Post-Graduate Degree Programme (CBCS)

in

ZOOLOGY

SEMESTER-IV

ELECTIVE THEORY PAPER

CYTOGEETICS AND MOLECULAR BIOLOGY

ZDSE(MJ)T-404

SELF LEARNING MATERIAL

DIRECTORATE OF OPEN AND DISTANCE LEARNING

UNIVERSITY OF KALYANI

KALYANI, NADIA

W.B. INDIA

Content Writer:

Dr. Subhabrata Ghosh, Assistant Professor, Department of Zoology (UG & PG), Asutosh College, Kolkata

---

May 2024

Directorate of Open and Distance Learning, University of Kalyani.

Published by the Directorate of Open and Distance Learning, University of Kalyani, Kalyani-741235, West Bengal.

Director's Message

ELECTIVE THEORY PAPER (ZDSE(MJ)T -404)

CYTOGENETICS AND MOLECULAR BIOLOGY

Module	Unit	Content	Credit	Page No.
ZDSE(MJ)T - 404 (CYTOGENETICS AND MOLECULAR BIOLOGY)	I	Genetic regulation during development: Gradients in early embryogenesis. in Drosophila. Cell fate & signaling pathways. Gap genes; segment polarity genes; axis formation; homeotic genes; homeo-domains; Hox genes & HOM-c genes	4	
	II	Infertility and its solutions		
	III	Teratogenesis, Stem cells and tissue engineering.		
	IV	Structural genomics: Genome sequencing, High resolution genome mapping radiation hybrid mapping		
	V	Physical mapping of genomes, FISH		

	VI	**Functional genomics: Study of gene interaction by the yeast two-hybrid system; Protein-DNA interaction, ChIP Assay,**		
	VII	**Mutagenesis, RNAi, Knockdown / knockout model**		
	VIII	**Comparative genomics: Homologous genes-Orthologous, paralogous; Sequence homology; Evolutionary relationships**		
	IX	**Allele frequencies and genotype frequencies: Hardy-Weinberg relationship**		
	X	**Haplotype frequencies and linkage disequilibrium, changing allele frequencies.**		
	XI	**Population structure and inbreeding**		
	XII	**Evolutionary genetics: Origin of species.**		
	XIII	**Phylogenetic trees, molecular evolution.**		
	XIV	**Comparative genomics of bacteria, organelles, and eukaryotes**		

Unit-I

Genetic regulation during development: Gradients in early embryogenesis in Drosophila. Cell fate & signalling pathways. Gap genes; segment polarity genes; axis formation; homeotic genes; homeo- domains; Hox genes & HOM-c genes

Objective: In this unit we will discuss about genetic control on development of Drosophila melanogaster. Different gene and their role in attern and organ formation will e discussed including Homeo Box genes.

Introduction:

It is the fly *Drosophila melanogaster* (Figure 21-23), more than any other organism, that has transformed our understanding of how genes govern the patterning of the body. The anatomy of *Drosophila* is more complex than that of *C. elegans*, with more than 100 times as many cells, and it shows more obvious parallels with our own body structure. Surprisingly, the fly has fewer genes than the worm—about 14,000 as compared with 19,000. On the other hand, it has almost twice as much noncoding DNA per gene (about 10,000 nucleotides on average, as compared with about 5000). The molecular construction kit has fewer types of parts, but the assembly instructions—as specified by the regulatory sequences in the non-coding DNA—seem to be more voluminous.

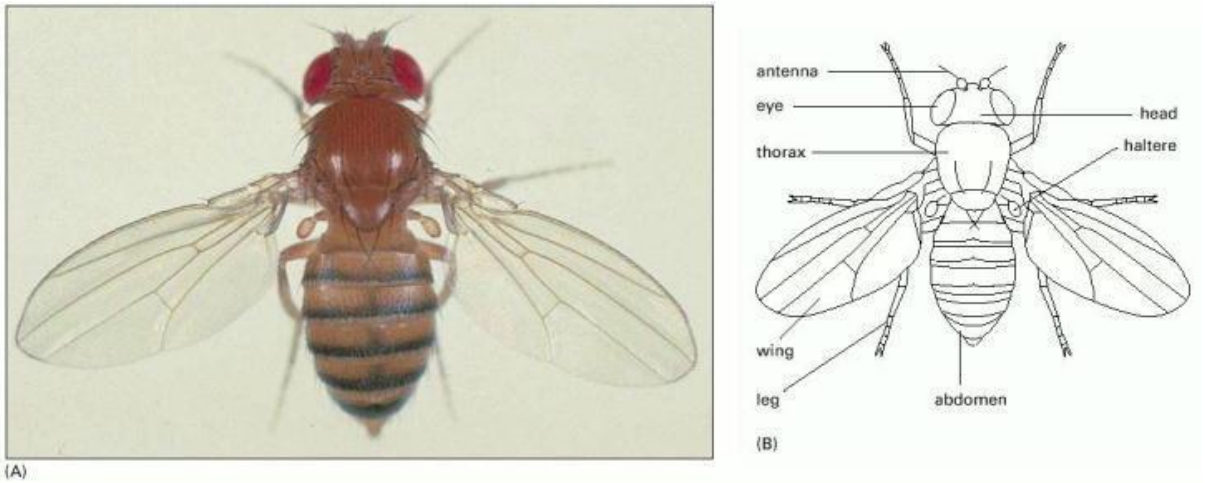


Figure 21-23 : Dorsal view of a normal adult fly. (A) Photograph. (B) Labelled drawing.

Decades of genetic study, culminating in massive systematic genetic screens, have yielded a catalogue of the developmental control genes that define the spatial pattern of cell types

and body structures of the fly, and molecular biology has given us the tools to watch these genes in action. By *in situ* hybridization using DNA or RNA probes on whole embryos, or by staining with labelled antibodies to reveal the distribution of specific proteins, one can observe directly how the internal states of the cells are defined by the sets of regulatory genes that they express at different times of development. Moreover, by analysing animals that are a patchwork of mutant and nonmutant cells, one can discover how each gene operates as part of a system to specify the organization of the body.

Most of the genes controlling the pattern of the body in *Drosophila* turn out to have close counterparts in higher animals, including ourselves. In fact, many of the basic devices for defining the body plan and patterning individual organs and tissues are astonishingly similar. Thus, quite surprisingly, the fly has provided the key to understanding the molecular genetics of our own development.

Flies, like nematode worms, are ideal for genetic studies: cheap to breed, easy to mutagenize, and rapid in their reproductive cycle. But there is a more fundamental reason why they have been so important for developmental geneticists. As emphasized earlier, as a result of gene duplications, vertebrate genomes often contain two or three homologous genes corresponding to a single gene in the fly. A mutation that disrupts one of these genes very often fails to reveal the gene's core function, because the other homologs share this function and remain active. In the fly, with its more economical gene set, this phenomenon of genetic redundancy is less prevalent. The phenotype of a single mutation in the fly therefore more often directly uncovers the function of the mutant gene.

The Insect Body Is Constructed as a Series of Segmental Units:

The timetable of *Drosophila* development, from egg to adult, is summarized in Figure 21-24. The period of *embryonic development* begins at fertilization and takes about a day, at the end of which the embryo hatches out of the egg shell to become a *larva.* The larva then passes through three stages, or *instars,* separated by moults in which it sheds its old coat of cuticle and lays down a larger one. At the end of the third instar it pupates. Inside the *pupa,* a radical remodelling of the body takes place—a process called *metamorphosis*. Eventually, about nine days after fertilization, an adult fly, or *imago,* emerges.

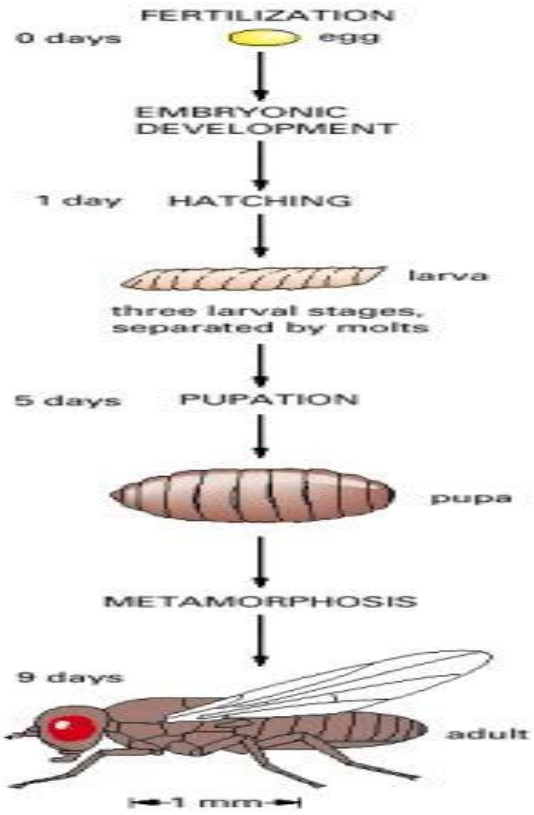


Figure 21-24: Synopsis of *Drosophila* development from egg to adult fly

The fly consists of a head, with mouth, eyes, and antennae, followed by three thoracic segments (numbered T1 to T3), and eight or nine abdominal segments (numbered A1 to A9). Each segment, although different from the others, is built according to a similar plan. Segment T1, for example, carries a pair of legs, T2 carries a pair of legs plus a pair of wings, and T3 carries a pair of legs plus a pair of halteres—small knob-shaped balancers important in flight, evolved from the second pair of wings that more primitive insects possess. The quasi-repetitive segmentation develops in the early embryo during the first few hours after fertilization (Figure 21-25), but it is more obvious in the larva (Figure 21-26), where the segments look more similar than in the adult. In the embryo it can be seen that the rudiments of the head, or at least the future adult mouth parts, are likewise segmental. At the two ends of the animal, however, there are highly specialized terminal structures that are not segmentally derived.

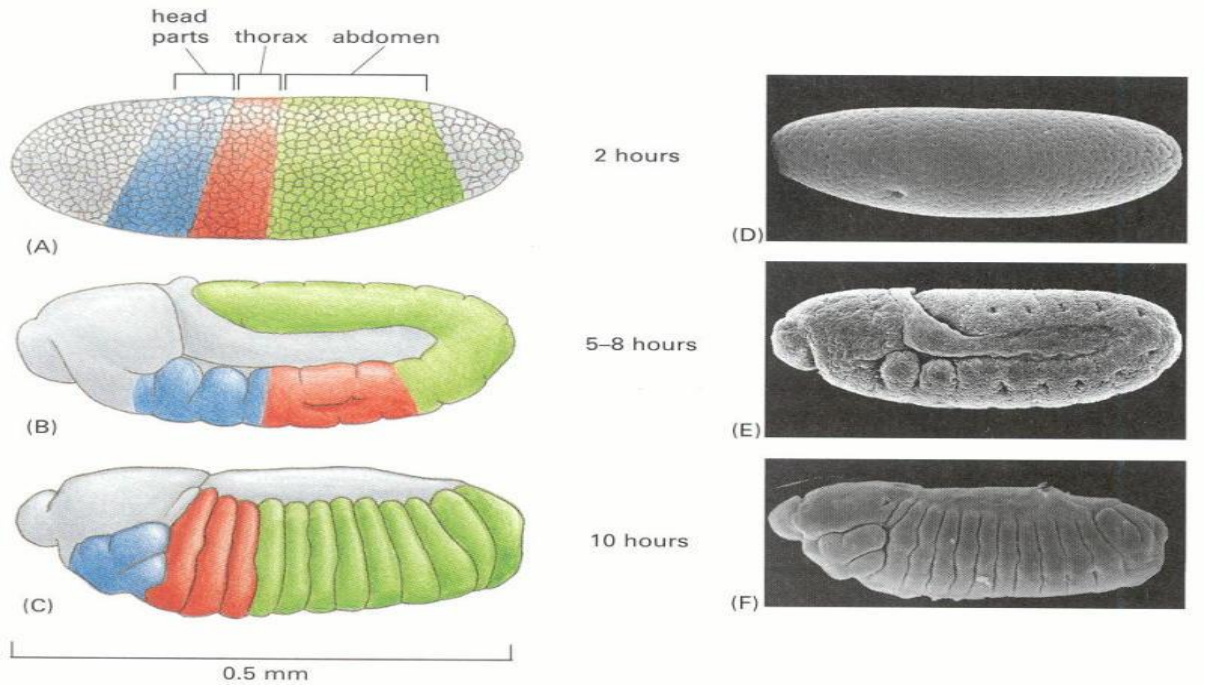


Figure 21-25 : The origins of the *Drosophila* body segments during embryonic development. **The embryos are seen in side view in drawings (A-C) and corresponding scanning** electron **micrographs (D-F). (A and D) At 2 hours the embryo is at the *syncytial blastoderm* stage (see** Figure 21-51**) and no segmentation is visible, although a fate map can be drawn showing the future segmented regions (*color* in A). (B and E) At 5–8 hours the embryo is at the *extended germ band* stage:** gastrulation **has occurred, segmentation has begun to be visible, and the segmented axis of the body has lengthened, curving back on itself at the tail end so as to fit into the** egg **shell. (C and F) At 10 hours the body axis has contracted and become straight again, and all the segments are clearly defined. The head structures, visible externally at this stage, will subsequently become tucked into the interior of the larva, to emerge again only when the larva goes through pupation to become an adult.**

