is large in size.
- It consists of diploid organisms that reproduce sexually and randomly and exist in non-overlapping generations.
- It is not under the influence of factors such as migration, mutation and selection.

The principle focused on genetic equilibrium with respect to the gene and genotype frequencies.

Hardy-Weinberg principle led to successful deduction of genotype proportions in succeeding generation if allele frequencies are known for one generation in a randomly mating population. According to Hardy-Weinberg principle, in a randomly mating population of diploid organisms, allele frequencies for the two alleles \( A \) and \( a \) at a biallelic locus can be represented as \( p \) and \( q \) respectively. If these allele frequencies are known then frequencies of possible genotypes, \( AA \), \( Aa \) and \( aa \) can be calculated by the formula \((p + q)^2\) where,

\[
AA = p^2, \quad Aa = 2pq \quad \text{and} \quad aa = q^2
\]

If the genotype proportions in the succeeding generations are found to be similar to those in the parental generation then the population is said to be in Hardy-Weinberg equilibrium. This interpretation was important also because it provided a solution to the problem of blending that troubled Darwin. Jenkin argued that with sexual reproduction, the variation in the population would be exhausted very rapidly. But the Hardy-Weinberg principle showed that this is not so. Sexual reproduction has no inherent tendency to destroy the genotypic variation present in the population as the genotypic proportions remain constant over generations. On the other hand, if the population is not found to be in equilibrium then this implies that one or more factors are operating in the population.

Works of Darwin and Mendel were thought to be incompatible for many years but it was R.A. Fisher, J.B.S. Haldane and Sewall Wright who had the farsightedness to see that works of these two visionaries were in fact complementary and could beautifully explain the process of speciation and evolution. And thus the two concepts were integrated in early 20th century to lay down the foundation of population genetics as it is understood today. The resulting synthesis combined the postulates of evolution by
natural selection presented by Darwin with the statistical rules of genetic transmission in populations to construct a mathematical theory of evolution. These three scholars conducted extensive mathematical studies on change in gene frequencies taking into account mutation, selection and genetic drift and concluded that selection is much more effective than mutation in changing gene frequencies. And thus was laid the theoretical foundation of neo-Darwinism theory of evolution.

In his 1918 paper, *The Correlation between Relatives on the Supposition of Mendelian Inheritance*, Fisher (1890-1962) demonstrated that if a given continuous trait was affected by a large number of Mendelian factors, each of which made a small difference to the trait, then the trait would show an approximately normal distribution in a population. This was an important step towards reconciling the Darwin’s idea of selection and Mendel’s laws of inheritance as Darwinian process was widely believed to work best on continuously varying traits and it was shown that the distribution of such traits was compatible with Mendelism. At this point in time, source of genetic variation was not known but Fisher proposed that evolution occurs by natural selection in presence of variation. Later when genetic variation was shown to be generated by spontaneous mutations, Darwin’s theory was transformed into neo-Darwinism or the synthetic theory of evolution (Mayr and Provine, 1980). Fisher in his book *The Genetical Theory of Natural Selection* (1930) discussed that mutations having small phenotypic effects would get incorporated into the genotype and contribute to evolution instead of mutations with large phenotypic effects.

Wright (1889-1988) was the propounder of the inbreeding coefficient and F-statistics, now standard tools in population genetics. He was also the chief developer of the mathematical theory of genetic drift, also known as the Sewall Wright effect which is cumulative stochastic changes in gene frequencies that arise from random births, deaths and Mendelian segregations in reproduction. He was convinced about the role of genetic drift and its interaction with other evolutionary forces for adaptation to take place. He described the relationship between phenotype or genotype and fitness as fitness landscapes. His early interest in mammalian genetics led him to create the method of path analysis as a staff scientist at the US Department of Agriculture. By 1921, he had developed his method of path coefficients to describe the effects of inbreeding, assortative mating and selection (Wright, 1921). By 1931, he had
articulated his shifting balance theory of evolution in his now classic paper *Evolution in Mendelian Populations*. According to this theory, in order to evolve to a higher adaptive peak, a species would first have to pass through a valley of maladaptive intermediate stages. Genetic drift could lead to this in small populations. If a species was divided into small populations, some could find higher peaks. If there was some gene flow between the populations, these adaptations could spread to the rest of the species. Wright had a long standing debate about this with Fisher, who believed that most populations in nature were too large for these effects of genetic drift to be important.

Haldane (1892-1964) also sought to analyze natural selection in terms of mathematical consequences of Mendelian genetics. Starting with a series of papers on *A Mathematical Theory of Natural and artificial Selection* (1924-1934) and then the book *The Causes of Evolution* (1932), he showed the direction and rates of changes in gene frequencies for the first time. He also addressed the issue of interaction of natural selection with mutation and migration for the first time. Starting from simple Mendelian models using two alleles at a single locus, he went on to consider selection with self fertilization, inbreeding, overlapping generations, incomplete dominance, isolation, migration, and fluctuating selection intensities (Provine, 1971). In the appendix of his book, he compared his views to those of Fisher and Wright. While largely agreeing to both, he differed from Fisher by placing greater emphasis on strong selection of single genes, migration and epistasis. He sided with Fisher, however, in thinking that Wright put too much emphasis on random genetic drift (Provine, 1971). He introduced many quantitative approaches in biology. He propounded the maximum likelihood method for estimating human linkage maps for the first time and also introduced several methods for estimating human mutation rates. He did pioneering work on inbreeding, mating systems and genetic drift as well.

Another important intellectual difference between the pioneers of mathematical models of population genetics was regarding their respective stand towards natural selection. Fisher and Haldane were both strong Darwinians whereas Wright believed that along with natural selection chance factor also played a crucial role in determining genetic composition of a population, as did migration between the constituent populations of a species. Wright also emphasized the role of epistasis, or
non-additive interactions between the genes within a single genome, to a much greater extent than Fisher or Haldane. Work of these visionaries marked a major step forward in understanding of evolution by enabling the consequences of various evolutionary hypotheses to be explored quantitatively rather than just qualitatively. The strategy of devising formal models to shed light on the process of evolution is still the dominant research methodology in contemporary population genetics.

Although de Vries’ ‘Mutation theory’ was rejected, the idea of speciation through some type of mutation was still appealing to quite a few. Goldschmidt (1940) worked with moth *Lymantria* and concluded that microdifferentiation of polymorphic traits is different from interspecific differences of traits. He talked about microevolution and macroevolution and observed that the former type works only at the intra species level. In order for speciation to take place, he proposed the concept of systemic or macromutation that alters the chromosomal composition. But such mutations were shown to be deleterious to the organism and not helpful for the process of evolution.

Evolutionist Ernst Walter Mayr (1904-2005) extended the knowledge as to how new species arise by putting forward the concept of allopatric speciation in his book *Systematics and the Origin of Species* (1942). According to it, geographically isolated sub-populations of a species slowly accumulate differences in their genome and under the influence of genetic drift and natural selection diverge so far over a period of time that reproductive isolation occurs. Today, it is accepted as the most common mode of speciation. He also introduced the biological species concept that defined a species as not merely a group of morphologically similar individuals but as a group of interbreeding or potentially interbreeding populations that were reproductively isolated from all other populations. He was often critical of early mathematical approaches to the study of evolution.

Mayr, along with Theodosius Dobzhansky and George Gaylord Simpson, is credited with formulating Synthetic Theory of Evolution or neo-Darwinism which came to be accepted by most biologists by 1960. Dobzhansky (1900-1975) was one of the first to apply genetics to natural populations and worked mostly with *Drosophila pseudoobscura*. His 1937 work, *Genetics and the Origin of Species*, was a key step in bridging the gap between population geneticists and field naturalists. It presented the
conclusions reached by Fisher, Haldane, and especially Wright in their highly mathematical papers in a form that was easily understandable to others and disseminated information about neo-Darwinism. It also emphasized that real world populations had far more genetic variability than the early population geneticists had assumed in their models, and that genetically distinct sub-populations were important. Dobzhansky argued that natural selection worked to maintain genetic diversity as well as drive change. His work complemented that of E.B. Ford, an experimental naturalist, who wanted to test natural selection in nature. He virtually invented the field of research known as ecological genetics. His work on natural selection in wild populations of butterflies and moths was the first to show that predictions made by R.A. Fisher were correct. He was the first to describe and define genetic polymorphism, and to predict that human blood group polymorphisms might be maintained in the population by providing some protection against disease (Ford, 1942).

