

INDIRA GANDHI NATIONAL TRIBAL UNIVERSITY

AMARKANTAK, M.P. 484886



SUBJECT- Biotechnology

TITLE- Gene, Genomics and Genetics

M. Sc Biotechnology 2nd Semester

Unit-IV Reference Notes

Dr. Parikipandla Sridevi

Assistant Professor

Dept. of Biotechnology

Faculty of Science

Indira Gandhi National Tribal University

Amarkantak, MP, India

Pin : 484887

Mob No: +919630036673, +919407331673

Email Id: psridevi@igntu.ac.in, devi.shri45@gmail.com

psridevi.igntu@gmail.com

Unit IV

Genetic variation: Mutations; kinds of mutation; agents of mutation; genome polymorphism; uses of polymorphism. Gene mapping and human genome project Physical mapping; linkage and association Population genetics and evolution Phenotype; Genotype; Gene frequency; Hardy Weinberg law; Factors distinguishing Hardy Weinberg equilibrium; Mutation selection; Migration; Gene flow; Genetic drift; Human genetic diversity; Origin of major human groups.

CONTENTS

1. Introduction
2. Types of polymorphism
3. Applications

Genome Polymorphisms

Introduction

How do we find a gene that contributes to a disorder or a behavior? The technology for such gene hunting has been revolutionized several times over the past thirty years. In fact, some techniques considered “hot” only a decade ago are now obsolete. The remarkable progress has been due to one major phenomenon— the ability to detect and cheaply genotype “spelling variations” in the human genome. To understand this, we must introduce (or refresh) some terminology.

Suppose that we gathered DNA from a large number of individuals and examined the first 1,000 nucleotides on the first chromosome. There would be some sections in which the base pair sequence is the same for all of the DNA strands. These are called *monomorphic* sections. There will be other sections in which the sequence of nucleotides differs. These are called *polymorphic* sections or, simply, *polymorphisms*. Using the analogy of a four-letter DNA alphabet, polymorphisms are really “spelling variations.” The term *allele* refers to a specific spelling variation.

The alleles at a polymorphic section are called either *mutants* or *common polymorphisms* depending on their frequency. A mutant allele has a

frequency of less than 1% in the general population. Alleles with a frequency higher than 1% are considered “common.”

Before the 1980s, finding the genotype of an individual usually involved various laboratory assays for the product of a gene—the protein or enzyme—but not the gene itself. The cases of the ABO and Rhesus blood groups are classic examples of how one infers genotypes from the reaction of gene products with certain chemicals. The actual number of known polymorphisms was probably in the low 100s. As a result, for most of the 20th century attempts to find the genes for many Mendelian disorders were unsuccessful.

In the mid 1980s, genetic technology took a great leap forward with the ability to genotype the DNA itself. The geneticist could now examine the DNA directly without going through the laborious process of developing assays to detect individual differences in proteins and enzymes. Direct DNA analysis had the further advantage of being able to identify alleles in sections of DNA that did not code for polypeptide chains. As a result of these new advances, the number of polymorphic regions increased exponentially.

Table 1.1 Types of polymorphisms

1. Protein/enzyme polymorphisms
2. DNA polymorphisms.
 - A. Single nucleotide polymorphisms (SNPs)
 - B. Tandem repeat polymorphisms
 - C. Structural polymorphisms (deletions, inversions, etc.)
 - D. Sequence polymorphisms

1.1 Protein/enzyme polymorphisms

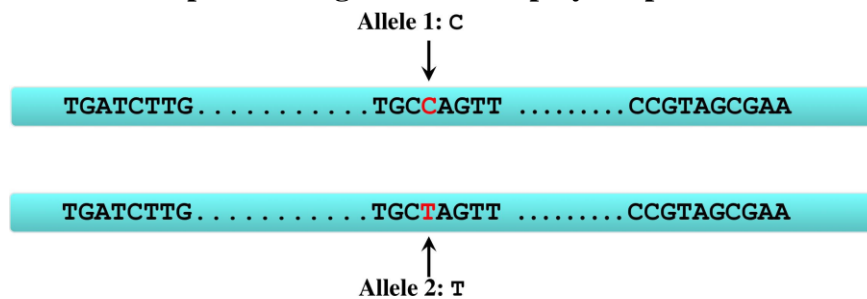
In the early days of human genetics, the majority of polymorphisms were those associated with proteins and enzymes. To detect the polymorphism and a person’s genotype, one performed assays for the gene product, i.e., the protein or enzyme produced by the genetic blueprint.

Most of these polymorphisms were detected in blood. When your blood is typed, you are informed that you are blood group O+ or AB- or A+, etc. The letter in this blood group gives your phenotype at the ABO locus, and the plus (+) or minus (-) sign denotes your phenotype at the Rhesus gene.

1.2 DNA polymorphisms

The other large class of polymorphisms are those that detect spelling variations at the level of DNA nucleotides. For our purposes, we can classify them into three types.

Example of a single nucleotide polymorphism.



Single nucleotide polymorphisms

A *single nucleotide polymorphism* or *SNP* is a sequence of DNA on which humans vary by one and only one nucleotide (see Figure 9.1). Because humans differ by one nucleotide per every thousand or so nucleotides, there are millions of SNPs scattered throughout the human genome.

The major advantage of SNPs, however, lies in the fact that they can be detected in a highly automated way using specialized DNA “chips” usually called *DNA arrays*.

Tandem repeat polymorphisms

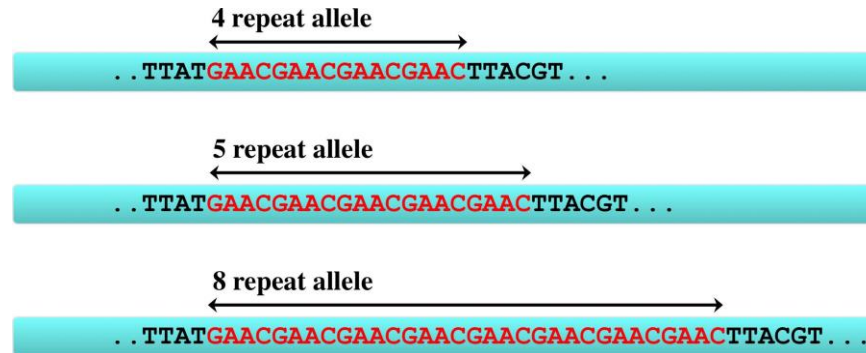
A tandem repeat polymorphism consists of a series of nucleotides that are repeated in tandem (i.e., one time after another). The polymorphism consists of the number of repeats. Figure 9.2 illustrates this type of polymorphism.

The repetitive nucleotide sequence is GAAC and the figure depicts three alleles—a four-repeat allele, a five-repeat allele, and an eight-repeat allele.

Tandem repeat terminology (graduate)

Unfortunately, even though the concept of the tandem repeat is quite simple, the terminology for referring to these polymorphisms can be confusing. When the number of repeats is small, usually five to six or fewer, then the polymorphism may be called a *microsatellite*, *simple sequence repeat* (SSR), or *short tandem repeat* (STR). When the number of repeated nucleotides is larger, then the polymorphism may be called a *minisatellite*, particularly when it is located in a telomere. Finally, the term *variable number of tandem repeats* (VNTR) polymorphism has been used equivocally. Sometimes it is generic and refers to any tandem repeat polymorphism. At other times, it refers to repeat with a large involving a large number of nucleotides.

Figure 1.2: A tandem repeat polymorphism.



Structural variants

Here, we use the term structural variants to refer to spelling variations that involve deletions or insertions of a nucleotide sequence, inversions, and translocations. When the structural variant is somewhat large (some geneticists define “large” as 1 kilobase or more, others 10 kb), the polymorphism is called a *copy number variant* or *CNV*.² There is considerable research being done on CNVs and medical disorders, including psychopathology (see Section X.X).

Insertion-deletion polymorphisms or an example of which is given in Figure 9.3, are intuitive. Whether an allele is called an insertion or deletion, however, depends on the consensus nucleotide sequence of the human genome. If an allele is missing a nucleotide sequence that is present in the consensus sequence, then the allele is called a *deletion*. If the allele contains a nucleotide sequence that is not in the consensus sequence, then it is an *insertion*. A particularly important type of insertion occurs when a section of DNA is duplicated and inserted into the same region. Remember pseudogenes from Section X.X? These are sections of DNA with a nucleotide sequence very similar to a known gene but the DNA does not produce a functional polypeptide. Most pseudogenes resulted from duplications.

An *inversion* polymorphism occurs when one allele has a nucleotide sequence that is reversed in another allele. Figure 9.4 presents an example. Assume that the spelling variation in the consensus sequence is the one on the top. The inverted allele has a section that has the same spelling but is “read in reverse” from right to left instead of the ordinary left to right order.

²Like many phenomena in molecular biology, the CNV has been defined partly by the laboratory techniques used to detect the polymorphism. A deletion of five nucleotides is difficult to detect with today’s technology, but deletions of several thousands of base pairs can be observed with current automated technology. Hence, the size limit for a CNV is a pragmatic issue not a biological one.

Figure 1.3: An insertion-deletion polymorphism.

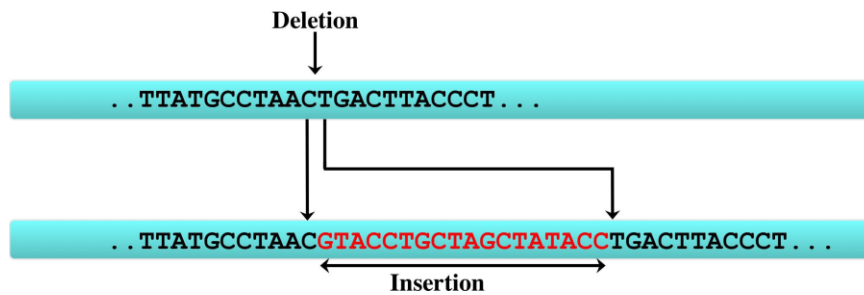
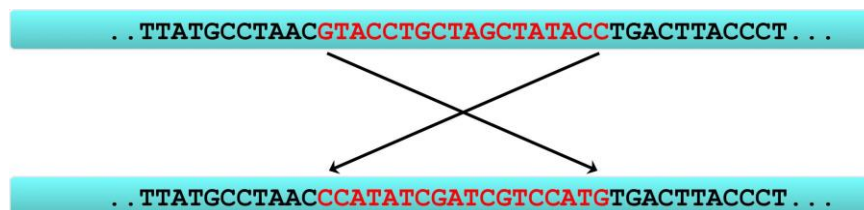


Figure 9.4: An inversion structural variant.



Inversions usually occur when a chromosome breaks in two places and DNA repair mechanisms mistakenly splice the middle fragment back but in reverse order. Typically they involve many thousands of base pairs.

A *translocation* occurs when a section of DNA is deleted from one chromosome and then inserted into another chromosome.

Sequence polymorphisms

The ultimate polymorphism is to actually have the whole sequence of nucleotides for a region for a large number of DNA strands and then examine all of the differences among the strands. Here, the DNA differences could be a SNP, a tandem repeat or a structural change. There is no accepted term for this phenomenon, so we call them *sequence polymorphisms*. In effect, sequence polymorphisms subsume all known DNA polymorphisms.

Basic techniques in molecular genetics

This section is short and merely defines several of the major techniques used in molecular genetics.

BASIC TECHNIQUES

Electrophoresis

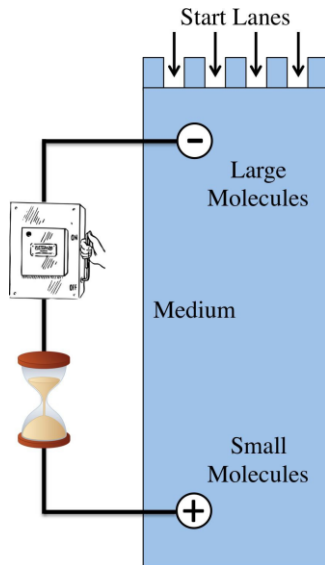


Figure 1.5: Schematic for electrophoresis

Electrophoresis is a generic chemical technique that separates molecules by their molecular weight and/or electronic charge (see Figure 9.5). One places purified biological material in a starting lane in a viscous liquid medium. An electric current is passed through the medium for a specified time. The molecules with a charge opposite to the electrode at the far end of the medium will migrate to the opposite end of the medium. The viscosity of the liquid, however, will impede the migration of large molecules more than small ones. Hence, at the end of a session, the smaller molecules will have moved further from the start lanes than the larger molecules. Current electrophoretic techniques are so sensitive that they can distinguish two DNA or RNA fragments that differ by only a single nucleotide.

Electrophoresis is used to detect tandem repeat polymorphisms and indel (insertion/deletion) polymorphisms. The logic is straightforward. If one allele in a tandem repeat polymorphism has 11 repeats while another has 16 repeats, then the 11 repeat DNA should move further on the electrophoretic medium than the 16 repeat fragment. The problem is that we require a technology to actually *see* the DNA. This is where our next tool—a DNA probe—comes into play.

Probes

A probe is a manufactured fragment of single-stranded DNA or RNA with a predetermined nucleotide sequence. It is introduced into a medium (such as the electrophoretic medium) so that it may bind to its complementary single-stranded DNA or RNA fragment. Usually the probe is comprised of nucleotides with specially colored fluorescent tags that will glow under appropriate lighting. To detect desired DNA fragments in electrophoresis, one “baths” the medium in probes, allowing enough time for them to bind to their complements. Remaining single-stranded probes that did not bind are then washed away and the medium is viewed under ultraviolet light. The result are visible bands in the electrophoretic medium. See the section on the U.S. Federal Bureau of Investigation’s CODIS system (Section X.X) for an example of how this technique is used in forensic applications.

Polymerase chain reaction

Imagine that you are a crime scene investigator who finds a tiny droplet of blood at a crime scene. How can you obtain enough DNA from such a small specimen to perform an analysis. The answer is the polymerase chain reaction or PCR. The technique involves a soup comprised of the DNA that you purified from the specimen, a large number of free nucleotides, some of those “replication stuff” enzymes that produce two copies of DNA from a single copy, and a number of *primers* (a DNA fragment with a nucleotide sequence specific to the DNA area you want to copy).

The first step in PCR is to heat this soup to just about the boiling point of water. This breaks the bonds for double-stranded DNA, making it single stranded. As the mixture cools, the primers in the soup will join with their complementary single-stranded DNAs from the specimen and the “replication stuff” will attach free nucleotides, making them double stranded. Hence, if your specimen had, say, 1,000 copies of the person’s DNA in the white blood cells, then after one round of PCR, you would have 2,000 copies of the desired region. Need more? Then do a second round bringing your total to 4,000 copies. By 15 rounds, you would have over 30 million copies—plenty for analysis.

Detecting polymorphisms

Methods used for detecting polymorphisms depend on the type of polymorphism. One technique genotypes SNPs while another detects tandem repeats. A second consideration is the purpose for genotyping. Some research studies require genotyping a million polymorphisms on many thousands of participants. Here, the cost of an individual genotype must be low. In a

clinical setting, however, the issue may be to confirm or rule out the diagnosis of a genetic or genetically influenced syndrome. Here, a more expensive—but also more discriminating—techniques may be used.

The following is a highly simplified overview of the major techniques used to detect polymorphisms. The purpose is to present the logic of the techniques. As a result, many important laboratory steps are overlooked and over simplified.

DNA is a very long molecule, so the first step in most procedures is *DNA fragmentation*. This “cuts” the DNA into short fragments that can then be used for the procedures. There are several ways to fragment DNA, and they range from the chemical (slice the DNA using enzymes) to physical (force the DNA through a nebulizer).

Tandem repeat polymorphisms

Traditionally, tandem repeat polymorphisms have been assayed with using electrophoresis and then probes. After fragmentation, the relevant loci are amplified through PCR and the PCR products are then separated by electrophoresis. The medium (actually, something that extracts the DNA fragments from the medium) is bathed in the relevant probes. Single-stranded probes that did not bind to their complementary PCR products are washed away. Electrophoresis will then separate the remaining strands according to size.

The nucleotides used in the PCR are special—they have fluorescent tags. Hence, the newly synthesized strands will be visible under the appropriate illumination. A laser sensitive to the fluorescence will scan the output of the electrophoresis and report “hits” to a computer. Which records these and saves the data.

SNPs

Today, detection of SNPs is done through large scale *DNA arrays* often termed *microarrays*. An illustration of how they work is presented in Figure 9.6. The SNP of interest has two alleles—T and C. The first step is to manufacture a single-stranded DNA section that is both unique to and complementary to the T allele. This, of course, will have an adenine (A) in the position complementary to the T. A second single-stranded DNA probe is manufactured that is unique to and complementary to the C allele; it will have a G. A technique like PCR is then used to make a very large number of these sections. Then, the A strands are glued onto a very tiny area of the array and the G strands onto a tiny adjacent area.

Next DNA is extracted from a biological specimen taken from the person. The DNA is purified, cut into millions of fragments and then amplified using PCR. But this amplification is done with a twist. Instead of an ordinary batch of free nucleotides, the nucleotides used in the PCR have fluorescent tags that will make them glow when viewed under a certain light. The whole DNA array is then batched in the fluorescently labeled, single-strand DNA from

the PCR. The DNA strands from the PCR will tend to bind with their complementary single stranded DNA fragments on the DNA array. After a suitable time, the remaining single stranded DNA is washed from the array. The array is exposed to the special light and a laser scans the array. As the array is being scanned, the areas that fluoresce are detected by the laser and recorded in a computer.