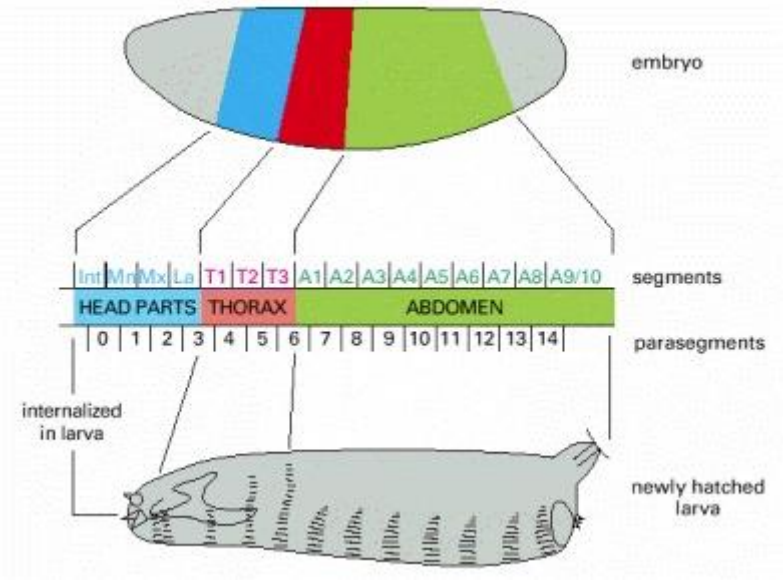


Figure 21-26 The segments of the *Drosophila* larva and their correspondence with regions of the blastoderm. The parts of the embryo that become organized into segments are shown in color. The two ends of the embryo, shaded *gray,* are not segmented and become tucked into the interior of the body to form the internal structures of the head and gut. (The future external, segmental structures of the adult head are also transiently tucked into the interior in the larva.) Segmentation in *Drosophila* can be described in terms of either segments or parasegments: the relationship is shown in the middle part of the figure. Parasegments often correspond more simply to patterns of gene expression**. The exact number of abdominal segments is debatable: eight are clearly defined, and a ninth is present vestigially in the larva, but absent in the adult. The boundaries between segments are traditionally defined by visible anatomical markers; but in discussing** gene expression **patterns it is often convenient to draw a different set of segmental boundaries, defining a series of segmental units called *parasegments*, half a segment out of register with the traditional divisions (see** Figure 21-26**).**

Drosophila Begins Its Development as a Syncytium:

The egg of *Drosophila* is about 0.5 mm long and 0.15 mm in diameter, with a clearly defined polarity. Like the eggs of other insects, but unlike vertebrates, it begins its development in an unusual way: a series of nuclear divisions without cell division creates a syncytium. The early nuclear divisions are synchronous and extremely rapid, occurring about every 8 minutes. The first nine divisions generate a cloud of nuclei, most of which migrate from the middle of the egg toward the surface, where they form a monolayer called the *syncytial*

blastoderm. After another four rounds of nuclear division, plasma membranes grow inward from the egg surface to enclose each nucleus, thereby converting the syncytial blastoderm into a *cellular blastoderm* consisting of about 6000 separate cells (Figure 21-27). About 15 of the nuclei populating the extreme posterior end of the egg are segregated into cells a few cycles earlier; these *pole cells* are the germ-line precursors (primordial germ cells) that will give rise to eggs or sperm.

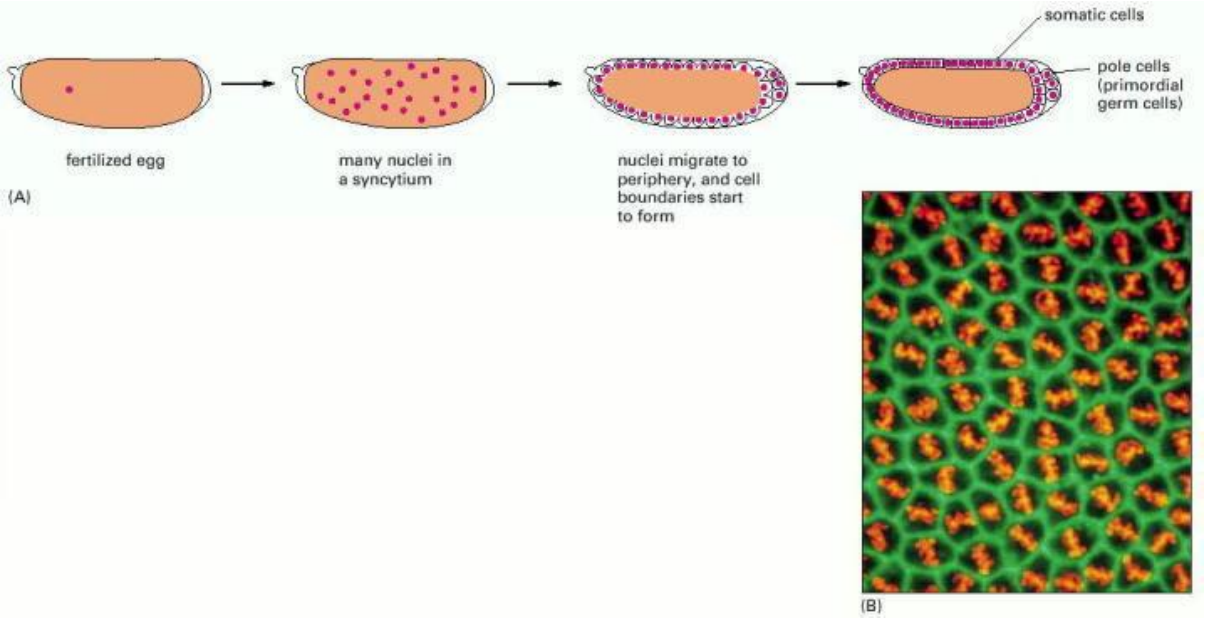


Figure 21-27 Development of the *Drosophila* egg from fertilization to the cellular blastoderm stage.(A) Schematic drawings. (B) Surface view—an optical-section photograph of blastoderm nuclei undergoing mitosis at the transition from the syncytial to the cellular blastoderm stage. Actin is stained *green,* chromosomes *orange.*

Up to the cellular blastoderm stage, development depends largely—although not exclusively—on stocks of maternal mRNA and protein that accumulated in the egg before fertilization. The frantic rate of DNA replication and nuclear division evidently gives little opportunity for transcription. After cellularization, cell division continues in a more conventional way, asynchronously and at a slower rate, and the rate of transcription increases dramatically. Gastrulation begins a little while before cellularization is complete, when parts of the sheet of cells forming the exterior of the embryo start to tuck into the interior to form the gut, the musculature, and associated internal tissues. A little later and in another region of the embryo, a separate set of cells move from the surface epithelium into the interior to form the central nervous system. By marking and following the cells through these various movements, one can draw a fate map for the monolayer of cells on the surface of the blastoderm (Figure 21-28).

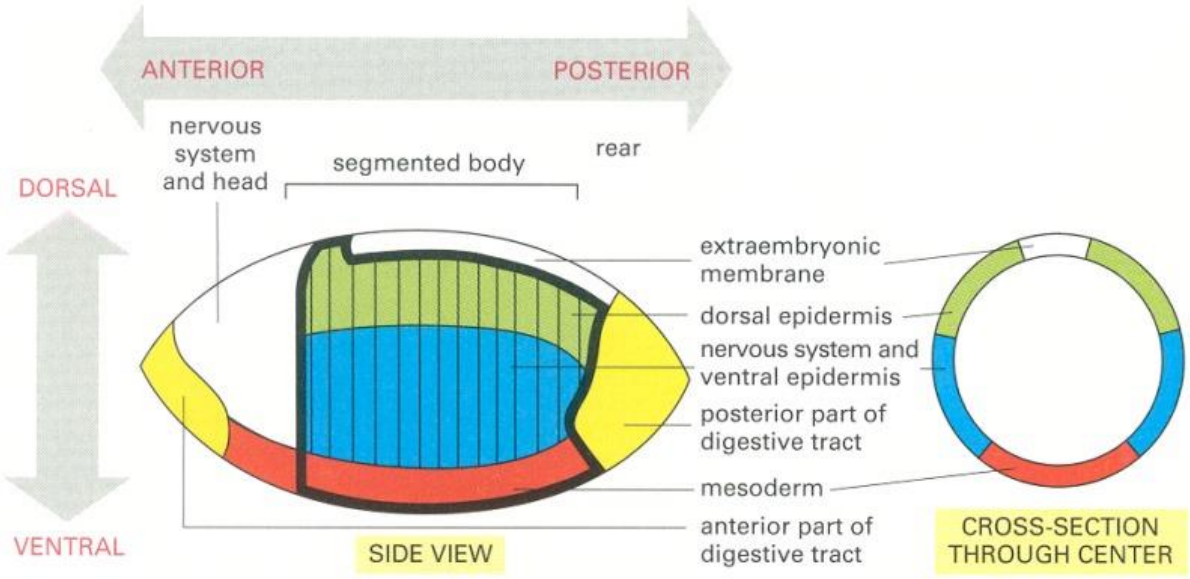


Figure 21-28 Fate map of a *Drosophila* embryo at the cellular blastoderm stage.The embryo is shown in side view and in cross section, displaying the relationship between the dorsoventral subdivision into future major tissue types and the anteroposterior pattern of future segments. A heavy line encloses the region that will form segmental structures. During gastrulation the cells along the ventral midline invaginate to form mesoderm, while the cells fated to form the gut invaginate near each end of the embryo.

As gastrulation nears completion, a series of indentations and bulges appear in the surface of the embryo, marking the subdivision of the body into segments along its anteroposterior axis (see Figure 21-25). Soon a fully segmented larva emerges, ready to start eating and growing. Within the body of the larva, small groups of cells remain apparently undifferentiated, forming structures called *imaginal discs*. These will grow as the larva grows, and eventually they will give rise to most of the structures of the adult body.

A head end and a tail end, a ventral (belly) side and a dorsal (back) side, a gut, a nervous system, a series of body segments—these are all features of the basic body plan that *Drosophila* shares with many other animals, including ourselves. We begin our account of the mechanisms of *Drosophila* development by considering how this body plan is set up.

Genetic Screens Define Groups of Genes Required for Specific Aspects of Early Patterning:

By carrying out a series of genetic screens based on saturation mutagenesis (see Chapter 8), it has been possible to amass a collection of *Drosophila* mutants that appears to include changes in a large proportion of the genes affecting development. Independent mutations in the same gene can be distinguished from mutations in separate genes by a complementation test, leading to a catalog of genes classified according to their mutant phenotypes. In such a catalog, a group of genes with very similar mutant phenotypes will often code for a set of proteins that work together to perform a particular function.