George Gaylord Simpson (1902-1984) was responsible for showing that the modern synthesis was compatible with paleontology in his book *Tempo and Mode in Evolution* (1944). Simpson's work was crucial because so many paleontologists had disagreed with the idea that natural selection was the main mechanism of evolution. It showed that the trends of linear progression (as seen in the evolution of the horse) that earlier paleontologists had used as support for neo-Lamarckian orthogenesis were not true. Instead the fossil record was consistent with the irregular, branching, and non-directional pattern predicted by the modern synthesis.

Several refinements have been made in the theory of neo-Darwinism over a period of several years (Dobzhansky, 1937, 1952; Muller, 1940; Huxley, 1942; Mayr, 1942, 1963; Simpson, 1944, 1949, 1953; Stebbins, 1950; Ford, 1964). It can be summarized as follows (Nei, 1987):

- Mutation occurs randomly with respect to gene function and recurs with a reasonably high frequency.
- Mutation is the primary source of variation but its effect on gene frequency change is so small that it plays a minor role in evolution.
- Because of mutations that have occurred in the past natural populations contain sufficient amount of genetic variability to respond to almost any kind of selection.
Introduction

- Evolution is determined mainly by environmental changes and natural selection. Since there is enough genetic variability, no new mutations are required for a population to evolve in response to an environmental change.
- Because mutations tend to recur at reasonably high frequencies, majority of advantageous mutations should have been fixed or reached their optimum frequencies in the population. Thus, the genetic structure of a population is almost always at or near its optimum for a given environment.
- Evolutionary change of a species occurs gradually by means of natural selection. Thus, macroevolution is nothing but accumulation of the effects of microevolution.

There were some differences in opinions of different scholars regarding some aspects of neo-Darwinism. One of the aspects was the maintenance of genetic polymorphism in the population. One group led by Dobzhansky (1955), Ford (1964) and Wallace (1968) believed that a large proportion of polymorphisms affecting population fitness are maintained by some types of balancing selection. This theory was termed by Dobzhansky as balance theory. Another view, termed classical theory by Dobzhansky, believed that these polymorphisms are maintained by mutation-selection balance. This group was led by Muller (1950) and Crow (Morton et al., 1956).

Following all these pioneering pieces of work in the first half of the 20th century that influenced how variations and evolution were looked at, molecular genetics entered into the picture in the later half and affected the modern synthetic theory. Several studies were going on in these times to understand the molecular evolution of macromolecules such as hemoglobins and cytochromes (Margoliash and Smith, 1965; Zuckerkandl and Pauling, 1965). It was observed that:

- For each protein or gene, the rate of evolution as measured by amino acid or nucleotide substitution is approximately constant per site per year for various evolutionary lineages, as long as the function of the gene remains the same.
- Also, functionally less important genes or their parts evolve faster than functionally more important ones.

It was tried to explain these observations in terms of neo-Darwinism. It was widely held at that time that a mutation cannot get fixed into the population without the
action of selection. But the contentions of neo-Darwinism such as that the rate of evolution depends on the frequency and speed of environmental changes, population size and generation time; and that most gene substitutions are supposed to occur by positive selection were contradictory to these and other observations made for the molecular evolution. Some of the other observations made regarding the molecular evolution were that:

- The mutations that impair the function of a protein occur less frequently than those that do not affect the function.
- A gene with a new function generally evolves from a duplicate gene.
- The genomes of higher organisms are highly flexible and the number of copies of a particular DNA sequence may vary rapidly under certain conditions.
- Functionally less important parts of genes are generally more polymorphic than functionally more important parts.
- The extent of genetic variability is generally higher in large populations than in smaller ones.

Motoo Kimura (1924-1994) proposed the neutral theory of molecular evolution in the late 1960s to fill the gaps in understanding of molecular evolution by neo-Darwinism (Kimura, 1983). Mutation, purifying selection and genetic drift are the corner stone pillars of the theory. Spontaneously occurring mutations can be beneficial to the organisms, increase their fitness, spread through the population overtime and become fixed. Or they may reduce the fitness and get eliminated from the population through the process of purifying selection. In third scenario, mutation may be selectively neutral i.e. may have little or no effect on the fitness. Fate of such mutations depends on genetic drift. Most are lost from the population shortly after they appear, some are lost after a few generations and a very small fraction spread throughout the population and get fixed. The rate of fixation of selectively neutral mutations and the rate of evolution is equal to the rate at which these occur in the population. Neutral theory provides an explanation for the near constancy in substitution rates of nucleotides and amino acids in lineages overtime. The theory is capable of explaining most observations on molecular evolution and variation. It has been modified by many and the currently followed theory can be summarised as follows (Nei, 1987):
The neutral theory is concerned with behaviour of majority of mutations that get incorporated into the population during evolution but does not rule out the existence of a small proportion of advantageous or overdominant mutations.

It is assumed that most new mutations are deleterious and are quickly eliminated from the populations and thus do not contribute much to genetic variation in populations.

Most nucleotide and amino acid substitutions are caused by random fixation of neutral mutations and the rate of substitution is thus equal to the rate of occurrence of these mutations.

Genetic polymorphism in a population represents a phase of the process of gene substitution. If the effects of mutation and genetic drift are balanced then expected heterozygosity per locus is given by $4Nu/ (1+4 Nu)$, where $N$ and $u$ are the effective population size and the mutation rate per locus, respectively.

Neutral alleles are not functionless genes but are of vital importance to the organism. A pair of alleles are called neutral if they are functionally equivalent and do not differentially effect the fitness of an organism. However, in population genetics, the definition of neutrality of a pair of alleles depends on the degree to which their behaviour in a population is dictated by genetic drift. A mutant gene which is advantageous in a large population may become neutral in a small population.

Parallel to the ground breaking work that was going on in the field of evolution, discoveries and inventions were being made in the field of genetics as well that were to play a complementary role in examining and understanding the aspects of human origin, adaptation, evolution and speciation; and also in developing techniques that will later help in analysing the variation. A few of these developments that are being used in population genetics in some way have been discussed below.

For many years protein was thought to be the carrier of genetic information but it was O.T. Avery (1877-1955) who demonstrated in 1944 along with his colleagues MacLeod and McCarty that Fred Griffith’s “transforming principle” was DNA i.e. DNA is the hereditary material. They conducted their experiments on R strain bacteria (Avery et al., 1944).
Barbara McClintock (1902-1992) did pioneering work in plant genetics and during the 1940s and 1950s discovered the mechanism for transposition while working on corn for which she won Nobel Prize for Physiology or Medicine in 1983. She showed how genes are responsible for controlling the physical characteristics. She worked with maize and identified its chromosomes and mapped them in linkage groups. She demonstrated that crossing over occurs in the chromosomes during meiosis, now a fundamental concept in genetics.

Linus Pauling (1901-1994) brought about a revolution in the way diseases were looked at by showing for the first time that a protein can be causally linked to a disease. Using electrophoresis, Pauling along with his colleagues, showed that individuals with sickle cell disease had a modified form of hemoglobin in their red blood cells (Pauling et al., 1949). Sickle cell anemia became the first disease to be understood at molecular level. This result also demonstrated for the first time that genes determine the structure of proteins, not simply their presence or absence.