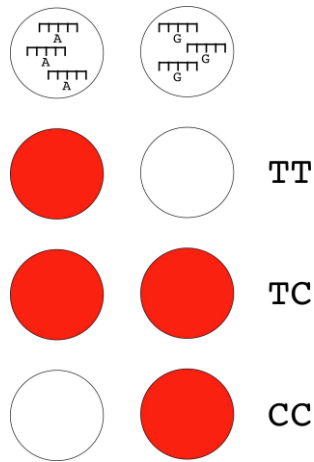


Figure 1.6: how DNA array detect SNPs.

Now, return to our polymorphism in Figure 9.6 and ask what the laser would “see” if the person’s genotype were TT. The single-stranded DNA with the T would bind to the complementary A strand on the chip, making the A area light up. Since the person lacks the C allele, there would be no strand complementary to the G section of the array. Hence, this area would not light up. This pattern is seen in the second row in Figure 9.6.

If the person were homozygous for the C allele, then we would observe the opposite pattern, i.e., the one on the last row of the figure. The area with the G DNA strand would fluoresce while that with the A strand would remain unlighted.

Finally, the DNA from a TC heterozygote would bind with both the A and the G sections, giving the pattern in the third row of the figure.

CNVs

There are many different ways to detect copy number variants (CNVs). Here, the purpose is paramount. Consider testing for a microdeletion in clinical cytogenetics. A *microdeletion* is a deletion involving many kb but is too small to detect using traditional karyotypes. Usually, the medical doctor suspects that an infant or young child may have a specific syndrome due to a microdeletion and requests tests to confirm or rule out that syndrome. Hence, the test is for one CNV and there is no need to use a method for cataloging all of the thousands of known CNVs.

There are many techniques used to detect CNVs in research designs intended to see which CNVs may be associated with a disorder or trait. One strategy is digital or virtual karyotyping (Wang et al., 2002). Here, one uses existing strategies for detecting polymorphisms and then applies software to look for CNVs. For example, a CNV deletion could be detected in a DNA array that assays SNPs by looking for a region where there is no heterozygosity.

Next generation sequencing

The Holy Grail for genotyping an individual is to obtain the complete nucleotide sequence of the person's genome. The Human Genome Project sequenced one human genome. It took about 10 years and cost three billion dollars. Today a variety of new technologies are emerging to sequence an individual's genome (Koboldt et al., 2010; Mardis, 2013). Collectively, they are called *next generation sequencing (NGS)* technologies or *massive parallel sequencing*. It is too early to predict which ones will prevail, but early results on the potential of NGS are striking. The current goal is the \$1K genome, i.e., a procedure to obtain an individual's genome for \$1,000 US.

Despite using very different laboratory methods, the logic of most NGS strategies is the same. The DNA is fragmented and then amplified. The PCR products are then sequenced in parallel. That is, millions to billions of the fragments are sequenced at the same time and the results stored into a computer.

Finally, computer algorithms are used to "align" the short segments into a long sequence. There are several varieties of NGS. Although we have spoke of sequencing a whole genome, *targeted sequencing* selects specific regions of the genome for sequencing (Koboldt et al., 2010). A particular type of targeted sequencing selects the exons and nearby regions, providing a sequence of what is called the *exome*. Another NGS strategy is to focus on the various types of RNA, particularly mRNA in the study of gene expression in animal brains (Hitzemann et al., 2013).

Because so much of the methodology for NGA is still in the development stage, there are some rough spots with the technology. Standard and protocols for NGS in research have been proposed (Goldstein et al., 2013), but it will take several years of data collection to come up with accepted standards.

Eventually, NGS will supersede all of the methods mentioned above for detecting polymorphisms. The major reason is that with a genome at hand, all one needs is software to search it, compare it to a genomic library of variants, and spit out the tandem repeats, SNPs, and structural variants. That said, such a technology is not readily available today. Currently, supercomputing resources are required to store the data, align the fragments, and arrive at an unambiguous sequence. Still, advances in technology—on both the chemistry and the data side—will make it possible for individuals to obtain their own genomic sequence in the future.

One current parameter underlying current sequencing is the *number of reads* denoted as X . Having fractionated and amplified the DNA, the number of reads is, roughly speaking, the number of times that this biological material is put through the sequencing step that “reads” the DNA sequence. A 1X sequence is the cheapest and does it once. A 25X sequence performs it 25 times. There is no gold standards for X . The number of reads all depends on the purpose at hand.

The 1,000 Genomes Project is the latest, large-scale, international attempt to catalog human genetic variation (The 1,000 Genomes Project Consortium, 2012). Using several NGS technologies, it has reported the nucleotide sequence of 1,092 people of different ethnic groups throughout the world. It estimates that the human genome contains 38 million single nucleotide polymorphisms, 1.4 million short indels (insertion/deletion polymorphisms), and 14,000 copy number variants.

NGS and personalized medicine

There is considerable speculation about the implications of the \$1K genome for personalized medicine. Personalized medicine involves customization of medical procedures and therapeutics so that they apply to the individual, not to the collection of individuals with a certain disorder. We have all experienced it to a certain degree. For example, hay fever (allergic rhinitis) sufferers respond differently to the antihistamines used to manage the problem. The typical course of treatment is to try this drug and then that one until, by chance, the patient arrives at one that controls the symptoms with a minimum of annoying side effects. The goal of personalized medicine is to develop tests that predict how a patient will respond to each drug and then start with the one likely to be the most efficacious.

References

Mardis, E. R. (2013). Next-generation sequencing platforms. *Annual review of analytical chemistry (Palo Alto, Calif.)*, 6:287–303.

Wang, T. L., Maierhofer, C., Speicher, M. R., Lengauer, C., Vogelstein, B., Kinzler, K. W., and Velculescu, V. E. (2002). Digital karyotyping. *Proceedings of the National Academy of Sciences of the United States of America*, 99(25):16156–16161.

Genetic Variation

CONTENTS

1. Genetic Variation
2. Mutation
3. Molecular basis of gene mutation
4. Types of Mutation
5. Agents

Genetic Variation

Genome is not a static entity. It is dynamic in nature it is subject to different types of heritable genetic changes. A genetic change in the genetic material of an organism that gives rise to alternate forms of any gene is called **mutation**. The process why which mutations is produced is called **mutagenesis**. This may occur spontaneously or be induced by mutagens. An organism exhibiting a novel phenotype as a result of the presence of a mutation is referred to as a **mutant**. In a broad sense, the term mutations includes all types of heritable genetic changes of an organism not explainable by recombination of pre-existing genetic variability.

General characteristics of mutation

- Mutations are generally recessive, but dominant mutations also occur.
- Mutations are generally harmful to the organisms.
- Mutations are random, occur at any time and in any cell of an organism.
- Mutations are recurrent i.e. the same mutation may occur again and again.

Role of mutation

- Ultimate source of all genetic variation and it provides the raw material for evolution.
- Organism would be able to evolve and adapt to environmental change.
- Mutation results into the formation of **alleles**. Any mutation occurring within a given gene will thus produce a new allele of that gene. Without mutation, all genes would exist in only one form. Genes containing mutations with no effect on phenotype or small effects that can be recognized only by special techniques are called **isoalleles**.

Figure: Tautomeric forms of the four common bases DNA. The shifts of hydrogen atoms between the number 3 and number 4 positions of the pyrimidines and between the number 1 and number 6 positions of the purines change the base-pairing potential of the bases.

However, if a base existed in the rare form at the moment that it was being replicated or being incorporated into a nascent DNA chain, a mutation might result. When the bases are present in their rare imino or enol states, they can form adenine-cytosine and guanine-thymine base pairs. The net effect of such an event, and the subsequent replication required to segregate the mismatched base pair, is an AT to GC or a GC to AT base pair substitution.

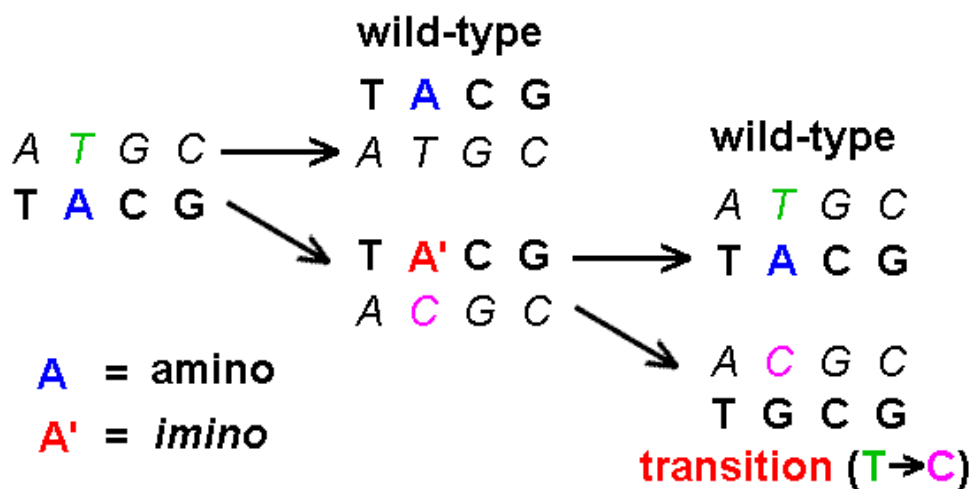


Fig: Mutation via tautomeric shifts in the bases of DNA

Error during replication may result in substitution mutation (also known as *point mutation*, *simple mutation* or *single site mutation*) or frameshift mutation.

Substitution mutation: A gene mutation that results from the substitution of one base pair for another (or one base for another in the case of single-stranded DNA genomes) is known as **substitution mutation**. It may be transition mutation as *mutation* or *transversion mutation*.

Transition: Transition, the most common *class*, comprising the substitution of one pyrimidine by the other or one purine by the other. This replaces a G.C pair with an A.T pair or vice versa.

Transversion: If purine is replaced by a pyrimidine, or pyrimidine is replaced by a purine then it is known as *transversion*.

Term	Mutation
Transition	G — A; A — G; C — T; T — C

Transversion G— T; G— C; A—T; A —C
 T— A; T— G; C—A; C— G

Most mispairing mutations are transitions. This is likely to be because an A•C or G•T mispair does not distort the helix as much as A•G or C-T base pairs do. However, transversions also can occur through mispairing.

Frameshift mutation: Aberrant replication can also result *in* small numbers of extra nucleotides being inserted into the polynucleotide being synthesized, or some nucleotides *in* the template not being copied. Addition or deletion of base pair that occurs within the protein-coding portion of a gene have the effect of shifting the translational reading frame. In majority of the cases, these results in a failure to synthesize a functional protein, thus, allowing the mutation to be identified by its phenotypic consequences. Because these mutations cause a shift in the translational reading frame, they are termed *frameshift* mutations. Frameshift mutations are particularly prevalent when the template DNA contains short repeated sequences. This is because repeated sequences can induce replication slippage, in which the template strand and its copy shift their relative position so that part of the template is either copied twice or missed out. The result is that the *new* polynucleotide has a larger or smaller number, respectively, of the repeat units.

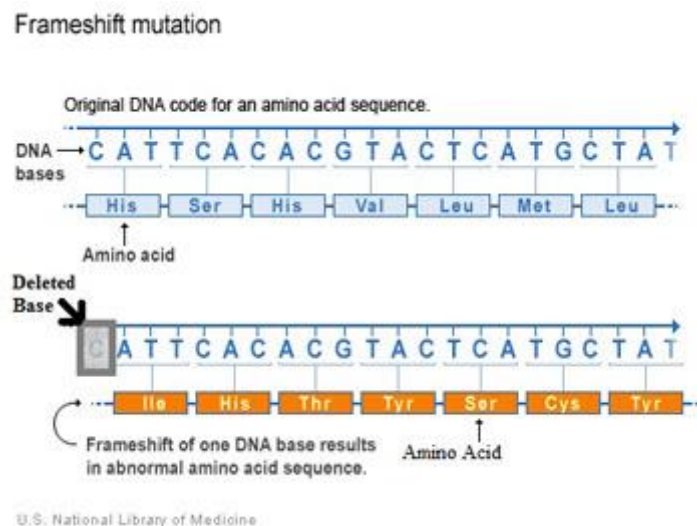
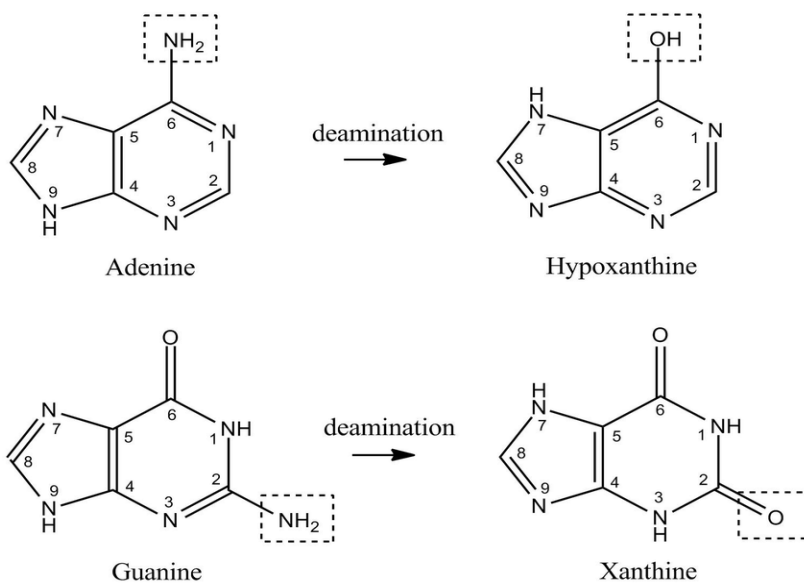
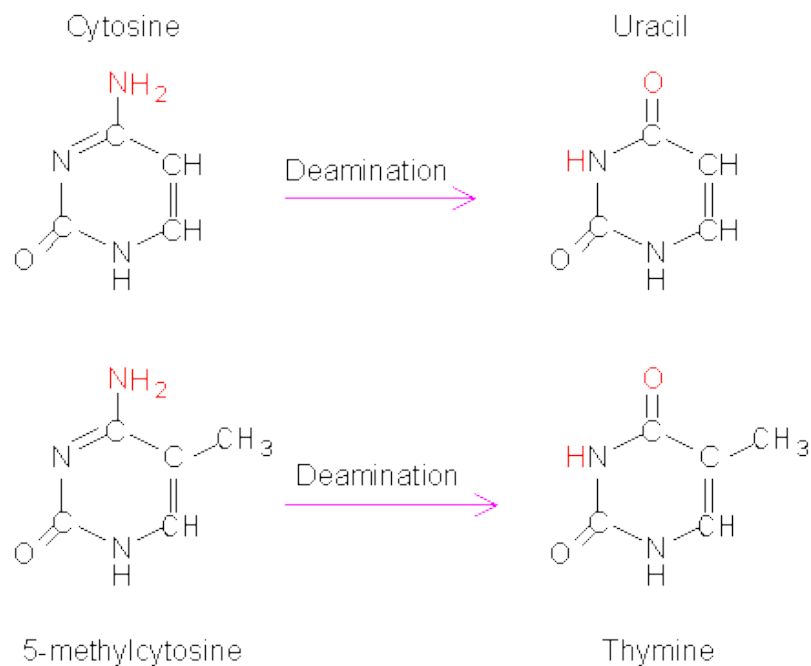


Fig:Frameshift mutation that results from the addition and deletion of a single base pair.

Spontaneous lesions

Naturally occurring damage to the DNA, called *spontaneous lesions*, also can generate substitution or frameshift mutations. Two of the most frequent spontaneous lesions are *depurination* and *deamination*, the former being more common.

Deamination: It is the removal of an amino group from a molecule. Three of the four nitrogenous bases normally present in DNA (cytosine, adenine and guanine) contain exocyclic amino group. Deamination of cytosine results in the formation of uracil. Similarly, deamination of adenine and guanine result in the formation of hypoxanthine and xanthine, respectively. Deamination of 5-methylcytosine gives thymine. Deamination changes the standard base pairing patterns. For example, xanthine results from deamination of guanine selectively base pairs with thymine instead of cytosine. Hypoxanthine selectively base pairs with cytosine instead of thymine.



Depurination and depyrimidation: depurination is the loss of purine due to breaking of glycosidic bond of nucleotides in DNA. Similarly, depyrimidation is the loss of pyrimidine base. Hydrolysis of the N-glycosidic bond (via base protonation by water) of purines and pyrimidines leading to the generation of base-free (abasic) sites or AP (for apurinic/apyrimidinic) sites. Depurination and depyrimidation are more common at acidic pH. However, the processes can also occur at appreciable rates at neutral or alkaline pH. The mechanism of depyrimidation is the same as for depurination, but at a substantially lower rate. Depurinated bases in single-stranded DNA undergoing replication can lead to mutations, because in the absence of information from the complementary strand, DNA polymerase can add an incorrect nucleotide at the apurinic site, resulting in either a transition or transversion mutation.

Oxidative damage: Besides deamination and depurination/depyrimidation, attack by *reactive oxygen species* must be considered as a major source of spontaneous damage to DNA. Radicals (like superoxide radicals and hydroxyl radicals), attack on DNA can produce a variety of products. For example, 8-oxo-deoxyguanosine (⁸-oxo-dG) and thymidine glycol are products of oxidative damage. The 8-oxo-dG (an oxidized derivative of deoxyguanosine) mispairs with A, if it base-pairs with A during replication, it gives rise to a G:C to T:A transversion.

Transposition

A *transposon* is a DNA sequence that is able to insert itself (or a copy of itself) at a new location in the genome, without having any sequence relationship with the target locus. Insertion of a transposable element into or near a functional gene can alter its expression by causing loss of gene function or by changing the gene's expression (*insertion mutations*).

Mutagen

Mutagens are chemical or physical agents that increases the occurrence of mutations. Biological agents, for example certain viruses or transposable genetic elements, can also increase the frequency of occurrence of mutations and thus act as mutagens. Mutagens induce mutations by at least three different mechanisms. They can replace a base in the DNA, alter a base so that it specifically mispairs with another base, or damage a base so that it can no longer pair with any base under normal conditions.