Sometimes the developmental functions revealed by mutant phenotypes are those that one would expect; sometimes they are a surprise. A large-scale genetic screen focusing on early *Drosophila* development revealed that the key genes fall into a relatively small set of functional classes defined by their mutant phenotypes. Some—*the egg-polarity genes* (Figure 21-29)—are required to define the anteroposterior and dorsoventral axes of the embryo and mark out its two ends for special fates, by mechanisms involving interactions between the oocyte and surrounding cells in the ovary. Others, the *gap genes*, are required in specific broad regions along the anteroposterior axis of the early embryo to allow their proper development. A third category, the *pair-rule genes*, are required, more surprisingly, for development of alternate body segments. A fourth category, the *segment polarity genes*, are responsible for organizing the anteroposterior pattern of each individual segment.

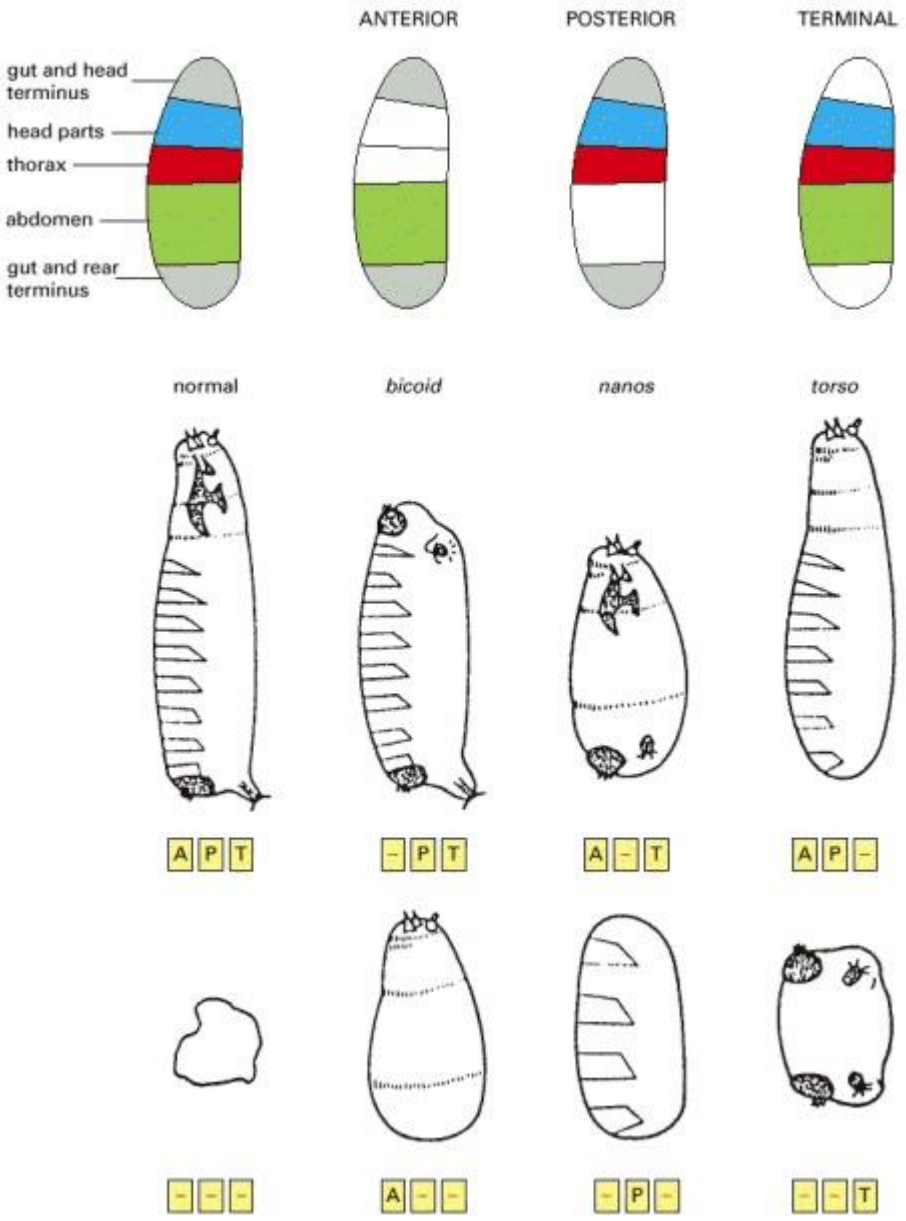


Figure 21-29 The domains of the anterior, posterior, and terminal systems of egg-polarity genes.**The *upper* diagrams show the fates of the different regions of the** egg**/early embryo and indicate (in *white)* the parts that fail to develop if the** anterior**,** posterior**, or terminal system is defective. The *middle* row shows schematically the appearance of a normal larva and of** mutant **larvae that are defective in a** gene **of the anterior system (for example, *bicoid),* of the posterior system (for example, *nanos),* or of the terminal system (for example, *torso).* The *bottom* row of drawings shows the appearances of larvae in which none or only one of the three gene systems is functional. The lettering beneath each larva specifies which systems are intact (A P T for a normal larva, -P T for a larva where the anterior system is defective but the**

posterior and terminal systems are intact, and so on). Inactivation of a particular gene **system causes loss of the corresponding set of body structures; the body parts that form correspond to the gene systems that remain functional. Note that larvae with a defect in the** anterior **system can still form terminal structures at their anterior end, but these are of the type normally found at the rear end of the body rather than the front of the head.**

The discovery of these four systems of genes and the subsequent analysis of their functions (an enterprise that still continues) was a famous tour-de-force of developmental genetics. It has had a revolutionary impact on all of developmental biology by showing the way toward a systematic, comprehensive account of the genetic control of embryonic development. In this section, we shall summarize only briefly the conclusions relating to the earliest phases of *Drosophila* development, because these are insect-specific; we dwell at greater length on the parts of the process that illustrate more general principles.

Interactions of the Oocyte With Its Surroundings Define the Axes of the Embryo: the Role of the Egg-Polarity Genes:

Surprisingly, the earliest steps of animal development are among the most variable, even within a phylum. A frog, a chicken, and a mammal, for example, even though they develop in similar ways later, make eggs that differ radically in size and structure, and they begin their development with different sequences of cell divisions and cell specialization events.

The style of early development that we have described for *C. elegans* is typical of many classes of animals. In contrast, the early development of *Drosophila* represents a rather extreme variant. The main axes of the future insect body are defined before fertilization by a complex exchange of signals between the unfertilized egg, or oocyte, and the follicle cells that surround it in the ovary (Figure 21-30). Then, in the syncytial phase following fertilization, an exceptional amount of patterning occurs in the array of rapidly dividing nuclei, before the first partitioning of the egg into separate cells. Here, there is no need for the usual forms of cell-cell communication involving transmembrane signaling; neighboring regions of the early *Drosophila* embryo can communicate by means of gene regulatory proteins and mRNA molecules that diffuse or are actively transported through the cytoplasm of the giant multinuclear cell.

n the stages before fertilization, the anteroposterior axis of the future embryo becomes defined by three systems of molecules that create landmarks in the oocyte (Figure 21-31). Following fertilization, each landmark serves as a beacon, providing a signal, in the form of

a morphogen gradient, that organizes the developmental process in its neighbourhood. Two of these signals are generated from localized deposits of specific mRNA molecules. The future anterior end of the embryo contains a high concentration of mRNA for a gene regulatory protein called Bicoid; this mRNA is translated to produce Bicoid protein, which diffuses away from its source to form a concentration gradient with its maximum at the anterior end of the egg. The future posterior end of the embryo contains a high concentration of mRNA for a regulator of translation called Nanos, which sets up a posterior gradient in the same way. The third signal is generated symmetrically at both ends of the egg, by local activation of a transmembrane tyrosine kinase receptor called Torso. The activated receptor exerts its effects over a shorter range, marking the sites of specialized terminal structures that will form at the head and tail ends of the future larva and also defining the rudiments of the future gut. The three sets of genes responsible for these localized determinants are referred to as the anterior, posterior, and **terminal** sets of **egg-polarity** genes.

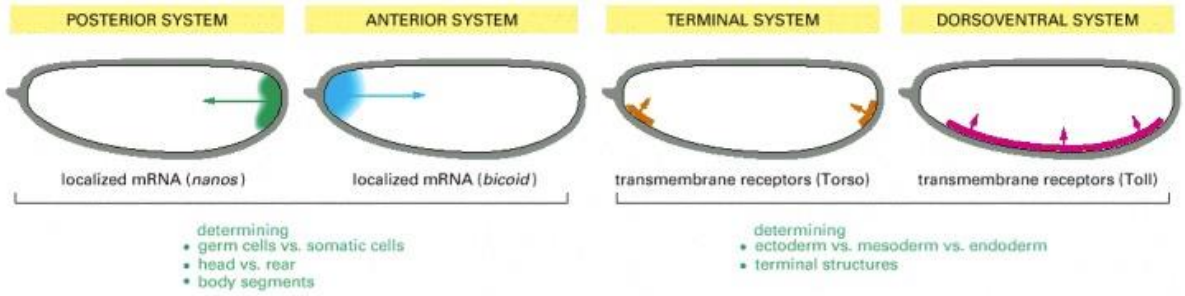


Figure 21-31 The organization of the four egg-polarity gradient systems.The receptors Toll and Torso are distributed all over the membrane; the coloring in the diagrams on the right indicates where they become activated by extracellular ligands.

A fourth landmark defines the dorsoventral axis (see Figure 21-31): a protein that is produced by follicle cells underneath the future ventral region of the embryo leads to localized activation of another transmembrane receptor, called Toll, in the oocyte membrane. The genes required for this function are called dorsoventral egg-polarity genes.

All the egg-polarity genes in these four classes are maternal-effect genes: it is the mother's genome, not the zygotic genome, that is critical. Thus, a fly whose chromosomes are mutant in both copies of the *bicoid* gene but who is born from a mother carrying one normal copy of *bicoid* develops perfectly normally, without any defects in the head pattern. However, if that daughter fly is a female no functional *bicoid* mRNA can be deposited into

the anterior part of her own eggs, and all of these will develop into headless embryos regardless of the father's genotype.

Each of the four egg-polarity signals—provided by Bicoid, Nanos, Torso, and Toll—exerts its effect by regulating (directly or indirectly) the expression of genes in the nuclei of the blastoderm. The use of these particular molecules to organize the egg is not a general feature of early animal development—indeed, only *Drosophila* and closely related insects possess a *bicoid* gene. And Toll has been coopted here for dorsoventral patterning; its more ancient and universal function is in the innate immune response.

Nevertheless, the egg-polarity system shows some highly conserved features. For example, the localization of *nanos* mRNA at one end of the egg is linked to, and dependent on, the localization of germ-cell determinants at that site, just as it is in *C. elegans*. Later in development, as the zygotic genome comes into play under the influence of the egg-polarity system, more similarities with other animal species become apparent. We shall use the dorsoventral system to illustrate this point.

The Dorsoventral Signalling Genes Create a Gradient of a Nuclear Gene Regulatory Protein:

Localized activation of the Toll receptor on the ventral side of the egg controls the distribution of Dorsal, a gene regulatory protein inside the egg. The Dorsal protein belongs to the same family as the NF-κB gene regulatory protein of vertebrates (discussed in Chapter 15). Its Toll-regulated activity, like that of NF-κB, depends on its translocation from the cytoplasm, where it is held in an inactive form, to the nucleus, where it regulates gene expression. In the newly laid egg, both the *dorsal* mRNA (detected by *in situ* hybridization) and the protein it encodes (detected with antibodies) are distributed uniformly in the cytoplasm. After the nuclei have migrated to the surface of the embryo to form the blastoderm, however, a remarkable redistribution of the Dorsal protein occurs: dorsally the protein remains in the cytoplasm, but ventrally it is concentrated in the nuclei, with a smooth gradient of nuclear localization between these two extremes (Figure 21-32). The signal transmitted by the Toll protein controls this redistribution of Dorsal through a signalling pathway that is essentially the same as the Toll-dependent pathway involved in innate immunity.