Double helical structure of DNA was determined in 1953 by James Watson (1928-) and Francis Crick (1916-2004) for which they received 1962 Nobel Prize in Physiology or Medicine along with Maurice Wilkins (Watson and Crick, 1953). Rosalind Franklin was also involved in the discovery but could not be honoured with the Nobel Prize because she died before the award was bestowed upon the team. Nature journal’s 1953 April issue featured five research papers by the team discussing the discovery. The discovery has been described as the defining moment in nucleic acids research. It was revealed that the hereditary molecule is made up of two chains coiled around each other. Each chain is made up of a sugar-phosphate backbone on the outside and each sugar molecule is attached to one of the four bases, Adenine, Guanine, Cytosine and Thymine on the inside. The two strands are bound to each other through non-covalent hydrogen bonds that form between the bases on the two strands. Adenine on one strand bonds with Thymine on the other strand and Guanine with Cytosine. This arrangement is known as complementary base pairing. The two strands run in opposite directions, one in 5′-3′ orientation and the other in 3′-5′ orientation i.e. are anti-parallel. The breakthrough was made possible by the previous developments such as X-ray crystallography that made possible studying organic macromolecules; the realisation that the hereditary material was DNA, not protein;
experimental finding of Erwin Chargaff that there are equal numbers of A and T bases and of G and C bases in DNA; and Linus Pauling’s discovery that the molecules of some proteins have helical shapes. This structural knowledge of DNA, as revealed by the double helical model, led to new understanding of heredity and hereditary diseases. It shed light on and paved way for understanding of processes such as replication, transcription etc. that will later play a major role in development of the field of biotechnology as well.

Frederick Sanger (1918- ) is credited with pioneering the technique of sequencing that led to determination of base sequences in DNA. His efforts in the Laboratory of Molecular Biology in Cambridge where Francis Crick, John Kendrew, Aaron Klug and others were all working on a DNA-related problem led to the development of dideoxy technique of sequencing, the one most commonly used today (Sanger et al., 1977). For this he won a Nobel Prize for Chemistry in 1980 sharing it with Walter Gilbert, for their contributions concerning the determination of base sequences in nucleic acids. He initially examined ways to sequence RNA which is a smaller molecule. Earlier he had won a Nobel Prize for Chemistry in 1958 for his work on the structure of protein. He was the first person to obtain a protein sequence and prove that proteins were ordered molecules and by analogy, the genes and DNA that make these proteins should have an order or sequence as well. In 1992, the Wellcome Trust and the Medical Research Council established the Sanger Centre which was one of the main sequencing centres of the Human Genome Sequencing Project and is one of the premier institutes in sequencing today.

Kary Mullis (1944- ) revolutionized DNA technology by conceiving the idea of polymerase chain reaction (PCR) in 1983, a technique that allows for exponential amplification of specific DNA sequences from very small amounts of genetic material (Mullis, 1990). He received a Nobel Prize in chemistry in 1993 for his contribution. He invented PCR while working at Cetus Corporation, San Francisco, California where he was carrying out research on synthesis of oligonucleotides - short DNA sequences of up to 20 nucleotide bases. Oligonucleotides later became an important component of PCR. The various components of PCR had been discovered/ invented by 1980 and were known to the scientific world. Product of PCR is an almost unlimited amount of highly purified DNA molecules suitable for analysis or manipulation. The technique has found
widespread applications, including the screening of genetic and infectious diseases, the reconstruction of phylogenetic trees, and forensic analysis.

Understanding of human genome and its functioning has improved over the years and it has been realised that genetic variations play a major role in disease etiology, including those of common diseases. Several such diseases have been observed to have population specific pattern of distribution and severity. This has led to application of principles of population genetics in health research as well. With increasing use of molecular tools and techniques in Anthropology to forward its goals and applications, a sub-branch known as Molecular Anthropology has emerged over the years. Its role is not limited to gaining understanding of human origin and evolution. Rather, the applications are far reaching. Two of the many areas of research and application can be broadly outlined as follows:

- **Understanding of population diversity and dynamics**: Use of various kinds of molecular markers to facilitate understanding of various questions pertaining to human origin, adaptation evolution and migration; intra- and inter-population and species relationships. The data so generated, in conjunction with archaeological and palaeoanthropological evidence, has facilitated this examination.

- **Epidemiological studies**: From an applied perspective perhaps the most important contribution of molecular anthropology has been in the field of epidemiology or biomedical research. Several population based studies have led to mapping of loci causing diseases. Several marker loci have also been identified that influence susceptibility to specific diseases in populations. With the realisation that populations exhibit genetic differences with respect to susceptibility towards specific diseases, molecular anthropology has gained immense importance as seen from rise in number of population based disease association studies. This is particularly relevant in context of rising occurrence of common diseases.

### 1.2. Tools of Population Genetics

It is well known that the branch of taxonomy was established on the basis of morphological traits as seen in the 18th century work of Carl Linnaeus when he used observable characters in plants to understand their systematic relationships. Visible
characters in plants were used by Gregor Mendel to propound his principles of inheritance. Their importance was equally evident in human population studies. In the beginning, anthroscopic and anthropometric techniques were made use of to describe variations in the extant population groups and to classify them into various races (Risley, 1915; Eickstedt, 1934; Guha, 1935; Sarkar, 1954; Bhasin et al., 1994). But these variables are known to be multi factorial and affected by environment. Also, such observations are restricted to only a few traits and these traits display a low degree of polymorphism. These limitations led to focus being shifted to genetic tools for the purpose of human evolutionary and diversity studies. With the advent of genetic tools, first the classical and then molecular genetics, the whole field has received an impetus; more so with molecular genetics techniques.

1.2.1. Population Genetics and the Classical Approach

Discovery of first serological (mainly blood cell polymorphisms) and then biochemical markers (protein and red cell enzyme polymorphisms) that could be easily scored by electrophoresis in the 20th century greatly facilitated anthropological studies at phenotypic level. It all started with the discovery of ABO blood group system in humans by Landsteiner in 1901 and subsequent discoveries of AB blood group (Decastello and Sturli, 1902) and further subgrouping of A group (Dungern and Hirszfeld, 1911; Friedenreich, 1936). Observation of polymorphism vis-a-vis this marker system in world populations (Hirszfeld and Hirszfeld, 1919) preceded understanding of its mode of inheritance (Bernstein, 1924; Thomsen et al., 1930; Friedenreich and Zacho, 1931) and paved the way for classical genetics and its application to differentiate populations. It also marked the advent of monogenic traits as tools of studying population history, diversity and evolution. Subsequent discoveries and associated polymorphisms of other blood group systems such as MN (Landsteiner and Levine, 1928), Rh (Landsteiner and Weiner, 1940); PTC tasting ability (Fox, 1932); Red cell enzyme systems such as Glyoxalase system (Komph et al., 1975), Glucose phosphate isomerase system (Lohmann, 1933) and many others brought about a revolution in the field of population genetics as now several polymorphic markers were available to make use of in studying human populations and differences between them. Elucidation of genetic basis of these marker systems
and further developments in the field strengthened their stand in human population studies. These markers were widely used to get insights into the origins and affinities of population groups as seen from innumerable studies carried out using these markers. Starting in 1950s, discovery of techniques like starch gel electrophoresis, polyacrylamide gel electrophoresis, western blotting, southern blotting and others greatly facilitated the analysis of serological and biochemical markers.

From examining the geographical distribution of gene frequencies at only a few loci to understand population relationships and dynamics, in later years focus shifted to studying several loci together with the realisation that each locus has a different distribution (Cavalli-Sforza and Edwards, 1964) and cannot give a clear picture by itself. It was also observed that all populations showed much greater intra-population variation than inter-population variability (Lewontin, 1972; Nei and Roychoudhury, 1972, 1974, 1982).

Affects of evolutionary forces on population structure were widely studied using these traditional markers. These studies on traditional markers mainly focused on finding pattern in allele frequency distribution in different population groups within the context of ethnicity, language and geography. They aimed at quantifying the similarities and differences between populations at local and regional level; and also understanding their inter-relation. Several such studies have been compiled by many (Mourant et al., 1976; Roychoudhury and Nei, 1988; Cavalli-Sforza et al., 1994). Bhasin et al. (1992) did an extensive compilation of studies on morphological and classical genetic markers carried out among the populations of Indian subcontinent. Quite a few hybrid populations were examined using these markers to quantitatively estimate admixture among them (Cerda-Flores et al., 1987, 1992; Kshatriya, 1995; Simmons et al., 2007; Gauniyal et al., 2008).

