UNIT 3  GENETIC POLYMORPHISM

Contents
3.1 Introduction
3.2 Balanced Polymorphism
3.3 Transient and Balanced Polymorphisms
3.4 Serological Markers
3.5 Biochemical Polymorphisms
3.6 Molecular Markers
  3.6.1 Repetitive DNA Sequence Variants
  3.6.2 Non-Repetitive DNA Sequence Variants
  3.6.3 Lineage Markers
3.7 Tools for Studying Polymorphisms
3.8 Genetic Markers and Disease
3.9 Genetic Mapping of Disease Gene on Human Chromosome Using Polymorphic Markers
3.10 Use of Polymorphic Markers in Forensic Testing
3.11 Use of Polymorphic Markers in Population Studies
3.12 Summary
  Suggested Reading
  Sample Questions

Learning Objectives

After reading this unit, you will be able to:

- define the concept of genetic polymorphism;
- explain genetic polymorphism with respect to serological, biochemical and molecular markers;
- explain the genetic markers in disease association; and
- discuss the use of polymorphic markers in population and forensic studies.

3.1 INTRODUCTION

Genetic polymorphism can be defined as the occurrence together in the same population two or more than two alleles such that the frequency of rare allele is always >1%, and is maintained in the population not merely by the recurrent mutation. Polymorphism can be in a coding region (coding region means the portion of DNA which code for a gene, it may be synonymous or non-synonymous) or more commonly, in the noncoding regions (which does not code for functional region), often vary by ethnicity. Basic information about the types, frequencies and distribution of common polymorphisms are essential not only for the understanding of pathological entities, but also to know our evolutionary past and provide guidance about our biological future. The most common polymorphism in our genome are single base pair sequence variation i.e. SNP
but other types like copy number changes, insertions, deletions, duplications and rearrangements also occur. The methods to assess this diversity is variable. Few examples of polymorphic markers are listed in table 3.1.

Table 3.1: Example of Genetic polymorphisms

<table>
<thead>
<tr>
<th>Type of marker</th>
<th>Year</th>
<th>No. of loci</th>
<th>Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blood groups</td>
<td>1910-1960</td>
<td>~20</td>
<td>May need fresh blood, rare antisera. Genotype cannot always be inferred from phenotype because of dominance. No easy physical localization.</td>
</tr>
<tr>
<td>Electrophoretic mobility variants of serum proteins</td>
<td>1960-1975</td>
<td>~30</td>
<td>May need fresh serum, specialized assays, no easy physical localization often limited polymorphisms</td>
</tr>
<tr>
<td>Human Leucocyte Antigens (HLA)</td>
<td>1970</td>
<td>1 (multi locus haplotype)</td>
<td>One linked set highly informative. Can only test for linkage to 6p21.3</td>
</tr>
<tr>
<td>DNA RFLPs</td>
<td>1975</td>
<td>&gt;105 (potentially)</td>
<td>Two allele markers, maximum heterozygosit y 0.5, initially required Southern blotting, now PCR. Easy physical localization</td>
</tr>
<tr>
<td>DNA VNTRs (minisatellites)</td>
<td>1985-</td>
<td>&gt;104 (potentially)</td>
<td>Many alleles, highly informative can be typed by southern blotting easy physical localization. Tend to cluster near ends of chromosomes.</td>
</tr>
<tr>
<td>DNA VNTRs (microsatellites)</td>
<td>1989-</td>
<td>105 (potentially)</td>
<td>Many alleles, highly informative Can be typed by automated multiplex PCR, easy physical localization. Distributed throughout genome</td>
</tr>
<tr>
<td>DNA SNPs</td>
<td>1998-</td>
<td>106 (potentially)</td>
<td>Less informative than microsatellites. Can be typed on a very large scale by automated equipment, without gel electrophoresis, etc.</td>
</tr>
</tbody>
</table>

3.2 BALANCED POLYMORPHISM

When natural selection favors heterozygotes over both homozygotes, the result is balanced polymorphism. It accounts for the persistence of an allele even though it is deleterious when homozygous. Some of the examples are given below:

Sickle Cell Disease

It is an autosomal recessive disorder that causes anemia, joint pain, a swollen spleen, and frequent, severe infections. It illustrates balanced polymorphism because carriers are resistant to malaria, an infection by the parasite *Plasmodium falciparum* that causes cycles of chills and fever. The parasite spends the first stage of its life cycle in the salivary glands of the mosquito *Anopheles gambiae*. When an infected mosquito bites a human, the malaria parasite enters the red
blood cells, which transport it to the liver. The red blood cells burst, releasing the parasite throughout the body.

