UNIT 3

GENES

Structure

3.1 INTRODUCTION

In Unit 2, you have learnt about the various types of repeat elements in the intergenic regions of genomes. The present Unit will focus on genes and their evolution. You will learn about the mechanisms by which genomes acquire genes with novel functions, while other genes become non-functional as a result of evolutionary processes. You will also learn more about how paralogs are classified into gene families based on sequence similarity. Finally, you will learn how bacteria acquire new genes through the process of horizontal gene transfer.

Objectives

By the end of this Unit you would be able to:

- describe gene duplication,
- differentiate between whole genome duplication and single gene duplication,
- explain genome evolution driven by gene duplication,
- explain acquiring of genes with new functions in genomes,
- describe the conversion of some genes to non-functional pseudogenes,
- differentiate between the different types of gene families, and
- describe the process of horizontal gene transfer and its role in bacterial evolution.

3.2 GENE DUPLICATION

Gene duplication is a mutation that produces duplicate copies of genes **(paralogs)** which serve as raw material for change during evolution. When a gene is duplicated, the extra copies are functionally unnecessary to the organism leading to functional redundancy. In some cases, the expression of the additional genes may even result in an imbalance of gene products. In many cases, the duplicate genes are, therefore, made non-functional by deleterious mutation or splicing errors, giving rise to non-functional genes called **pseudogenes**. In other cases, the functional constraints over the duplicate copies are removed, and they are free to mutate and produce genes with similar functions to the original, such as the *Hox* family of genes in humans, or novel genes with new functions, such as the evolution of the stomach lysozyme by duplication of the gene for non-stomach lysozyme. Thus, gene duplication gives rise to **gene families** of structurally and functionally related genes.

Gene duplication can occur by whole-genome or single-gene duplication (Fig. 3.1). Whole genome duplication or polyploidization involves the duplication of the entire genome of an organism. It is an extreme mechanism of gene duplication that leads to a sudden increase in both genome size and the entire gene set, and is rampant in the plant kingdom, but not among animals. For example, the content of duplicate genes is 16% in barley (*Hordeum vulgare*), 49% in rice (*Oryza sativa*), and 75% in soybean (*Glycine max*). A large number of duplicate genes in crop species such as wheat (*Triticum aestivum*), cotton (*Gossypium hirsutum*), and soybean have contributed to important agronomic traits, such as grain quality, fruit shape, and flowering time.

Single gene duplication includes different types, namely:

(a) **Tandem duplication:** The tandem duplication usually takes place by homologous recombination between repeat sequences such as *Alu* sequences involving unequal sister-chromatid exchange. It gives rise to gene copies which insert in tandem at a chromosomal location. For example, tandem gene duplication gave rise to the paralogous α- and β-globin genes, which led to the evolution of multimeric hemoglobin composed of two subunits ($\alpha_2\beta_2$).

(b) **Transposon-mediated duplication or DNA-transposed duplication:** It is a results of DNA-based replicative transposition in which a gene remains in its original location, while the duplicate copy is inserted at a new chromosomal location.

(c) **Segmental duplications:** The segmental duplications or "duplicons" are regions of more than 1 kb length with greater than 90% similarity, and are present more than once in a genome, either on the same chromosome (intrachromosomal), or on different chromosomes (interchromosomal). Approximately 5% of the human genome comprises of segmental duplications such as LINEs.

(d) **Retrotransposed duplication:** It is the duplication of a gene *via* reverse transcription of processed mRNA into a complementary DNA (cDNA) copy which is inserted back into the genome, and is known as a **retrogene**. For example, the *PTEN1* gene is a retrogene arising by duplication of the *PTEN* tumor suppressor gene.