Physical mutagens

Physical mutagens include ionizing (e.g. X-ray, γ -ray) and non-ionizing (e.g. UV-ray) radiations. H.J. Muller first demonstrated that X-rays can cause mutations. He found that X-rays treatment markedly increased the frequency of sex-linked recessive lethal mutation in *D. melanogaster*. Radiation damage to DNA is ascribed to both direct and indirect effects. Direct effects result from the direct interaction of the radiation energy with DNA. Indirect effects result from the interaction of DNA with reactive species formed by the radiation. Ionizing radiation has various effects on DNA depending on the type of radiation and its intensity.

Ionizing radiation mainly causes strand breakage (i.e. acts as a clastogenic agent that causes chromosomal breakage). Some types of ionizing radiation act directly on DNA, while others act indirectly by stimulating the formation of reactive molecules such as hydroxyl radicals in the cell. Because of the aqueous nature of biological systems, many different types of reactive oxygen species generated by the effects of ionizing radiation on water cause the most damage. These species can damage bases and cause different adducts and degradation products.

Non-ionizing radiations such as UV radiation also acts as a potent physical mutagen and generates a number of photoproducts in DNA. The UV radiation spectrum has been subdivided into three wavelength bands designated UV-A (400 to 320 nm), UV-B (320 to 280 nm) and UV-C (280 to 100 nm). UV radiation of 260 nm (which corresponds to the DNA absorption peak) induces dimerization of adjacent pyrimidine bases, especially if these are both thymines. Adjacent pyrimidines become covalently linked by the formation of a four-membered cyclobutane ring structure, resulting from the saturation of their respective 5, 6 double bonds. The structure formed by this photochemical cycloaddition is referred to as a **cyclobutyl dimer**.

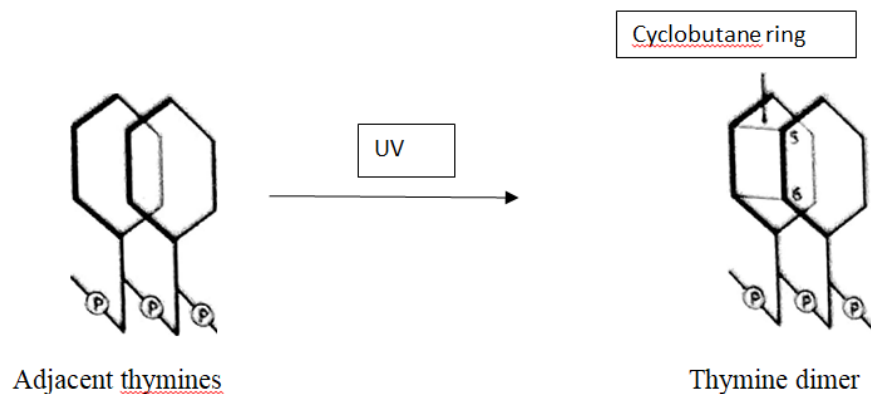
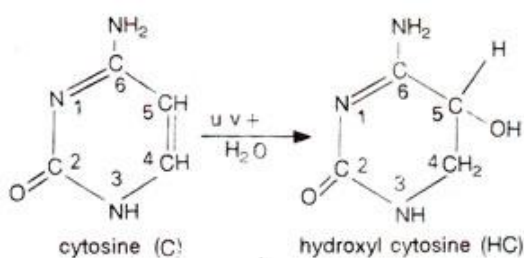


Fig: Pyrimidine dimer, UV-radiation stimulates the formation of a 4-membered cyclobutane ring between two adjacent pyrimidines on the same DNA strand.

Thymine dimer is the most common form of pyrimidine dimer. Other pyrimidine combinations also form dimers, in the order of frequency being 5'-CT-3' > 5'-TC-3' > 5'-CC-3'. Purine dimers are much less common. Another type of UV-induced photoproduct is the (6-4) lesion in which carbon number 4 and 6 of adjacent pyrimidines become covalently linked. UV-induced dimerization usually results in a deletion mutation when the modified strand is copied. Ultraviolet irradiation of DNA also produces cytosine hydrate. It is the addition of a molecule of water across the 5, 6 double bond to form a hydroxy derivative.



Heat stimulates the water-induced cleavage of the N-glycosidic bond that attaches the base to the sugar component of the nucleotide. This occurs more frequently with purines than with pyrimidines and results in an AP (apurinic/apyrimidinic) site. The sugar-phosphate that is left is unstable and rapidly degrades, leaving a gap if the DNA molecule is double-stranded. This reaction is not normally mutagenic because cells have effective systems for repairing nicks.

Chemical mutagens

A large number of chemicals act as mutagens. So, it is very difficult to devise a classification that includes all. Broadly, chemical mutagens can be classified into three categories - base analogs, base modifiers and intercalating agents. These chemicals may cause substitution or frameshift mutations.

Base analogs

Certain bases that are not normally present in DNA but bear a strong structural resemblance to normal nitrogenous bases can be incorporated from the appropriate triphosphate precursor during DNA synthesis. These compounds are called base analogs. These chemicals are mutagenic only to replicating DNA.

For example, 5-bromouracil (5-BU) is an analog of thymine that has bromine at the C-5 position in place of CH₃ group found in thymine. 5-BU has the same base-pairing properties as thymine, and nucleotides containing the base can be added to the daughter polynucleotide at positions opposite as in the template. The common keto form of 5-BU pairs with adenine. But 5-BU can frequently change to the enol form. The mutagenic effect arises because the equilibrium between the two tautomers of 5-BU is shifted more towards the rare enol form than is the case with thymine.

It means that during the next round of replication, there is a relatively high chance of the polymerase encountering enol-5BU, which (like enol-thymine) pairs with G rather than A. This results in a point mutation.

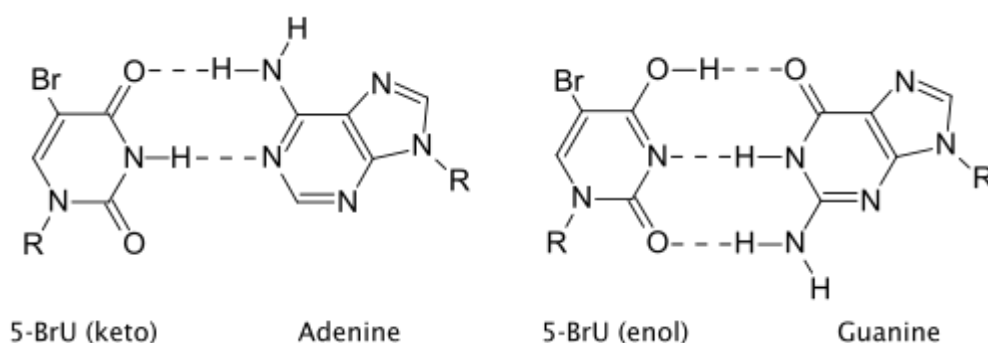


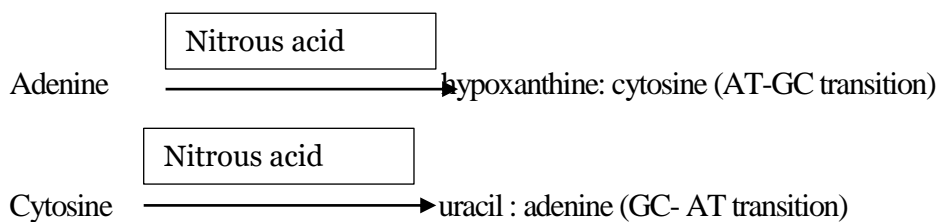
Fig: Mode of action of 5-bromouracil (AT-GC Transition)

2-Aminopurine (2-AP) acts in the similar way. It is an analog of adenine with an amino-tautomeric form pairs with thymine and an imino tautomeric form pairs with cytosine. But the imino form of 2-AP being more common than imino form of adenine and hence inducing T-to-C transitions during DNA replication (AT - GC transition).

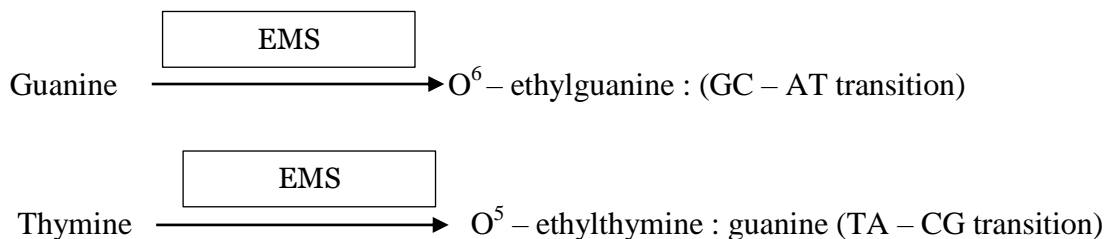
Base modifiers

Some mutagens are not incorporated into the DNA but instead modify or alter a base, causing specific mispairing. These mutagens are called base modifiers. These mutagens can be deaminating agents, alkylating agents and others.

Deaminating agents cause point mutations by removal of an amino group from bases. Such chemicals are nitrous acid, which deaminates adenine, cytosine and guanine (thymine has no amino group and so cannot be deaminated), and sodium bisulfite, which acts only on cytosine. Deamination of adenine gives hypoxanthine, which pairs with C rather than T, and deamination of cytosine gives uracil, which pairs with A rather than G.



Alkylating agents cause alkylation of nitrogenous base by covalently linking an alkyl group. Chemicals such as Ethyl Methane Sulfonate (EMS, also termed as *nitrogen mustard gas*), Methyl Methane Sulfonate (MMS), Nitrosoguanidine (NTG) and dimethylnitrosamine add alkyl groups to nucleotides in DNA molecules and are called alkylating agents.



Hydroxylamine also acts as mutagen but its mode of is different from alkylating agents. It causes GC-AT transition. It preferentially hydroxylates the amino nitrogen of cytosine, creating N-4-hydroxycytosine, which can mispair with adenine.

Intercalating agents are usually associated with single 'nucleotide-pair insertions and deletions (frameshift mutation). This group of compounds includes Proflavin, acridine orange, ethidium bromide and a class of chemicals termed SCR compounds (ICR⁻170, ICR⁻191 and so on).

Types of mutation

Somatic versus germinal mutations

In multicellular organisms, genes can mutate in either somatic or germinal tissue, and the changes are called somatic mutations and germinal mutations, 'respectively. A germinal mutation arises in the germ line, a special tissue that is set aside during development to form gametes. If a mutant gamete participates in fertilization, then the mutation will be passed on to the next generation.

Hereditary versus acquired mutations

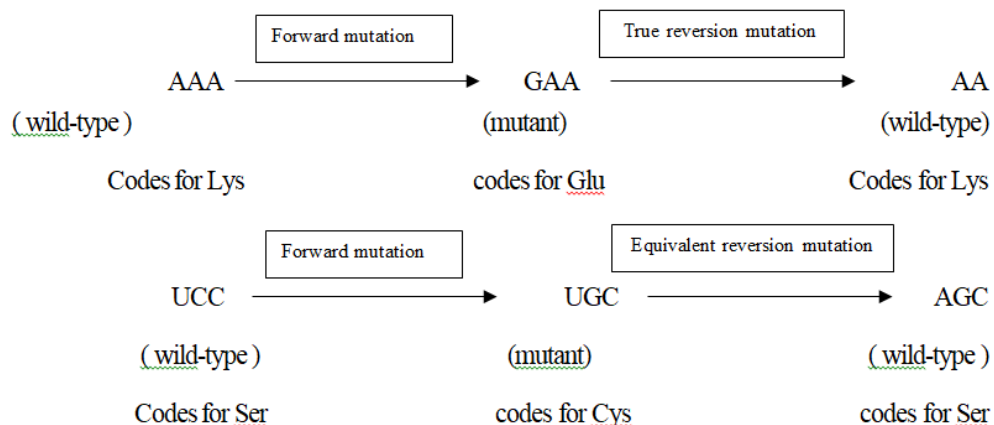
Gene mutations occur in two different ways: they can be inherited from a parent or acquired during a person's lifetime. Mutations that are passed from parent to offspring's are called hereditary mutations. This type of mutation is present throughout a person's life in virtually every cell in the body.

Mutations that occur in the DNA of a cell at some time during a person's life are termed acquired mutations. These changes can be caused by environmental factors such as ultraviolet radiation, or can occur if a mistake is made as DNA copies itself during cell division. Acquired mutations in somatic cells (cells other than sperm and egg cells) cannot be passed on to the next generation.

Forward versus back mutation

A mutation that changes the phenotype from wild-type to a mutant phenotype is said to be a forward mutation, (reversion a mutation that causes a change of the phenotype from mutant to wild-type is said to be a back mutation (reversion or reverse mutation). Restoration of the original phenotype by reversion may occur:

By a true back mutation at the same site in the gene as the original mutation, restoring the wild-type nucleotide sequence (change in same codon and same nucleotide). By an equivalent reversion which does not generate original condition at the gene level. However, it restores the wild-type phenotype due to formation of synonymous codon.



Suppressor mutations

A suppressor mutation is a second mutation that restores a function lost by the first mutation. Mutations of this kind are called suppressor mutations because they suppress the effects of the first mutations. True back mutation restores the original wild-type nucleotide sequence of the gene, whereas a suppressor mutation does not. Suppr_{ess} mutations may occur at distinct sites in the same gene as the original mutation or in different genes, e_{ven} On different chromosomes. So, a suppressor mutation can be intragenic (if a mutation occurs at distinct sites With_{in} the same gene) or intergenic (if a mutation occurs in a different gene).

Intragenic suppressor mutations can restore the activity of a mutant protein by many means. For example, the original mutation may have made an unacceptable amino acid change that inactivated the protein, but changing another amino acid somewhere else in the polypeptide could restore the protein activity.

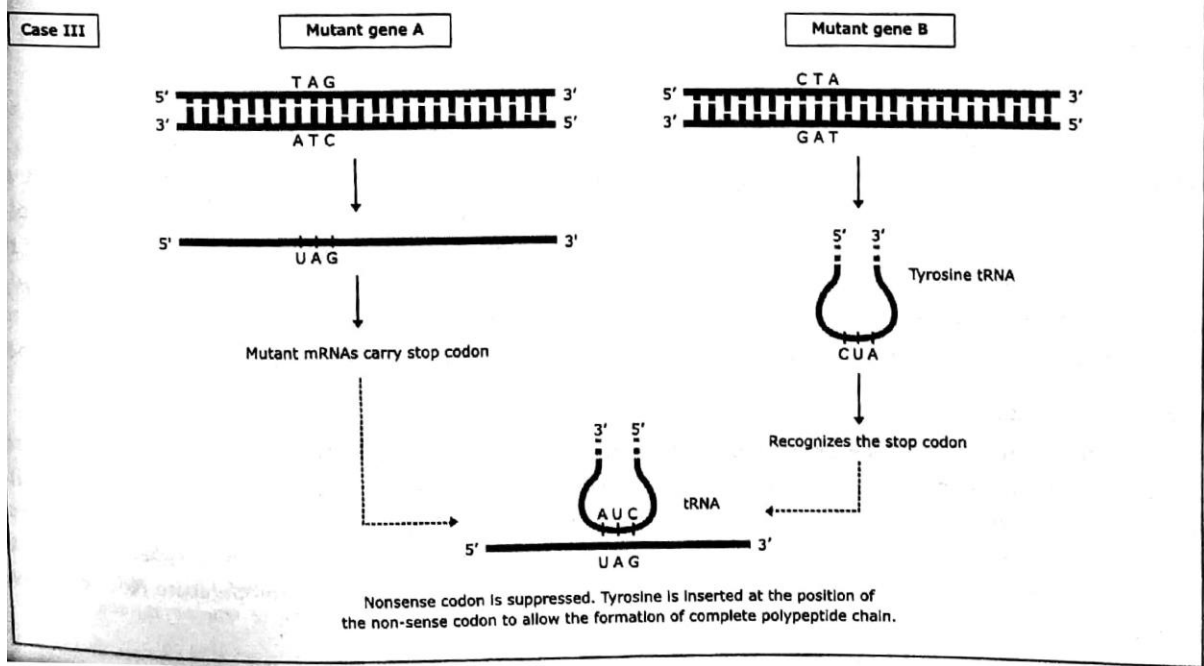
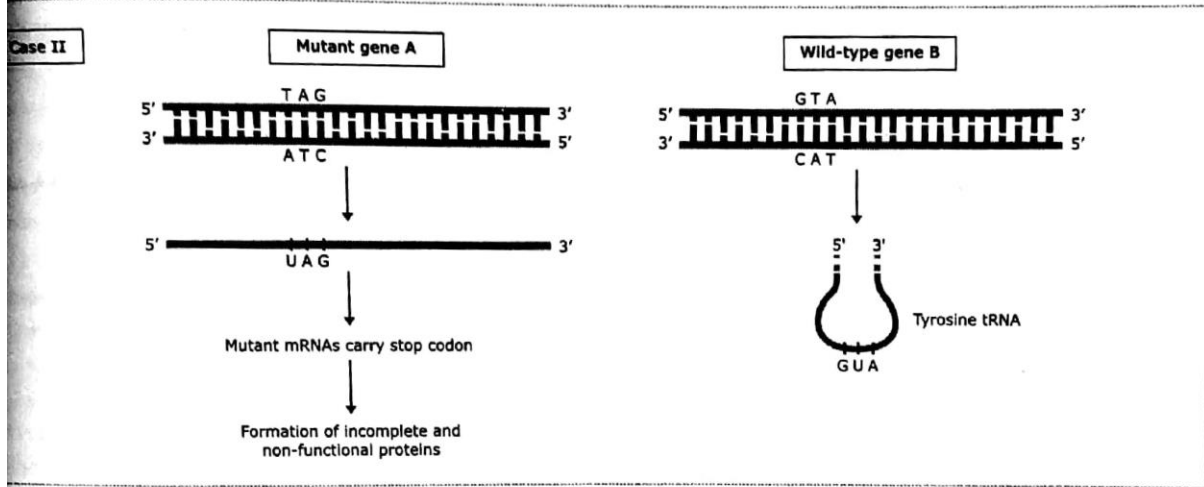
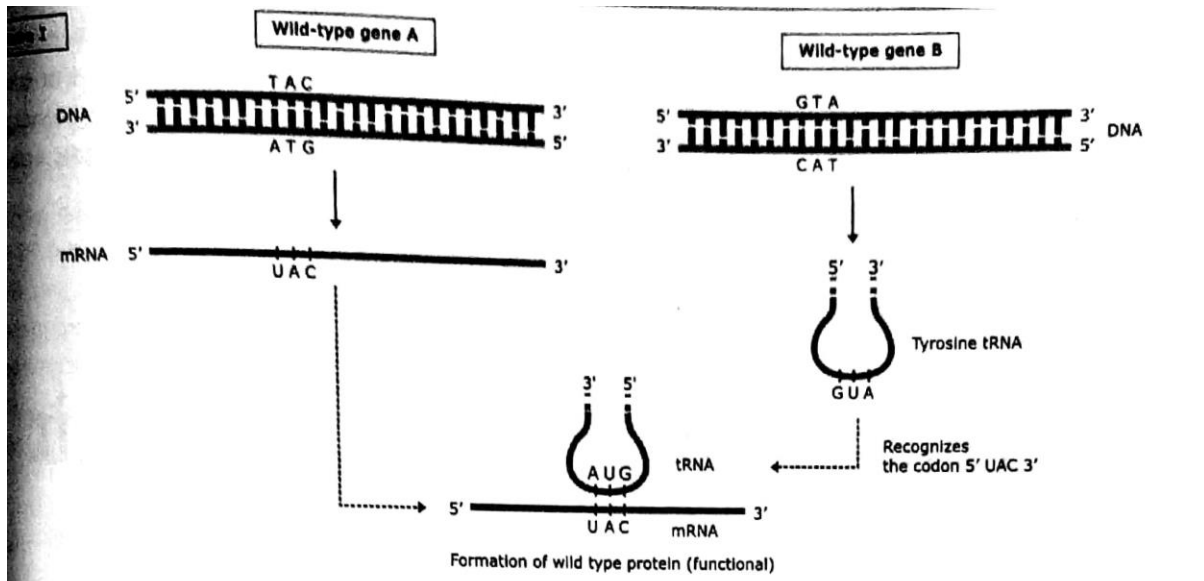
The suppression of *one* frameshift mutation by another frameshift mutation in the same gene is another example of intragenic suppression. If the original frameshift resulted from the removal of a base pair; the addition of another base pair close by can restore correct reading frame.