It is known since long that malaria is a quite common in the tropical regions of Africa. Sickle shape red blood cells provide selective advantage as malarial parasite cannot grow in these cells. Therefore, along with malaria the sickle cell anemia also increased in these parts of Africa. The sickle cell disease is less common in Caucasians due to the less frequency of malaria. This shows the heterozygous advantage of sickle cell as it provides protective effect.

The rise of sickle cell disease goes hand in hand with the cultural development with the advent of cultivation of crops gave a breeding ground to *Anopheles* mosquitoes as the malaria rose the selective pressures gave rise to the change in the shape of the RBCs from elliptical to sickle shaped and when it occurred in homozygous condition the disease was caused otherwise it had selective advantage. The spread of sickle cell disease is associated to the migratory events. Africa by people migrating from Southern Arabia and India, or it may have arisen by mutation directly in East Africa.

Settlements with large numbers of sickle cell carriers escaped devastating malaria. They were therefore strong enough to clear even more land to grow food-and support the disease-bearing mosquitoes. Even today, sickle cell disease is more prevalent in agricultural societies than among people who hunt and gather their food.

**G6PD Deficiency**

It is a sex-linked enzyme deficiency. It affects 400 million people throughout the world. It results into hemolytic anemia which is life-threatening. It is under the influence of certain environmental conditions like eating fava beans, inhaling certain types of pollen, taking certain drugs, or catching certain infections. It has been seen in Africa that hemizygous males and heterozygous males for this enzyme deficiency are at less risk for malaria again revealing a selective advantage for heterozygotes. Therefore, natural selection acts in two directions hence it could be one of the example of balanced polymorphism.

### 3.3 TRANSIENT AND BALANCED POLYMORPHISMS

Polymorphism occurs when two or more clearly different phenotypes exist in the same population of a species-in other words, the occurrence of *more than one form* or *morph*.

A transient polymorphism is one that is changing in frequency over time. In transient polymorphism, one form is gradually being replaced by another. As the name implies, it represents a temporary situation as a by-product of directional natural selection.

The phenomenon of industrial melanism occurs in a number of moth species in Europe and the United States. The British ecological geneticist, E. B. Ford, first called attention to this phenomenon as a way of demonstrating the effect of natural selection in nature (as opposed to artificial selection experiments which have long enjoyed success in the lab). Ford noted that a light colored moth species,
Biston betularia, occasionally undergoes mutation at a single locus to produce a dark or melanic individual. Since the mutant allele is dominant, any gamete containing this mutant will produce a melanic individual upon syngamy. The first melanic specimen in this species was found in a collection from Manchester, England dated 1848, but by 1895 about 95% of all collected specimens were dark morphs, referred to as the form carbonaria. In a series of 12 observations and mark-recapture experiments during the 1950s, H. B. D. Kettlewell demonstrated that the two forms (light and dark) were differentially preyed upon by birds. He found that the birds selectively caught and ate more individuals of the form that did not match its background as compared to the one that was masked. In industrialized areas of England where the substrate (walls and tree trunks) upon which the moths rested were darkened by pollutants in the smoke poured out by factories, the carbonaria form possessed a selective advantage. Rural areas, unaffected by pollutants, afforded the light form an adaptive advantage. The environmental change brought on by the industrial revolution did not produce the carbonaria form (which presumably appeared from time to time due to recurrent mutation); it only protected the dark moths from bird predation (the agent of natural selection). The fact that the light form still exists in rare numbers in industrialized areas testifies to the amount of time selection requires to eliminate a recessive allele.

Mendelian Population

A population is a group of individuals who share a common gene pool where the characters are transmitted in a Mendelian fashion from one generation to the next generation. A group of individuals within which marriages are performed is called a Mendelian population. In a given Mendelian population, which is under Hardy-Weinberg equilibrium, the resultant genotype and phenotype frequencies are more or less permanently established.