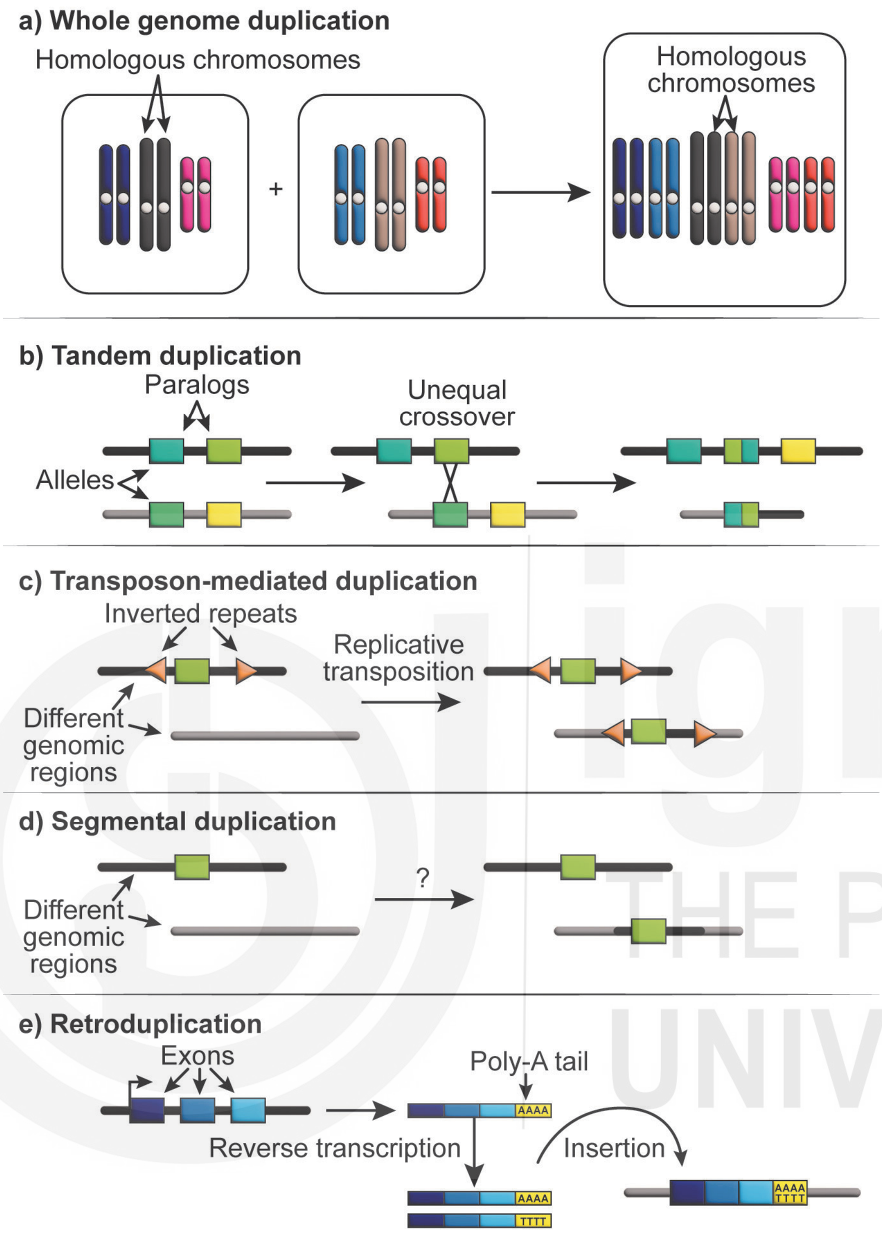

Fig. 3.1: Mechanisms of gene duplication

3.2.1 The Globin Genes

The evolution of the vertebrate globin gene family is an excellent example of the role of gene duplication in promoting the emergence of novel genes with specialised functions. The diversification of the vertebrate globins occurred through a combination of both whole-genome duplication and tandem duplication events. Vertebrates have two major types of globin proteins namely haemoglobin (Hb) and myoglobin (Mb). Hb is a tetrameric protein found in red blood cells. It is composed of two α-chains and two β-chains ($\alpha_2\beta_2$). Each of these subunits contain an iron heme group which reversibly binds a single O_2 molecule. Hb transports O_2 from the respiratory exchange surfaces such as lungs, gills, or skin to the cells of respiring tissues. On the other hand, the Mb protein consists of a single polypeptide chain with a prosthetic heme group which stores O_2 and facilitates intracellular O_2 diffusion in cardiac and skeletal muscles. The common ancestors of vertebrates were small animals with only a single-chain, myoglobin-like ancestral globin for oxygen storage. These animals probably did not require a protein for O_2 transport because of their small size.

The progenitor genes of Hb and Mb proteins referred to as the proto *Hb* and *Mb* genes originated via whole-genome duplication in the common ancestor of vertebrates which is estimated to have occurred about 500 million years ago (mya). Two different progenitor genes made way for the division of labour between two different O_2-binding proteins, one of which would function as an O_2 carrier (Hb) and the other in O_2 storage functions (Mb). Jawless fish called lampreys are among the most primitive living vertebrates and possess a primitive form of Hb which is monomeric when oxygenated and associates into dimers and tetramers upon deoxygenation. Subsequent duplication of the proto *Hb* gene occurred approximately 450 mya at about the same time as the divergence of bony fish and sharks, giving rise to the progenitors of the α- and β-type globins which ultimately led to the formation of tetrameric *Hb* (Fig. 3.2). The proto α- and β-globin genes then evolved different functions through a series of gene duplication events giving rise to the α-globin and β-globin gene clusters. Gene duplications occurring about 350 mya and then 165 mya gave rise to the α-globin gene cluster, and those occurring 165 mya, 105 mya and more recently around 40-45 mya resulted in the formation of the β-globin gene cluster (Fig. 3.2).