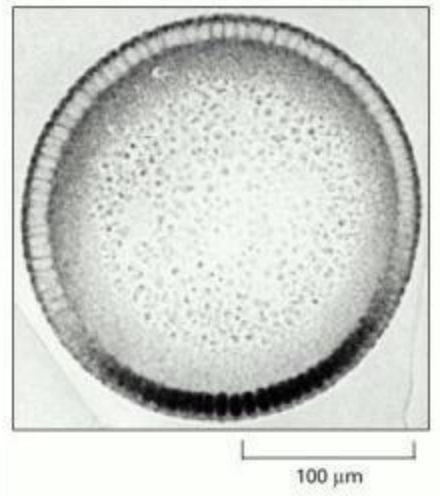


Figure 21-32 The concentration gradient of Dorsal protein in the nuclei of the blastoderm, as revealed by an antibody. Dorsally, the protein is present in the cytoplasm and absent from the nuclei; ventrally, it is depleted in the cytoplasm and concentrated in the nuclei.

Once inside the nucleus, the Dorsal protein turns on or off the expression of different sets of genes depending on its concentration. The expression of each responding gene depends on its regulatory DNA—specifically, on the number and affinity of the binding sites that this DNA contains for Dorsal and other regulatory proteins. In this way, the regulatory DNA can be said to *interpret* the positional signal provided by the Dorsal protein gradient, so as to define a dorsoventral series of territories—distinctive bands of cells that run the length of the embryo (Figure 21-33A). Most ventrally—where the concentration of Dorsal protein is highest—it switches on, for example, the expression of a gene called *twist* that is specific for mesoderm (Figure 21-34). Most dorsally, where the concentration of Dorsal protein is lowest, the cells switch on *decapentaplegic (dpp)*. And in an intermediate region, where the concentration of Dorsal protein is high enough to repress *dpp* but too low to activate *twist,* the cells switch on another set of genes, including one called *short gastrulation (sog).*

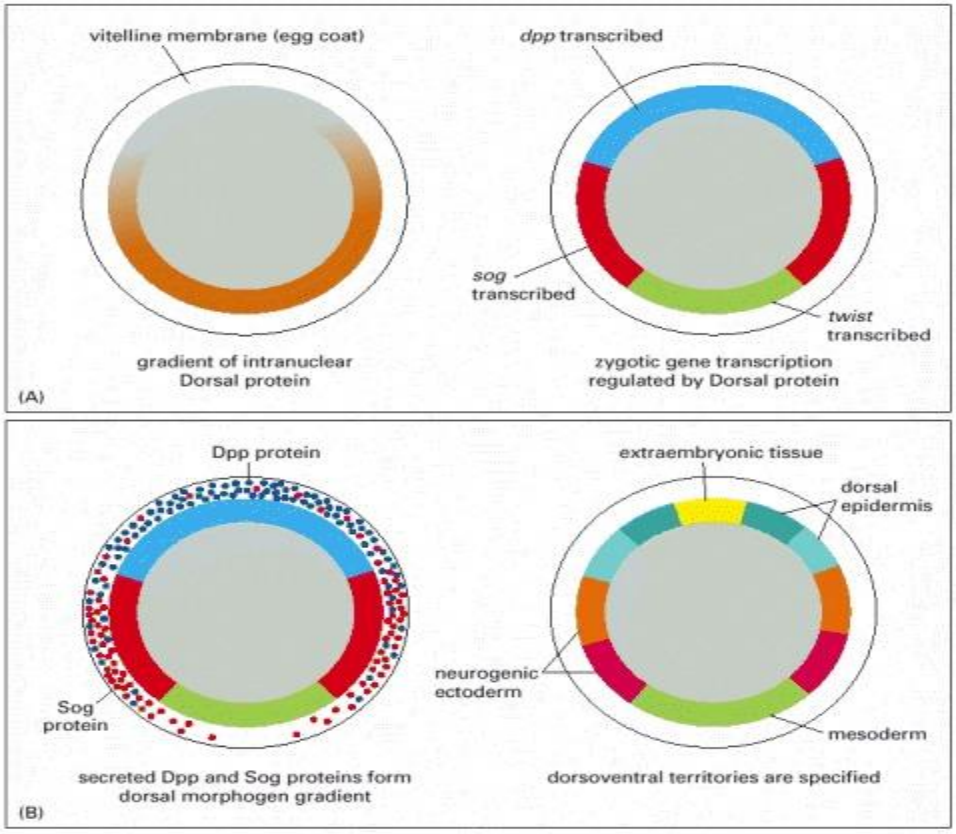


Figure 21-33 Morphogen gradients patterning the dorsoventral axis of the embryo.(A) The gradient of Dorsal protein defines three broad territories of gene expression, marked here by the expression of three representative genes—dpp, sog, and twist. (B) Slightly later, the cells expressing dpp and sog secrete, respectively, the signal proteins Dpp (a TGFβ family member) and Sog (an antagonist of Dpp). These two proteins diffuse and interact with one another (and with certain other factors) to set up a gradient of Dpp activity that guides a more detailed patterning process.

Dpp and Sog Set Up a Secondary Morphogen Gradient to Refine the Pattern of the Dorsal Part of the Embryo:

Products of the genes directly regulated by the Dorsal protein generate in turn more local signals that define finer subdivisions of the dorsoventral axis. These signals act after cellularization, and take the form of conventional extracellular signalling molecules. In particular, *dpp* codes for the secreted Dpp protein, which forms a local morphogen gradient in the dorsal part of the embryo. The gene *sog*, meanwhile, codes for another secreted protein that is produced in the neurogenic ectoderm and acts as an antagonist of Dpp. The

opposing diffusion gradients of these two proteins create a steep gradient of Dpp activity. The highest Dpp activity levels, in combination with certain other factors, cause development of the most dorsal tissue of all—extraembryonic membrane; intermediate levels cause development of dorsal ectoderm; and very low levels allow development of neurogenic ectoderm (Figure 21-33B).

Three Classes of Segmentation Genes Refine the Anterior- Posterior Maternal Pattern and Subdivide the Embryo:

After the initial gradients of Bicoid and Nanos are created to define the anteroposterior axis, the **segmentation genes** refine the pattern. Mutations in any one of the segmentation genes alter the number of segments or their basic internal organization without affecting the global polarity of the embryo. Segmentation genes are expressed by subsets of cells in the embryo, so their products are the first components that the embryo's own genome, rather than the maternal genome, contributes to embryonic development. They are therefore called *zygotic-effect genes* to distinguish them from the earlier maternal-effect genes.

The segmentation genes fall into three groups according to their mutant phenotypes and the stages at which they act (Figure 21-36). First come a set of at least six **gap genes**, whose products mark out coarse subdivisions of the embryo. Mutations in a gap gene eliminate one or more groups of adjacent segments, and mutations in different gap genes cause different but partially overlapping defects. In the mutant *Krüppel,* for example, the larva lacks eight segments, from T1 to A5 inclusive.

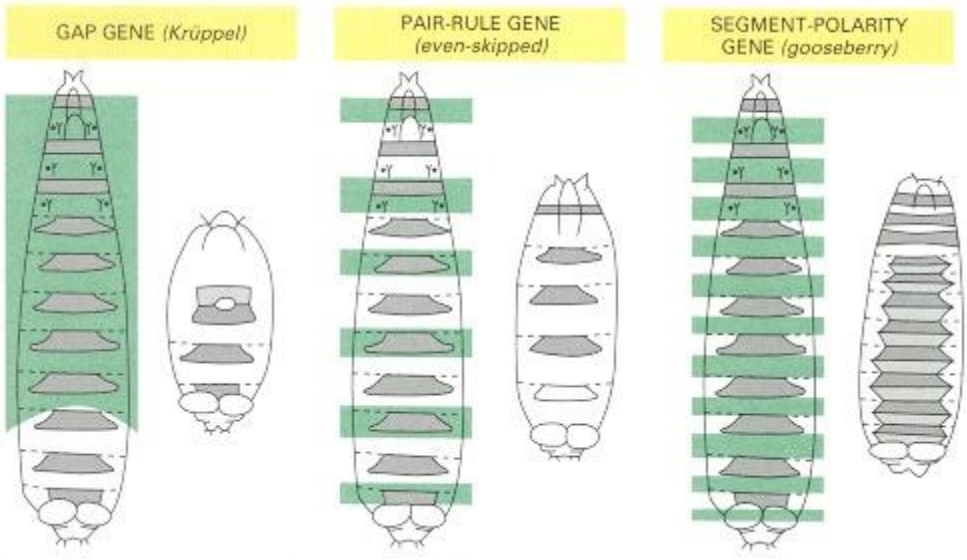


Figure 21-36 Examples of the phenotypes of mutations affecting the three types of segmentation genes. In each case the areas shaded in *green* on the normal

larva *(left)* are deleted in the mutant or are replaced by mirror-image duplicates of the unaffected regions. By convention, dominant mutations are written with an initial capital letter and recessive mutations are written with a lower-case letter. Several of the patterning mutations of *Drosophila* are classed as dominant because they have a perceptible effect on the phenotype of the heterozygote, even though the characteristic major, lethal effects are recessive—that is, visible only in the homozygote.

The next segmentation genes to act are a set of eight **pair-rule genes**. Mutations in these cause a series of deletions affecting alternate segments, leaving the embryo with only half as many segments as usual. While all the pair-rule mutants display this two-segment periodicity, they differ in the precise positioning of the deletions relative to the segmental or parasegmental borders. The pair-rule mutant *even-skipped (eve),* for example, which is discussed in Chapter 9, lacks the whole of each odd-numbered parasegment, while the pair-rule mutant *fushi tarazu (ftz)* lacks the whole of each even-numbered parasegment, and the pair-rule mutant *hairy* lacks a series of regions that are of similar width but out of register with the parasegmental units.

Finally, there are at least 10 **segment-polarity genes**. Mutations in these genes produce larvae with a normal number of segments but with a part of each segment deleted and replaced by a mirror-image duplicate of all or part of the rest of the segment. In *gooseberry* mutants, for example, the posterior half of each segment (that is, the anterior half of each parasegment) is replaced by an approximate mirror image of the adjacent anterior half-segment (see Figure 21-36).

We see later that, in parallel with the segmentation process, a further set of genes, the *homeotic selector genes,* serve to define and preserve the differences between one segment and the next.

The phenotypes of the various segmentation mutants suggest that the segmentation genes form a coordinated system that subdivides the embryo progressively into smaller and smaller domains along the anteroposterior axis, distinguished by different patterns of gene expression. Molecular genetics has helped to reveal how this system works.

The Localized Expression of Segmentation Genes Is Regulated by a Hierarchy of Positional Signals:

About three-quarters of the segmentation genes, including all of the gap genes and pair-rule genes, code for gene regulatory proteins. Their actions on one another and on other genes can therefore be observed by comparing gene expression in normal and mutant embryos. By using appropriate probes to detect the gene transcripts or their protein products, one can, in effect, take snapshots as genes switch on and off in

changing patterns. Repeating the process with mutants that lack a particular segmentation gene, one can begin to dissect the logic of the entire gene control system.

The products of the egg-polarity genes provide the global positional signals in the early embryo. These cause particular gap genes to be expressed in particular regions. The products of the gap genes then provide a second tier of positional signals that act more locally to regulate finer details of patterning through the expression of yet other genes, including the pair-rule genes (Figure 21-37). The pair-rule genes in turn collaborate with one another and with the gap genes to set up a regular periodic pattern of expression of segment-polarity genes, and the segment-polarity genes collaborate with one another to define the internal pattern of each individual segment. The strategy, therefore, is one of sequential induction (see Figure 21-15). By the end of the process, the global gradients produced by the egg-polarity genes have triggered the creation of a fine-grained pattern through a hierarchy of sequential, progressively more local, positional controls. Because the global positional signals that start the process do not have to directly specify fine details, the individual cell nuclei do not have to be governed with extreme precision by small differences in the concentration of these signals. Instead, at each step in the sequence, new signals come into play, providing substantial localized differences of concentration to define new details. Sequential induction is thus a robust strategy. It works reliably to produce fly embryos that all have the same pattern, despite the essential imprecision of biological control systems, and despite variations in conditions such as the temperature at which the fly develops.