Intergenic (or extragenic) suppressors do not occur in the same gene as the original mutation. There are many ways in which intergenic suppression can occur. The suppressing mutation may restore the activity of the mutated gene product or provide another gene product to take its place.

One of the best known examples of intergenic suppressor mutation is a mutant tRNA gene that suppresses the effects of a nonsense mutation in a protein-coding gene (*nonsense suppression*). Let us take one wild-type gene 'A', encoding a tRNA that recognizes a 5' UAC 3' codon in the mRNA and inserts tyrosine into the growing polypeptide chain. A mutation in the gene changes the anticodon so that it recognizes the stop codon 5' UAG 3' in the mRNA and, instead of terminating, Inserts a tyrosine at that position in the polypeptide chain.

Missense and nonsense mutation

The mutation that alters the codon so that it specifies a different amino acid is known as *missense* (non- synonymous) mutation. One common example of missense mutation is *sickle cell hemoglobin*. Hemoglobin is the oxygen transporting macromolecule present in the red corpuscles of chordate animals. Hemoglobin A (adult hemoglobin contains two identical alpha-chains and two identical Beta-chains. Each alpha polypeptide consists of a specific sequence of 141 amino acids. The Beta-chains are 146 amino acids long.



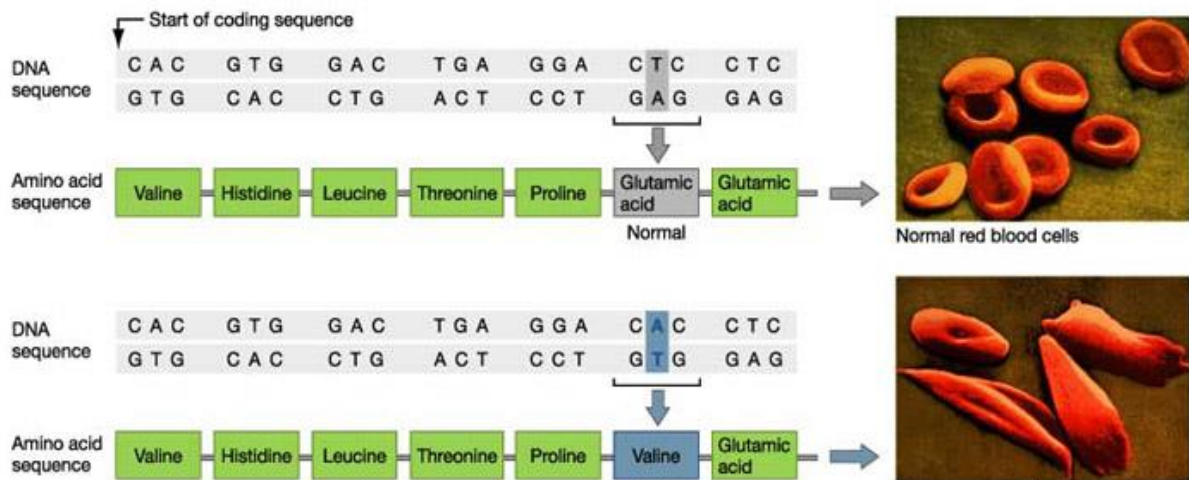


Figure 1.5 Mutational origin of sickle-cell hemoglobin (hemoglobin 5). The observed substitution of valine at amino acid position 6 in the 13-chain of hemoglobin 5, one can deduce that the mutation gMng rise to hemoglobin S gene is the substitution of an adenine for a thymine in the transcribed strand of DNA.

In sickle cell hemoglobin (Hemoglobin 5) sixth amino acid i.e. glutamic acid (a negatively charged amino acid) from the amino terminal end of the p-chain is replaced by valine (no charge at neutral pH). This changes the shape of hemoglobin. Mutation that changes a codon in a gene to one of the three termination codon (UAA, UGA or UAG) is described as *nonsense mutation*. Nonsense mutation results in a shortened protein because the translation of the mRNA stops at this new termination codon.

Silent and neutral mutation

Not all mutations in DNA lead to a detectable change in the phenotype. Mutations without apparent effect are called *silent mutation*. If a mutation in which the new codon specifying the same amino acid as the unmutated codon then it a case of silent or *synonymous mutation*. Because it has no effect on the coding function of the genome: the mutated gene codes for exactly the same protein as the unmutated gene. For example, the change of ACG into CGG, both code for an arginine. Codon specifies different but functionally equivalent amino acid and not ,alter protein function called *neutral mutation* (for example, AAA-AGA : changing basic lysine to bask arginine).

Loss- and gain of function mutations

In principle, mutation of a gene might cause a phenotypic change in either of two ways:

- Loss of function (*null*) mutation : the product may have reduced or no function.
- Gain of function mutation : the product may have increased or new function.

Because mutation events introduce random genetic changes, most of the time they result in loss of function. Generally, loss of *function* mutations are found to be recessive. In a wild-type diploid cell, there are two wild-type alleles of a gene, both making normal gene product. In heterozygotes, the single wild-type allele may be able to provide enough normal gene product to produce a wild-type phenotype. In such cases, loss of function mutations are recessive. However, some loss of function mutations are dominant. In such cases, the single wild-type allele in the heterozygote cannot provide the enough amount of gene product needed for the cells to be wild-type. Gain of function mutations usually cause dominant phenotypes, because the presence of a normal allele does not prevent the mutant allele from behaving abnormally.

References: Gardner, BD Singh, Wikipedia (images).

Population Genetics and Evolution

Contents:

Introduction

Hardy Weinberg Equilibrium (HWE)

Importance and implications of Hardy Weinberg Equilibrium

Applications in human population genetics

Departure from HWE

Factors affecting change in gene frequency:

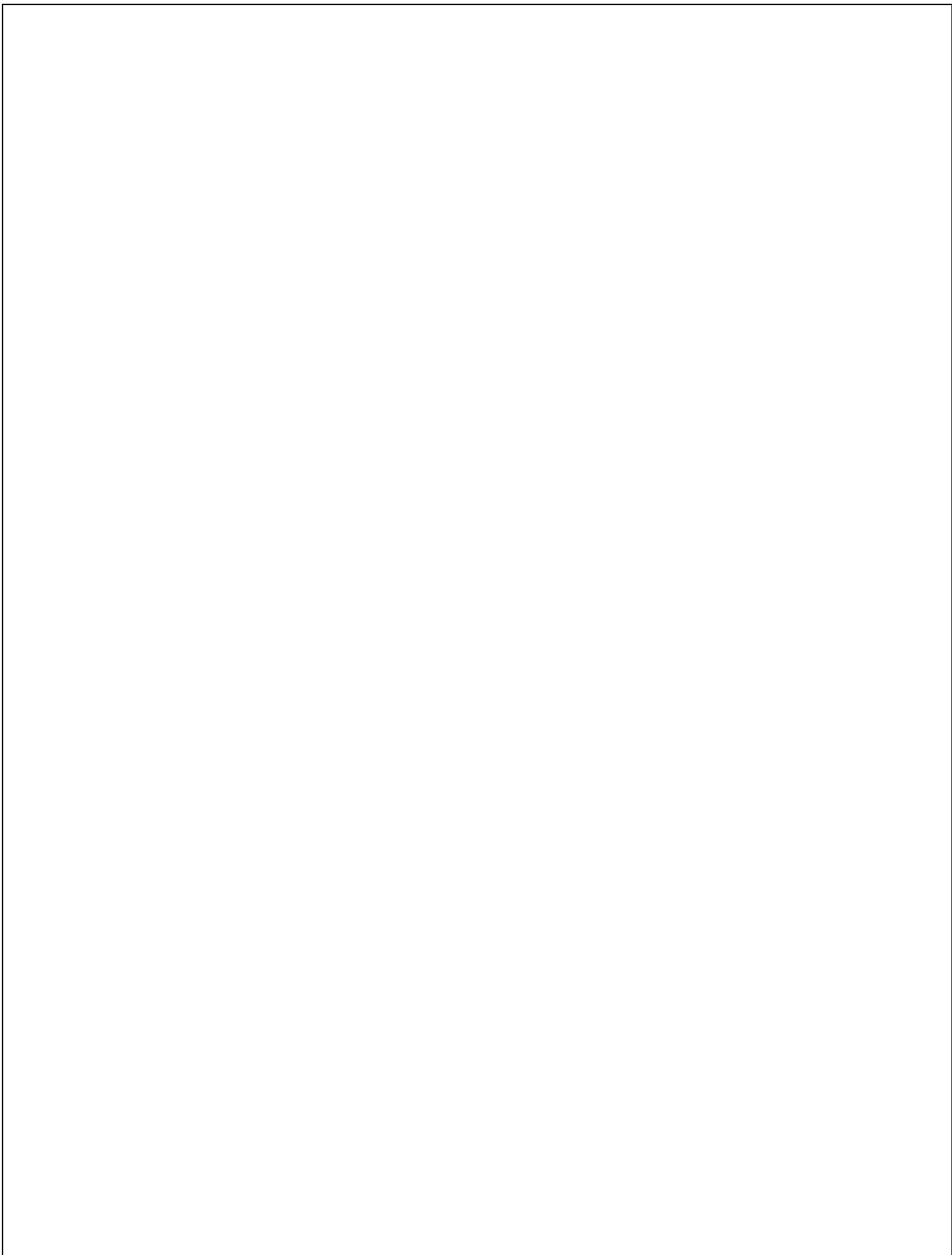
Mutation, Genetic Drift,

Natural Selection and

Gene Flow

Genetic Equilibrium

Summary



Introduction

Living organisms are endowed with unique abilities, traits that allow them to survive in a given environment. These traits or abilities may show or exhibit enormous variations within species and across species. Some of these traits are unique to that species; some traits are common within and across species with little variation, these are adaptive characters and gives survival advantage.

These traits are the '*phenotypic*' forms that can be observed as a quantitative trait (or measurable) or classified as types or categories. These traits are hereditary and transmitted across generations: either in the same form or in slight variable form. At times some new traits or variations of the trait appear among the offspring. Some of these traits are governed by '*genes*' or located in the '*genome*' of an organism. The nature of heredity of some of these traits could be complex and/or it could follow some simple principles of transmission.

Human population genetics deals with how these traits or variation change in a population over space and time (generations)? What are the factors that influence the variation of these traits in the population? To what extent these traits are hereditary and are influenced by environment? Can we understand them by simple theoretical models? Can we study how different forces operate differentially in different populations to give a characteristics distribution of gene and genotype frequencies?

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.

2.2 Hardy-Weinberg Equilibrium

During the early 1900s people were interested to validate the Mendelian laws of genetics to other organisms, including Man. Are there Mendelian traits in Man?

a) Mendelian genetics in Man

At Cambridge one research scholar was studying ‘brachydactyly’ – a trait characteristic of small or short digital length (*‘brachy’* and *‘dactyl’* in Greek means ‘short’ and ‘digit’ respectively) than the normal type. The trait runs in some families. Does ‘brachydactyly’ follow Mendelian principles? The results of the study showed that ‘brachydactyly’ is dominant Mendelian trait and the pedigrees showed 3:1 ratio of brachydactyly to normal offspring. This has invoked an important and interesting question? *If it is a dominant trait, there will be more and more brachydactyly individuals in the population, but normal individuals are more frequent than brachydactyly individuals (See Box 2.1)*

BOX 2.1

Brachy dactyly in a population: All the mating types are:

The mating types include individuals Normal and Brachydactyly (a Dominant Mendelian trait)

N – Normal and B – Brachydtly

- Both parents are normal -- All the offspring are N
- One parent is N and the other B (heterozygous)
- One parent is N the other B
- Both parents are brachydactyly B (heterozygous)
- Both parents are brachydactyly B (homozygous)

Of the 5 possible combinations of parental mating types

4 types of matings results brachydactyly offspring

Therefore, B are more frequent than N in a population

as per Mendelian expectation. However Normal

individuals are more frequent than Brachydactyly in a population.

there is apparent contradiction between what is observed and what is expected (Mendelian)!

a NN – NN = All the offspring N

b NB – NN = ½ : ½ = N : B

c BB – NN = all are B

d NB – NB = 1 N : 1 B: 2 B

e BB – BB = All the offspring B

GH Hardy has solved the puzzle theoretically and published the theorem in Science (Hardy, 1908). **GH Hardy's proof illustrates that the gene (or allele) frequency**, -- here in this case, frequency of brachydactyly individuals in a population, -- **will not increase over generations, but remain the same, under equilibrium conditions or in the absence of confounding variables. In 1908, Dr. W. Weinberg independently also published similar results (Weinberg, 1908) and is called as HWEquilibrium.** (See Box 2.2)

BOX 2.2

Historical anecdotes: HWEquilibrium/Law

In 1908, a German physicist Dr. Weinberg published similar results on Mendelian genetics in a German journal. It was discovered by Dr. Curt Stern (publication in Science 1943), and the Hardy theorem was rightfully referred as HW Law or HWEquilibrium. However, in 1903, at least five years earlier, two scientists have considered similar such possibility of change in gene frequency. They are: WE Castle 1903 in America and Karl Pearson 1903 in England. These two papers considers the question of equilibrium state of gene frequency and change in gene frequency partially with respect some factors.

a) What is H-W EQUILIBRIUM/LAW?

HWE states that *in a randomly mating population of sufficiently large size, and in the absence of the influencing factors such as; mutation, migration, selection, genetic drift and inbreeding, the gene and genotype frequencies will remain constant from generation to generation.*

The mathematical proof of invariance of gene frequency under given assumptions, require:

- a) simple knowledge of school algebra and
- b) basic concepts of Mendelian genetics (See Box 2.3).

The proof in case of autosomal '*biallelic*' trait is given in Box 2.4. (for further reading see references)

BOX 2.3 -- Basic concepts

Phenotype: A trait or a character that is observed as types or measurable and is transmitted from parents to offspring. Some phenotypes are complex with unknown genotypes, and some are directly governed by hereditary units (genes)

Gene: The causative factor of hereditary transmission of traits (phenotypes) and are located in the chromosomes (the hereditary materials in cell nucleus and in mitochondria)

Allele: Genes, the causative factor of hereditary transmission and can exist or express in different forms and are referred as 'alleles'

Codominant: Where both the alleles are equally expressive in the offspring.

Recessive: the alleles whose expression is suppressed at phenotypic level. The heterozygote offspring of a recessive allele will express the phenotype of the dominant allele

Haploid: Organisms which carry one set of chromosomes

Haplotype: It is short form of '**Haploid genotype**'. Refers to genetic markers located on one chromosome. A haplotype can be identified by **SNP (single nucleotide polymorphism)**.

Diploid: Organisms which carry two sets of chromosomes, each set derived from either of the parent. Man is diploid and carries two sets of chromosome (2N)

A diploid individual can carry two copies (alleles) of the gene in each of the chromosome that he or she gets from his or her parents. The two copies could be of the same type (form/status) or of different type (form/status)

Homozygous: the two alleles that an individual carries are of the same or identical types

Heterozygous: the two alleles that an individual carries are of different type

Genotype: Is the combination of alleles that a diploid individual can carry in each of the chromosomes.