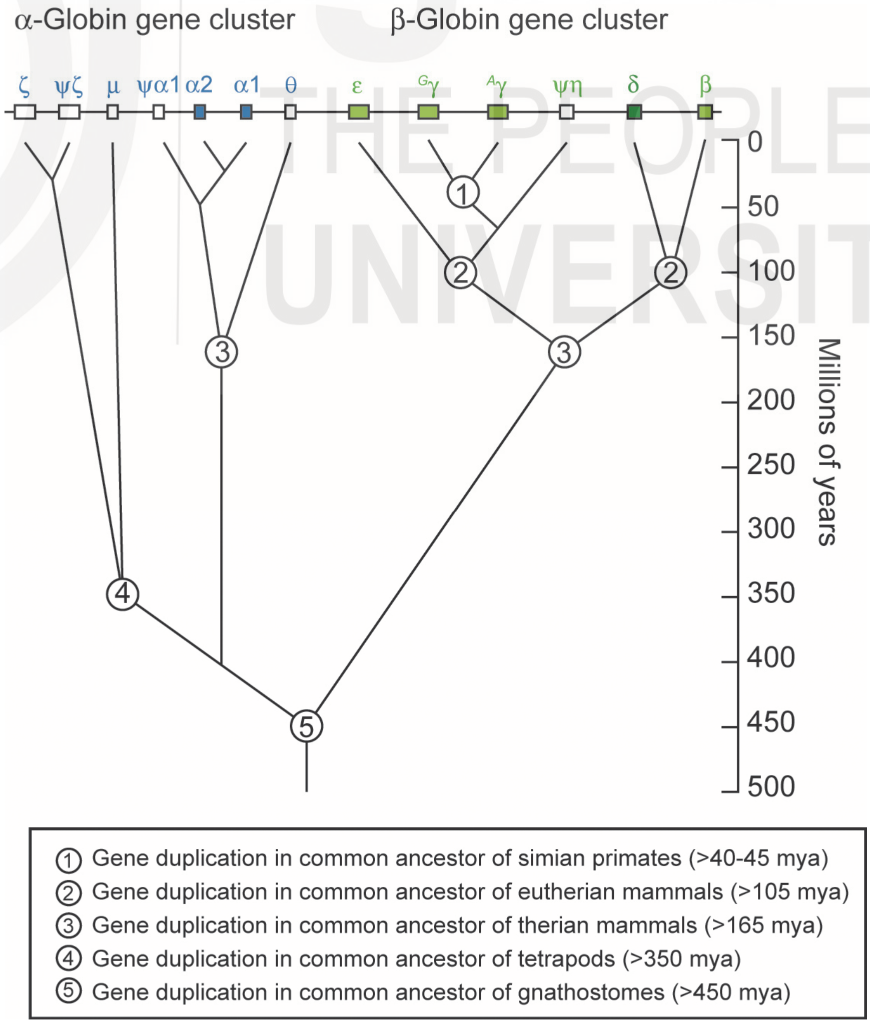

Fig. 3.2: Diversification of globin genes in vertebrates

Expression of different genes from the α and β clusters during different stages of development leads to the formation of hemoglobins differing from one another in properties such as O_2 affinity. This confers advantages during the specific developmental stage. Thus, $\zeta_2\gamma_2$ (Hb Portland-1), $\zeta_2\varepsilon_2$ (Hb Gower-1), and $\alpha_2\varepsilon_2$ (Hb Gower-2) are formed in the first few weeks of life during the embryonic period; $\alpha_2\gamma_2$ (fetal Hb or HbF) is formed during the fetal stage and degrades rapidly in newborns; and $\alpha_2\beta_2$ (adult Hb or HbA) as well as small amounts of $\alpha_2\delta_2$ (HbA_2) are formed in adults. The ancestral linkage arrangement of the proto α- and β-globin genes is still retained in the genomes of some modern-day amphibians and teleost fish. However, in the human genome, the α-globin gene cluster is located on chromosome 16, and the β-globin gene cluster is located on chromosome 11 indicating that the ancestral β-globin gene underwent transposition to a new chromosomal location in the lineage leading to modern reptiles, birds and mammals.

SAQ 1

Fill in the blanks:

a) A duplicated gene provides .... for change during evolution.

b) Gene duplication can occur by .... duplication or .... duplication.

c) Polyploidization leads to a sudden increase in .... size and the entire ....

d) .... are living animals which have primitive haemoglobin.

e) Evolution of the multimeric haemoglobin gene involved in the .... duplication of alpha and beta globin genes.

f) Fetal haemoglobin has two .... chains and two .... chains.

g) Segmental duplications or .... are regions of more than 1 kb length with greater than 90% similarity.

h) A .... is formed when a cDNA copy formed by transcription of processed mRNA is inserted back into a genome.

3.3 GENE FAMILIES

A gene family is a group of genes which have been derived from a common ancestral gene by gene duplication. Members of a gene family share the homology of sequence and similarity of function. However, the extent of sequence identity and similarity of the function differs, depending upon the degree to which they have diverged from the ancestral gene during the process of evolution. Gene families can be classified into four categories on the basis of the extent of sequence identity:

a) **Gene families with closely related members:** These families comprise of genes having a high degree of sequence homology over most of the length of the gene. For example, the α-globin gene family comprises of 7 genes clustered together at chromosomal location 16p13 (Fig. 3.3). All these genes are highly similar in sequence to one another.

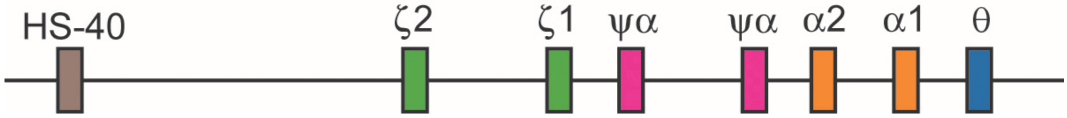

Fig. 3.3: The alpha-globin genes on chromosome 16. The two pseudogenes are denoted by Ψ

b) **Gene families with members having low sequence homology but encoding similar protein domains:** The genes in such gene families have low sequence homology over their entire length, except for short sequences of high similarity that encode one or more protein domains which confer the family with specific functions. For example, the helix–loop–helix DNA-binding domain family comprises of a number of eukaryotic DNA-binding proteins that act as transcription factors. These proteins share a region coding for a conserved domain of 40-50 amino acid residues called the 'helix-loop-helix' (HLH) domain which mediates protein dimerization. Most of these proteins have an extra basic region of about 15 amino acid residues that is adjacent to the HLH domain and specifically binds to DNA (Fig. 3.4).

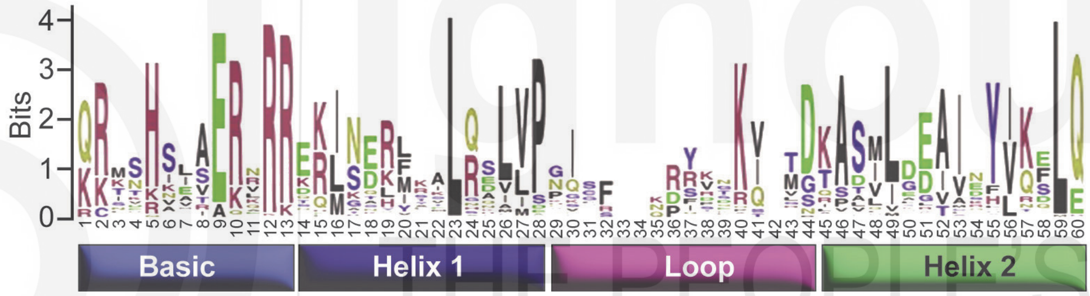

Fig. 3.4: The DNA binding domain of the helix-loop-helix family

c) **Gene families with members having low sequence homology but encoding similar protein motifs:** Members of these gene families may be dissimilar to one another in sequence except for short stretches coding for similar motifs. A motif is a short conserved sequence pattern associated with distinct functions of a protein or its encoding DNA. It is shorter in length than a domain and is often associated with a distinct structural site performing a particular function. For example, the DEAD-box proteins encoded by a family of genes, play important roles in RNA metabolism, are highly specific and cannot mutually be replaced. They comprise nine highly conserved motifs including the DEAD box made up of the amino acids D-E-A-D (Asp-Glu-Ala-Asp), after which the family is named (Fig. 3.5).

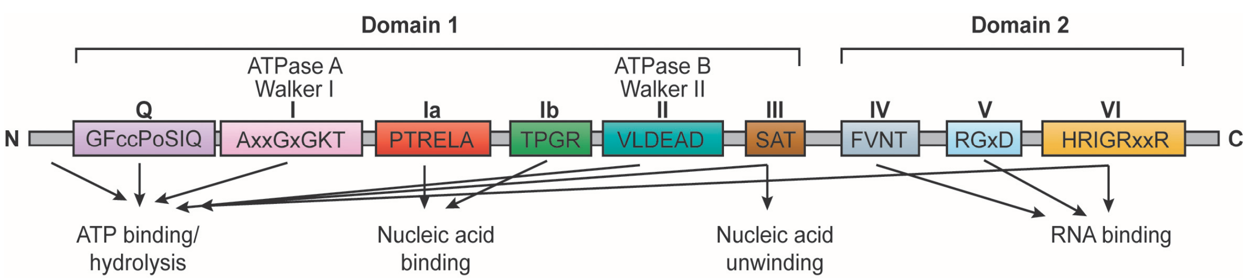

Fig. 3.5: Conserved sequence motifs of the DEAD box containing RNA helicases

d) **Gene superfamilies:** Superfamilies comprise genes which originated from a common ancestral gene, but have evolved to acquire different functions. For example, the immunoglobulin superfamily (IgSF) represents a large group of proteins with the common feature that they all contain one or more extracellular Ig domains. In addition to immunoglobulins, this superfamily includes a variety of cell surface proteins and soluble proteins involved in the recognition, binding, or adhesion processes of cells. For example, CD2, CD3, CD4, CD7, CD8, CD28, T cell receptor (TCR), MHC class I and MHC class II molecules.