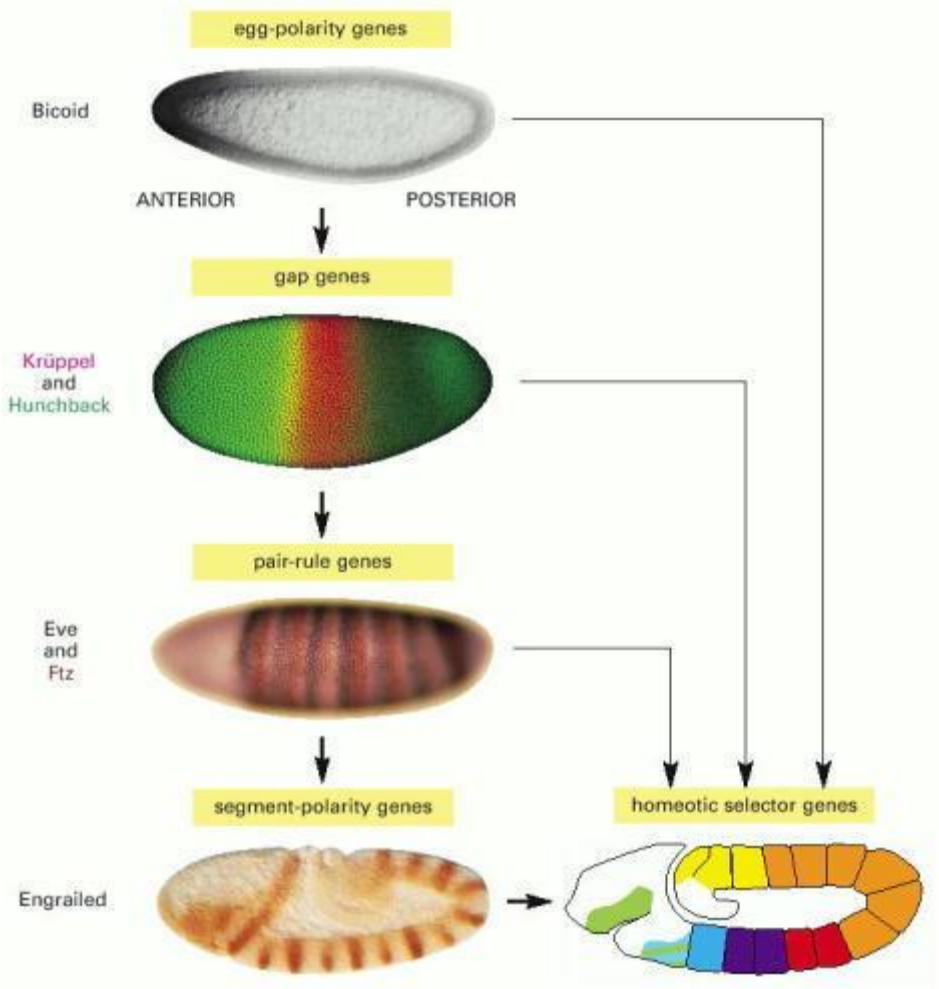


Figure 21-37 The regulatory hierarchy of egg-polarity, gap, segmentation, and homeotic selector genes. The photographs show expression patterns of representative examples of genes in each category, revealed by staining with antibodies against the protein products. The homeotic selector genes, discussed below, define the lasting differences between one segment and the next.

Egg-Polarity, Gap, and Pair-Rule Genes Create a Transient Pattern That Is Remembered by Other Genes:

Within the first few hours after fertilization, the gap genes and the pair-rule genes are activated one after another. Their mRNA products appear first in patterns that only approximate the final picture; then, within a short time—through a series of interactive adjustments—the fuzzy initial distribution of gene products resolves itself into a regular, crisply defined system of stripes (Figure 21-39). But this system itself is unstable and

transient. As the embryo proceeds through gastrulation and beyond, the regular segmental pattern of gap and pair-rule gene products disintegrates. Their actions, however, have stamped a permanent set of labels—positional values—on the cells of the blastoderm. These positional labels are recorded in the persistent activation of certain of the segment-polarity genes and of the homeotic selector genes, which serve to maintain the segmental organization of the larva and adult. The segment-polarity gene *engrailed* provides a good example. Its RNA transcripts are seen in the cellular blastoderm in a series of 14 bands, each approximately one cell wide, corresponding to the anteriormost portions of the future parasegments (Figure 21-40).

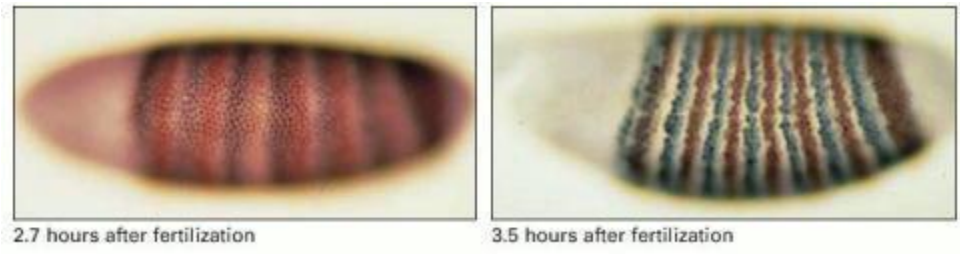


Figure 21-39 The formation of *ftz* and *eve* stripes in the *Drosophila* blastoderm ***ftz*** **and** ***eve*** **are both pair-rule genes. Their expression patterns (shown in** ***brown*** **for** ***ftz*** **and in** ***gray*** **for** ***eve*****) are at first blurred but rapidly resolve into sharply defined stripes.**

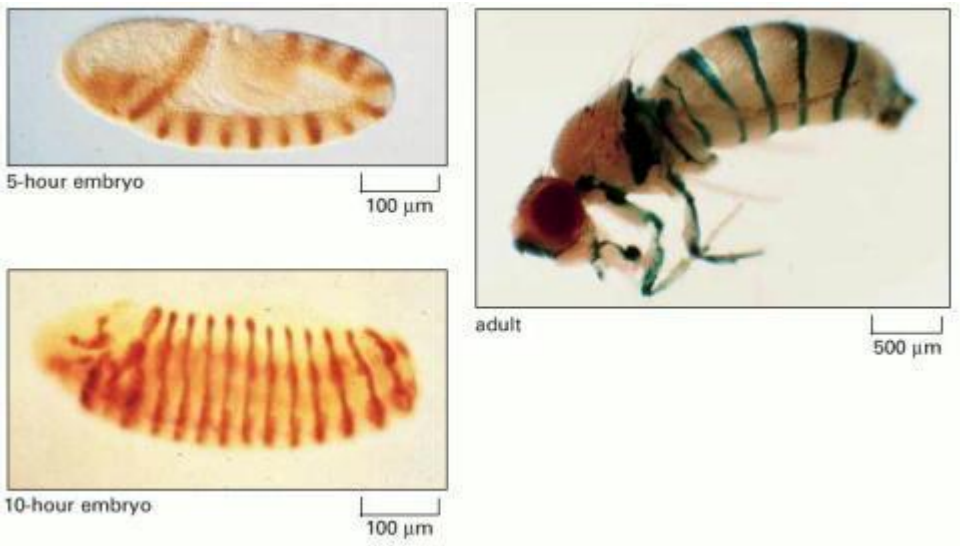


Figure 21-40 The pattern of expression of *engrailed,* a segment-polarity gene. The *engrailed* pattern is shown in a 5-hour embryo (at the extended germ-band stage), a 10-hour embryo, and an adult (whose wings have been removed in this preparation). The pattern is revealed by an antibody *(brown)* against the Engrailed protein **(for the 5- and 10-hour embryos) or (for the adult) by constructing a strain of *Drosophila* containing the control sequences of the *engrailed*** gene **coupled to the coding sequence of the reporter *LacZ*, whose product is detected histochemically through the *blue* product of a** reaction **that it catalyses. Note that the *engrailed* pattern, once established, is preserved throughout the animal's life.**

The segment-polarity genes are expressed in patterns that repeat from one parasegment to the next, and their bands of expression appear in a fixed relationship to the bands of expression of the pair-rule genes that help to induce them. However, the production of this pattern within each parasegment depends on interactions among the segment-polarity genes themselves. These interactions occur at stages when the blastoderm has already become fully partitioned into separate cells, so that cell-cell signalling of the usual sort has to come into play. A large subset of the segment-polarity genes code for components of two signal transduction pathways, the Wnt pathway and the Hedgehog pathway, including the secreted signal proteins Wingless (a Wnt family member) and Hedgehog. These are expressed in different bands of cells that serve as signalling centers within each parasegment, and they act to maintain and refine the expression of other segment-polarity genes. Moreover, although their initial expression is determined by the pair-rule genes, the two signalling proteins regulate one another's expression in a mutually supportive way, and they proceed to help trigger expression of genes such as *engrailed* in precisely the correct sites.

The *engrailed* expression pattern will persist throughout life, long after the signals that organized its production have disappeared (see Figure 21-40). This example illustrates not only the progressive subdivision of the embryo by means of more and more narrowly localized signals, but also the transition between the transient signalling events of early development and the later stable maintenance of developmental information.

Besides regulating the segment-polarity genes, the products of pair-rule genes collaborate with the products of gap genes to cause the precisely localized activation of a further set of spatial labels—the homeotic selector genes. It is the homeotic selector genes that permanently distinguish one parasegment from another. In the next section we examine these selector genes in detail and consider their role in cell memory.

Evolution of Homeotic genes; Homeo domains; Hox genes & HOM-c genes:

In 1894, William Bateson coined the word **homeosis** to describe the situation in which "something has been changed into the likeness of something else" (Lewis 1994). Bateson was attempting to provide evidence in support of Darwin's theory of evolution and homeotic variations seemed to Bateson to be the kind of dramatic changes that could explain how evolution occurred. E. B. Lewis (1994) concluded that homeosis provided a rich legacy: "Besides giving us the homeobox, it has opened up a completely new approach to the study of development. And over the past 15 years, it as led to the realization that the body plan of most animals, and presumably of plants as well, is controlled by a set of master regulatory genes, first identified by their homeotic mutations."

The periodic pattern of body segments generated by segmentation genes (gap genes, pair-rule genes, and segment-polarity genes) has to be converted into segments with wings, legs, and antennae (Figure 4.7). Thus, in insects, thoracic segment 2 is different from thoracic segment 3 and abdominal segment 2 will be different from the terminal abdominal segments, which typically have genital structures. This fine-tuning is determined by **homeotic** or **Hox genes**.

The homeobox consists of ≈180 bp that is translated into a 60-amino-acid domain. The sequences of the different homeoboxes are nearly identical and they mediate the binding of homeotic proteins to specific DNA sequences and thus regulate the expression of many downstream genes. It has been proposed that just two homeotic genes, *even-skipped*$^+$ and *Fushi tarazu*$^+$, directly control the expression of the majority of genes in the *Drosophila* genome (Mannervik 1999). Homeodomain proteins occur in all eukaryotes, where they perform important functions during development.