Several studies were also undertaken to examine the relation of these traditional/classical marker systems with infectious, digestive, immunological and other diseases (Weiner and Peters, 1940; Aird et al., 1953; Giblett et al., 1972; Mourant, 1972; Chen et al., 1974; Mourant et al., 1978; Mourao and Salzano, 1978; Murty and Padma, 1984; Vogel and Motulsky, 1997). While some studies were conclusive as to their findings others were merely speculative. Nonetheless, for a very long time these
classical genetic markers helped the population geneticists in gaining an understanding of human populations and the dynamics operating within them.

Malhotra and Vasulu (1993) summarised the findings from the anthropometric and classical genetic studies carried out on Indian populations as follows:

- There is a lot of variation in the Indian populations cutting across linguistic and geographic boundaries; although geographically contiguous populations tend to be genetically more similar than those with linguistic affiliations.
- Apart from Africa, India harbours more genetic diversity than other comparable regions in the world.
- Tribal populations are morphologically and genetically different from non-tribal populations. Tribal groups of south India are different from those of central and north-eastern India.

1.2.2. Population Genetics and the Molecular Approach

In the 1960s attention shifted from studying population structure and affinities using classical genetic markers to studying variation at the most fundamental level, DNA. Such a transition was natural as it was realised that phenotypic expressions observed in terms of classical genetic markers are in fact manifestations of genes. The progression from phenotypic level to genotypic level has made it possible to not only study the variation but also the underlying factors that result in particular genetic constitution. Use of molecular tools and techniques to address the research issues in anthropology has not made the inputs from the field of palaeoanthropology and classical genetics redundant. Rather, they have given higher degree of resolution and have expanded and diversified the scope of studies investigating origin and evolution, both at micro and macro levels; and understanding migration patterns in quantitative, spatial and temporal frameworks. They have given unprecedented clarity and opportunity for testing various anthropological hypotheses. The findings from palaeoanthropology and classical genetics laid down the foundation and direction for the endeavours on understanding of human origins and existence.

Use of molecular techniques for studying phylogenetics started with the exploration of amino acid sequences in proteins. Relation of number of amino acid substitutions
for a protein between species and the concept of molecular clock propelled the evolutionary studies (Zuckerkandl and Pauling, 1962; Dayhoff, 1969). Analyses of variation at level of amino acids substitutions was the driving force behind the evolutionary studies at intra and inter-species levels.

DNA (Deoxyribose Nucleic Acid) became the next molecule of interest in population genetic studies. It can broadly be categorised into genomic or nuclear and mitochondrial DNA (mtDNA). Several classes of markers are now known to be present in the human genome, each with their own characteristics and use for population genetics and forensic studies.

1.2.2.1. Molecular Markers

Although ~99% DNA is known to be similar between individuals but still sequence differences exist between individuals in non-coding regions of the genome and such polymorphic regions are useful for various kinds of analyses in population genetic studies. A ‘marker’ is defined as a nucleotide sequence with known physical location in the genome and associated with a gene or a trait. It can be described as an observable variation caused due to mutation or alteration in the locus. A genetic marker can be a nucleotide sequence of variable length, varying from a single base pair to several hundred base pairs. Selection of markers for any study is dictated by the nature and purpose of the study. The more commonly used markers in population genetics studies can broadly be grouped as follows:

1.2.2.1.1. Repetitive DNA Sequence Variants

a. Tandem Repeats

Besides the interspersed repeats (SINEs and LINEs), Tandem repeats are the other kind of repeated elements found in the genome. These are highly variable tandemly repeated arrays of 2 or more base pair core units in the non-coding regions of the genome and are located adjacent to each other. On the basis of size of the core unit, they are categorised into minisatellites (10-60 bp\(^1\), Short tandem Repeats (STRs) or

\(^1\) Base pairs
Introduction

Genetic Variation in the Populations of Western India

Microsatellites (<10 bp). When the number of nucleotides in the core unit is not known or is variable then it is called Variable Number Tandem Repeats (VNTRs). Because of enormous intra- and inter-individual variability in the number of times these core units are represented in a specific sequence and their high mutation rate, these markers are useful in paternity testing, other forensic analysis, genetic mapping and evolutionary studies especially to explore recent population history and have been extensively studied (Bamshad et al., 2001; Kashyap et al., 2004, 2006; Reddy et al., 2005; Watkins et al., 2005).

b. Insertion/ Deletion Polymorphisms

An InDel or Insertion-Deletion polymorphism refers to insertion or deletion of a DNA sequence of variable length in the genome. The concerned DNA sequence may vary in length from a single nucleotide to several hundred nucleotides. They are widely spread across the genome and constitute around 1.5 million of more than 10 million polymorphisms known in humans (Mills et al., 2006). Indels can usually be genotyped with simple methods based on size separation. Also, the chances that two indel mutations of exactly the same length will occur at the same genomic location is very less and hence may indicate identity-by-descent, a characteristic especially useful for phylogenetic studies.

_Alu_ InDels – _Alu_ Insertion/ Deletion polymorphisms (_Alu_ InDels) involve _Alu_ sequences that are characterized by the cleavage action of _Alu_I restriction endonuclease (Houck et al., 1979) that was first identified in and isolated from the bacteria _Arthrobacter luteus_. _Alu_ sequences, ~300 bp in length, are present in introns and are the most abundant subtype in the category of Short Interspersed Nuclear Elements (SINEs) and account for ≥ 10% of the human genome (Deininger and Batzer, 1993; Smit, 1996). There are over 1 million _Alu_ sequences interspersed throughout the human genome (Rinehart et al., 1981) however, less than 0.5% are polymorphic. _Alu_ sequences are thought to be derived from the 7SL RNA gene which encodes the RNA component of the signal recognition particle that functions in protein synthesis. They have the necessary sequence to be transcribed by RNA Polymerase III and the LINE L1 supplies the retrovirus enzyme reverse transcriptase needed for _Alu_ to jump. This entitles this “jumping gene” to be classified as
The origin and amplification of Alu elements are evolutionarily recent events coinciding with the radiation of primates in the past 65 million years (Deininger and Daniels, 1986). Since then they have undergone amplification from a few master genes (Deininger et al., 1992) to their current repetitive status at a rate of approximately $8 \times 10^3$ Alu elements per year with about $5 \times 10^{-7}$ nucleotide substitutions per year (Zuckerkandl et al., 1989). Most of the Alu elements are not identical and can be classified into different families (Deininger et al., 1993) that were generated post mutation within the ‘master’ or ‘source’ gene. Retroposition has taken place at different rates and time periods of primate evolution (Deininger et al., 1992) and the present rate has witnessed a 100-fold reduction from its peak early in primate evolution (Shen et al., 1991). Categorisation of Alu sequences into families and subfamilies is dependent on the rate of divergence between them; 20-30% or less leads to grouping into same families and 4% or less leads to enlisting in the same subfamily.

Most human Alu sequence insertions can be found in the corresponding positions in the genomes of other primates, but about 2,000 Alu insertions are human specific (Batzer et al., 1991) and are actively retrotransposing (Wallace et al., 1991). A few of these have inserted recently, within the last 1 million years, during evolution and dispersion of modern humans and are not fixed in the human species (Batzer et al., 1991) and hence are polymorphic and useful for diversity studies and for reconstructing human prehistory. Most of these recently integrated Alu elements belong to closely related ‘young’ subfamilies such as Ya5 and Yb8 (Batzer et al., 1994, 1995).

Along with other transposable elements, Alu elements are thought to have played an important role in evolution by creating new mutations and gene combinations. Barbara McClintock hypothesized that transposable elements provide a mechanism to rapidly reorganize the genome in response to environmental stress.