SAQ 2

Fill in the blanks:

a) Gene families consist of members which have evolved from a common .... gene by duplication.

b) The members of the HLH gene family are functionally .... They all act as .... factors.

c) A .... is a conserved sequence shorter in length than a domain which is associated with a specific function of proteins.

d) The DEAD box proteins function in .... metabolism.

e) Gene superfamilies comprise of members which share a common ancestral gene but have .... functions.

f) Members of the α-globin gene family are located on chromosome .... and have .... sequence homology.

3.3.1 Organization of Gene Families in Genomes

Eukaryotic genomes are less compact than prokaryotic genomes, and the genes are organized into ‘gene rich’ or ‘gene poor’ regions, separated by stretches of intergenic DNA of variable length. The location of genes in a genome can be identified by computational methods or experimental analysis.

Computational methods for locating gene: There are two approaches commonly used for locating genes in a genome sequence by computational analysis:

(a) **Locating genes by genome analysis:** Computer-based gene prediction softwares like GENSCAN and GrailEXP are used to locate genes in a genome sequence by identifying the distinctive features of a gene. These softwares search open reading frame (ORF) of a protein-coding gene, codon bias, identifying exon-intron boundaries, upstream regulatory elements and CpG islands.

- **ORF Scanning:** This involves scanning a genome sequence for ORFs. An ORF is the sequence of codons which specify the amino acids of a protein. It begins with the start codon ATG (GTG and

ATA in the mitochondrial genome of some species), and ends with a stop codon (TAA, TAG, TGA). The length of the ORF is crucial in correct identification, as the average length of a gene is 317 codons for genes of *Escherichia coli*, 483 codons for genes of *Saccharomyces cerevisiae*, and approximately 450 codons for genes of humans.

- **Codon bias:** The genetic code is degenerate hence multiple codons can code for particular amino acid, not all codons are used at an equal rate in the genes of an organism, and some codons are found to be more favoured than others. For example, while the amino acid may be coded by four different codons- GCA, GCC, GCU and GCG, the preferred codon for alanine in human genes is GCC. This feature should be kept in mind while locating the genes of a species in a genome.
- **Exon-intron boundaries:** Exon-intron boundaries are the sequences which signify the end of an exon and the beginning of an intron. Locating them in series can indicate the presence of a gene in a eukaryotic genome. Exon-intron boundaries are indicated by the sequence 5'-AG↓GT/A-3' at the exon-intron junction.
- **CpG islands:** CpG islands are 500-1500 bp long stretches of DNA rich in CpG dinucleotides (p indicates a phosphodiester bond), and are characterised by a CG: GC ratio of more than 0.6.They are frequently associated with the promoters of vertebrate genes, and are located in the vicinity of about 40% of promoters of mammalian genes. Thus, CpG islands are good indicators of the location of genes in vertebrate genomes.

(b) **Locating genes by homology with known proteins:** Computer-based gene prediction softwares may also be used to search databases for sequences of known proteins or segments of proteins which match with the protein predicted from a putative gene sequence. A match (homology) indicates that the putative gene sequence is likely to be evolutionarily related to the gene coding for the known protein, and both genes probably have similar functions.

Experimental approaches to locating genes on chromosomes: In addition to the above computational approaches, there are a number of experimental approaches through which genes can be located. These include hybridization of a gene-specific probe to a chromosome or a transcript of a gene. Fluorescence *in situ* hybridization (FISH) is a cytogenetic technique used to detect and locate a gene or specific DNA sequence on a chromosome. This method is also called chromosome painting. The full set of metaphase chromosomes of an organism are fixed on a glass slide and probed with a piece of purified single-stranded DNA which is labelled with a fluorescent dye. The fluorescent labelled probe hybridizes to a complementary region of DNA in the target gene and can be visualized on the chromosome with the help of a fluorescence microscope. A number of genes or DNA sequences on the same or different chromosomes can be located at the same time using differently coloured probes.

FISH played an important role in mapping genes on human chromosomes during the annotation phase of the Human Genome Project when the location and functions of genes in the human genome sequence were being determined. It continues to remain an important method for mapping the genes on chromosomes of genomes which are still in the process of being sequenced. In humans, FISH is now commonly used as a technique for the clinical diagnosis of chromosome abnormalities like chromosomal translocations involved in producing the BCR-ABL fusion gene in acute lymphoblastic leukemia (ALL).

Gene family members are arranged in clusters: Availability of the gene sequences and gene locations of different organisms revealed that the members of gene families are not located randomly on chromosomes, and are mostly organized in clusters on the same or different chromosomes. A **gene cluster** is a group of a few to several hundred genes that encode products with similar sequence and/or function, and are often located near to each other on a genome. The different types of gene clusters formed are as follows:

a) **Clustered gene families:** Genes that produce structurally and functionally similar proteins are often located near to one another at the same location called **gene clusters**. Several examples such as the cluster of 16S rRNA genes located on chromosome 1q42 are **tandem clusters** with multiple copies of genes inserted in tandem. On the other hand, **interspersed gene clusters** comprise of genes from the same family located in small clusters on different chromosomal locations. For example, the olfactory receptor gene family encodes a diverse range of receptors which enable us to sense a large collection of smells. These are located in large clusters at multiple different locations on chromosomes **9, 17, 19, 11, 16, 5, 1, 6 and X.**

b) **Dispersed gene families:** The individual members of the dispersed gene families are located on different chromosomal locations. Many of the dispersed gene families are thought to have been formed as a result of duplication by retrotransposition followed by subsequent integration of the retrogene into the genome. In many cases, the retrogene becomes non-functional and is converted to a pseudogene. For example, the aldolase gene family comprises three functional genes and two pseudogenes. The functional human aldolase A gene is located on chromosome 16, aldolase B on chromosome 9, and aldolase C on chromosome 17.