For example, in case of a 'biallelic' gene say A, B two forms (alleles) of the gene that occur in each of the two sets of the chromosomes. There could be three different genotypes: AA, AB, BB

AA and BB : two different homozygotes (genotype)

AB = BA : heterozygote (genotype)

The box shows the "Punnet's square" – method of scoring different combination of genotypes based on the male and female gametes or mating types. This can be extended to multiple alleles.

Polymorphism: If a gene exists in more than one form or morph (alleles) and that occurs in stable frequency in a population.

Punnet's square			
		Male Gamete	
		A	B
Female gamete	A	AA	BA
	B	AB	BB
		Genotype	

BOX 2.4

Hardy-Weinberg theorem or principle: Proof

In case of genetic trait that has positive family history in a population, let us assume that the gene is biallelic and therefore the two alleles are: B1 and B2 and let

‘p’ is the frequency of ‘B1 allele’ and

‘q’ is the frequency of B2 allele,

N is the total individuals and

So that $(p + q = 1)$ or $p = (1 - q)$ or $q = (1 - p)$

An individual in the population can have three types of genotypes: B1B1, B1B2, and B2B2.

And let the frequency of the above three genotypes in the parental population are: P, H and Q respectively.

	Gene (alleles)		Genotypes		
	B1	B2	B1B1	B1B2	B2B2
Frequencies	p	q	P	H	Q

If there are total N individuals in the sample, there will be

P individuals with genotype B1B1 type,

H individuals with genotype B1B2 type and

Q individuals with genotype B2B2 type

So that sum of $(P + H + Q) = N$,

Assuming all the individuals of the three genotypes are equally fertile, then given the genotypes, one can calculate the frequencies ‘p’ and ‘q’ in the population, by gene counting method:

The gene (allele) frequency ‘p’ = $[P + \frac{1}{2}(H)]/N = (B1B1)/N + \frac{1}{2}(B1B2)/N$, and

The gene (allele) frequency ‘q’ = $[Q + \frac{1}{2}(H)]/N = (B2B2)/N + \frac{1}{2}(B1B2)/N$

This is the gene (allele) frequencies of ‘p’ and ‘q’, which are also the gametes produced in the population.

Only some of the gametes form zygotes that will eventually become individuals in the next generation. The allele (gene) frequency in the zygote is unchanged provided there is no reproductive advantage of either of the allele and the zygotes formed represent a large sample of the parental gametes.

Random mating between individuals is equivalent to random union among their gametes. Therefore, in the next generation, the genotype frequencies among the zygotes (fertilized eggs) are the result of random union of two types of gametes. The genotype frequencies among the progeny are therefore can be worked out by Punnet’s square. Or it is the multiplication of the frequencies of the gametic types produced by the parents. Viz.,

<table style="width: 100%; border-collapse: collapse; text-align: center;"> <thead> <tr> <th style="border-top: 1px dashed black; border-bottom: 1px dashed black;"></th> <th colspan="3" style="border-top: 1px dashed black; border-bottom: 1px dashed black;">Genotype</th> </tr> <tr> <th style="border-bottom: 1px dashed black;"></th> <th style="border-bottom: 1px dashed black;">B1B1</th> <th style="border-bottom: 1px dashed black;">B1B2</th> <th style="border-bottom: 1px dashed black;">B2B2</th> </tr> </thead> <tbody> <tr> <th style="border-bottom: 1px dashed black;">Frequency</th> <td style="border-bottom: 1px dashed black;">p^2</td> <td style="border-bottom: 1px dashed black;">$2pq$</td> <td style="border-bottom: 1px dashed black;">q^2</td> </tr> </tbody> </table>		Genotype				B1B1	B1B2	B2B2	Frequency	p^2	$2pq$	q^2	<table style="width: 100%; border-collapse: collapse; text-align: center;"> <thead> <tr> <th style="border-bottom: 1px solid black;">Allele</th> <th style="border-bottom: 1px solid black;">B1</th> <th style="border-bottom: 1px solid black;">B2</th> </tr> </thead> <tbody> <tr> <th style="border-bottom: 1px solid black;">Genotype</th> <td style="border-bottom: 1px solid black;">B1B1</td> <td style="border-bottom: 1px solid black;">2B1B2</td> <td style="border-bottom: 1px solid black;">B2B2</td> </tr> <tr> <th style="border-bottom: 1px solid black;">Frequency</th> <td style="border-bottom: 1px solid black;">p^2</td> <td style="border-bottom: 1px solid black;">$2pq$</td> <td style="border-bottom: 1px solid black;">q^2</td> </tr> <tr> <th style="border-bottom: 1px solid black;">Absolute freq.</th> <td style="border-bottom: 1px solid black;">P</td> <td style="border-bottom: 1px solid black;">H</td> <td style="border-bottom: 1px solid black;">Q</td> </tr> </tbody> </table>	Allele	B1	B2	Genotype	B1B1	2B1B2	B2B2	Frequency	p^2	$2pq$	q^2	Absolute freq.	P	H	Q
	Genotype																											
	B1B1	B1B2	B2B2																									
Frequency	p^2	$2pq$	q^2																									
Allele	B1	B2																										
Genotype	B1B1	2B1B2	B2B2																									
Frequency	p^2	$2pq$	q^2																									
Absolute freq.	P	H	Q																									

BOX 2.4 (Contd.)

Hardy-Weinberg theorem or principle: Proof

In a population there will be three different types of genotypes among males and females, who will mate randomly and they will give rise to their offspring who will represent the same genotypes in the next generation. We will have to work out the frequencies of offspring genotypes given the three genotypes among the male and female parents. This is worked out easily by Punnet's square: the frequencies of different mating types among the male and female genotypes in the population and different possible genotypes among the offspring are.

Frequency of different mating types and the offspring genotypes

Female Parent		Male Parent -- Genotype		
		B1B1	B1B2	B2B2
Freq.		p^2	$2pq$	q^2
B1B1	p^2	p^4	$2p^3q$	p^2q^2
B1B2	$2pq$	$2p^3q$	$2p^2q^2$	$2pq^3$
B2B2	q^2	p^2q^2	$2pq^3$	q^4

Once we know the possible offspring genotypes as a result of random mating among the three parental genotypes we can calculate the expected frequencies among the offspring genotypes for different combination of mating types in the population. Given the three genotypes six possible mating types are possible in the population and each mating type will give rise to offspring of different possible combination of genotypes. These are worked out in the following table and this gives the allele frequencies in the offspring population in the next generation:

Frequency of different mating types and the offspring genotypes

Parent		Offspring -- Genotype		
Female X Male	Freq.	B1B1	B1B2	B2B2
(Genotype)				
B1B1 X B1B1	p^2	p^4	--	--
B1B1 X B1B2	$4p^3q$	$2p^3q$	$2p^3q$	--
B1B2 X B1B1				
B1B1 X B2B2	$2p^2q^2$	--	$2p^2q^2$	--
B2B2 X B1B1				
B1B2 X B1B2	$4p^2q^2$	p^2q^2	$2p^2q^2$	p^2q^2
B1B2 X B2B2	$4pq^3$	--	$2pq^3$	$2pq^3$
B2B2 X B1B2				
B2B2 X B2B2	q^4	--	--	q^4
1		$p^2 (p^2 + 2pq + q^2)$	$2pq (p^2 + 2pq + q^2)$	$q^2 (p^2 + 2pq + q^2)$
		p^2	$2pq$	q^2

Thus the genotypic frequencies in the offspring remain the same in two successive generations, assuming that allele frequencies are not influenced by selection, mutation and mating is random and there is no differential fertility and mortality and the population is large.

The above is true for *autosomal loci* and can be extended for multiple loci. It is also true for sex-linked trait. Here the gene frequencies will oscillate (by 1/2) between two sexes in successive generation and will soon reach to equilibrium.

2.2.1 Importance and implications of HW

What are the implications and why it is so important? In brief, it is the fundamental theorem of population genetics.

- **Methodology:** tells us how to calculate (or estimate) the allele frequency or genotype frequency from observed phenotypes in an empirical situation. It can help us to investigate how many alleles are governed by a phenotypic trait.
- **Evolution:** It is quantitative way of understanding the mechanism of evolutionary factors and its influences. Evolution is a dynamic and complex phenomenon and it is hardly possible to study evolution in the laboratory conditions. It gives insights into the inter-relationship between the forces and how to study the effects of each of these forces and the gene frequency. (See the box 2.5 for the relationship between gene frequencies and genotype frequencies).
- It is the **benchmark criterion** to test whether a new trait is in equilibrium or if not how to test the reasons for the deviations.
- It helps us in **genetic counselling** to expect the likelihood of a child being homozygous for a recessive deleterious trait given the parental genotype. It helps in forensic science in cases like identification of suspects, parent-offspring disputes etc.
- **Quantitative Genetics:** HWE helps us to investigate complex genetic traits, to estimate the role of environment and genetic components, spatial distribution of gene frequency etc

Further implications of this principle are as under:-

- In case in a population a particular trait or character is in HWE, (the converse) it does not mean that the assumptions are satisfied. (The theoretical proof is complicated and it is available)
- The allele frequencies remain constant from generation to generation. This means that hereditary mechanism itself does not change allele frequencies. It is possible for one or more assumptions of the equilibrium to be violated and still not produce deviations from the expected frequencies that are large enough to be detected by the goodness of fit test,

- When an allele is rare, there are many more heterozygotes than homozygotes for it. Thus, rare alleles will be impossible to eliminate even if there is selection against homozygosity for them,
- For populations in HWE, the proportion of heterozygotes is maximal when allele frequencies are equal ($p = q = 0.50$), and when this happens the heterozygote frequency will be 0.50 ($2 \times 0.50 \times 0.50$). Unless HWE is violated (as in selective loss of homozygotes), heterozygosity can never be more than 0.50 at any biallelic locus. The relationship between gene frequency and genotype frequency is illustrated in Box 2.5.
- An application of HWE is that when the frequency of an autosomal recessive disease (e.g., sickle cell disease, hereditary hemochromatosis, congenital adrenal hyperplasia) is known in a population and unless there is reason to believe HWE does not hold in that population, the gene frequency of the disease gene can be calculated. Likewise, the carrier rate may be calculated for autosomal recessive disorders if the disease gene frequency is known. For example, phenylketonuria (**PKU**) occurs in 1/11,000 (q^2), which gives a heterozygote carrier frequency of approximately 1/50 [$2 \times q(1-q)$]. If the diseased individuals (q^2) are deducted from the whole population, the carrier rate in normal individuals approximates to [$2q/1+q$].
- It has to be remembered that when HWE is tested, mathematical thinking is necessary. When the population is found in equilibrium, it does not necessarily mean that all assumptions are valid since there may be counterbalancing forces. Similarly, a significant deviance may be due to sampling errors (including **Wahlund effect**), misclassification of genotypes, measuring two or more systems as a single system, population substructure, failure to detect rare alleles and the inclusion of non-existent alleles. The Hardy-Weinberg laws rarely holds true in nature (otherwise evolution would not occur). Organisms are subject to mutations, selective forces and they move about, or the allele frequencies may be different in males and females. The gene frequencies are constantly changing in a population, but the effects of these processes can be assessed by using the Hardy-Weinberg law as the starting point.
- The direction of departure of observed from expected frequency cannot be used to infer the type of selection acting on the locus even if it is known that

selection is acting. If selection is operating, the frequency of each genotype in the next generation will be determined by its relative fitness (W). Relative fitness is a measure of the relative contribution that a genotype makes to the next generation. It can be measured in terms of the intensity of selection (s), where $W = 1 - s$ [$0 < s < 1$]. The frequencies of each genotype after selection will be $p^2 W_{AA}$, $2pq W_{Aa}$, and $q^2 W_{aa}$. The highest fitness is always 1 and the others are estimated proportional to this. For example, in the case of heterozygote advantage (or overdominance), the fitness of the heterozygous genotype (Aa) is 1, and the fitnesses of the homozygous genotypes negatively selected are $W_{AA} = 1 - s_{AA}$ and $W_{aa} = 1 - s_{aa}$. It can be shown mathematically that only in this case a stable polymorphism is possible. Other selection forms, underdominance and directional selection, result in unstable polymorphisms. The weighted average of the fitnesses of all genotypes is the mean fitness. It is important that genetic fitness is determined by both fertility and viability. This means that diseases that are fatal to the bearer but do not reduce the number of progeny are not genetic lethal and do not have reduced fitness (like the adult onset genetic diseases: Huntington's chorea, hereditary hemochromatosis). The detection of selection is not easy because the impact on changes in allele frequency occurs very slowly and selective forces are not static (may even vary in one generation as in antagonistic pleiotropy).

- All discussions presented so far concerns a simple biallelic locus. In real life, however, there are many loci which are multiallelic, and interacting with each other as well as with the environmental factors. The Hardy-Weinberg principle is equally applicable to multiallelic loci but the mathematics is slightly more complicated. For multigenic and multifactorial traits, which are mathematically continuous as opposed to discrete, more complex techniques of quantitative genetics are required.

2.3 Application in human population genetics

- The behaviour of HW principle under different assumptions is the discipline of 'population genetics', which describes, primarily, the changes in gene frequency that are influenced by demographic factors, population structure variables, historical, random events, sampling fluctuations and evolutionary

factors of selection and mutation In simple the four main factors that influence the gene frequency in a population are: mutation and genetic drift

BOX 2.5

The relationship between gene frequency and genotype frequency

It is interesting to know the relation between the gene and genotype frequency for a biallelic loci which is under H-W equilibrium. The graph shows the changes in the three genotype frequencies as against the change in allele frequencies A1 and A2 from 0 to 1 in Cartesian coordinates (drawn on x and y axis).

A1A1/A2A2 – homozygotes, A1A2 – heterozygotes

**Relationship between genotype frequency and gene frequency
- biallelic GENE**

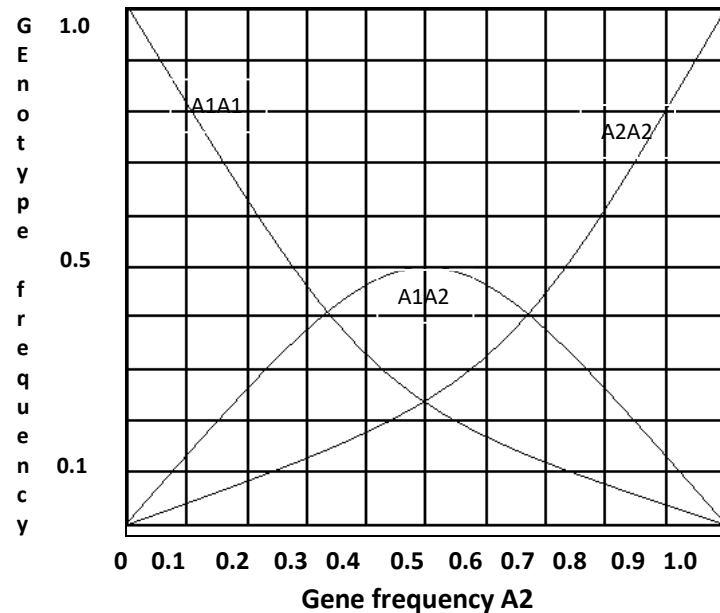


Fig: Left top curve – A1A1, Right top curve – A2A2, Lower curve – A1A2

The graph shows two interesting properties of the HWE:

- ✓ The frequency of the heterozygotes can reach to a maximum of 50%
- ✓ And this can occur when the gene frequency of 'p' = 'q' = 50%
- ✓ When one of the gene frequencies of an allele is low, the rare allele predominantly occurs as heterozygotes and there will be few heterozygotes

(non-systematic factors), migration and selection (systematic factors). The genetic drift is effective, more specifically, in populations whose size is small or

limited e.g., an isolate or an island population or a small endogamous population. These are described in detail below.

- For example, HWE has helped us to find out to investigate the number of alleles of ABO locus and how to calculate the gene frequency of ABO locus (e.g., Bernstein has given the method of correction) (see Box 2.6)
- We were able to understand how HbS despite its deleterious effect it maintains its equilibrium in the population.
- HWE helps to understand the some of the health problems in some isolated populations, whose propagation is the result of genetic drift, genetic drift and selection or inbreeding etc.
- HWE has forensic applications in solving problems related to disputed paternity, to provide evidence in case of crime to detect the culprit, property or biological inheritance cases.
- It helps in understanding the complex genetic disorders, to be able to estimate the contribution of genetic versus environmental effects.
- HWE helps to understand to investigate the human origins, the role of selection versus demographic effects on the genetic diversity in a population.

BOX 2.6

HWE – Gene frequency estimation: Gene counting method

Given the information about the genotypes, HWE helps us to estimate the allele frequency by ‘gene counting method’ (how many alleles a genotype contains). For example,

- As each homozygote carries two alleles and each
- Heterozygote carries one allele, therefore, estimate of an allele frequency in a population of size N individuals (or 2N alleles) will be
- $= (2 \text{ homozygotes} + \text{heterozygote}) * (1/2N)$
- In a population there will be three genotypes and their absolute frequency will be say N₁, N₂ and N₃ (where N₁ + N₂ + N₃ = N). If there are two alleles say ‘A’ and ‘a’ with a frequency ‘p’ and ‘q’ respectively (where p + q = 1).
- By gene counting method assuming HW law the gene (allele) frequencies of $p = (1/2N) * (2N_1 + N_2)$, $q = (1/2N) * (2N_3 + N_2)$ and $p = (1 - q)$

2.4 Departure from HWE

In general, the factors that are assumed to be non-operative under HWE are hardly realised in the living systems. The living system (populations or organisms) are structured (non-random entities) and are influenced by multiple and interactive factors that operate through space and time. With the help of HW equilibrium it is possible to investigate and estimate the effect of these individual forces that change gene frequency in human populations..