Alu InDels have been widely used in human population genetics for studying evolution and diversity and forensic analysis because of their known ancestral state which is mostly absence of insertion and direction of mutational change is thus known. Since the likelihood of independent insertion at the exact same chromosomal site is practically zilch (Batzer et al., 1991), all loci carrying a particular polymorphic
insertion are derived from a unique event and are thus identical by descent (Batzer et al., 1996; Stoneking et al., 1997). This property of \textit{Alu} sequences facilitate accurate rooting of population networks as they would reflect population relationships better than other markers that may be identical by state. Also, once inserted into the genome they are invariably maintained and transmitted to the next generations i.e. are highly conserved and stable through evolutionary time.

Most of the \textit{Alu} insertions occur in non-coding regions of the genome and are selectively neutral but are sometimes disruptive and can result in inherited disorders. \textit{Alu} sequences inserted into exons; or unequal homologous recombination events between \textit{Alu} repeats such as those in NF-1, TPA and ACE genes have been implicated in several inherited human diseases (~0.4%; Deininger and Batzer, 1999), including various forms of cancer, Familial hypercholesterolemia, Haemophilia, Neurofibromatosis, Diabetes mellitus type II. However, most \textit{Alu} insertions act like markers since they may segregate with a disease allele.

In a nutshell, properties of Alu sequences such as their known ancestral state, identity by descent, wide occurrence and stability make them ideal markers for human evolutionary and diversity studies.

1.2.2.1.2. Non-Repetitive DNA Sequence Variants

\textbf{a. Single Nucleotide Polymorphisms (SNPs or Snips)}

SNP or Single Nucleotide Polymorphism is a single nucleotide (base pair) change in a DNA sequence. As with all polymorphisms, for an alteration to be considered a snip it must be present in \( \geq 1\% \) of the population being considered. They make up about 90\% of all the human gene sequence variation. SNPs may be present in coding regions (exons) or non-coding regions (introns) or intergenic regions. They can be \textit{synonymous} or \textit{non-synonymous}. The former ones do not lead to change in resultant amino-acid but the latter ones do thereby leading to an abnormal phenotype. \textit{Snips} can lead to diseases (especially the ones in coding regions) and even the ones in intronic regions can affect gene-splicing and transcription-factor binding. They are most commonly examined using the method of Restriction Digestion. Novel SNPs are detected using sequencing. They have implications for both evolutionary (Brumfield et al., 2003) and clinical studies.
**Restriction Fragment Length Polymorphisms (RFLPs)** are the characteristic pattern of fragments of DNA produced when a DNA sequence is cleaved by specific enzymes belonging to endonuclease class of enzymes. The property of these enzymes that enables them to cleave DNA segment only at specific locations known as restriction sites have led to their use in detecting genetic differences on the basis of absence or presence of restriction sites. Genetic differences such as SNPs leading to absence or presence of restriction site cause changes in length of fragments generated that are then detected using electrophoresis. Initially, they were analysed using tedious methods such as Southern blotting and hybridisation (Botstein et al., 1980). But now other forms of electrophoresis such as agarose gel electrophoresis are commonly used. Single nucleotide changes in DNA sequences can thus be detected by using restriction enzymes. Polymorphism of this kind was first observed with \textit{HpaI} enzyme in the region near human β-globin gene (Kan and Dozy, 1978). RFLP analysis was the first DNA profiling technique and has played an important role in genome mapping, localization of genes for genetic disorders, and paternity testing. Several hundred polymorphic restriction sites are now known to exist in human genome that have been extensively used in population genetics studies.

1.2.2.1.3. Lineage Markers

a. Mitochondrial Markers

Maternally inherited mitochondrial genome consists of multiple copies of circular mitochondrial DNA or mtDNA. Markers present on this haploid genome are primarily used for tracing maternal ancestral lineage(s) in populations because of their uniparental inheritance. Because of this difficulties in analysis because of recombination during sexual reproduction do not come up. mtDNA is known to evolve about five to ten times faster than nuclear DNA (Brown et al., 1982) and it thus accumulates molecular changes faster allowing for analysis of recent evolutionary events. Hyper variable regions (HVRI and II) are especially useful because of their highly polymorphic nature. Some of the diseases such as Leber’s Optic Neuropathy and others are maternally inherited and can thus be characterized using mtDNA. mtDNA, thus, have implications for both disease and evolutionary studies.
b. Y-chromosomal Markers

Like mtDNA, Y-chromosome has a uniparental inheritance but in the male line and can thus be used for tracing paternal ancestral lineages. In absence of recombination, Y-chromosome is more or less transmitted unchanged from one generation to next and the few changes that may occur usually do not have any effect as around 98% of the DNA is in non-coding region. More than half the chromosome is made up of heterochromatin and thus does not contain genes. The chromosome has a variety of markers such as SNPs, STRs, Alu InDels etc. Y chromosomal markers have found use in examining migration patterns and genealogical histories. They are also useful for paternity testing and in forensics such as in sexual assault and missing person cases.

Since mtDNA and Y chromosome markers have uniparental mode of inheritance, they play an important role in understanding migration patterns that have contributed to the present day population scenario. Also, they are useful for studying inter-population differences and relationships (Bamshad et al., 1996, 1998; Basu et al., 2003; Kivisild et al., 2003; Metspalu, et al., 2004; Quintana-Murci et al., 2004; Thangaraj et al., 2005; Sahoo and Kashyap, 2006; Reddy et al., 2007).

Documentation of diversity using molecular markers has been facilitated by use of certain techniques that have evolved overtime to give greater resolution and have become more user friendly, prolific, and accurate.

1.2.2.2. Laboratory Techniques

Foremost requirement for carrying out molecular analysis of any kind is the availability of the genetic material. As mentioned earlier, DNA is the focal point of human diversity and disease-association studies by virtue of the fact that it is the blueprint of our existence. As almost all cells in our body contain DNA, it can be extracted from a number of sources such as cheek cells, body fluids of various kinds, tissues etc. Blood is one of the most commonly used source material for obtaining high molecular weight DNA. There are several techniques for isolating DNA such as manual methods (like Phenol Chloroform, Salting-out) and kits. The technique of DNA isolation or extraction varies depending on the starting material. The isolated DNA contains the entire genome of the organism and this is problematic when only
certain specific regions of the genome are to be examined. Also, for the purpose of analysis, target genomic region is required in sufficient amounts and thus comes the technique of PCR in to the picture.

**Polymerase Chain Reaction (PCR):** Invention of PCR is one of those inventions that have had an unprecedented impact, with its rapid incorporation into varied research fields and its wide applications. It has developed into one of the most essential techniques used in molecular work of any kind by enabling *in vitro* availability of genetic material by domesticating the process of DNA replication. It involves cycling DNA sample through a series of heating and cooling cycles with the required raw materials and enzymes to achieve its exponential amplification. Knowledge and understanding of properties of DNA molecule (such as complementary base pairing, its synthesis), availability of primers and the discovery of polymerases laid the foundation for invention of the technique in 1980s by Kary Mullis and others. The technique has come a long way since its invention. Instead of having to manually maintain the heating and cooling cycles, automated thermal cyclers are now available; and instead of having to add fresh polymerase (earlier derived from *E. coli*) after every cycle because of its denaturation due to heating, thermally stable DNA polymerases such as *Taq* DNA Polymerase are now made use of. There are several variations available today in the basic type of PCR that can be chosen depending upon the kind of analysis desired. Some of them are Allele-specific PCR, Hot-start PCR, Asymmetric PCR, Helicase-dependent amplification (Isothermal amplification), Strand –displacement amplification, Long-range amplification etc.