SAQ 3

Fill in the blanks:

a) Genes encoding structurally and functionally similar products are often located near one another on chromosomes in the form of gene ....

b) 16S rRNA genes on chromosome 1q42 are organized in .... clusters.

c) The members of gene families arranged in interspersed gene clusters are located on .... chromosomes.

d) The olfactory receptor gene family is involved in the sense of ....

e) Many of the .... gene families are thought to have originated by insertion of a retrogene into a genome.

f) The .... is an experimental technique for locating genes and DNA sequences on chromosomes.

g) The .... are regions rich in CG dinucelotides associated with the promoters of vertebrate genes.

h) AG↓GT is the sequence found at ....

3.4 PSEUDOGENES

The term 'pseudogene' was coined in 1977 to refer to a non-functional 5S rRNA gene in *Xenopus laevis*. They are important constituents of non-coding DNA found in plant, insect, mammalian and bacterial genomes. A DNA segment that structurally resembles a gene but is unable to code for a protein is known as a **pseudogene**. Most often, pseudogenes are derived from genes that have accumulated mutations during evolution to the stage that they can no longer code for proteins. They have high sequence homology to known functional genes, but do not code for proteins. They are inheritable genetic elements and are often members of gene families containing functional genes. For example, the beta-globin gene family located on chromosome 11 consists of five functional genes and two pseudogenes (Fig. 3.6).

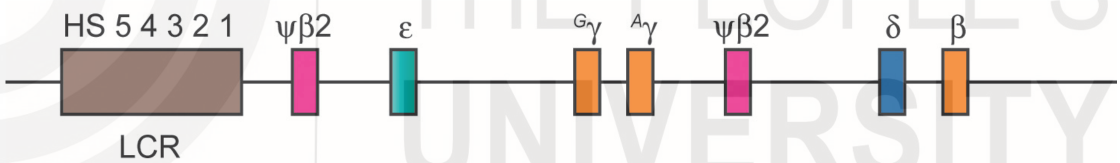

Fig. 3.6: The Beta-globin genes on chromosome 11. The two pseudogenes are denoted by Ψ

Pseudogenes are of two types:

(a) **Unprocessed pseudogenes:** The unprocessed pseudogenes are formed when genes are duplicated by one or more copies underwent deleterious genetic alterations such as frameshift mutations and creation of a premature stop codon which prevents the gene from encoding a functional protein. Some pseudogenes such as the *OCT4*-pseudogene 1 (*OCT4*-pg 1), *OCT4*-pseudogene 2 (*OCT4*-pg 2), *OCT4*-pseudogene 3 (*OCT4*-pg 3) and *OCT4*-pseudogene 4 (*OCT4*-pg 4) are transcribed to produce unstable proteins in various human cancer cell lines and are thought to have a role in tumor cell growth and tumor progression. Unprocessed pseudogenes are found in proximity to the genes from which they are derived, such as the pseudogenes in the previous example of the beta-globin gene family.

(b) **Processed pseudogenes:** Processed pseudogenes are produced by a retrotransposition event involving the synthesis of complementary DNA (cDNA) from an RNA product of a gene by reverse transcription, followed by insertion of the cDNA in a different region of the genome. Unlike the functional gene, the cDNA copy lacks a promoter, since the promoter sequence itself is not transcribed during RNA synthesis. Thus, the cDNA copy forms a pseudogene which is not expressed. A processed pseudogene also often lacks introns, since these are absent from the mature mRNA template, and instead contain a poly A-tail. For example, a processed pseudogene, *Mgst1-psi,* is found in the GST supergene family in *Drosophila melanogaster.* This pseudogene contains a poly-A tail towards the 3′ end, has no introns, a premature stop codon, and a direct repeat at both ends. It is presumed to have been derived from the reverse transcription and subsequent integration of *Mgst1* mRNA into the *Drosophila melanogaster* genome.

Pseudogenes were once referred to as 'genomic fossils' and were regarded as junk DNA because they are not able to produce functional proteins. However, approximately 20% of pseudogenes are transcribed into RNAs that control expression of other genes by producing natural small interfering RNA (siRNAs) and long non-coding RNAs (lncRNAs). A few pseudogenes occurring in the human genome have been found to have oncosuppressive functions, that is, they can prevent the growth of cancer cells and the progression of cancer. The *PTEN* pseudogene, *PTEN1*, is a processed pseudogene of *PTEN*. It contains a mutated start codon and is unable to produce a protein. However its mRNA acts as a 'sponge' by complementary binding to a number of miRNAs such as miR-17, miR-19, miR-21, miR-26, and miR-214 which normally target and inhibit *PTEN*. The sponging effect of *PTEN1* relieves the inhibitory effects of the miRNAs on PTEN, restoring its normal tumor suppressor activities and thus leading to a decrease in the growth of many cancers.

SAQ 4

Fill in the blanks:

a) A non-functional gene found within a family of other functional genes is called a ....

b) The .... pseudogenes are produced as a result of deleterious mutations which cause genes to become non-functional.

c) Pseudogenes that have a poly-A tail but lack introns and a premature stop codon are called .... pseudogenes and are produced as a result of .... events.

d) Approximately 20% of pseudogenes encode .... and ....

e) *Oct4* pseudogenes are examples of .... pseudogenes.

f) *PTEN1* is a pseudogene with .... functions.

g) Insertion of a .... copy of a gene into a genome leads to the formation of a processed pseudogene.

3.5 SEGMENTAL DUPLICATIONS

Segmental duplications (see section 3.2) are long DNA sequences with high similarity which are repeated multiple times in a genome. They exist as low copy number repeats and are associated with genetic diversity, and genome evolution. Segmental duplications have been found to contain rapidly evolving genes in the human and great ape lineages. These repeats also constitute significant paralogous regions in the human genome which act as sites for non-allelic homologous recombination (NAHR) leading to genetic rearrangements which play an important role in human health and disease. Red-green colour blindness is an example of a genetic disorder resulting from NAHR between paralogous genes. The red and green pigment genes (*OPN1LW* and *OPN1MW*) are located in tandem on chromosome Xq28, and share 98% identity. Individuals with normal colour vision have at least one copy of the red pigment gene, *OPN1LW*, and more than one copy of the green pigment gene, *OPS1MW*. However, NAHR between *OPN1LW* and *OPN1MW* leads to deletion of one of the genes, resulting in red or green colour blindness in males (Fig. 3.7).