2.4.1 Factors affecting change in gene frequency

The four aspects of the H-W assumptions are:

- i. Demographic:
 - a. Size, mating, fertility and mortality, and migration
- ii. Evolutionary:

Mutation, selection, gene flow
- iii. Population structure:
 - a. Social and Cultural factors
 - i. Matings and Marriage specifications that regulate the marriage or mating type in a population.
 - ii. Non-random mating – Sexual selection of mates
- iv. Ecological:
 - a. Population bottle-neck events:
 - i. Pandemic: disease, earthquake etc.
 - ii. Historical: wars etc.,

Of the above factors, for the present academic purpose, we will be dealing a few factors and examine how these factors change or influence the gene frequency in a population and how to estimate them in empirical situation.

2.4.1.1 Mutation

Mutation is a random change in phenotypic or genotypic forms that occur once a while in a population. The probability or likelihood of occurrence, in a population, in general, is of the order of one over several lakhs or tens of thousand of individuals.

For example, several of the Mendelian syndromes and disorders that have been discovered in human populations are the result of mutation. In general, it is observed to be a single mutation or point mutation. At molecular level, mutation primarily refers to changes in the DNA sequences (or SNPs -- Single Nucleotide

Polymorphism) in the genome of an individual (population) with phenotypic manifestations resulting to non-normal cases, some of them are clinically or medically identified as diseases or a syndrome. If one can search web resources, there is a data base created by Hopkins institute and or by NIH (America) on a list of Mendelian syndromes, which can be found by a search criterion OMIM (Online Mendelian Inheritance in Man. One can also find such data bases from a variety of sources.

Some examples will help us to get an idea of mutation and its effects. Sickle cell anaemia (or HbS condition), is a disease related to Haemoglobinopathies. It is a inability to synthesize Oxygen (O_2) to its full capacity by an individual who is suffering from the disease or the trait, which results a risk to survival liability. This is identified as due to a point mutation or single mutation at the 6th position of the β -globin chain of the haemoglobin gene. The single aminoacid substitution (β_6 Glu to Val) changes the haemoglobin structure, which is phenotypically identified as sickle cell shaped form (or half moon shaped form) of the RBC.

Mutation is an important factor or ingredient leading to the appearance of new characters in the population. The fate of the mutation as a new character in a population depends on its advantage or disadvantage that it can impinge to the survival fitness of the population. For example, mutation is a significant evolutionary force which can change allele frequency variation in a population under a favourable environment. How we can know the relation between the mutation and allele frequency change from HWE.

a) Change in gene frequency due to mutation (μ):

Random changes that happen at the DNA sequence, especially at the coding region of the gene can create an allele which can alter the gene frequency in the population in successive generations. This can be investigated theoretically given the mutation rate per generation in the population. This has been shown in Box 2.7 for bi allelic loci.

The theoretical results suggest that the change in gene frequency of a mutant allele, after 't' generations, depends on the initial allele (gene) frequency before mutation and the mutation rate of the allele per generation in the population.

This is an important result and can help to calculate change in gene frequency after 't' generations given the mutation rate (μ) per generation and the initial gene frequency in the population.

b) Rate of mutation (μ):

Though the mutation is random, but the rate of mutation varies. It is site specific – there are 'hot-spots' where mutation rate is more frequent than in other parts of the genome. In general, the coding part of the gene does not support mutation to occur, as a result of proof reading process and functional importance of the codons. However, mutations occur at higher rate in the intronic region, and in the repeat sequences than in exons or codons. Also the mitochondrial non-coding parts, viz., hyper variable regions HV1, HV2 in the D-loop has higher rate of mutation than in nuclear genome.

BOX 2.7

Change in gene frequency due to mutation (μ)

If there are two alleles 'A' and 'a' with its frequencies ' p_0 ' and ' q_0 ' at the initial stage (say at time ' t_0 ') in a population and ' μ ' is the mutation rate that changes allele 'A' to 'a' per generation, then gene frequency (g.f.) of 'A' will decrease by an amount ' μp_0 ' in the first generation. Therefore the g.f. of 'A' allele in the first generation after mutation will be:

$$P_1 = p_0 - \mu p_0 = (1 - \mu) P_0$$

In the (next) second generation the gene frequency is expected to be:

$$\begin{aligned} P_2 &= P_1 - \mu P_1 = (1 - \mu) P_0 - \mu(1 - \mu) P_0 \\ &= (1 - \mu) P_0 \quad [(1 - \mu)] \\ &= (1 - \mu)^2 P_0 \end{aligned}$$

After 't' generations the g.f. of 'A' is expected to be

$$P_t = (1 - \mu) P_{t-1} = (1 - \mu)^2 P_{t-2} \dots = (1 - \mu)^t P_0$$

When μ is very small $(1 - \mu)^t$ can be approximately equated to $= e^{-\mu t}$, (where e is natural logarithm to base e), therefore, gene frequency after 't' generations will be

$$P_t \approx P_0 e^{-\mu t}$$

Therefore, the mutations that occur at HV1 and HV2 regions of mitochondrial genome help us to investigate the short-term evolution or micro-evolutionary trends in sub-populations. This has helped us to address some of the questions of human origins or to verify the Darwin's hypothesis that the Africa is the origin of Man. This also helps to enquire the antiquity and past genetic history of diverse populations and their diversity and relationship with other human populations.

2.4.1.2 Genetic drift

Genetic drift is an important non-systematic evolutionary force. To understand the concept of genetic drift, let us know what the word ‘drift’ conveys, in general. One of the descriptions for the word ‘drift’ in the English Dictionary is: “move aimlessly from one place or activity to another’ – this is more with reference to things or events that we experience with practical world e.g., drifting by air, wind and water or ocean. Similar phenomena can also happen with respect to gene frequency in a small population. In small populations, as a result of population-events such as pandemic diseases, earthquakes etc, the population size is drastically reduced which can have significant effect on the genetic diversity and gene frequency: for example, the gene frequency can drift from one generation to generation randomly leading to either loss or fixation of alleles over generations (in the absence of other interfering factors). In small populations or due to demographic and ecological effects the population size drastically reduced to a fraction (or a random sample) of the original population with allelic representation different from the original population. In these cases, there will be random changes in gene frequency, which appear to drift at varying frequencies in successive generations in an erratic manner. For example, the studies on the origins of Man, suggest that decreasing heterozygosity and linkage disequilibrium levels away from Africa are supportive of the role of genetic drift among human populations.

To understand how the genetic drift can happen or possible, one can investigate and/or understand by attempting some simple examples or simulation exercises. These are available on the online resources. One such example is illustrated in Box 2.8.

a) Bottle-neck effect

Genetic Drift can happen in a variety of ways due to different events that populations experiences in empirical situation. These have been referred as part of ecological factors that disturb the population size (see 2.4.1). Historically the world has experienced several pandemic diseases in the past: e.g. Syphilis, Plague, leprosy, malaria, HIV infection, etc which has killed or wiped out bulk of the population. The natural geographical events like earthquakes, tsunami etc. had killed vast majority of the populations. Even the political and man interfering events like explosion of atomic bombs, world wars etc. have affected the demographic size of the populations. Each such event is followed by a drastic reduction in population sizes. In genetic terms it means reduction in

genetic diversity (at the time of the event), and those survived will have different allelic profile or gene frequency and the stability of a particular allele over generations depends on the demographic structure of the population.

‘Breeding individuals’ part of the demographic structure of a population is of particular genetic importance. They are capable of mating and producing children. They will be a fraction (of the total population) who contribute to the next generation or gene pool and is referred as ‘effective size’ (N_e).

BOX 2.8

Simple exercises to understand the genetic drift

There are different ways to replicate to illustrate the random drift phenomena. One such simple example could be the following:

- Start with a jar that contain with N number of blue, red, yellow balls.
- At the first step blindly or randomly take out (say e.g., by hand) some balls and put them in the second bottle.
- Then from the second bottle, take some balls (e.g., by hand) and put in the third bottle.

If you have started with large sample (N) of mixed coloured balls you can repeat the same. Otherwise, at the third/fourth bottle you can count how many of red, blue and yellow balls. Compare the outcome with the original number of red, blue and yellow balls at the start. They will differ from the original number at the start. You may also find the absence of a particular colour at the fourth (or nth) bottle.

In case of Genetic Drift, similar such random sampling of gene frequency changes happen over successive generations in a small population. One can search several such simple examples on the online resources on genetic drift – bottleneck effect, founder effect etc.

Genetic drift can alter the ‘effective size’ of a population and change the genetic diversity. After successive generations, the gene frequency in the population will be significantly different from the gene frequency before genetic drift. This is similar to the bottle neck, where the narrow neck of the bottle restricts the flow and this event is referred as ‘bottle neck-effect’ in population genetics. Such bottle neck effect resulting to sudden population size reduction had been experienced by several human populations in the past historical times affecting the genetic structure: genetic diversity, gene frequency changes.

b) Founder effect

The word ‘founders’ refers to the ancestors or the earliest settlers who colonised or founded the new population in alien territories. It could be an historical adventure of warfare, or exploration to a new island or new area or it could also be due to chance factors like surviving from a sudden calamities like

ship wreck, etc. or it could be serial migration of people at different timing to other places: in all the cases, a few founders start living and establishing a new subpopulation.

In genetic scenario, the few founders represent a random sample of the genes from the original population or gene pool from which they got separated. It is possible that, some of the rare alleles that are in the large population, by chance, may not be present in the founder individuals. It could be that, among the founders, especially if the founders are related, by chance, some of alleles may be of a higher frequency than the original population. Therefore, in the new colony after generations the gene pool will have either absence of the allele or higher frequency of the rare allele than when compared to the original population.

c) Serial founder effect

It is possible that people or organisms migrate repeatedly over time or waves of migration from a region to found new colonies. Such repeated waves of migration at different time periods produce successive subpopulations or gene pools whose genetic profile will be different. There appears to be waves of out of Africa migration to other continents that had happened at different time periods in the past, whose genetic signature can now be traced among the extant populations in South Asia, Europe, and America etc. The mitochondrial, X and Y chromosomal haplogroup distribution of continental populations can be explained as a result of founder effect of out of Africa hypothesis of human origins.

d) Empirical studies of founder effect in Man

The importance of 'Founder effect' as significant evolutionary factor has been outlined by German evolutionary biologist Ernst Mayr (1942). Founder effect is the *"The establishment of a new population by a few original founders (in an extreme case, by a single fertilized female) which carry only a small fraction of the total genetic variation of the parental population."* This is sampling effect especially the genetic composition and evolution of the successive generations entirely depends upon the few founders. A few examples illustrating the role of genetic drift in the gene frequency changes are shown in Box 2.9

BOX 2.9

Studies on genetic drift

- **Tristan da Cunha is an island**; the few hundred individuals (<300) living on the island are mostly the (15) descendants (8 males and 7 females) who had founded the island in 1816-1908. Three of the founders were Asthma sufferers and there is high incidence of Asthma in the population. In a study of the 9 Y-chromosome haplotypes of the island, seven of them are traced to its 7 male founders.
- **Amish population, USA**: All most all the Amish population (~249K) descended from about 200 founders from German during 18th century. The population is endogamous, they show high frequency of genetic disorders as a result of founder effect that include dwarfism, metabolic disorders, unusual distribution of blood types, metabolic disorders etc.
- **'Blue Fugates' of Appalachian, Kentucky, USA**: In 1800, Martin Fugate and his wife settled in trouble some creek in Kentucky. They carried recessive gene methemoglobinemia (met-H). Due to deficiency of an enzyme diaphorase (NADH methemoglobin reductase) met-H levels rise and this gives raise to reduced oxygen-carrying capacity. This gives a tinge of blue skin of the homozygous condition. Isolation and inbreeding has caused to increase of blue people which are traced to the founders Fugates.
- **India**. In the northeast populations, some of them live in geographical isolation, practice endogamy show unusual frequency of a few genetic traits which are expected to be due to genetic drift and founder effect. Some of them include
 - Complete lack of A2, cde, K, pc, and AK2 genes, lack of isozyme ALDH-1 (Roychoudhury and Nei 1997), a high prevalence (about 50%) of lactase malabsorption (Flatz 1987),
 - low frequency of AIBG*2 allele (Juneja et al. 1989), high frequency of G6PD deficiency in Naga (Seth and Seth 1971), absence of 'Gd_' variant in Adi and Hmar and high frequency of this variant in Bodos (Saha et al. 1990).
 - Continuing from classical genetic observations, unique and rare allele frequency of microsatellite loci among the Adi subpopulations (Krithika et al. 2005). High frequency of susceptibility of tuberculosis in some clans of tribes, stomach cancer, high incidence of cardio deaths etc.
 - Absence of attached ear lobe among the Nandiwalas in Maharashtra,
 - Population size reduction and allele frequency changes among Ahmedias of Kashmir population.

2.4.1.3 Natural selection

Charles Darwin (and Wallace) has described natural selection as one of the important factor (key mechanism) of evolution. Natural selection happens where there is differential rate of reproductive success among different genotypes (underlying the phenotype, or trait or observed character). How selection operates at the molecular (genome) level for example, especially change in gene frequency considered, theoretically, in population genetics.

Due to differential reproductive success involving these variant of the trait, there will be more offspring with the variant than those individuals with other variant of the trait. In Darwinian sense '*fitness*' ('*Darwinian Fitness*') refers to ability to contribute successfully to the next generation. This is also referred as 'adaptive value' or 'selective value'. Therefore, if the differences of fitness are in a way associated with the presence or absence of a particular allele (or gene) in the individual's genotype then selection operates at the genetic level.

When a gene is subjected to selection (or under selective pressure), its frequency in the offspring is not the same as in the parents (or in the previous generation) as parents with different genotypes pass on their genes *unequally* to the next generation. This leads to change in gene frequency and consequently also of genotype frequency, as a result of selection (of a particular gene). The theoretical investigation of change in gene frequency of an allele under selection pressure is more complex, than factors like mutation, migration. There could be different situations under which selection can operate in a population and different situations need to be incorporated in theoretical models. Here we will consider a few of those situations (types of selection) in a more descriptive way, rather than theoretically, which is beyond the scope of the present purpose.

Theoretically, **selection** is measured by '*fitness*' ('*W*') or by selection coefficient ('*s*'). *Fitness* refers to '*relative rate of survival*'. The selection coefficient ('*s*') is defined as ($1-W$) and the value varies between 0 and 1. Once the fitness is quantified and defined the different types of dominance can be taken as degrees of dominance with respect to fitness (this is different from the dominance effect of the gene). In general, most mutant genes are completely recessive compared to the wild type as can be observed from phenotypic form of the trait. This does not imply that the heterozygotes are equally fit when compared to homozygote.

Before we get to know the effect of selection on gene frequency, it is required to know different types of selection and its fitness values. Some of the known selection types are: no dominance, partial dominance, complete dominance, over dominance. The fitness values for the four types are shown in below (See Box 2.10). The change in gene frequency with respect the four types of selection (with fitness values) are given in Box 2.11.

BOX 2.10			
Types of dominance or degree of dominance and fitness			
a. No dominance:	A_2A_2	A_1A_2	A_1A_1
	----- -----		
	$1 - s$	$1 - 1/2s$	1
b. Partial dominance:	A_2A_2	A_1A_2	A_1A_1
	----- -----		
	$1 - s$	$1 - hs$	1
c. Complete dominance:	A_2A_2	A_1A_2, A_1A_1	

	$1 - s$	1	
d. Over dominance:	A_2A_2	A_1A_1	A_1A_2
	----- -----		
	$1 - s_2$	$1 - s_1$	1

a) Types of selection

Selection is a systematic force and operates in different ways. Selection takes place when there is differential fitness of a heritable trait. Based on the effect on the allele frequencies, the selection can be seen operating into three types.

Directional selection: occurs one extreme value or allele is selected. In case if one of the allele of a variety of the trait has greater fitness and producing more offspring of that allele or a variety, then the selection is said to be directional. The effect of directional selection is fixation of allele with greater fitness and the loss of the allele with least fitness. For example: well known cases come from the parasitic world, especially resistance to antibiotics in case of some of the vector-borne diseases. Initially as a result of antibiotic the parasite growth comes down to zero, but the parasites develops some mutant or new variant which gets resistance against the antibiotics or better fitness in the presence of antibiotics, in due course, the less fit variant is replaced by new variant which can survive against antibiotics. This can be illustrated as a shift in the mean of the character of a distribution (See box 2.12)

BOX 2.11: Change in gene frequency under selection

First we will consider the basic formulae for the change in gene frequency that is achieved in one generation of selection. Under the similar notation that has been used above for other factors (p = gene freq of A_1 and q = gene freq. of A_2), the below table shows the genotype frequencies under HWE before selection to the allele for the three genotypes (first line).

	Genotypes			Total
	A_1A_1	A_1A_2	A_2A_2	
Initial frequency	p^2	$2pq$	q^2	1
Coefficient of selection	0	0	s	
Fitness	1	1	$1 - s$	
Genetic contribution	p^2	$2pq$	$q^2(1-s)$	$(1 - sq^2)$

Here we consider selection acting on the recessive genotype A_2A_2 with a selection coefficient: 's' acting against it. This will have a differential fitness to the genotypes that will be as given in the second line. By multiplying the initial frequency by the fitness values gives the frequency of each genotype after selection. This is the third line – the genetic contribution to allow to selection to operate over life cycle. Therefore, after selection, there will be a loss of fitness that is proportional to an amount $(1 - sq^2)$. From this we can calculate the frequency of A_2 gametes produced (frequency of A_2 genes in the progeny). The new g.f. is (where $p = (1 - q)$)

$$q_1 = [q^2(1 - s) + pq] / (1 - sq^2)$$

$$= [q - sq^2] / (1 - sq^2)$$

The change in gene frequency Δq , resulting from one generation of selection is

$$\Delta q = q_1 - q = sq^2(1 - q) / (1 - sq^2)$$

This tells us that the effect of selection on gene frequency depends not only on the intensity of selection s , but also on the initial gene frequency (of the recessive allele).