3.4 SEROLOGICAL MARKERS

Blood groups are the best cited examples of serological markers. Both ABO and Rh are quite important serological markers as they can be used to study population diversity. These blood groups cause newborn hemolytic diseases. They also have a role in blood transfusion and also solid organ transplantation. They follow mendelian inheritance. ABO blood groups were discovered by Landsteiner in 1900 and are cited as a best example of triallelic inheritance. Blood groups can be tested by using antisera and red blood cells using simple agglutination techniques. Presently more advanced molecular techniques are also used.

The ranges of phenotypes in humans are a direct result of genetic variations which act together with environmental and behavioral factors to produce diversity. The identification of gene polymorphisms, which control the blood group antigen expression, contributes to the understanding of the biological significance of blood group systems. In addition to assisting in the characterization of allelic variations, the identification of gene polymorphisms allows us to estimate the processes involved in the formation of different populations (the founder effect, genetic drift, migration, etc.). Thus, blood group gene polymorphisms are valuable predictors of genomic ethnic ancestry.
3.5 BIOCHEMICAL POLYMORPHISMS

There is marked difference between individuals on the basis of biochemical markers like G6PD, human enzymes and proteins etc. This has been explained in the above section. However, here we would like to throw some light on the molecular basis of G6PD variants.

Molecular Basis of G6PD Variants

The G6PD gene, located on chromosome Xq28 region, is 18 Kb long consisting of 13 exons transcribed to a 2.269 Kb messenger RNA with 1.545 Kb of coding regions. The commonest variant in South China, G6PD Canton, has been sequenced and was found to be due to a mutation at nucleotide (nt) position 1376 of cDNA, G to T, resulting in a missense mutation in amino acid position 459, Arg to Leu. With improved DNA technology, the whole cDNA sequence can be amplified and screened for mutation directly. PCR technique and restriction analysis has been used.

World Incidence and Distribution of G6PD Deficiency

G6PD deficiency in male subjects can be detected easily by a number of screening tests. The simplest one is the fluorescent spot test developed by Beutler and Mitchell which relied on the fluorescence of NADPH, generated by an adequate amount of G6PD enzyme. This test can also be done on blood sample dried on filter paper similar to the Guthrie cards. In Hong Kong, the routine screening of newborns have included test for G6PD deficiency.

3.6 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 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:

3.6.1 Repetitive DNA Sequence Variants

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), Short Tandem Repeats (STRs) or 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).

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.

*Alu InDels* – *Alu* Insertion/ Deletion polymorphisms (*Alu* InDels) involve *Alu* sequences that are characterized by the cleavage action of *Alu*I restriction endonuclease.

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.

### 3.6.2 Non- Repetitive DNA Sequence Variants

#### 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.

#### 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.

### 3.6.3 Lineage Markers

#### 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.

#### 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.

### Hardy-Weinberg equilibrium

The Hardy-Weinberg equilibrium predicts that in the absence of evolutionary forces, both allelic and genotypic frequencies remain constant in a population and that if the equilibrium is disturbed a new equilibrium will be reached within one generation based on the allelic frequencies of the remaining population. The conditions that must be met for the predictions of the Hardy-Weinberg equilibrium to be valid are described below:

1) Random mating: Mating patterns must randomly reflect the entire breeding population, with no dependence on genotype or closeness of relationship (either positive or negative).
2) No sex bias in allelic frequencies: The distribution of alleles must be the same in both sexes.

3) All genotypes equally viable and fertile: There must not be any selective advantages or disadvantages. This is seldom true in a real population, and often must be taken into account in terms of evolutionary pressures.

4) Mutation rate too low to alter ratios: The basic assumption is that alleles are stable through many generations and are not altered or degraded significantly by mutation. In practice this is generally not a serious problem.

5) Closed population (no in or out migration): The “population” that is being considered must be a constant one. Introduction of new genes into the breeding pool or loss of genes from the breeding pool by migration between “populations” can distort trends.

6) Population must be large: The population must be large enough so that there are no confounding effects due to genetic drift (random events altering allelic frequencies by pure chance) or due to “founder” effects, where a recessive gene becomes fixed in a population because too many of its members are descendants of a single individual.