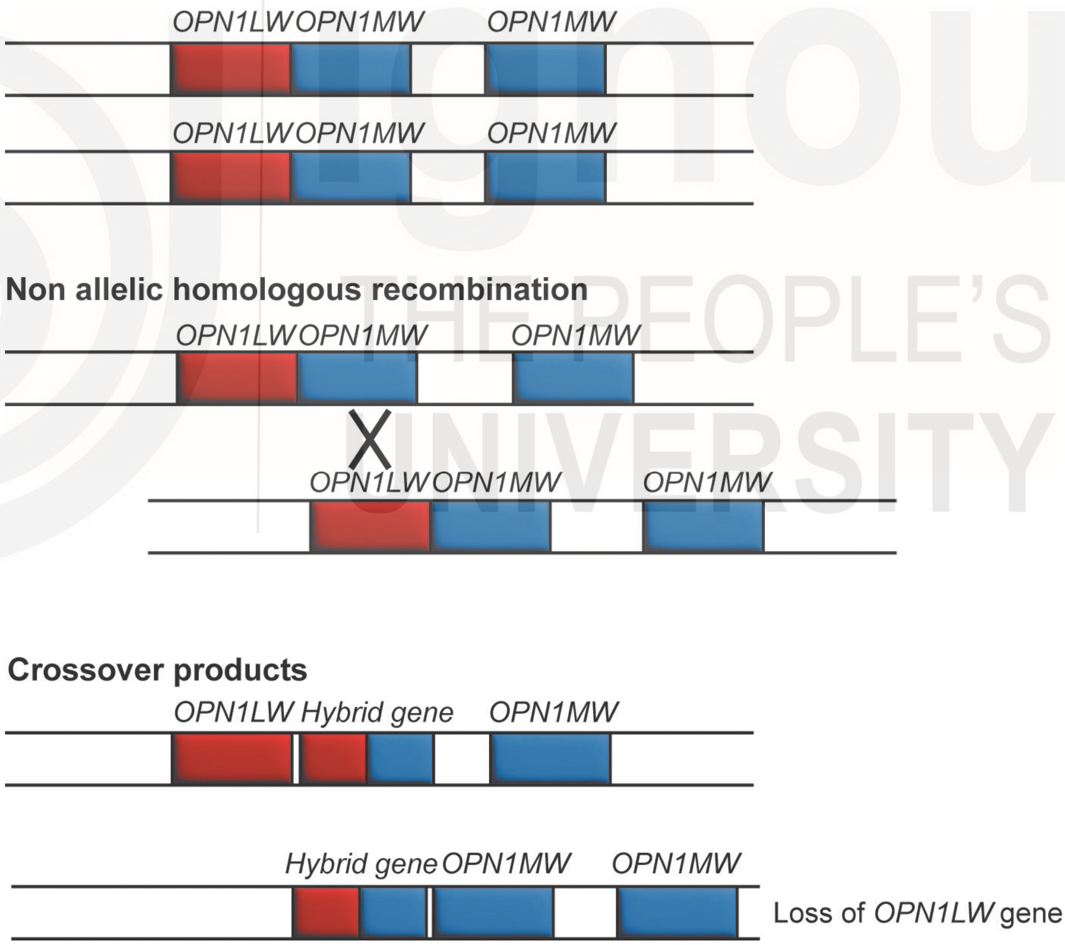

Fig. 3.7: Non allelic homologous recombination (NAHR) between the genes for red and green pigment leads to loss of gene for red pigment

Alpha-thalassemia is another disorder resulting from recombination between paralogous genes. NAHR between two highly homologous copies of α-globin genes (α1 and α2) separated by a distance of approximately 4 Kb on chromosome 16p13.3 results in the deletion of one functional copy, leading to α-thalassemia.

3.6 HORIZONTAL GENE TRANSFER

Horizontal gene transfer (HGT), also called **lateral gene transfer** is the transfer of genetic material between two organisms that are not in a parent-offspring relationship. It is mostly observed in bacteria and some of multicellular eukaryotes. For example, the moss *Physcomitrella patens* has acquired genes from various sources, including fungi, bacteria, viruses, and aquatic animals through HGT. HGT plays a significant role in the development of bacterial antibiotic resistance, pathogenicity and evolution. HGT was initially proposed to take place by three major mechanisms, namely, transformation, conjugation, and transduction. A number of other mechanisms such as cell fusion and gene transfer agents (GTAs) have been described more recently (Fig. 3.8).

a). **Transformation:** Transformation involves the uptake, integration, and expression of naked fragments of extracellular DNA from the environment by bacteria, and was first demonstrated by Frederick Griffith in 1928. Bacteria may use transformation to evade antibiotics, by exchanging antibiotic resistance genes. For example, Alexander and coworkers demonstrated the intra- and inter-species of streptomycin resistance between *Haemophilus influenzae, Haemophilus parainfluenzae,* and *Haemophilus suis*. In order for transformation to take place, DNA must be present in the extracellular environment, the recipient bacteria must be in a state of competence; and, the translocated DNA must be stabilized.

b). **Conjugation:** Conjugation is the transfer of genetic material from one bacterial cell (donor) to another (recipient) through cell surface pili which form a physical connection between the cells. Antibiotic resistance genes are usually located on plasmids which are transferred primarily through conjugation. The transfer of antibiotic-resistance genes occurs between bacteria in insects, soil, and water environments to various food and healthcare-associated pathogens, resulting in the development of multidrug resistance. Conjugation is restricted to bacteria with the exception of *Agrobacterium tumefaciens* which transfers its genetic material to plants, causing crown gall disease in the latter.

c). **Transduction:** During transduction, bacteriophages (bacterial viruses) are instrumental in moving genes from one cell to another. The transfer of antibiotic resistance genes by bacteriophages is well documented for various bacterial species, such as the transfer of tetracycline and gentamicin resistance between enterococci and the transfer of β-lactamase genes by bacteriophages in *Escherichia coli* and *Salmonella*. There are two types of transduction: generalized transduction in which a random fragment of DNA from a lysed donor cell is carried by the bacteriophage to a recipient cell; and, specialized transduction in which the genome of a temperate bacteriophage first integrates into the genome of the host cell (donor) by lysogeny, following which it is excised as a prophage carrying only the specific fragments of host cell genome flanking it, which are subsequently transferred to a recipient.

d). **Cell fusion:** Cell fusion involves the bidirectional transfer of genetic material between two different species through a bridge formed between the cells.

e). **Gene transfer agents:** Gene transfer agents (GTAs) are phage like particles which are produced by some bacteria and archaebacteria. They carry random fragments of the host DNA and mediate HGT.