Different type of selection

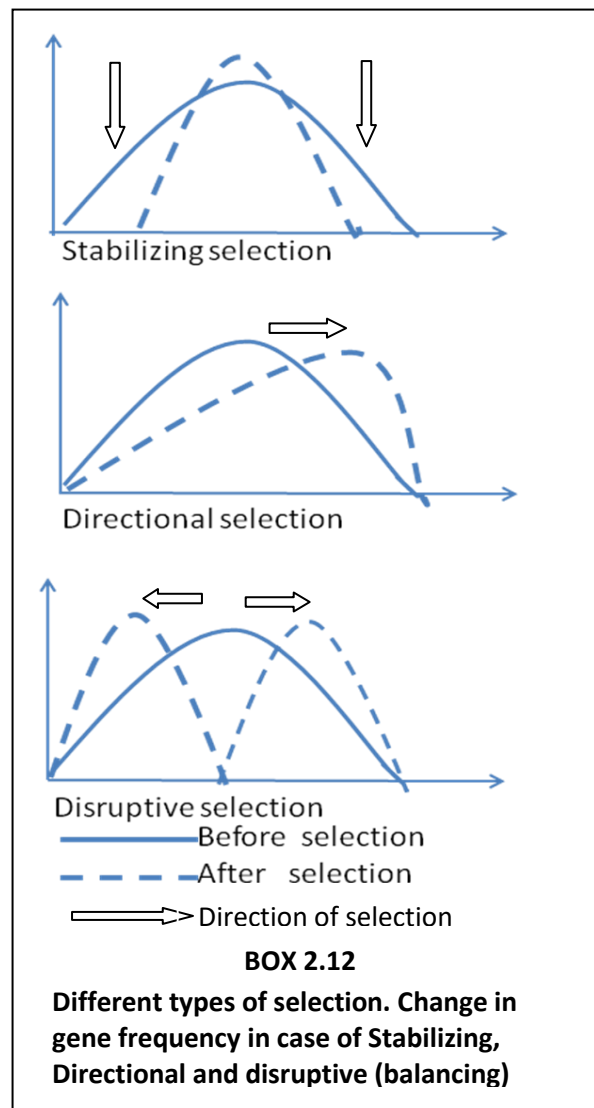
What we have considered above is in general selection with respect to recessive allele q under selective pressure. But there are variety (or types) of selection that can act on the allele frequency. Depending upon the type of selection the change in gene frequency will consequently change. These are shown in the following table.

Initial freq. & fitness of Genotypes			New gene frequency due to selection at q	Change in gene frequency
A_1A_1	A_1A_2	A_2A_2	q_1	$\Delta q = q_1 - q$
p^2	$2pq$	q^2		
1	$1 - \frac{1}{2}s$	$1 - s$	$(q - \frac{1}{2}sq - \frac{1}{2}sq^2) / (1 - sq)$	$-\frac{1}{2}sq(1 - q) / (1 - sq)$
2	$1 - hs$	$1 - s$	$(q - hspq - sq^2) / (1 - 2hspq - sq^2)$	$-spq[q + h(p - q)] / (1 - 2hspq - sq^2)$
3	1	$1 - s$	$(q - sq^2) / (1 - sq^2)$	$-[sq^2(1 - sq)] / (1 - sq^2)$
4	$1 - s$	$1 - s$	$(1 - sq + sq^2) / (1 - s(1 - q^2))$	$+ [sq^2(1 - sq)] / [1 - s(1 - q^2)]$
5	$1 - s_1$	1	$(q - s_2q^2) / (1 - s_1p^2 - s_2q^2)$	$+ [pq(s_1p - s_2q)] / (1 - s_1p^2 - s_2q^2)$

Above the different types of dominance are:

1. No dominance, selection against A_2
2. Partial dominance of A_1 : selection against A_2
3. Complete dominance of A_1 : selection against A_2
4. Complete dominance of A_1 : selection against A_1
5. Over dominance: Selection against A_1A_1 and A_2A_2 (Applicable to any degree of dominance with fitnesses expressed relative to A_1A_2)

Stabilizing selection: The extremes are selected in favour of the middle. In case of stabilizing selection, the two extreme values of a trait or alleles will have lower fitness than the intermediate value or the heterozygote alleles of a trait. One of the well known examples includes birth weight. The average birth weight of offspring ranges between 2500g (5.5pounds) to 4500g (10 pounds). Offspring with weight less than 2500g are low birth weight and greater than 4500g are the heavy babies and both have less chance of survival. As a result the selection favours the offspring with the average birth weight. Stabilizing selection is also the reason in case of height distribution in a population. This can be illustrated as a change in mean values of the distribution (See Box 2.12)



Balanced Selection: In case of balanced selection, the heterozygotes have higher fitness than either of the homozygotes. This is also called heterozygous advantage or over-dominance. The best example is the sickle cell anaemia. In non-malarial environment the homozygote state of the sickle cell anaemia will have low fitness and as a result the allele gets lost in the population in due course of time. However, in malarial environment, Homozygote sickle cell anaemic individuals have the better fitness as equal to the normal homozygote individuals; as such both the alleles will be maintained in the population. (See Box 2.12)

Disruptive selection: both the extreme value (alleles) of a trait gets selected. It is one form of balanced selection. In case of disruptive selection, the extreme values or the alleles (low and high) of a trait will have a higher fitness when compared to the average value. As a result of disruptive selection the extreme values will increase as against the average values of the trait. This can be explained as leading to bimodal distribution (See Box 2.12).

b) Opportunity for natural selection

In general, to investigate natural selection in human populations is complex. Since natural selection operates on fertility and mortality, it can help us to get an overall idea of operation of natural selection. Indeed, Crow (1958) has formulated an index (Crow's Index) to examine the maximum intensity of (natural) selection that is more applicable for human populations; the index is based on the demographic components of fertility and mortality rates. According Crow "*there can be selection only if, through differential survival and fertility, individuals of one generation are differentially represented by progeny in succeeding generations. The extent to which this occurs is a measure of 'total selection intensity'. It sets an upper limit on the amount of genetically effective selection.*"

The total selection intensity (as defined by Crow) has two components: A fertility component (I_f) and mortality (I_m). The fertility and mortality patterns depend on several factors that vary across populations such as age at marriage, menarche, and survival to reach to fertility age, variation in fertility and age death etc. Likelihood of these occurring needs to be calculated based on age-sex structure.

The fertility and mortality also include embryonic development and birth; these have been incorporated to make it more rigorous and efficient estimate by Johnson and Kensinger (1971). More details of the Crow Index and the relationship are given in Box 2.13.

The estimates of ‘total intensity of selection’ have been studied in wide diverse populations. In Indian scenario, tribal populations show larger ‘Index of mortality than fertility components. There is also an overall all declines in I_m and I_t among urban communities as a result of socio-economic and public health facilities. More details of the trends of the Crow’s Index in Indian populations are described by Gautam (2009).

BOX 2.13

Index of opportunity for selection
Crow (1958) and Jhonston & Kensinger (1971)

The total selection intensity (I_t), is computed based on

I_m = index of opportunity for natural selection due to pre-reproductive mortality
(mortality from birth to reproductive age, i.e. below 15 years)

I_f = index of opportunity for natural selection due to fertility

X = average number of live births per women who have completed their reproductive life span
(aged 45 years and above)

V_f = variance (average deviation from mean) of number of live births

P_d = proportion of pre-reproductive deaths

P_s = proportion of survivors from birth to reproductive ages

The proportion of pre-reproductive deaths (P_d) is calculated from children ever-born to mothers aged 45 years and above (who have completed their fertility) and pre-reproductive deaths.

The proportions of survivors were calculated by subtracting P_d from 1:

$$I_t = I_m + I_f/P_s$$

$$I_m = P_d/P_s \quad \& \quad P_s = 1/P_d$$

$$I_f = V_f / X^2$$

The crow’s Index of opportunity for selection was modified by Johnston and Kensinger (1971) to account for the survival and mortality component during conception, before the birth of an infant.

This include I_{me} = the selection due to prenatal mortality, P_{ed} = the probability to die before birth,

P_b = the probability to survive till birth, I_{mc} the index of total selection due to postnatal mortality,

P_d = the probability to die before reaching reproductive age,

P_s = the likelihood to survive til reproductive age, I_f = selection due to fertility,

V = variance due to fertility among women who had completed their fertility,

X is the mean number of births, P_d and P_s are proportion of deaths and survivors.

The modified total intensity index I_t is:

$$I_t = I_{me} + (I_{mc}/P_b) + (I_f/P_b) P_s$$

$$I_{me} = P_{ed}/P_b, P_b = 1 - P_{ed}$$

$$I_{mc} = P_d/P_s$$

$$P_s = (1 - P_d)$$

$$I_f = V/X^2$$

2.4.1.4 Gene flow

a) Migration

Migration or gene flow is an important factor that can change the gene frequency. Emigration or immigration of individuals between populations can alter or change in the gene frequency. In genetic terms it is either loss of genetic diversity due to emigration or increase of genetic diversity due to immigration of individuals. There is loss of gene flow from a gene pool or gain of gene flow into a subpopulation from other gene pool. The quantitative estimate of the effect of migration in case of an allele at a single locus has been estimated by Bernstein and it has been shown (Box 2.14).

BOX 2.14

Change in gene frequency due to migration (m) / gene flow or genetic admixture

Suppose if migration is unidirectional from mainland to a nearby island and is random, then suppose

'm' is the rate of migration per generation from mainland to island

a) p_i be the frequency of gene A in immigrating individuals

' p_0 ' is the frequency of gene A in the island

b) The gene freq of A in the *island after migration* is

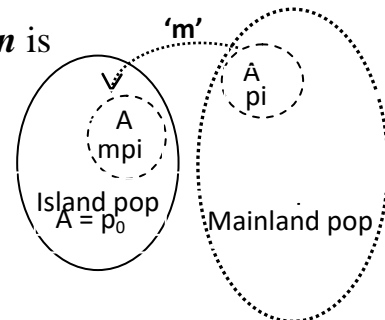
$$p_{am} = m p_i + (1 - m) p_0$$

The change in gene freq in one generation is

$$p_{am} - p_0 = [m p_i + (1 - m) p_0] - p_0$$

$$= m (p_i - p_0) + p_0 - p_0$$

$$m = (p_{am} - p_0) / (p_i - p_0)$$



NB: This is based on Bernstein's formula for an allele at a single locus

The effect of migration rate (m) on allele frequency in a population is the proportion of differences of allele frequency in the island population (p_i) before and after migration (p_{am}) to the difference between allele frequency in the migrant population (p_0) and the island population (p_i). The above formula can be extended for a multiple loci by using least square or maximum likelihood estimate procedures. It can also be worked out based on gene identity method.

b) Genetic admixture

Gene flow can happen between two subpopulations through random mating or admixture or marriages. The American Blacks, Anglo Indians, are

examples of genetic admixed populations. The Latin American countries are populated by admixed populations contributed by native tribes, African, European and other settlers. The estimates of admixture proportion can be estimated for a gene located at a specific locus of interest or for a set of genetic markers located at different loci. The above formula (Box 2.14) can be used to estimate the 'm' the admixture in a hybrid population. It is also possible to estimate the admixture proportions based on genetic distances and from principal component analysis (for multilocus allele frequencies).

c) **Barriers to gene flow**

Human populations live over wide geographical regions forming local subpopulations; these subpopulations are formed as a result of endogamy which is promoted by geographical, cultural, linguistic, political and other factors. The same factors form barriers for gene flow and restrict the admixture, intermarriages etc between the local populations. In India caste, geographical isolation, cultural, linguistic, political factors play a major role in restricting the gene flow or admixture or intermarriages between groups.

d) **Theoretical Models of gene flow**

These factors are important to consider estimating or modelling the gene flow between populations. In population genetic point of view, there is a decrease in genetic diversity with the increasing distance or geographical location of the populations. This gives spatial pattern of gene frequency clines, which help us to understand the geographic variation of genetic markers across populations and regions.

Since gene flow can occur in different scenarios, there are a variety of theoretical models to account for different situations of spatial gene exchange or flow. For example, Sewall Wright has proposed 'island', 'neighbourhood' by 'isolation by distance' models and 'steppingstone' models by Kimura and Weiss

Island model: It is the simple situation similar to island population. Suppose the population is distributed among a few close (equi distance) islands, each of population size N. The people tend to marry within each of the islands and gene flow is restricted, in the sense that there is equal immigration between islands, hence 'island model'. Suppose the mating takes place at random in each of such island or insular populations. The gene frequency in each of the island will differ with respect to total population (of all the islands). The theoretical results show that the deviation in such island model is exactly the

variance in allele frequency among the islands. The number of homozygotes in the total population is always larger than expected from HW proportions in that population. The result is known as '*Wahlund principle*'. For a two-allele polymorphism, the genotypic proportions in the total population are:

$$AA : p_0^2 + V, \quad Aa : 2p_0q_0 - 2V \quad \text{and} \quad aa : q_0^2 + V.$$

These proportions are similar to those population practicing inbreeding with inbreeding coefficient 'F' ($F = V/p_0q_0$). Where P_0 is the gene frequency of allele A and V is the variance of the gene frequency among the islands. "*The change in heterozygote frequency is twice the covariance among populations in the frequency of the allele in the heterozygote, and this may be positive or negative.*" (Christiasen and Feldman, 1986). One other model proposed by Sewall Wright is ***Neighbourhood model***.

Steppingstone model: The island model is too realistic to realise, therefore other models have been proposed which is more close to geographically structured populations. Kimura and Weiss (1964) proposed the '*stepping stone model*'. In 'one-step-linear (one dimensional) stepping stone model, the populations are arranged, rather in a linear fashion, on a long chain. The migration occurs between the neighbouring populations. This situation allows the distant populations with least migration between them are expected to behave differently than the neighbouring populations that are expected to change the gene frequency of the extreme populations as against the neighbouring populations. Kimura and Weiss (1964) have shown that the correlation in gene frequencies (r) between demes decreases approximately exponentially as a function of the number of steps (x) between deems.

This is expected to lead to clines in the gene frequency or geographical clines of the allele frequency. ***Isolation by distance model***: This was proposed by Sewall Wright, which is in a similar to the stepping stone model in a continuously distributed population.

2.4.1.5 GENETIC EQUILIBRIUM

The evolutionary forces of mutation, selection, and drift may oppose each other to create a dynamic equilibrium in which allele frequencies no longer change.

In a randomly mating population without selection or drift to change allele frequencies, and without migration or mutation to introduce new alleles, the Hardy-Weinberg genotype frequencies persist indefinitely. Such an idealized population is in a state of genetic equilibrium. In reality, the situation is much more complicated; selection and drift, migration and mutation are almost at work changing the population's genetic composition. However, these evolutionary forces may act in contrary ways to create a dynamic equilibrium in which there is no net change in allele frequencies. This type of equilibrium differs fundamentally from the equilibrium of the ideal Hardy-Weinberg population. In a dynamic equilibrium, the population simultaneously tends to change in opposite directions, but these opposing tendencies cancel each other and bring the population to a point of balance. In the ideal Hardy-Weinberg equilibrium, the population does not change because there are no evolutionary forces at work. However, opposing evolutionary forces can create a dynamic equilibrium within a population.

Box 2.15

Calculating Equilibrium Allele Frequencies with Balancing Selection

Genotypes:	AA	Aa	aa
Relative fitnesses:	1 - s	1	1 - t
Frequencies:	p^2	$2pq$	q^2
Average-relative fitness:	$\bar{w} = p^2 \times (1-s) + 2pq \times 1 + q^2 \times (1-t)$		
Frequency of A in the next generation after selection:	$p^1 = [p^2 (1 - s) + (1/2) 2pq] / \bar{w} = p(1-sp) / \bar{w}$		
Change in frequency of A due to selection:	$\Delta p = p^1 - p = pq(tq-sp) / \bar{w}$		
At equilibrium, $\Delta p = 0$	$p = t / (s + t)$ and $q = s / (s + t)$		

Balancing Selection

One type of dynamic equilibrium arises when selection favors the heterozygotes at the expense of each type of homozygote in the population. In this situation, called *balancing selection* or *heterozygote advantage*, one can assign the relative fitness of the heterozygotes to be 1 and the relative fitness of the two types of homozygotes to be less than 1:

Genotype:	AA	Aa	Aa
Relative fitness	1 - s	1	1 - t

In this formulation, the terms $1 - s$ and $1 - t$ contain selection coefficients that are assumed to lie between 0 and 1. Thus, each of the homozygotes has a lower fitness than the heterozygotes. The superiority of the heterozygotes is sometimes referred to as '*overdominance*'.

In cases of heterozygote advantage, selection tends to eliminate both the A and 'a' alleles through its effects on the homozygotes, but it also preserves these alleles through its effects on the heterozygotes. At some point these opposing tendencies balance each other, and a dynamic equilibrium is established. To determine the frequencies of the two alleles at the point of equilibrium, one must derive an equation that describes the process of selection, and then solve this equation for the allele frequencies when the opposing selective forces are in balance that is, when the allele frequencies are no longer changing. (Box 2.15).

At the balance point, the frequency of A is $p = t/(s + t)$,
and the frequency of a is $q = s/(s + t)$

As an example, let's suppose that the AA homozygotes are lethal ($s = 1$) and that the aa homozygotes are 50 percent as fit as the heterozygotes ($t = 0.5$). Under these assumptions, the population will establish a dynamic equilibrium when $p = 0.5/(0.5 + 1) = 1/3$ and
 $q = 1/(0.5 + 1) = 2/3$.