Amplification of DNA by PCR has found applications in a variety of fields ranging from forensics to archaeology; study of variation and evolution to mutation detection; gene mapping and cloning and DNA sequencing to epidemiology among several others. It facilitates observation and analysis of ancient DNA and also aid in examination of sequence differences associated with genetic disorders. It allows isolation of a particular region of genome and its further examination. Small quantity of available material can be exponentially amplified using PCR making it especially useful in forensic analysis. It can also be used to quantify the DNA present in a sample because the time required to amplify DNA to a particular level depends upon the initial amount present. It has also found immense use in recombinant DNA technologies.
Restriction Digestion: It is the method of cutting DNA sequences into fragments using restriction endonucleases or enzymes that cut at specific recognition sites. This generates DNA fragments of varying lengths producing a variation pattern known as Restriction Fragment Length Polymorphisms (RFLPs). The variation may be produced in response to absence or presence of particular SNP(s) or an insertion or deletion event in that region and is recognised in the form of banding pattern. Resulting fragments are separated according to molecular size using gel electrophoresis. There are several classes of endonucleases- Type I, Type II, Type III and Type IV but the most commonly used restriction enzymes are of type II and they cleave DNA fragment at specific sites within or close to the recognition sequence. Most of these enzymes cut palindromic sequences.

The technique is useful in detection of mutations/ SNPs. It is also used to detect VNTRs. The technique has been widely used for constructing physical maps of the genome, genetic linkage maps; in forensic testing; and in epidemiological and evolutionary studies. Restriction enzymes, in combination with ligases, have also been widely used in the field of genetic engineering and recombinant DNA technology. Werner Arber, Daniel Nathans, and Hamilton Smith received the 1978 Nobel Prize in Medicine for discovering the restriction endonucleases.

Electrophoresis: It is one of the few techniques that has been in use since the beginning of study of classical genetic markers and is still in use for molecular markers. It is the method of separating macromolecules (both proteins and nucleic acids) on the basis of size, electric charge or other physical properties under the influence of electric field. This electrokinetic phenomenon was observed for the first time in 1809 by Reuss. There are several forms of electrophoresis such as agarose gel, cellulose acetate membrane, polyacrylamide gel, SDS-PAGE, electrofocusing, capillary electrophoresis etc. From primary dependence on various forms of electrophoresis to distinguish between different kinds of macromolecules, availability of other advanced techniques has enabled generation of data at a much larger scale and greater depth and resolution.

Sequencing: DNA sequencing refers to establishing the exact sequential arrangement of bases in a stretch of DNA. Knowledge of exact sequence of bases in a gene is crucial especially in ascertaining the function of genes. This is also important as the
disease-causing alterations in the genes can then be identified. Once considered implausible, scientists have been able to sequence genomes of several organisms. Beginning with complete sequencing of phage \( \phi X174 \) in 1976, genomes of several viruses, plasmids, bacteria and several eukaryotes including human have been sequenced. The technique has seen tremendous development in a short span of time. Sequencing using automated sequencers has replaced the tedious and slow manual sequencing of the past. Use of automated sequencers that can sequence long stretches of DNA in short durations has revolutionised the research field. The techniques have also become comparatively more economical than before. Developments in sequencing technologies have made possible the execution of large scale genome projects and genome wide studies that have played crucial role in understanding of human genome and have had phenomenal impact on applications in both evolutionary and disease association studies.

1.3. Statistical Measures

Generation of data using various tools and techniques of population genetics, whether classical or molecular, will not serve the purpose if the data cannot be quantitatively assessed to yield meaningful and comparable measures. Over a period of time several statistical measures and models have been used to estimate genetic variation within and between population groups and its implications for the studies on population structure and human evolution. These statistical measures have evolved overtime and become more sophisticated to incorporate the increasingly complex data being generated with advanced techniques.

Gene or allele frequency is the most basic statistical measure that is estimated using the data generated from laboratory analysis. In early days, it was attempted to study population affinities using allele frequency data on a few loci but with the realisation that different loci have different frequency distributions, large number of loci came to be analysed in order to get a clearer picture and to detect even small differences. All other statistical measures are built up on allele frequency data. Depending on the research questions, different measures are estimated. Genetic distance is commonly used to detect population affinities by measuring differences between a pair of populations. There are several ways in which genetic distance can be measured. It was
first used by Sanghvi (1953). Based on number of gene and codon substitutions per locus between two populations, Nei also devised a genetic distance measure (1971, 1972). These differences can then be used to study population relationships by making use of dendrograms, principal component analysis, principal coordinate analysis etc. (Cavalli-Sforza and Bodmer, 1971; Sneath and Sokal, 1973; Nei, 1987). Several measures have been proposed to study genetic differentiation. One of the most commonly used and appropriate measure of genetic variation is average heterozygosity or gene diversity (Nei, 1973). Wright’s Fixation Index (1921) is an important index of genetic differentiation as it allows an objective comparison of the overall effect of population substructure among different organisms without getting into details of allele frequencies, observed levels of heterozygosity etc. (Hartl et al., 1997). Wright (1943, 1951) proposed to measure the deviations of genotype frequencies in a subdivided population in terms of three parameters or fixation indices, F_{IS}, F_{IT} and F_{ST}. These fixation indices or F-statistics are useful for understanding the breeding structure of populations or the pattern of selection associated with polymorphic alleles (Nei, 1987). With increasing use of linked markers in population studies, other parameters that are being used for examining population histories and localise genetic variations underlying complex diseases are haplotype frequency and Linkage Disequilibrium (LD). Different measures of LD have been proposed, each with its inherent assumptions. Most commonly used measures of LD are D’ (Lewontin, 1964) and r^2.

Use of different statistical measures is essentially dependant on the type of data available and research questions. In order to maximise the gains, there should be compatibility between the research questions, type of markers selected and statistical parameters used.

1.4. Evolutionary and Genome Diversity Studies

One of the pertinent questions regarding human evolution is when and where did ‘modern’ humans originate. This has been the core area of speculation, research and debate in anthropology since the beginning. Earlier the answers were searched in the fossil records, but lately attention is being given to genetic data as well, not just from living populations but also from fossils. While it is generally agreed that the earliest
Introduction

members of the genus *Homo* and the preceding hominins evolved in Africa, the debate is still on as to the origins of anatomically modern humans or *Homo sapiens*.

There are currently two models popular for explaining the evolution of modern *H. sapiens*. According to one of the models, the Multiregional Evolution or Regional Continuity Model, *H. erectus* radiated out of Africa between Early to Middle Pleistocene to inhabit different parts of the Old World. These sub populations then evolved to have regional morphological characteristics that persist today also but at the same time maintained gene flow amongst each other to have evolved into the same species, *sapiens* (Wolpoff et al., 1984; Frayer et al., 1993). On the other hand, the second model, Out of Africa Model or Recent African Origin Model, postulates that *H. sapiens* evolved in Africa to attain characteristics of modern humans, differentiated regionally there and then migrated and replaced other populations of *H. erectus* or other archaic species of *Homo* residing in other parts of the Old World. The model assumes that very little, if any, hybridisation took place between the archaic and the modern species.

Variations of these two extreme models have also been suggested (Aiello, 1993). One model, The Hybridisation and Replacement Model, is slightly different from the Recent African Origin Model in that it allows for some degree of hybridisation between the migrating modern humans and the native pre-modern humans. The other model, Assimilation Model, accepts the African origin for modern humans but emphasises the importance of gene flow, admixture, changing selection pressures and resulting directional morphological changes (Smith, 1992). With passage of time and accumulation of data, propounders of Multiregional Evolution have altered their stance and begin to incorporate the importance of African origin in their model (Stringer, 2002). The original Multiregional Evolution Model no longer seems tenable. Increasingly, the archaeological and genetic evidence is favouring African origin of *sapiens* followed by population bottleneck(s) that led to dispersal of the modern humans outside Africa.