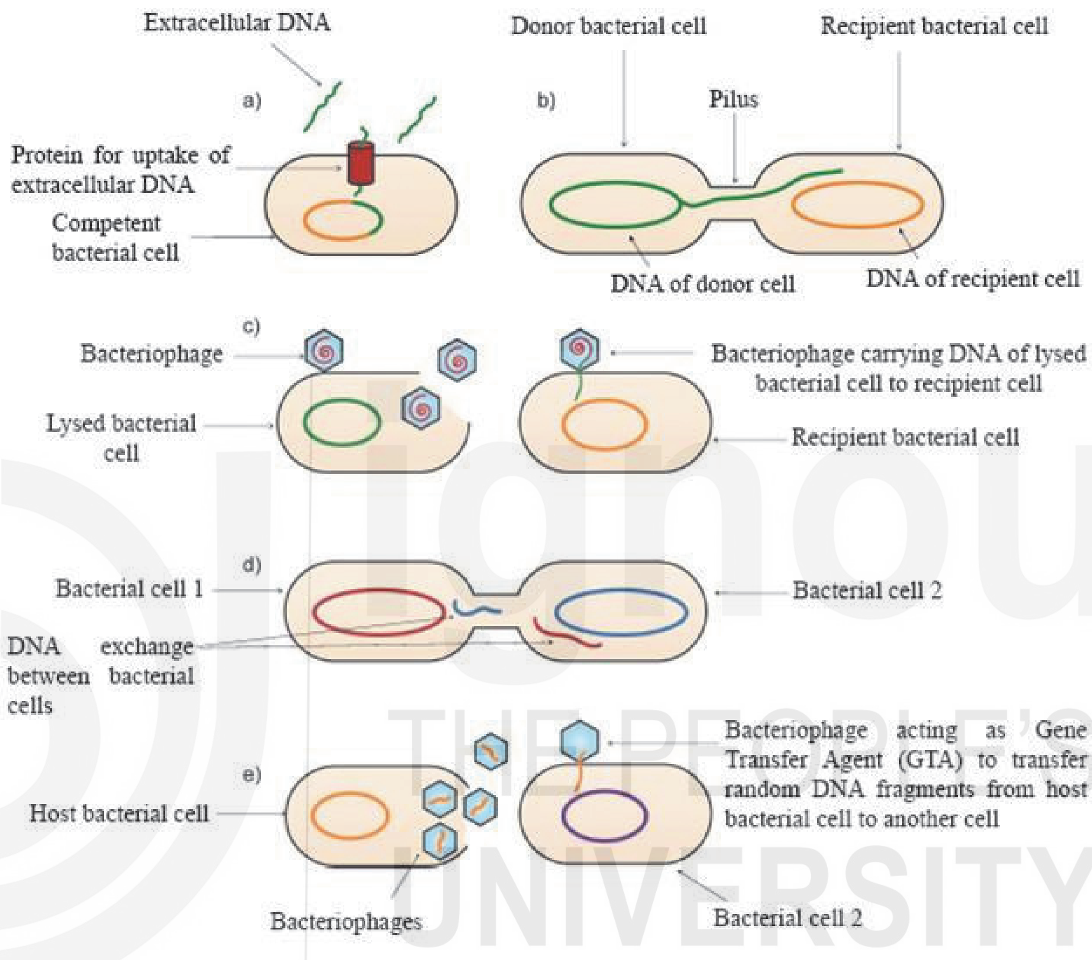

Fig. 3.8: Mechanisms of horizontal gene transfer (HGT) in bacteria by: (a) transformation (b) conjugation (c) generalized transduction (d) cell fusion (e) gene transfer agents (GTAs).

HGT plays an important role in a number of processes. These are:

(a) **Pathogenesis:** HGT plays a crucial role in the emergence of new pathogens by transferring the genes encoding virulence factors such as toxins, adhesions, capsules and invasion properties, etc. between different organisms. For example, virulent *Agrobacterium* species are Gram negative soil bacteria harbouring either a root-inducing Ri plasmid (*Agrobacterium rhizogenes*) or tumor inducing Ti plasmid (*Agrobacterium tumefaciens*). In addition, these plasmids also carry the *vir* genes which regulate the transfer of a 20 kb fragment of plasmid DNA from the *Agrobacterium* to the plant and *tra* genes which regulate the transfer of the entire Ti plasmid to from one *Agrobacterium* to another through conjugation. Agrobacteria are drawn to the site of plant wounds where they bind to target plant cells and activate the transcription of a number of plasmid encoded *vir* genes. A subset of *vir* genes encode proteins responsible for the processing of the plasmid DNA, leading to the excision and transfer of a fragment of plasmid DNA called T-DNA (transferred DNA) into the host plant cell. T-DNA becomes integrated into the host plant genomic DNA and its expression leads to the disruption of the normal balance of plant hormones, which results in abnormal root proliferation in the case of *A. rhizogenes* or crown gall tumor formation in case of *A. tumefaciens.*

(b) **Evolution:** HGT plays a vital role in evolution by introducing new genetic elements into a recipient. These genetic elements once incorporated into the genome of the recipient, provide additional material for the development of new functions and increase in complexity. Genetic elements like introns are commonly involved in HGT. These are associated with a variety of genes such as antibiotic resistance genes, toxin resistance genes, metabolic genes and virulence factor, which can confer an evolutionary advantage on the recipient.