The Hardy-Weinberg law can also be applied to multiple alleles and X-linked alleles. The genotypic frequencies expected under Hardy-Weinberg equilibrium will differ according to the situation.

3.7 TOOLS FOR STUDYING POLYMORPHISMS

Both conventional and advanced techniques are used to study polymorphisms. Conventional techniques are blood groups by carrying out simple agglutination techniques or protein electrophoresis for studying the protein polymorphisms etc. Under advanced techniques are the tools for studying molecular markers, and the 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. 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, but, it is the technique of PCR which is the most useful for DNA analysis.

Polymerase Chain Reaction (PCR): It involves cycling of DNA sample through a series of heating and cooling cycles with the required raw materials and enzymes to achieve its exponential amplification. 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.

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.
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.

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.

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.

The selection of technique and markers depends upon the purpose of study. In the following section we have discussed the uses of polymorphic markers.

Uses of polymorphisms: All the markers listed in table 3.1 can be used for population diversity studies. Now a days most extensively studied markers are Single nucleotide polymorphisms. Genomics and specially SNP research can be used to improve health care through gene therapy, to yield new targets for drug discovery, to renew the process of drug development and to discover new diagnostics.

3.8 GENETIC MARKERS AND DISEASE

Understanding the genetic basis of complex human diseases (like hypertension, cardiovascular disease, diabetes etc.) has been increasingly emphasized as a means of achieving insight into disease pathogenesis, with the ultimate goal of improving preventive strategies, diagnostic tools, and therapies. Genetic approaches to complex disorders thus offer great potential to improve our understanding of their pathophysiology, but they also offer significant challenges. These can be studied either using linkage analysis. In linkage analysis we use families and try to find out which polymorphic marker is near to the disease gene and then try to map the gene on the human genome. The other approach is where we study populations of both types of individuals. One would be those suffering with a disease and the other would be who are not suffering with the disease. We take different polymorphic markers and study in these two sets of samples. Then we
compare both the groups and if both the groups differ significantly at these markers we propose that these markers may be associated with the disease.

Association studies can be a very powerful approach for finding genetic determinants of a complex disorder. It has been suggested that if hundreds of thousands of single nucleotide polymorphisms (SNPs) were identified across the genome, then it would be possible to perform genome-wide association studies to identify the regions of linkage disequilibrium around disease susceptibility genes. In addition, they noted that much smaller sample sizes would be required to detect association than to detect linkage. The SNP Consortium is rapidly identifying single nucleotide polymorphisms, and within next several years, genome-wide association studies may become a reality.

These association studies can result into positive association or negative association. Some time they result into false positive or false negative results. The following general guidelines, summarized in Table 3.2, may be useful for genetic association studies. First, are the candidate gene(s) under study should be biologically reasonable. Several factors can determine the appropriateness of a candidate gene. If human genetic linkage studies have identified a chromosomal region linked to a disease, or if an animal model for a disease is influenced by a particular gene or syntenic chromosomal region, positional candidate genes in such genomic regions warrant strong consideration. In addition, the biologic plausibility of a candidate gene for involvement in disease pathogenesis is important. However, obvious limitations of this candidate approach are the large number of potential candidate genes for complex diseases and the reality that only known genes can be investigated. Although candidate genes can be selected for study on this basis, they should not be ruled out on the basis of our current understanding of disease pathophysiology- important new insights may be missed if potential candidate genes must fit into current pathophysiologic models.

<table>
<thead>
<tr>
<th>Issue</th>
<th>Key Questions</th>
<th>Possible Solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection of candidate</td>
<td>Is candidate gene biologically reasonable?</td>
<td>Demonstration of biologically functional effect</td>
</tr>
<tr>
<td>Gene polymorphism</td>
<td>Is the candidate gene a positional candidate?</td>
<td>Within linked region in man or systemic from animal model</td>
</tr>
<tr>
<td>Population stratification</td>
<td>Are cases and controls matched?</td>
<td>Matching on ethnicity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Family-based association designs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Negative results with multiple unlinked markers</td>
</tr>
<tr>
<td>Hardy-Weinberg (H-W) equilibrium</td>
<td>Is control group in H-W equilibrium?</td>
<td>Calculation of H-W equilibrium with goodness-of-fit test (2 alleles) or simulation (multiple alleles)</td>
</tr>
<tr>
<td>Multiple comparisons</td>
<td>How many alleles were tested?</td>
<td>Bonferroni correction</td>
</tr>
<tr>
<td></td>
<td>How many genetic loci were tested?</td>
<td>Estimation of empirical P values</td>
</tr>
</tbody>
</table>