Pioneering work on reconstructing human evolutionary history using genetic data started with the works of Cavalli-Sforza and Bodmer (1971) and Nei and Roychoudhury (1982) but beginning of 1980s saw greater resolution with use of
genetic systems such as beta-globins (Wainscoat et al., 1986). They made use of RFLPs in a population based approach and showed that African populations are quite different from non-African ones. Cann et al. (1987) followed this up with their study on mtDNA and showed that mtDNA diversity in modern humans is relatively low as compared to other primates and put forward their haplotype data as evidence for Out of Africa Model. Many studies have focussed on aspects such as estimating coalescent times (Ingman et al., 2000), reconstructing ancient demographic patterns (Harpending and Rogers, 2000) and mapping ancient dispersal events (Richards et al., 2000; Underhill et al., 2001). Several studies have made use of different genetic systems and come up with findings that support Out of Africa Model. Studies on mtDNA (Vigilant et al., 1991), Y- chromosome DNA (Hammer et al., 1997; Underhill et al., 2000; Jorde et al., 2000, Ke et al., 2001), autosomal microsatellites (Bowcock et al., 1994; Jorde et al., 1997; Calafell et al., 1998; Deka et al., 1999), autosomal Alu polymorphisms (Watkins et al., 2001), nuclear DNA sequences (Nickerson et al., 1998; Halushka et al., 1999; Rieder et al., 1999; Labuda et al., 2000; Wall and Przeworski, 2000; Yu et al., 2001) have found that maximum genetic diversity is found in Africa. This is consistent with the theory that a subset of African population, with consequently a subset of African genetic variation, left Africa to colonise the rest of the world. But it should be kept in mind that genetic diversity is also dependent on the effective population size. It has also been found that the root of the human phylogenetic tree falls nearest or within African populations (Penny et al., 1995; Zischler et al., 1995; Nei and Takezaki, 1996; Ingman et al., 2000; Thomson et al., 2000; Underhill, 2000; Watkins et al., 2001). Also, linkage disequilibrium values have been found to be usually lower in African populations (Tishkoff et al., 1996, 1998, 2000; Kidd et al., 1998, 2000; Jorde et al., 2000; Gabriel et al., 2002). Lower values of linkage disequilibrium signify that population has been in existence for time sufficient enough for recombination to act and break the LD, indicating longer population history. It has also been observed that haplotype variation is higher in Africa and haplotypes found outside are a subset of the same; and geographical areas across the world share many haplotypes (Armour et al., 1996; Tishkoff et al., 1996, 2000; Hammer et al., 1997; Calafell et al., 1998; Fay and Wu, 2000; Underhill et al. 2000). Most ancestral forms of nuclear genes have also been found to persist in Africa (Takahata et al., 2001; Satta and Takahata, 2004). Findings from a few studies do not
support the recent African origin (Templeton, 1997; Hawks et al., 2000a, b; Wolpoff et al., 2000; Zhao et al., 2000). Templeton (2002) proposed a model in which out of Africa expansion occurred more than once along with gene flow and not replacement. In this model also, impact of African population on human gene pool is found to be greater as the ancestral haplotype lineage can be traced back to Africa for most of the studied loci.

Archaeological and genetic data has shown that migrants from Africa increased in number in succeeding generations and constituted modern humans (Oestigaard et al., 2004; Ramachandran et al., 2005; Rosenberg et al., 2005). Evidence has been found for the peopling scenario where the migrants out of Africa inhabited an area, underwent expansion and then a subgroup moved further to occupy new area (Ramachandran et al., 2005). In this scenario it is also expected that evolutionary accumulation of new genetic variations and founding of new populations will lead to increasing distance between the source population and descendant population, which will increase with geographical distance. This has been shown to be true by Lahr and Foley (1994). Geographical barriers, in conjunction with local cultural norms, led to further differentiation between populations.

Synthesis of studies based on classical genetic markers by Cavalli-Sforza et al. (1994) remains till date the most comprehensive analysis of allele frequencies across the world. In their analysis they found that south Asia is positioned between west and southeast Asia; and is characterised by high degree of geographic microdifferentiation. This variation is partly a result of influx of genes into the subcontinent at various points in time; and partly an outcome of practice of endogamy, characteristic of the subcontinent populations, that results in relative isolation and genetic drift within the groups (Majumder, 2008, 2010). Cavalli-Sforza et al. (1994) also observed that in the Indian context language difference accounts for most of the genetic variation but at the same time there is confounding between language and geography.

South Asia, including India, because of its geographical location was probably the first area to be inhabited by the early migrants and also served as a corridor for dispersal to other regions (Cann et al., 2001). This is based on the finding that India harbours genetic diversity that is next only to Africa and thus may have been
inhabited during one of the early waves of migration. Various studies conducted in the Indian subcontinent have favoured this view. By virtue of the fact that higher diversity was observed among the tribals than the caste populations, Basu et al. (2003) suggested that tribals are the native populations of India. This finding has been reiterated by many studies, most recently by Indian Genome Variation Consortium (2008).

1.5. A Look Inside the Genome

Past two decades have witnessed unprecedented advances made at large scales in cataloguing and understanding various aspects of human genome. The major landmark attempts that have been made are discussed below.

1.5.1. Human Genome Project (HGP)

A National Institute of Health (NIH, US) initiative started in 1990, HGP was a multinational collaborative project aimed at identifying all the genes in the human DNA and determining the sequence of about 3 billion nucleotide pairs that constitute the human DNA to understand the species’ genetic makeup. Genomes of a few model organisms such as fruit fly, E. coli and others were also investigated for comparative purposes. Findings from this mega project have been released in several parts over the years. First draft was released in 2001 followed by the complete draft in 2003. In 2003, entire April month issues of journals Science and Nature featured insights from the findings of the project and were titled Building on the DNA revolution and Double helix at 50 respectively. The project also aimed at transferring technology to private sector and funding innovative research in order to foster the biotechnology sector. The project made the crucial contribution in terms of generating information on location and probable functions of the genes that paved the way for future investigations. Some of the main findings from the draft sequence are as follows:

- Total number of genes was estimated at 30,000.
- The average gene was found to consist of 3000 bp but sizes vary greatly.
- Almost all (99.9%) nucleotide bases are exactly the same in all people.
- The functions are unknown for over 50% of discovered genes.
• Less than 2% of the genome codes for proteins.
• Repeated sequences that do not code for proteins ("junk DNA") make up at least 50% of the human genome.
• Chromosome 1 has the most genes (2968), and the Y chromosome has the fewest (231).
• About 1.4 million locations with SNPs were identified.

Findings from HGP are already having profound impact on diverse areas of research including molecular medicine (improved diagnosis of disease, earlier detection of genetic predispositions to disease, rational drug design etc.), bioarchaeology, anthropology, evolution and human migration, DNA forensics (identification), agriculture, livestock breeding etc.

1.5.2. Human Genome Diversity Project (HGDP)

HGDP was formally organised in 1993 under Stanford University’s Morrison Institute. Laboratories from around the world contributed cell lines from the studied populations (Cann et al., 2002). The project attempted to map the portion of genome that varies between humans, which is less than 1% of the entire genome, in samples collected from across the world. The project aimed at understanding the diversity patterns worldwide, the contributing factors and the implications of the observed diversity patterns. This was a crucial endeavour as diversity plays an important role in human traits; characteristics specific to ethnic groups; evolutionary studies and biomedical research (susceptibility and adaptability of different populations towards certain diseases). Findings from the project could also shed light on the origins and migration patterns of the entire human species. HGDP could also aid in understanding the role played by environmental factors in complex human diseases.

1.5.3. HapMap Project

The International HapMap Consortium is an international collaborative venture between Japan, the United Kingdom, Canada, China, Nigeria, and the United States aimed at developing haplotype map of the human genome in a bid to identify genetic
determinants of complex diseases. In the initial phase of the project, four populations with African, Asian and European ancestries were studied. The project aimed to identify and document genetic variations in humans using the most common type of variation found in the genome, SNPs. To reduce the cost of detection of genes of interest (i.e. the genes related to a disease phenotype), knowledge about existing common haplotypes in the genome and the associated tag SNPs generated through the project would come in handy. The information made available through the HapMap project is helping researchers find genes that affect health, disease, and individual responses to medications and environmental factors. The project was started in October 2002 with expected timeline of three years (The International HapMap Consortium, 2003). Findings from the project have been disseminated through several publications (The International HapMap Consortium, 2005, 2007, 2010).