(c) **Development of key metabolic processes:** HGT enables additive gene transfer, which is the process of the integration of novel genetic material from the donor into the genome of the host genetic recombination. An additive transfer from an organism with a gene that has an ortholog in the recipient genome leads to two similar copies of the gene in the recipient genome, as is observed after a gene duplication event. This can lead to the evolution of new capabilities which have been found to expand the metabolic capabilities of recipients and help them in adapting to unfavourable conditions such as highly saline, acidic or toxic environments. For example, the methylaspartate cycle is a novel pathway for carbon assimilation in haloarchaea which are found in water almost saturated with salts, such as the Dead Sea. This pathway combines genes from several bacterial metabolic pathways that were transferred to the ancestral species of haloarchaea from different bacterial donors.

SAQ 5

Fill in the blanks:

a) Red-green colour blindness is a genetic disorder resulting from .... between paralogous genes produced by segmental duplications.

b) .... is the transfer of genetic material between two organisms which are not in a parent-offspring relationship.

c) The three mechanisms by which horizontal gene transfer takes place are ...., ... and ....

d) Bacteria take up naked DNA from their environment by ....

e) During conjugation, genetic material is transferred between bacteria through a ....

f) .... mediate the transfer of genetic material during transduction.

g) Horizontal gene transfer leads to the emergence of new .... by transferring virulence genes between organisms.

h) .... is the process of the integration of novel genetic material from the donor into the genome of the host genetic recombination.

i) .... is a mechanism of horizontal gene transfer which involves bidirectional transfer of genetic material.

j) The .... genes regulate transfer of genetic material.

3.7 SUMMARY

- Gene duplication is a type of mutation that results in duplicate copies of genes, or paralogs, which are used as building blocks for evolutionary change. Functional redundancy results from duplicate genes that the organism does not need for its normal functioning.
- As a result, harmful mutations or splicing errors frequently render the duplicate genes non-functional, giving rise to non-functional genes known as pseudogenes.
- The two main globin protein types found in vertebrates are myoglobin (Mb) and haemoglobin (Hb). Whole-genome duplication and tandem duplication events combined to cause the diversification of vertebrate globins.
- A gene family is a collection of genes resulting from gene duplication of a common ancestral gene. Gene family members exhibit sequence homology and functional similarity, determined by how far apart they diverged from their ancestral genes during the evolutionary process.
- Computer-based gene prediction programs such as GENSCAN and GrailEXP are used to locate genes inside a genome sequence by determining a gene's unique characteristics such as the open reading frame (ORF) of a protein-coding gene, codon bias, exon-intron boundaries, upstream regulatory elements, and CpG islands.
- A pseudogene is a segment of DNA that structurally resembles a gene but is not capable of coding for a protein. Pseudogenes are most often derived from genes that have lost their protein-coding ability due to accumulated mutations that have occurred over the course of evolution.
- Segmental duplications (or low copy repeats) are copied from DNA at one site and move to other sites through evolution.
- The flow of genetic information across organisms is known as horizontal gene transfer (HGT), and it is this process that drives pathogen evolution. HGT occurs by well-understood genetic mechanisms *viz.* Transformation, Conjugation, Transduction, Cell fusion, and *via* gene transfer agents.
- HGT plays an important role in a number of processes such as pathogenesis, evolution and development of key metabolic processes.

3.8 TERMINAL QUESTIONS

1. What is gene duplication and its role in the formation of pseudogenes? Explain different types single gene duplication.

2. Explain the role of gene duplication in promoting the emergence of novel genes with specialised functions with reference to the evolution of the vertebrate globin gene family.

3. What are Gene families? Describe their classification into different categories on the basis of the extent of sequence identity.

4. How the location of genes in a genome can be identified by computational methods or experimental analysis? Explain.

5. Define Pseudogenes and explain their types.

6. Describe the mechanisms of horizontal gene transfer in bacteria with a help of suitable diagram.

7. Explain the important roles played by horizontal gene transfer in the following processes:

   (i) Pathogenesis

   (ii) Evolution

   (iii) Development of key metabolic processes

3.9 ANSWERS

Self-Assessment Questions

1. a) raw material, b) whole genome; single gene, c) genome; gene set, d) Lampreys, e) tandem, f) alpha; gamma, g) duplicons, h) retrogene.

2. a) ancestral, b) similar; transcription, c) motif, d) RNA metabolism, e) different, f) 16; high.

3. a) clusters, b) tandem, c) different, d) smell, e) dispersed, f) Fluoresecence *in situ* hybridization (FISH), g) CpG islands, h) exon-intron boundaries.

4. a) pseudogene, b) Non-processed, c) Processed; retrotransposition, d) siRNAs; lncRNAs, e) processed, f) oncosuppressive, g) cDNA.

5. a) non-allelic homologous recombination, b) Horizontal gene transfer, c) transformation, conjugation; transduction, d) transformation, e) pilus, f) Bacteriophages, g) pathogens, h) Additive gene transfer, i) Cell fusion, j) *tra*.

Terminal Questions

1. Refer to Section 3.2.

2. Refer to Subsection 3.2.1.

3. Refer to Section 3.3.
4. Refer to Subsection 3.3.1.
5. Refer to Section 3.4.
6. Refer to Section 3.6.
7. Refer to Section 3.6.

SUGGESTED READINGS

1. Burmeister, A.R. (2015). Horizontal gene transfer. Evolution, Medicine & Public Health 2015(1): 193-194.
2. Das, R., and Sharma, P. (2020) Disorders of abnormal haemoglobin. In: Clinical molecular medicine, Academic Press.
3. Panchy, M., Lehti-Shiu, M., and Shiu, S-H. (2016). Evolution of gene duplication in plants. Plant Physiology 171: 2294-2316.
4. Soucy, S.M., Huang., J., and Gogarten, J.P. (2015) Horizontal gene transfer: building the web of life. Nature Reviews Genetics 16: 472-482.
5. Storz, J.F. (2016). Gene duplication and evolutionary innovations in haemoglobin-oxygen transport. Physiology 31: 223-232.
6. Strachan, T., and Read, A.P. (2004) Human Molecular Genetics 3. Garland Science.
7. Tuteja, N., Vashisht, A., and Tuteja, R. (2008). Translation initiation factor 4A: A prototype member of DEAD-box protein family. Physiology and Molecular Biology of Plants 14: 101-107.
8. Zhao, K., Li, S., Yao, W., Zhou, B., Li, R., and Jiang, T. (2018). Characterization of the basic helix-loop-helix gene family and its tissue differential expression in response to salt stress in poplar. PeerJ 6: e4502.