Since the first homeobox sequence was isolated from the *Antennapedia*$^+$ gene in late 1983, it has been used as a probe to identify and isolate previously unknown homeotic genes from *Drosophila*. Furthermore, because the homeobox is evolutionarily conserved, this *Drosophila* sequence was used as a probe to identify homeotic genes from other species, including humans (Gehring 1985).The products of Hox genes are Hox proteins. Hox proteins are a subset of the homeodomain-containing transcription factors, which are proteins that are capable of binding to specific nucleotide sequences on the DNA called enhancers where they either activate or repress genes. The same Hox protein can act as a repressor at one gene and an activator at another. The ability of Hox proteins to bind DNA is conferred by a part of the protein referred to as the homeodomain. The homeodomain is a 60-amino-acid-long DNA-binding domain (encoded by its corresponding 180-base-pair DNA sequence, the homeobox). This amino acid sequence folds into a "helix-

turn-helix" (i.e. homeodomain fold) motif that is stabilized by a third helix. The consensus polypeptide chain is:

```
       Helix 1       Helix 2       Helix 3/4
     ___________   _________   ________________
RRRKRTAYTRYQLLELEKEFLFNRYLTRRRRIELAHSLNLTERHIKIWFQNRRMKWKKEN
....|....|....|....|....|....|....|....|....|....|....|
      10      20      30      40      50      60
```

In Drosophila, like all insects, has eight Hox genes. These are clustered into two complexes, both of which are located on chromosome 3. The Antennapedia complex (not to be confused with the *Antp* gene) consists of five genes: labial (*lab*), proboscipedia (*pb*), deformed (*Dfd*), sex combs reduced (*Scr*), and Antennapedia (*Antp*). The Bithorax complex, named after the Ultrabithorax gene, consists of the remaining three genes: Ultrabithorax (*Ubx*), abdominal-A (*abd-A*) and abdominal-B (*abd-B*).

a. Labial

The *lab* gene is the most anteriorly expressed gene. It is expressed in the head, primarily in the intercalary segment (an appendageless segment between the antenna and mandible), and also in the midgut. Loss of function of *lab* results in the failure of the *Drosophila* embryo to internalize the mouth and head structures that initially develop on the outside of its body (a process called head involution). Failure of head involution disrupts or deletes the salivary glands and pharynx. The *lab* gene was initially so named because it disrupted the labial appendage; however, the lab gene is not expressed in the labial segment, and the labial appendage phenotype is likely a result of the broad disorganization resulting from the failure of head involution.

b. Proboscipedia

The *pb* gene is responsible for the formation of the labial and maxillary palps. Some evidence shows *pb*interacts with *Scr*.

c. Deformed

The *Dfd* gene is responsible for the formation of the maxillary and mandibular segments in the larval head.The mutant phenotypes of *Dfd* are similar to those of labial. Loss of function of *Dfd* in the embryo results in a failure of head involution (see labial gene), with a loss of larval head structures. Mutations in the adult have either deletions of parts of the head or transformations of head to thoracic identity.

d. Sex combs reduced

The *Scr* gene is responsible for cephalic and thoracic development in *Drosophila* embryo and adult.

e. Antennapedia

The second thoracic segment, or T2, develops a pair of legs and a pair of wings. The *Antp* gene specifies this identity by promoting leg formation and allowing (but not directly activating) wing formation. A dominant *Antp* mutation, caused by a chromosomal inversion, causes *Antp* to be expressed in the antennal imaginal disc, so that, instead of forming an antenna, the disc makes a leg, resulting in a leg coming out of the fly's head.

Figure: Wild type (left), Antennapedia mutant (right)

Ultrabithorax:

The third thoracic segment, or T3, bears a pair of legs and a pair of halteres (highly reduced wings that function in balancing during flight). *Ubx*patterns T3 largely by repressing genes involved in wing formation. The wing blade is composed of two layers of cells that adhere tightly to one another, and are supplied with nutrient by several wing veins. One of the many genes that *Ubx* represses is blistered, which activates proteins involved in cell-cell adhesion, and spalt, which patterns the placement of wing veins. In *Ubx* loss-of-function mutants, *Ubx* no longer represses wing genes, and the halteres develop as a second pair of wings, resulting in the famous four-winged flies. When *Ubx* is misexpressed in the second thoracic segment, such as occurs in flies with the "Cbx" enhancer mutation, it represses wing genes, and the wings develop as halteres, resulting in a four-haltered fly.

Abdominal-A

In *Drosophila*, *abd-A* is expressed along most of the abdomen, from abdominal segments 1 (A1) to A8. Expression of *abd-A* is necessary to specify the identity of most of the abdominal segments. A major function of *abd-A* in insects is to repress limb formation. In *abd-A* loss-of-function mutants, abdominal segments A2 through A8 are transformed into an identity more like A1. When *abd-A* is ectopically expressed throughout the embryo, all segments anterior of A4 are transformed to an A4-like abdominal identity.[7] The *abd-A gene* also affects the pattern of cuticle generation in the ectoderm, and pattern of muscle generation in the mesoderm.

Abdominal-B

Gene *abd-B* is transcribed in two different forms, a regulatory protein, and a morphogenic protein. Regulatory *abd-B* suppress embryonic ventral epidermal structures in the eighth and ninth segments of the *Drosophila* abdomen. Both the regulatory protein and the morphogenic protein are involved in the development of the tail segment.

Classification of Hox proteins:

Proteins with a high degree of sequence similarity are also generally assumed to exhibit a high degree of functional similarity, i.e. Hox proteins with identical homeodomains are assumed to have identical DNA-binding properties (unless additional sequences are known to influence DNA-binding). To identify the set of proteins between two different species that are most likely to be most similar in function, classification schemes are used. For Hox proteins, three different classification schemes exist: phylogenetic inference based, synteny-based, and sequence similarity-based. The three classification schemes provide conflicting information for Hox proteins expressed in the middle of the body axis (*Hox6-8*and *Antp, Ubx* and *abd-A*). A combined approach used phylogenetic inference-based information of the different species and plotted the protein sequence types onto the phylogenetic tree of the species. The approach identified the proteins that best represent ancestral forms (*Hox7* and *Antp*) and the proteins that represent new, derived versions (or were lost in an ancestor and are now missing in numerous species).

Genes regulated by Hox proteins:

Hox genes act at many levels within developmental gene hierarchies: at the "executive" level they regulate genes that in turn regulate large networks of other genes (like the gene pathway that forms an appendage). They also directly regulate what are called realisator genes or effector genes that act at the bottom of such hierarchies to ultimately form the tissues, structures, and organs of each segment. Segmentation involves such processes as morphogenesis (differentiation of precursor cells into their terminal specialized cells), the tight association of groups of cells with similar fates, the sculpting of structures and

segment boundaries via programmed cell death, and the movement of cells from where they are first born to where they will ultimately function, so it is not surprising that the target genes of Hox genes promote cell division, cell adhesion, apoptosis, and cell migration.

Examples of targets

Organism	Target gene	Normal function of target gene	Regulated by
Drosophila	distal-less	activates gene pathway for limb formation	ULTRABITHORAX[14] (represses distal-less)
	distal-less	activates gene pathway for limb formation	ABDOMINAL-A[14] (represses distal-less)
	decapentaplegic	triggers cell shape changes in the gut that are required for normal visceral morphology	ULTRABITHORAX[15] (activates decapentaplegic)
	reaper	Apoptosis: localized cell death creates the segmental boundary between the maxilla and mandible of the head	DEFORMED[16] (activates reaper)
	decapentaplegic	prevents the above cell changes in more posterior positions	ABDOMINAL-B[15] (represses decapentaplegic)

Enhancer sequences bound by homeodomains:

The DNA sequence bound by the homeodomain protein contains the nucleotide sequence TAAT, with the 5' terminal T being the most important for binding. This sequence is conserved in nearly all sites recognized by homeodomains, and probably distinguishes such locations as DNA binding sites. The base pairs following this initial sequence are used to distinguish between homeodomain proteins, all of which have similar recognition sites. For instance, the nucleotide following the TAAT sequence is recognized by the amino acid at position 9 of the homeodomain protein. In the maternal protein Bicoid, this position is occupied by lysine, which recognizes and binds to the nucleotide guanine. In Antennapedia, this position is occupied by glutamine, which recognizes and binds to adenine. If the lysine

in Bicoid is replaced by glutamine, the resulting protein will recognize Antennapedia-binding enhancer sites.
However, all homeodomain-containing transcription factors bind essentially the same DNA sequence. The sequence bound by the homeodomain of a Hox protein is only six nucleotides long, and such a short sequence would be found at random many times throughout the genome, far more than the number of actual functional sites. Especially for Hox proteins, which produce such dramatic changes in morphology when misexpressed, this raises the question of how each transcription factor can produce such specific and different outcomes if they all bind the same sequence. One mechanism that introduces greater DNA sequence specificity to Hox proteins is to bind protein cofactors. Two such Hox cofactors are Extradenticle (Exd) and Homothorax (Hth). Exd and Hth bind to Hox proteins and appear to induce conformational changes in the Hox protein that increase its specificity.

Regulation of Hox genes:

Just as Hox genes regulate realisator genes, they are in turn regulated themselves by gap genes and pair-rule genes, which are in their turn regulated by maternally-supplied mRNA. This results in a transcription factor cascade: maternal factors activate gap or pair-rule genes; gap and pair-rule genes activate Hox genes; then, finally, Hox genes activate realisator genes that cause the segments in the developing embryo to differentiate. Regulation is achieved via protein concentration gradients, called morphogenic fields. For example, high concentrations of one maternal protein and low concentrations of others will turn on a specific set of gap or pair-rule genes. In flies, stripe 2 in the embryo is activated by the maternal proteins Bicoid and Hunchback, but repressed by the gap proteins Giant and Kruppel. Thus, stripe 2 will only form wherever there is Bicoid and Hunchback, but *not* where there is Giant and Kruppel.

MicroRNA strands located in Hox clusters have been shown to inhibit more anterior hox genes ("posterior prevalence phenomenon"), possibly to better fine tune its expression pattern. Non-coding RNA (ncRNA) has been shown to be abundant in Hox clusters. In humans, 231 ncRNA may be present. One of these, HOTAIR, silences in trans (it is transcribed from the HOXC cluster and inhibits late HOXD genes) by binding to Polycomb-group proteins (PRC2). The chromatin structure is essential for transcription but it also requires the cluster to loop out of the chromosome territory.

Homeotic mutants may have segments that are transformed dramatically. For example, antennal segments may be transformed into leg-like structures, and metathoracic segments with halteres may be transformed into mesothoracic segments with a set of wings. The four-winged *D. melanogaster* is the result of combining three separate mutated genes in one fly! Normally, of course, a pair of wings is found on the second thoracic segment and a

pair of balancing organs, called halteres, is on the third. However, this fly has two essentially normal second thoraces (and no third thoracic segment) because the combined effect of the three mutations is to transform the third thoracic segment into the second without affecting any other parts of the fly.

Lewis, E., (1978) proposed a combinatorial model that assumes each insect segment is specified by a unique combination of homeotic genes that are expressed in that particular segment. Thus, the fewest number of homeotic genes would be required in thoracic segment 2, which would be the prototypical segment, and progressively more genes would be active in the more-posterior segments. Although this model has been modified, it provided a useful conceptual framework for investigating *Drosophila* development.

Homeotic genes have some unusual characteristics. First, several homeotic genes seem to be very large relative to most other genes in *Drosophila*. For example, the *Antennapedia*$^{+}$ primary gene transcript is ≈100 kb long and the *Ultrabithorax*$^{+}$ transcript is ≈75 kb. However, after the introns are spliced out, the remaining sequences are only a few kilobases. Many of the exons in homeotic genes seem to encode protein domains with distinct structural or enzymatic functions. As a result, alternative splicing patterns in large genes, such as the *Antennapedia*$^{+}$ and *bithorax*$^{+}$gene complexes, may allow organisms to adapt one basic protein structure to different, but related, developmental uses. By adding or subtracting functional protein domains encoded by optional exons, the structural and enzymatic properties of the homeotic gene product can be modified and the ability of the protein to interact with other cellular components can be altered as development proceeds.

Homeotic Gene Cluster (HOM-c)

Drosophila has two homeotic gene clusters: the ANT-C (*Antennapedia* complex) is responsible for segmental identity in the head and anterior thorax and the BX-C (*Bithorax* complex) which is responsible for segmental identity in the posterior thorax and abdomen.

These two gene clusters found on chromosome 3 are found in one cluster in more primitive insects, called the HOM-C (homeotic gene complex). The general case is that there is only one homeotic gene cluster in insects and in the evolution of *Drosophila* it was separated into two clusters.

Summary:

The fly *Drosophila* has been the foremost model organism for study of the genetics of animal development. Like other insects, it begins its development with a series of nuclear divisions generating a syncytium, and a large amount of early patterning occurs in this single giant multinucleate cell. The pattern originates with asymmetry in the egg, organized both by localized deposits of mRNA inside the egg and by signals from the follicle cells around it. Positional information in the multinucleate embryo is supplied by four intracellular gradients that are set up by the products of four groups of maternal-effect genes called egg-polarity genes. These control four distinctions fundamental to the body plan of animals: dorsal versus ventral, endoderm versus mesoderm and ectoderm, germ cells versus somatic cells, and head versus rear.

The egg-polarity genes operate by setting up graded distributions of gene regulatory proteins in the egg and early embryo. The gradients along the anteroposterior axis initiate the orderly expression of gap genes, pair-rule genes, segment-polarity genes, and homeotic selector genes. These, through a hierarchy of interactions, become expressed in some regions of the embryo and not others, progressively subdividing the blastoderm into a regular series of repeating modular units called segments. The complex patterns of gene expression reflect the modular organization of the regulatory *DNA*, with separate enhancers of an individual gene responsible for separate parts of its expression pattern.