1.5.4. Indian Genome Variation (IGV) Consortium

IGV Consortium was the first large scale effort to document and understand the genomic structure of enormously varied Indian populations. This was a Council of Scientific and Industrial Research (CSIR), Government of India initiative to study more than 15000 individuals from across India for variations at the levels of SNPs and repeat elements in genes with probable functional and locational significance with respect to common diseases. The populations in the study were representative of the vast ethnic, linguistic and geographical diversity of India. The study found high degree of genetic differentiation among the different ethnic groups. Study’s results also enable identification of population groups from which cases and controls may be sampled, and their data analysed in genome wide association studies without the disadvantage of confounding effects of population stratification, thereby increasing the power of association studies (Indian Genome Variation Consortium, 2008). Findings from the project are expected to address a variety of questions related to ethnic diversity, migrations, founder populations, predisposition to complex disorders or pharmacogenomics.

All these projects have led to formation of vast and highly informative databases. They have immensely contributed to the ever expanding knowledge of human genome and have also shed light on probable applications of this knowledge in different fields.
1.6. Rationale of the Study

Indian subcontinent is a reservoir of immense cultural, ethnic, linguistic, demographic and genetic diversity (Karve, 1961; Beteille, 1998; Majumder, 1998) with populations culturally stratified into endogamous tribal (8.2% of the total population of India according to Census of India, 2001) and non-tribal groups (Hindu caste populations and other religious groups) and classified under one of the four language families: Austro-Asiatic, Dravidian, Tibeto-Burman and Indo-European. It harbours more genetic diversity than other comparable global regions (Majumder, 1998). The subcontinent has served as a major corridor for the dispersal of modern humans (Cann, 2001). Facilitated by its strategic geographic location and biological affluence, there have been multiple waves of migrations and gene flow from different parts of the world in prehistoric and historic times (Ratnagar, 1995; Thapar, 1995) that have led to the assimilation of culturally, linguistically, ethnically and genetically disparate migrants with existing population substratum to varying degrees and have shaped the contemporary population scenario.

Of the several views prevalent as to the entry of various ethnic and linguistic elements in the subcontinent, one pertains to the entry of Indo-European linguistic element and its retreating effect on Dravidian speakers that are considered to be the older inhabitants. Dravidian speakers were possibly widespread throughout India before the arrival of Indo-European speakers from central Asia some 3500 ybp (Lal, 1974; Poliakov, 1974; Renfrew, 1987; Thapar, 2003) when they were forced to retreat southwards due to Indo-European dominance, after an initial period of language adoption and genetic admixture in the population groups that did not withdraw to regions completely out of reach of the Indo-Europeans (Majumder, 1998).

The tribal population groups are possibly the original inhabitants of India (Thapar, 1966; Ray, 1973) and their dialects have mainly been classified under Austro-Asiatic (AA), Dravidian (DR) and Tibeto-Burman (TB) linguistic families (Kosambi, 1991). However, several tribes in the western and north-western regions of India (including

---

2 Years before present
the tribes of Gujarat) speak languages belonging to Indo-European (IE) linguistic family. It is interesting that despite the late entry of Indo-European linguistic family there are populations that are culturally tribal groups but come under the Indo-European linguistic fold. It is said that such tribal groups of Gujarat are probably of proto-Australoid racial affinity, belonging to the earliest group of settlers in India who must have interacted with the people of various racial affinities such as Mediterraneans, Alpines, Dinarics and others passing through their habitats while entering India. This would have resulted in change in racial constitution of the tribes of Gujarat to some degree (Fuchs, 1964). Being vitally located at one of the entry points to the mainland, it is probable that the people residing in this region faced maximum impact of the numerous incoming waves of migrants and underwent changes in their cultural and genetic makeup, although to varying degrees. This is critically important as the tribal populations of north-western and western regions of India constitute 29.31% of the total tribal population of India (Census of India, 2001).

Despite its crucial location at the entrance to the subcontinent, accessibility both by land and sea; and its connectivity to the mainland that have attracted numerous ebbs and flows of people from around the world at different points in time, potential of the western region of the subcontinent for understanding the linguistic and genetic affinities of the tribal groups inhabiting the area has hitherto not been explored much. There are several studies that have been conducted to understand the genetic composition and affinities of the different ethnic and linguistic lineages extant in India. Most of these studies have stated that linguistic diversity concurs with genetic differences among populations but majority of such studies have not taken into consideration Indo-European speaking tribal groups (Cavalli-Sforza et al., 1994). A few studies have excluded Indo-European tribes as they may not reflect pre-Indo-European genetic composition of India (Cordaux et al., 2004).

In light of the above stated facts, it would thus be interesting to study the genetic makeup of these population groups. Equally important would be to study the genetic heterogeneity and pattern of genetic differentiation of the populations of southern Gujarat. Moreover, it will also be important to understand whether the Indo-European speaking tribes of southern Gujarat have greater genetic similarities with the
immigrant groups that brought Indo-European languages to the subcontinent; or the native groups only adopted/ were forced to adopt the new language while retaining their original genetic composition which is expected to be similar to other native groups of the mainland. Also, the present study will lead to generation of comparable baseline data that can be used in the future for genome diversity analysis of populations of Indian origin.

Keeping this in mind, two Indo-European-speaking tribal groups (Dhodia and Dubla) from Valsad district, part of ‘tribal belt’ of Gujarat, were selected for the present study in order to understand the dynamics of genetic versus linguistic affinities.

Of the several molecular markers used to study the genetic variation in the study groups, a set of markers belong to β- globin gene cluster that aid in haplotype reconstruction. This set of markers was especially chosen because of high prevalence of a mutant allele of the β- globin gene, Hb*S, in the populations under study (Negi, 1968). Analysis of these markers would assist in assessing the diversity in general population as well as in affected members of the populations. Restriction-site polymorphisms in the chromosomes bearing Hb*S have been studied by many in the Indian population groups (Kulozik et al., 1986; Labie et al., 1989; Majumder et al., 1999a; Mukherjee et al., 2004; Uma Mahesh et al., 2011). In this study also an attempt is being made to gain an understanding of haplotypes associated with this mutation.

**1.7. Objectives of the Study**

- To study the genetic heterogeneity among the Dhodia and Dubla tribes of southern Gujarat.
- To study pattern of gene differentiation among the Dhodia and Dubla tribes of southern Gujarat.
- To study genetic and linguistic affinities of the tribes under study.
- To find out the prevalence of HbS gene in the two populations understudy.
- To find out the most common haplotype(s) associated with the sickle cell mutation.

Keeping in view the above objectives, following molecular markers were selected and analysed in the study populations.
Introduction

- **Alu Insertion Deletion** (*Alu* InDel) markers
  - PV92, FXIIIB, D1, APO, ACE, CD4, PLAT
- **Restriction Fragment Length Polymorphisms** (RFLPs)
  - **Unlinked**: ESR (*Pvu*II), NAT (*Kpn*I), PSCR (*Taq*I), T2 (*Msp*I), LPL (*Pvu*II), ADH2 (*Rsa*I) and ALAD (*Rsa*I)
  - **Linked**: These permit haplotype reconstruction and have been examined for two gene systems:
    - **DRD2 Locus**: *Taq* I ‘A’, *Taq* I ‘B’ and *Taq* I ‘D’

Apart from analysing these molecular markers, sickle cell status of the collected samples was also determined using Dithionite Tube Turbidity test and Cellulose Acetate Membrane Electrophoresis.

In order to extract meaning from seemingly incomprehensible raw data obtained from laboratory analysis, several statistical parameters were made use of. Starting with the most basic measure, allele frequency, various other parameters such as test for Hardy-Weinberg Equilibrium, heterozygosity, gene diversity, haplotype frequencies, linkage disequilibrium, *AMOVA* among others were computed.