Table 3.2: Evaluation of candidate gene case-control association studies
A second criterion in evaluation of case-control association studies is the careful selection of cases and control subjects. Do the case subjects meet appropriate criteria for disease affection? Are control subjects free from symptoms of disease, associated intermediate phenotypes, and potential confounders? Have control subjects been exposed to relevant environmental influences involved in disease pathogenesis while remaining clearly unaffected? Were the cases and controls matched on demographic and environmental factors? Was consideration of population stratification included, either by attempting to match ethnicity or by typing unlinked markers.

A third criterion in the evaluation of case-control studies is assessment of Hardy-Weinberg equilibrium in the markers studied within the control group. Hardy-Weinberg equilibrium indicates that the genotype frequencies can be determined directly from the allele frequencies; failure to demonstrate Hardy-Weinberg equilibrium could result from genotyping errors, inbreeding, genetic drift, mutation, or population substructure. Hardy-Weinberg equilibrium can be readily assessed with a goodness-of-fit chi square test for biallelic markers; for markers with multiple alleles (such as short-tandem repeat markers), more accurate determination of Hardy-Weinberg equilibrium can be obtained with Markov Chain Monte Carlo methods. Significant deviations from the expected proportions of homozygote and heterozygote classes in a population of case subjects may be caused by association with the disease allele. Lack of consistency with Hardy-Weinberg equilibrium among control subjects should prompt investigation for potential complications, including genotyping errors and population stratification. A final criterion for evaluation of a case-control study is correction for multiple comparisons. This remains a problematic topic requiring additional statistical genetic research. However, an effort to correct for spurious associations, which can result from testing a large number of alleles, is warranted. The multiple comparison issue is especially problematic with markers that have multiple alleles like short-tandem repeat polymorphisms; the conservative Bonferroni approach to use a corrected significance value calculated by multiplication of the observed $P$ value by the number of alleles tested. Bonferroni corrections for the total number of alleles at all loci are probably too conservative because the alleles at one locus are not independent of each other and closely linked loci are probably not independent either. A less conservative but more computationally intensive approach is to estimate empirical significance values using simulation approaches.

**Genome Wide Studies**

Unlike the direct approach of case-control association with candidate genes, genome scanning (screening) is an indirect strategy that does not rely on conjecture. Basically, either affected individuals, usually siblings, from a number of families or families with two or more affected individuals are genotyped with polymorphic DNA markers that cover the entire chromosome complement. A set of about 400 short tandem repeat polymorphic markers that are spaced at about every 10cM is used for most genome scans. This level of resolution has been enhanced with the assembly of about 3000 simple sequence repeat polymorphic markers that are about 1.5cM apart. Single-nucleotide polymorphic sites (SNPs) are preferred for genome scans because they are uniformly distributed about every 300 bases throughout the genome and easily identified with automated equipment. Eventually, sets of SNPs will supersede short tandem repeat polymorphic sequence marker systems.
Furthermore, major landmark attempts that have also been made to study various aspects of human genome, and few are listed below.

**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.

First draft was released in 2001 followed by the complete draft in 2003. 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.
- Repeated sequences that do not code for proteins (“junk DNA”) make up at least 50% of the human genome.
- 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.

**Human Genome Diversity Project (HGDP):** HGDP was formally organised in 1993 under Stanford University’s Morrison Institute, and was aimed at understanding the diversity patterns worldwide, the contributing factors and the implications of the observed diversity patterns. 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.

**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. 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.

**Indian Genome Variation (IGV):** IGV was the first large scale effort to document and understand the genomic structure of enormously varied Indian populations. The study found high degree of genetic differentiation among the different ethnic groups.