The segment-polarity genes come into play toward the end of the segmentation process, soon after the syncytium has become partitioned into separate cells, and they control the internal patterning of each segment through cell-cell signalling via the Wnt (Wingless) and Hedgehog pathways. This leads to persistent localized activation of genes such as engrailed, giving cells a remembered record of their anteroposterior address within the segment. Meanwhile, a new cell-cell signalling gradient is also set up along the dorsoventral axis, with the TGFβ family member Decapentaplegic (Dpp) and its antagonist, Short gastrulation, acting as the morphogens. This gradient helps to refine the assignment of different characters to cells at different dorsoventral levels. Homologous proteins are also known to control the patterning of the ventrodorsal axis in vertebrates.

Probable questions:

1. Describe different stages of life cycle of Drosophila.
2. How dorsoventral axis is determined in Drosophila ?
3. How anterior posterior side is determined in Drosophila?
4. State the function of dorsal protein.
5. What is the role of pair rule gene in Drosophila development?
6. What is the role of gap gene in Drosophila development?
7. What is the role of segment polarity gene in Drosophila development?
8. What is Hox gene? What is Homeo domain?
9. What is the role of Dpp and Sog protein in determination of pattern of the dorsal part of the Drosophila embryo ?
10. How Hox genes are regulated ?
11. What is Homeotic gene cluster?

Suggested Readings:

1. Developmental biology. Gilbert, S. F., & Barresi, M. J. F. (2016).

2. Molecular Biology of the Cell – by Bruce Alberts

3. Molecular Cell Biology by Lodish, Fourth Edition.

UNIT-II
Infertility and its solutions

Objective: In this unit we will discuss causes of male and female infertility and also discuss about the different treatments by which we can solve infertility problems.

Infertility:

Infertility is a term that is defined as the inability of couples to become pregnant after one year of regular intercourse without contraception, or 6 months if a woman is 35 years or older. Infertility can occur in both females and males, and there are many causes. In order to get pregnant, a woman's body must go through three main steps: ovulation, fertilization, and implantation. In ovulation, an egg is released from one of her ovaries, which then travels down to one of the fallopian tubes and into the uterus. The egg must then be fertilized by the sperm during the process of fertilization. Finally, in implantation, the fertilized egg implants itself in the uterine wall. When there are problems in any of these steps, infertility can be diagnosed.

Types of Infertility: Infertility can be primary or secondary.

Primary infertility is when a couple has not conceived after trying for at least 12 months without using birth control

Secondary infertility is when they have previously conceived but are no longer able to.

Causes of Infertility:

The causes of infertility may be physical, congenital, disease, drug, immunological or even psychological. In India, when a couple is childless, the female is usually blamed. But more often, the males are detected to be responsible. However, now, specialized health care units known as infertility clinics are available. They could identify the cause of infertility and take up treatment to remove the disorder.

A. Male Infertility:

Infertility in males may be due to the following causes:

a. Azoospermia:

Absence of sperms in the semen is known as azoospermia. This may occur because of lack of sperm production or because of blocked tubes which does not permit the sperms to appear in the semen. Blockage can occur due to an infection or injury.

Failure of the ejaculation mechanism is another possible reason of azoospermia. Failure to produce sperms may result because of injury to the testes or as a result of infection such as mumps virus or due to hormonal reasons (Fig. 3).

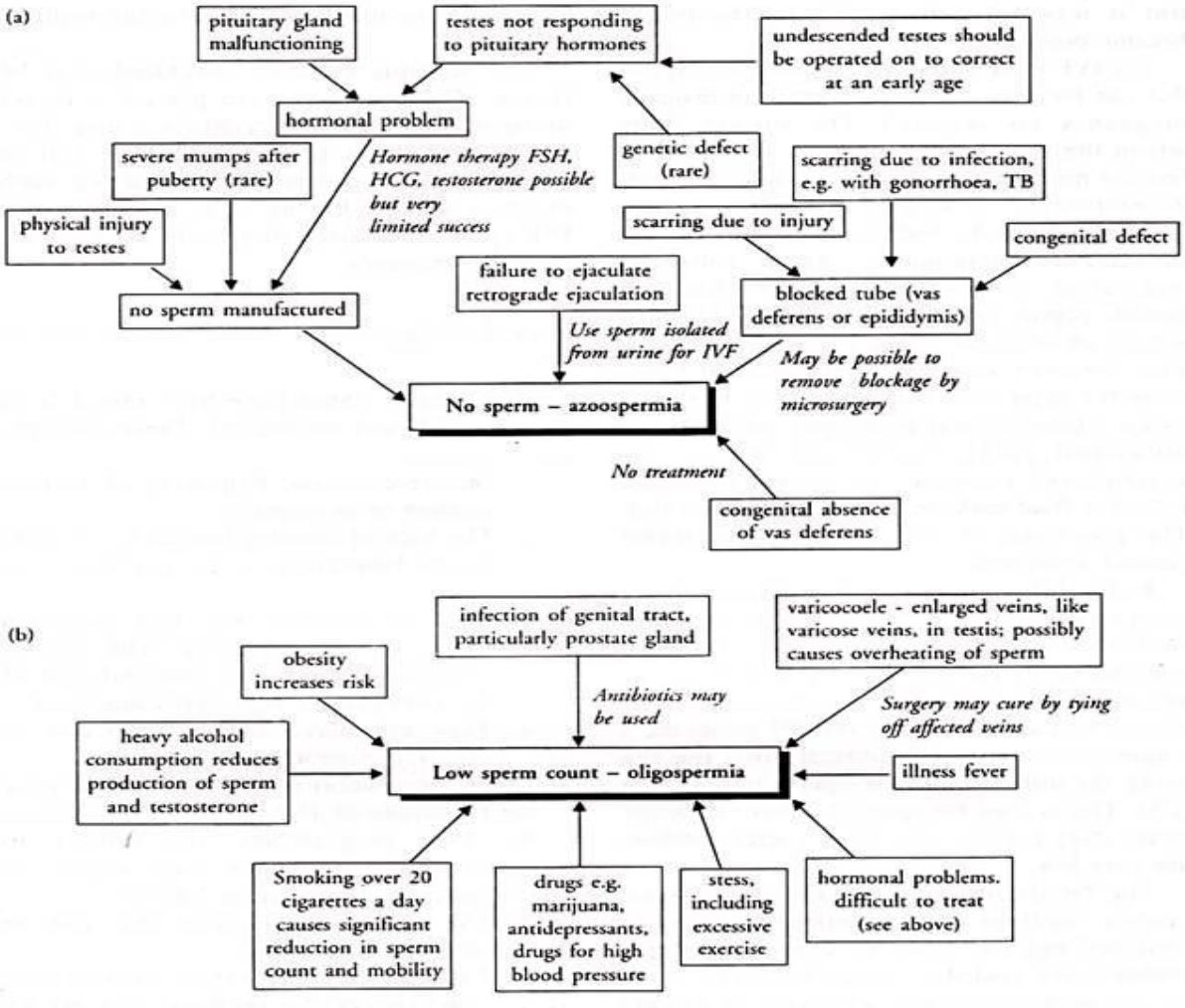


Fig. 3 The causes and treatment of azoospermia and oligospermia.

b. Oligospermia:

Low sperm count is known as oligospermia. More than 90% males suffer from infertility due to low sperm count. The reasons of oligospermia is summarised in Fig. 3.

c. Abnormal Sperms:

Abnormal sperms may possess two heads, or no tail or may have abnormal shapes (Fig. 4). The reasons are not known and may be because of hormonal malfunctions.

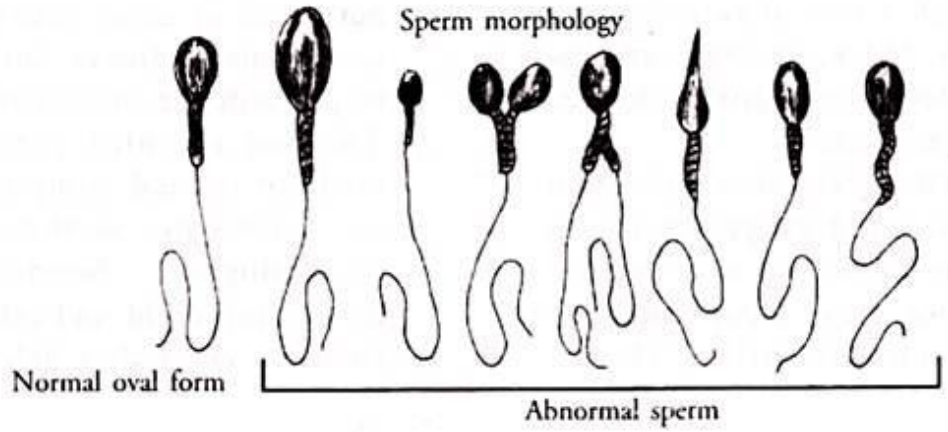


Fig. 4 Semen analysis.

d. Autoimmunity:

In some males, the immune system may attack the sperms and reduce the sperm numbers. Treatment is not usually possible.

e. Impotence and Premature Ejaculation:

The inability to achieve an erection of the penis is known as impotence. Psychological counselling may help in some cases. Premature ejaculation is a condition where the man releases the semen even before penetration into the vagina. This condition is treatable with psychological treatment.

f. Immotile cilia:

Absence of tail in sperm makes it immotile. Hence, sperms cannot move from vagina to upper portions of genital tract of female.

g. Absence of Y-chromosome:

Sometimes, deletion of Y-chromosomes in primordial germ cells leads to sperm production without Y-chromosome. Such sperms cannot form viable zygole.

h. Tubular blockage:
Blockage of vasa deferentia and vasa efferentia stops sperm transport.

i. Antisperm antibodies:
Such antibodies are IgG, IgM and IgA. Sometimes IgG is found in cervical mucous, serum and semen.

j. High scrotal temperature:
Due to development of dilated veins in testis (varicocela) scrotal temperature is raised and sperm production is minimized leading to oligospermia.

k. Low fructose content and high prostaglandin in seminal fluid lead to sperm destruction.

l. Vasectomy leads to irreversible infertility in males.

Other causes of male sterility may include:
a. Genetic factors: A man should have an X and Y chromosome. If he has two X chromosomes and one Y chromosome, as in Klinefelter's syndrome, the testicles will develop abnormally and there will be low testosterone and a low sperm count or no sperm.

b. Mumps: If this occurs after puberty, inflammation of the testicles may affect sperm production.

c. Hypospadias: The urethral opening is under the penis, instead of its tip. This abnormality is usually surgically corrected in infancy. If the correction is not done, it may be harder for the sperm to get to the female's cervix. Hypospadias affects about 1 in every 500 newborn boys.

d. Cystic fibrosis: This is a chronic disease that results in the creation of a sticky mucus. This mucus mainly affects the lungs, but males may also have a missing or obstructed vas deferens. The vas deferens carries sperm from the epididymis to the ejaculatory duct and the urethra.

e. Radiation therapy: This can impair sperm production. The severity usually depends on how near to the testicles the radiation was aimed.

f. Some diseases: Conditions that are sometimes linked to lower fertility in males are anaemia, Cushing's syndrome, diabetes, and thyroid disease.

Some medications increase the risk of fertility problems in men:

a. **Sulfasalazine:** This anti-inflammatory drug can significantly lower a man's sperm count. It is often prescribed for Crohn's disease or rheumatoid arthritis. Sperm count often returns to normal after stopping the medication.

b. **Anabolic steroids:** Popular with bodybuilders and athletes, long-term use can seriously reduce sperm count and mobility.

c. **Chemotherapy:** Some types may significantly reduce sperm count.

d. **Illegal drugs:** Consumption of marijuana and cocaine can lower the sperm count.

e. Age: Male fertility starts to fall after 40 years.

f. Exposure to chemicals: Pesticides, for example, may increase the risk.

g. Excess alcohol consumption: This may lower male fertility. Moderate alcohol consumption has not been shown to lower fertility in most men, but it may affect those who already have a low sperm count.

h. Mental stress: Stress can be a factor, especially if it leads to reduced sexual activity.