Both alleles will be maintained at appreciable frequencies by selection in favour of the heterozygotes – a condition known as a balanced polymorphism.

In humans, the disease sickle-cell anaemia is associated with a balanced polymorphism. Individuals with this disease are homozygous for a mutant allele of the β -globin gene, denoted Hb^S , and they suffer from a severe form of anaemia in which the haemoglobin molecules crystallize in the blood. This crystallization causes the red blood cells to assume a characteristic sickle shape. Because sickle-cell anaemia is usually fatal without medical treatment, the fitness of $Hb^S Hb^S$ homozygotes has historically been 0. However, in some parts of the world, particularly in tropical Africa, the frequency of the Hb^S allele is as high as 0.2. With such harmful effects, why does the Hb^S allele remain in the population at all?

The answer is that there is moderate selection against homozygotes that carry the wild-type allele Hb^A . These homozygotes that carry the wild-type allele Hb^A . These homozygotes are less fit than the $Hb^S Hb^A$ heterozygotes because they are more susceptible to infection by the parasites that cause malaria, a fitness-reducing disease that is widespread in regions where the frequency of the Hb^S allele is high.

We can schematize this situation by assigning relative fitness to each of the genotype of the β -globin gene:

Genotype:	$Hb^S Hb^S$	$Hb^S Hb^A$	$Hb^A Hb^A$
Relative fitness:	$1 - s$	1	$1 - t$

If one assumes that the equilibrium frequency of Hb^S is $p = 0.1$ – a typical value in West Africa – and if one notes that $s = 1$ because the $Hb^S Hb^S$ homozygotes die, one can estimate the intensity of selection against the $Hb^A Hb^A$ homozygotes because of their greater susceptibility to malaria:

$$p = t / (s + t)$$

$$0.1 = t / (1+t)$$

$$t = (0.1)/(0.9) = 0.11$$

This result tells us that the $Hb^A Hb^A$ homozygotes are about 11 percent less fit than the $Hb^S Hb^A$ heterozygotes. Thus, the selective inferiority of the $Hb^S Hb^S$ and $Hb^A Hb^A$ homozygotes compared to the heterozygotes creates a balanced

polymorphism in which both alleles of the β -globin gene are maintained in the population.

Various other mutant Hb alleles are found at appreciable frequencies in tropical and subtropical regions of the world in which falciparum malaria is – or was – endemic. It is plausible that these alleles have also been maintained in human populations by balancing selection.

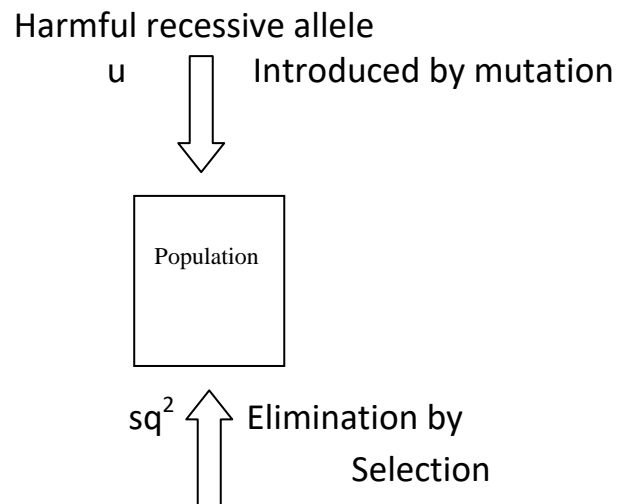
Mutation-Selection Balance

Another type of dynamic equilibrium is created when selection eliminates deleterious alleles that are produced by recurrent mutation. For example, let's consider the case of a deleterious recessive allele a that is produced by mutation of the wild-type allele A at rate u . A typical value for u is 3×10^{-6} mutations per generation. Even though this rate is very low, over time, the mutant allele will accumulate in the population, and, because it is recessive, it can be carried in heterozygous condition without having any harmful effects. At some point, However, the mutant allele will become frequent enough for aa homozygotes to appear in the population, and these will be subject to the force of selection in proportion to their frequency and the value of the selection coefficient s . Selection against these homozygotes will counteract the force of mutation, which introduces the mutant allele into the population.

If one assumes that the population mates randomly, and if one denotes the frequency of A as p and that of a as q , then one can summarize the situation as follows:

Mutation: Produces a $A \rightarrow a$ rate= u		Selection: eliminates a		
	Genotype:	AA	Aa	aa
	Relative fitness:	1	1	$1-s$
	Frequency:	p^2	$2pq$	q^2

Mutation introduces mutant alleles into the population at rate u , and selection eliminates them at rate sq^2



Mutation-selection balance for a deleterious recessive allele with frequency q . Genetic equilibrium is reached when the introduction of the allele into the population by mutation at rate u is balanced by the elimination of the allele by selection with intensity s against the recessive homozygotes.

When these two processes are in balance, a dynamic equilibrium will be established. We can calculate the frequency of the mutant allele at the equilibrium created by mutation – selection balance by equating the rate of mutation to the rate of elimination by selection:

$$u = sq^2$$

Thus, after solving for q , we obtain

$$q = \sqrt{u/s}$$

For a mutant allele that is lethal in homozygous condition, $s = 1$, and the equilibrium frequency of the mutant allele is simply the square root of the mutation rate. If one uses the value for u that was given above, then for a recessive lethal allele the equilibrium frequency is $q = 0.0017$. If the mutant allele is not completely lethal in homozygous condition, then the equilibrium frequency will be higher than 0.0017 by a factor that depends on $1/\sqrt{s}$. For example, if s is 0.1, then at equilibrium the frequency of this slightly

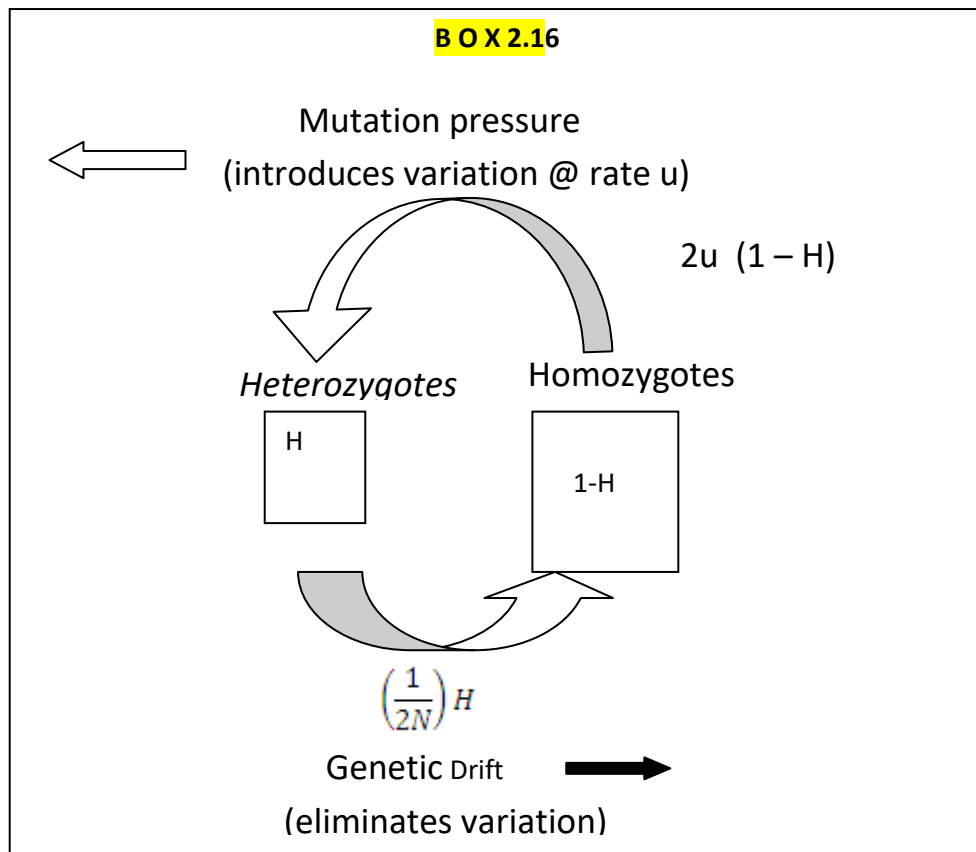
deleterious allele will be $q=0.0055$, or 3.2 times greater than the equilibrium frequency of a recessive lethal allele.

Studies with natural population of *Drosophila* have indicated that lethal alleles are less frequent than the preceding calculations predict. The discrepancy between the observed and predicted frequencies has been attributed to partial dominance of the mutant alleles—that is, these alleles are not completely recessive. Natural selection appears to act against deleterious alleles in heterozygous condition as well as in homozygous condition. Thus, the equilibrium frequencies of these alleles are lower than one would otherwise predict. Selection that acts against mutant alleles in homozygous or heterozygous condition are sometimes called **purifying selection**.

Mutation-Drift Balance

The random genetic drift eliminates variability from a population. Without any counteracting force, this process would eventually make all populations completely homogeneous. However, mutation replenishes the variability that is lost by drift. At some point, the opposing forces of mutation and genetic drift come into balance and a dynamic equilibrium is established. The genetic variability can be quantified by calculating the frequency of heterozygotes in a population— a statistic called the heterozygosity, which is symbolized by the letter H . The frequency of homozygotes in a population— often called the *homozygosity*— is equal to $1-H$. Over time, genetic drift decreases H and increases $1-H$, and mutation does just the opposite as shown in the figure below (Box 2.16).

Let's assume that each new mutation is selectively neutral. In a randomly mating population of size N , the rate at which drift decreases H is $(\frac{1}{2N})H$. The rate at which mutation increases H is proportional to the frequency of the homozygotes in the population $(1-H)$ and the probability that one of the two alleles in a particular homozygote mutates to a different allele, thereby converting that homozygote into a heterozygote. This probability is simply the mutation rate μ for each of the two alleles in the homozygote; thus, the total probability of mutation converting a particular homozygote into a heterozygote is 2μ . The rate at which mutation increases H in a population is therefore equal to $2\mu(1 - H)$.

BOX 2.16

When the opposing forces of mutation and drift come into balance, the population will achieve an equilibrium level of variability denoted by \hat{H} . This equilibrium value of H can be estimated, by equating the rate at which mutation increases H to the rate at which drift decreases it:

$$2\mu(1 - H) = \left(\frac{1}{2N}\right)H$$

By solving for H , the equilibrium heterozygosity at the point of mutation-drift balance is obtained as :

$$\hat{H} = 4N\mu / (4N\mu + 1)$$

Thus, the equilibrium level of variability (as measured by the heterozygosity) is a function of the population size and the mutation rate.

If one assumes that the mutation rate is $\mu = 1 \times 10^{-6}$, one can plot \hat{H} for different values of N . For $N < 10,000$, the equilibrium frequency of heterozygotes in the population will be quite low; thus, drift dominates over mutation in small populations. For N equal to $1/\mu$, the reciprocal of the mutation rate, the equilibrium frequency of heterozygotes would be 0.8, and for even greater values of N , the frequency of heterozygotes increases asymptotically towards 1. Thus, in large populations, mutations dominate over drift; every mutational event creates a new allele, and each new allele contributes to the

heterozygosity because the large size of the population protects the allele from being lost by random genetic drift.

Values of \hat{H} in natural populations vary among species. In the African cheetah, for example, \hat{H} is 1 percent or less among a sample of loci, suggesting that over evolutionary time, population size in this species has been small. In humans, \hat{H} is estimated to be about 12 percent, suggesting that evolutionary time population size has averaged about 30,000 to 40,000 individuals. Estimates of population size that are derived from heterozygosity data are typically much smaller than estimates obtained from census data. The reason for this discrepancy is that the estimates based on heterozygosity data are *genetically effective* population sizes- sizes that take into account restrictions on mating and reproduction, as well as temporal fluctuations in the number of mating individuals. The genetically effective size of a population is almost less than the census size of a population.

(Source: Principles of Genetics (2006) by D. Peter Snustad and Michael J. Simmons. John Wiley & Sons (Asia Edition) PP. 750-754.

2.5 Summary

1. Understanding of Population genetics principles, requires the basic concepts of Mendelian genetics: the result of segregation, the concept 'gene', 'phenotype', 'genotype', 'dominant', 'recessive' traits, 'allele' etc. Parental mating types and expected distribution of genotypes among the offspring.
2. Hardy-Weinberg equilibrium is the solution to an intriguing question: what happens to gene frequency of a dominant character over generations in a population. With three times more frequent than normal does this will increase over generations?
3. HWE law states that under the absence of intervening factors, especially in a large population, given random mating, no selection of any sort, no mutation and absence of demographic factors like migration, differential fertility and mortality etc., the allele frequency remain constant over generations. This can be proved theoretically, easily, for a 'biallelic locus and it can be extended to multilocus as well.
4. The importance of HWE: it gives a methodology to estimate the allele frequency in a population based on phenotypic/genotypic information of the parental mating types. It helps us to investigate the relationship between

change in gene frequency with respect to mutation, migration, selection, genetic drift etc. The entire investigation is the kernel of a branch of biomathematics or the new field: 'population genetics' and 'quantitative genetics'.

7. HWE is the bench mark of qualitative test to check whether a trait, an allele, SNP, is in equilibrium. It tells how to distinguish between the effects of evolutionary forces from the demographic factors.

8. Mutation is a non-systematic and random, but rate of mutation is site specific. Mutations are more frequent at hot-spots and are rare at the 'conserved region'. The mitochondrial non-coding genome has a higher frequency of mutations than the nuclear genome.

9. Genetic drift is a non-systematic force which can lead to significant changes in gene frequency in a small population. If an allele is rare in a small population, it can get lost or get fixed in the population over generations.

10. Founder effect is one form of genetic drift. The founders are a sample (represent a fraction of the genetic diversity) of original populations. The descendents of a few founders have the gene frequency that is dependent on the genetic composition and genetic structure of the founders. It can also happen as bottleneck effect, especially as a result of sudden population size reduction in a population, due to reasons such as natural causes or man-made causes or socio-cultural regulations. There could be serial founder effect as a result of waves of migration at different times. The mitochondrial investigation of human origins suggests that the human origins and migration to other continents appears as a result of serial founder effect from Africa.

11. Natural selection is one of the complex systematic forces that can influence significant changes in gene frequency. Selection can operate in multitude ways and it is a slow process than to the effect of migration or admixture etc.

12. Selection basically operates at differential fertility and mortality levels. It is measured as 'fitness' the ability to leave offspring and refers to 'relative rate of survival'. It is measured by 'selection coefficient' ('s') which is a function of fitness (W). The fitness or selection coefficient differs with respect to the type of dominance: complete, partial, over etc.

13. The effect of 'directional selection' to shift the mean allele frequency towards its extremes. Or it could be stabilizing selection that shifts the allele frequency of extreme alleles as a result the heterozygote frequency will

increase. Or it could be disruptive selection where the extreme allele frequency increases as against the heterozygote frequency.

14. Selection can also be measured based on demographic factors of fertility and mortality trends. Crow's Index of opportunity for selection measures total selection intensity that a population can experience which depend on two components, fertility and mortality.

15. Gene flow (migration/admixture) is a systematic factor which can bring rapid changes in gene frequency within a short period. In general, human populations follow a variety of restrictions or regulations that restrict gene flow between and within populations. The barriers for gene flow could be because of culture or due to geographical, political, religious and linguistic etc.

16. There are theoretical models to investigate the effect of spatial gene flow or population structure between populations. Island model, stepping stone model, neighbourhood model help us to investigate the spatial gene flow in different situations of population structure.

Over view

Overall, this unit gives an account of what is population genetics, what are its basic concepts, HWE and its importance, what are the deviating forces of HWE. What is the theoretical expectation of change in gene frequency when there is mutation, migration, genetic drift, gene flow in populations? The empirical examples help us to examine the operation of these forces among human populations.

Suggested reading

Crow, J.F. and Kimura M. 1970. An introduction to population genetics theory

Li, C.C. 1976. First course in population genetics

Cavalli Sforza L.L. and Bodmer W.F. 1971. The genetics of human populations.

W.H. Freeman, Sanfrancisco. USA

Falconer DS, 1980. Introduction to quantitative genetics. Longman, London and New York, Second edition.

Christiansen B Freddy and Feldman W Marcus 1986. Population genetics.

Blackwell Scientific Publications (Australia) Pvt. Ltd., Victoria.

Gautam RK (2009). Opportunity for natural selection among the Indian population: secular trend, covariates and implications. *J. Biosoc. Sci.* 41:705-745.

Kimura M and GH Weiss 1964. The steppingstone model of population structure and the correlation with distance. *Genetics* 49:561-576.

Majumder, PP. 1993. *Human Population Genetics. A centennial tribute to JBS Haldane.* Ed. PP Majumder. Plenum Press. New York.

Malhotra KC. 1988. *Statistical Methods in Human Population Genetics.* Ed. KC Malhotra. Eka Press. Calcutta 700035.

Dobzhansky, T. 1951. *Genetics and the origin of species.* Columbia University Press. NY.

Sample questions

1. A total of 120 individuals were tested for M, N blood group and the observed genotype frequencies of MM, MN and NN are 34, 62 and 24 respectively. Calculate the gene (allele) frequencies?
2. If ' μ ' is the mutation rate ($\mu = 10^{-5}$) per generation for a gene frequency of A then how many generations are required to reduce the gene frequency by a factor of $\frac{1}{2}$.
3. What is Hardy-Weinberg equilibrium? Explain why HWE is important in genetic of populations?
4. In case in a population the observed gene frequencies of a particular bi-allelic locus are in HW equilibrium for the locus, does this imply the population satisfies the assumptions of the HW equilibrium? Explain?
5. What is genetic drift and how it operates in populations? Explain with Examples.