**Genetic Testing and Counseling**

Frequently the question may arise as to whether the patient has a certain disease for which there is a genetic basis. Often among the 10,000 conditions for which a genetic basis has been identified, the diagnosis can be made from evaluation of personal and family history, physical examination, and conventional laboratory tests. A useful database for identifying these conditions is available on Online Mendelian Inheritance in Man (OMIM) (www.ncbi.nlm.nih.gov/omim). This
catalog is updated regularly and can be searched using multiple terms. The entries provide information about the clinical signs as well as the genetic basis for the condition, if known, including mutations that have been found to cause the condition. To determine whether genetic testing is available for a given condition and to find a laboratory, a useful link is GeneTests, a free online service (www.genetests.org). The entries in this catalog indicate the test menus and contact information for the laboratories, as well as whether the testing is provided on a routine or research basis. A very useful adjunct in the GeneTests Website is GeneReviews, which provides succinct summaries about many genetic conditions and the ways the genetic testing can be used for diagnosing these conditions, including prediction of natural history.

The clinician is likely to encounter many situations in which a genetic test may be useful. Sometimes genetic testing is required from diagnosis when it cannot be made by clinical criteria alone. The fragile X syndrome is the most common genetic form of mental retardation. Although the diagnosis may be suggested by the presence of the characteristic signs—large ears, protruding chin, and large testes—the only way to diagnose fragile X is by genetic testing. For the various forms of spinocerebellar ataxia, there is considerable overlap. Yet, these can be readily distinguished by their specific mutations. Patients with atypical forms of certain diseases may have a negative gold standard test, but positive genetic test. For most patients with cystic fibrosis, the diagnosis can usually be made by a sweat chloride test. However, a number of individuals have been described with pulmonary disease suggestive of this condition for whom the sweat chloride test is normal. For these patients, the diagnosis has been based on observation of mutations in both copies of their CFTR genes.

For some conditions, the signs of disease may not yet have developed, yet on the basis of one’s family history, one may want to know about the risk of developing disease. This is true for the person whose parent(s) may have died from Huntington’s disease, a progressive neurodegenerative disease or for the person whose mother and sister may have died from breast or ovarian cancer, suggesting a heritable risk. For these individuals, a positive genetic test result will indicate an increased, although not necessarily absolute, risk for developing the disease.

Genetic testing is used for assessing reproductive risks—by testing the parents for carrier status and by testing the fetus. Individuals with a positive family history of genetic disease (usually autosomal recessive or X-linked) or who come from ethnic groups with an increased prevalence of autosomal recessive or X-linked diseases are candidates for carrier screening. Currently, carrier screening for cystic fibrosis, fragile X syndrome, and spinal muscular atrophy is recommended in the United States. For people of Mediterranean, African, or South Asian ancestry, hemoglobinopathy screening is recommended. For individuals of Ashkenazi Jewish ancestry, screening for Tay-Sachs disease, Canavan disease, cystic fibrosis, Gaucher disease, Bloom syndrome, Fanconi anemia, Niemann-Pick disease, familial dysau-tonomia, maple syrup urine disease, glycogen storage disease, and familial hyperinsulinism is available. An individual who is a carrier for a certain condition may choose not to marry another individual who is a carrier for the same condition. Alternatively, if a carrier couple is identified, they may choose to have prenatal diagnosis to determine whether their fetus is affected with this condition. This can be performed either at 10-11 weeks using the procedure of chorionic villus sampling where a bit of placenta is obtained under ultrasound
As another option, an amniocentesis can be performed at 15-18 weeks of pregnancy to obtain cells from the amniotic fluid. These couples might also choose to have pre-implantation genetic diagnosis with selection implantation of only those embryos that are deemed unaffected.

Not all genetic testing involves looking for heritable mutations. Sometimes it is used to look for genetic alterations that are confined to a specific population of cells. These alterations may cause certain cells to become cancerous, or if cancerous, to progress to a more aggressive stage. Genetic testing can be used to identify chromosomal translocations between two non-homologous chromosomal segments and in the process diagnose a specific form of leukemia. For example, the translocation between chromosomes 1 and 19 in leukemic cells is diagnostic of the acute promyelocytic form of this disease and the translocation between chromosomes 9 and 22 is diagnostic of the chronic myelogenous form. The expression patterns of RNA transcribed from many genes can be assessed to predict the natural history of the disease. This approach has been used to predict breast cancer outcome and whether more or less aggressive therapies should be used to treat patients.

Individuals might also have genetic tests of identity. These might be voluntary and selected to test specific questions, such as whether they are members of a known patrilineal lineage, such a people with a specific surname. These tests analyze a series of polymorphic genetic markers on the Y chromosome. On the basis of the general pattern of markers, or “haplogroup,” they may be told of the geographic region where their Y chromosome originated. According to the number of markers that match with people who are suspected to be of the same lineage, individuals may be advised about the common ancestors or other people in that lineage. Such testing is also possible for matrilineal lineages by testing mitochondrial DNA markers.

### 3.9 GENETIC MAPPING OF DISEASE GENE ON HUMAN CHROMOSOME USING POLYMORPHIC MARKERS

In genetic mapping the diseased gene polymorphic markers play a very important role. These markers could be short tandem repeats, variable number of tandem repeats, blood groups, restriction fragment length polymorphism etc. Mapping can be done in a step wise manner.

Collect all the pedigrees where the disease is found. Analyse all the members against various polymorphisms and perform linkage analysis.

Linkage study entails collecting blood cells from members of several two – and three – generation families or from individuals of a large multiple generation family with a specific genetic disorder. The blood can be cultured and cell lines can be maintained large number of polymorphic markers (probes), representing sites from all parts of all autosomes, are used. A two –point (two – locus) LOD score is calculated for each polymorphic locus and the site of the genetic disease from all informative parent offspring combinations and finally the linkage is established. However, genotyping errors can give –ve or +ve LOD score. Hence perfect genotyping is must to get the correct results.
3.10 USE OF POLYMORPHIC MARKERS IN FORENSIC TESTING

Polymorphic markers have great utility in personal identification. As mentioned above no two individuals are alike. These differences are at both phenotypic and genotypic levels. The genetic differences can be identified by testing these markers. This testing is provided by commercial firms that market directly to consumers. Identity genetic analysis may also be involuntary and used for paternity testing of children or fetuses or for identification of forensic samples in murder, assault or rape cases, in which the perpetrator of the crime left a tissue sample of blood, semen, hair, or other tissue type from which DNA can be extracted and the test can be performed. However, it must be kept in mind that there are ethnic differences in the distribution of these markers. Hence every population should have its own genetic profile.

3.11 USE OF POLYMORPHIC MARKERS IN POPULATION STUDIES

Population Diversity Studies

Human genome varies from individual to individual and therefore no two individuals look alike. This was noted long back. Historically, individual variation was studied on the basis of conventional somatoscopic markers. However, with the advancement of technology various genetic markers were discovered and the gene frequency data for studying the evolution of human races was analyzed using these markers. Initially, the classical serological and biochemical markers have played important roles in various types of human population genetic studies. One of the problems that limited their practical utility results from the limited number of possible genotypes at each of such loci. The discovery of hyper variable DNA loci offers the opportunity to ameliorate this problem. It was later realized that comparison of gene frequencies for one or two loci are not reliable since each locus has a different geographical distribution, hence the differences observed may be because of chance factor. Only when a large number of loci are used, the genetic relationship among populations could be drawn successfully. Recent analysis based upon polymorphic markers reveal that inter and intragroup genetic variation may be of a lesser magnitude and may not be of significance if proper markers are not selected and more so if statistical tools used are not highly powerful. However, it is important to record population variation because it is helpful to know the various mechanisms involved in causing variation and it further enhances our knowledge about the molecular basis of disease susceptibility.

3.12 SUMMARY

It is difficult to attribute any functional significance to genetic polymorphisms. However, the non-coding sequences of the genes which are located far away from the functional region of the gene may affect the function of the gene. However, these sequences are otherwise useful in studying population diversity, disease gene mapping, forensic investigations etc. Recently after the advent of microarray genes for many complex disorders have been found by using genome wide association studies.
Suggested Reading


Sample Question

1) Define polymorphism with few examples.

2) What are the evolutionary forces that affect gene frequency of polymorphic markers?

3) Give some uses of polymorphic markers.

4) What is law of Hardy Weinberg?

5) What is genetic testing?

6) Describe the utility of studying molecular markers in anthropological genetics.