B. Female Infertility:

A woman may be infertile due to several causes.

Some important reasons are as follows:

a. Failure to Ovulate:

Failure to ovulate is one common cause of infertility in females. This is because the pituitary or hypothalamus fails to produce the FSH which is required for follicle development or LH required for release of the egg from the ovary. It may also be because the ovaries fail to produce oestrogen or progesterone. Hormonal imbalances may be corrected by administering synthetic hormones to the affected individual.

The most commonly used drug is Clomiphene, a synthetic oestrogen like drug which stimulates ovulation. Tamoxifen is another drug used. These pills are taken orally for five days soon after the menstrual cycle starts. Injection of HCG, which is chemically similar to LH is given at the middle of the cycle to stimulate ovulation. 'Fertility drugs' which contains FSH and LH or only FSH is also used. But these have the danger of multiple egg release and consequently multiple pregnancies. Advance techniques include small implants in the upper arm which releases small amounts of GnRH mimicking the activity of the hypothalamus.

b. Damage to Oviducts:

The fallopian tubes may be blocked or narrowed in some women. This interferes with the movement of the eggs and fertilisation. This can be treated by laser surgery.

c. Damage to Uterus:

In about 5-10% cases, infertility problems are due to a damaged uterus. The uterus is unable to maintain pregnancy, i.e., the fertilised zygote does not get implanted. Sometimes large non-malignant tumours called fibroids or smaller growths known as polyps which grow in the walls of the uterus can cause infertility. These can be surgically removed. IUCD or PID also causes inflammation in the uterus and cause problems. This can be treated by using antibiotics. Adhesion in the uterus, i.e. sticking of parts of the uterus which occurs as a result of an abortion is another reason for infertility.

d. Damage to the Cervix:

The cervix is the neck of the uterus. The cervix may become damaged because of the abortion or difficult birth. A narrow cervix may interfere with sperm movement.

e. Antibodies to Sperm:

In some rare cases, women may produce antibodies against sperms. These are found in the cervix, uterus and oviducts. These may be treated using immunosuppressant drugs, but IVF is a better method of treatment.

f. Ovarian problem:

There may not be normal ovulation in ovary. Sometimes there is failure of corpus luteum formation.

g. Hormonal cause:

Decreased level of FSH and LH, drug induced ovulation may not allow fertilization and development of the foetus.

h. Uterine factor:

Unfavourable endometrium for implantation, chronic endometritis, fibroid uterus etc. may be the cause of infertility.

i. Cervical factor: In effective sperm penetration, chronic cervicitis, presence of anti sperm antibody and elongation of cervix may be the cause of infertility.

j. Fimbriae:

Fimbriae of Fallopian tube may not pick up secondary oocyte from ovary.

k. Dyspareunia:
Painful sexual intercourse experienced by female may be another cause of infertility.

l. Macrophages:
Increased sperm phagocytosis by macrophages may be the cause of infertility.

m. Miscarriage:
Early miscarriage before complete development of foetus due to various gyaenic problems may be also the reason of infertility.

n. Tubectomy:
Like vasectomy in males, tubectomy in females causes permanent infertility.

Other causes of female sterility may include:
a. Age: The ability to conceive starts to fall around the age of 32 years.

b. Smoking: Smoking significantly increases the risk of infertility in both men and women, and it may undermine the effects of fertility treatment. Smoking during pregnancy increases the chance of pregnancy loss. Passive smoking has also been linked to lower fertility.

c. Alcohol: Any amount of alcohol consumption can affect the chances of conceiving.

d. Being obese or overweight: This can increase the risk of infertility in women as well as men.

e. Eating disorders: If an eating disorder leads to serious weight loss, fertility problems may arise.

f. Diet: A lack of folic acid, iron, zinc, and vitamin B-12 can affect fertility. Women who are at risk, including those on a vegan diet, should ask the doctor about supplements.

g. Exercise: Both too much and too little exercise can lead to fertility problems.

h. Sexually transmitted infections (STIs): Chlamydia can damage the fallopian tubes in a woman and cause inflammation in a man's scrotum. Some other STIs may also cause infertility.

i. Exposure to some chemicals: Some pesticides, herbicides, metals, such as lead, and solvents have been linked to fertility problems in both men and women. A mouse study has suggested that ingredients in some household detergents may reduce fertility.

j. Mental stress: This may affect female ovulation and male sperm production and can lead to reduced sexual activity.

Some medical conditions can affect fertility:

Ovulation disorders appear to be the most common cause of infertility in women. Ovulation is the monthly release of an egg. The eggs may never be released or they may only be released in some cycles.

Ovulation disorders can be due to:

a. **Premature ovarian failure:** The ovaries stop working before the age of 40 years.

b. Polycystic ovary syndrome (PCOS): The ovaries function abnormally and ovulation may not occur.

c. Hyperprolactinemia: If prolactin levels are high, and the woman is not pregnant or breastfeeding, it may affect ovulation and fertility.

e. **Poor egg quality:** Eggs that are damaged or develop genetic abnormalities cannot sustain a pregnancy. The older a woman is, the higher the risk.

f. **Thyroid problems:** An overactive or underactive thyroid gland can lead to a hormonal imbalance.

g. Chronic conditions: These include AIDS or cancer.

Problems in the uterus or fallopian tubes can prevent the egg from traveling from the ovary to the uterus, or womb. If the egg does not travel, it can be harder to conceive naturally.

Causes include:

a. Surgery: Pelvic surgery can sometimes cause scarring or damage to the fallopian tubes. Cervical surgery can sometimes cause scarring or shortening of the cervix. The cervix is the neck of the uterus.

b. **Submucosal fibroids:** Benign or non-cancerous tumours occur in the muscular wall of the uterus. They can interfere with implantation or block the fallopian tube, preventing sperm from fertilizing the egg. Large submucosal uterine fibroids may make the uterus' cavity bigger, increasing the distance the sperm has to travel.

c. **Endometriosis:** Cells that normally occur within the lining of the uterus start growing elsewhere in the body.

d. Previous sterilization treatment: In women who have chosen to have their fallopian tubes blocked, the process can be reversed, but the chances of becoming fertile again are not high.

Diagnosis of Infertility:

Most people will visit a physician if there is no pregnancy after 12 months of trying. If the woman is aged over 35 years, the couple may wish to see a doctor earlier, because fertility testing can take time, and female fertility starts to drop when a woman is in her 30s. A doctor can give advice and carry out some preliminary assessments. It is better for a couple to see the doctor together. The doctor may ask about the couple's sexual habits and make recommendations regarding these. Tests and trials are available, but testing does not always reveal a specific cause

I. Infertility tests for men:

The doctor will ask the man about his medical history, medications, and sexual habits and carry out a physical examination. The testicles will be checked for lumps or deformities, and the shape and structure of the penis will be examined for abnormalities.

a. **Semen analysis:** A sample may be taken to test for sperm concentration, motility, colour, quality, any infections, and whether any blood is present. Sperm counts can fluctuate, so that several samples may be necessary.

b. **Blood test:** The lab will test for levels of testosterone and other hormones.

c. **Ultrasound:** This may reveal issues such as ejaculatory duct obstruction or retrograde ejaculation.

d. Chlamydia test: Chlamydia can affect fertility, but antibiotics can treat it.

II. Infertility tests for women:

A woman will undergo a general physical examination, and the doctor will ask about her medical history, medications, menstruation cycle, and sexual habits.
She will also undergo a gynaecologic examination and a number of tests:

a. Blood test: This can assess hormone levels and whether a woman is ovulating.

b. Hysterosalpingography: Fluid is injected into the woman's uterus and X-rays are taken to determine whether the fluid travels properly out of the uterus and into the fallopian tubes. If a blockage is present, surgery may be necessary.

c. Laparoscopy: A thin, flexible tube with a camera at the end is inserted into the abdomen and pelvis, allowing a doctor to look at the fallopian tubes, uterus, and ovaries. This can reveal signs of endometriosis, scarring, blockages, and some irregularities of the uterus and fallopian tubes.

Other tests include:

a. ovarian reserve testing, to find out how effective the eggs are after ovulation
b. genetic testing, to see if a genetic abnormality is interfering with fertility
c. pelvic ultrasound, to produce an image of the uterus, fallopian tubes, and ovaries
d. Chlamydia test, which may indicate the need for antibiotic treatment
e. thyroid function test, as this may affect the hormonal balance.

Ovarian hyperstimulation syndrome:

The ovaries can swell, leak excess fluid into the body, and produce too many follicles, the small fluid sacs in which an egg develops.
Ovarian hyperstimulation syndrome (OHSS) usually results from taking medications to stimulate the ovaries, such as clomifene and gonadotrophins. It can also develop after IVF.
Symptoms include:

- bloating
- constipation
- dark urine
- diarrhoea
- nausea
- abdominal pain
- vomiting

They are usually mild and easy to treat. Rarely, a blood clot may develop in an artery or vein, liver or kidney problems can arise, and respiratory distress may develop. In severe cases, OHSS can be fatal.

Ectopic pregnancy:

This is when a fertilized egg implants outside the womb, usually in a fallopian tube. If it stays in there, complications can develop, such as the rupture of the fallopian tube. This pregnancy has no chance of continuing.

Immediate surgery is needed and, sadly, the tube on that side will be lost. However, future pregnancy is possible with the other ovary and tube. Women receiving fertility treatment have a slightly higher risk of an ectopic pregnancy. An ultrasound scan can detect an ectopic pregnancy.

Treatment:

Treatment will depend on many factors, including the age of the person who wishes to conceive, how long the infertility has lasted, personal preferences, and their general state of health.

Frequency of intercourse:

The couple may be advised to have sexual intercourse more often around the time of ovulation. Sperm can survive inside the female for up to 5 days, while an egg can be fertilized for up to 1 day after ovulation. In theory, it is possible to conceive on any of these 6 days that occur before and during ovulation. However, a survey has suggested that the 3 days most likely to offer a fertile window are the 2 days before ovulation plus the 1 day of ovulation. Some suggest that the number of times a couple has intercourse should be reduced to increase sperm supply, but this is unlikely to make a difference.

Fertility treatments for men: Treatment will depend on the underlying cause of the infertility.

a. Erectile dysfunction **or premature ejaculation:** Medication, behavioural approaches, or both may help improve fertility.

b. **Varicocele:** Surgically removing a varicose vein in the scrotum may help.

c. **Blockage of the ejaculatory duct:** Sperm can be extracted directly from the testicles and injected into an egg in the laboratory.

d. **Retrograde ejaculation:** Sperm can be taken directly from the bladder and injected into an egg in the laboratory.

e. **Surgery for epididymal blockage:** A blocked epididymis can be surgically repaired. The epididymis is a coil-like structure in the testicles which helps store and transport sperm. If the epididymis is blocked, sperm may not be ejaculated properly.

Fertility treatments for women: Fertility drugs might be prescribed to regulate or induce ovulation. They include:

a. **Clomifene (Clomid, Serophene):** This encourages ovulation in those who ovulate either irregularly or not at all, because of PCOS or another disorder. It makes the pituitary gland release more follicle-stimulating hormone (FSH) and luteinizing hormone (LH).

b. **Metformin (Glucophage):** If Clomifene is not effective, metformin may help women with PCOS, especially when linked to insulin resistance.

c. Human menopausal gonadotropin, or hMG (Repronex): This contains both FSH and LH. Patients who do not ovulate because of a fault in the pituitary gland may receive this drug as an injection.

d. Follicle-stimulating hormone (Gonal-F, Bravelle): This hormone is produced by the pituitary gland that controls oestrogen production by the ovaries. It stimulates the ovaries to mature egg follicles.

e. Human chorionic gonadotropin (Ovidrel, Pregnyl): Used together with clomiphene, hMG, and FSH, this can stimulate the follicle to ovulate.

f. Gonadotropin-releasing hormone (Gn-RH) analogs: These can help women who ovulate too early—before the lead follicle is mature—during hmG treatment. It delivers a constant supply of Gn-RH to the pituitary gland, which alters the production of hormone, allowing the doctor to induce follicle growth with FSH.

g. Bromocriptine (Parlodel): This drug inhibits prolactin production. Prolactin stimulates milk production during breastfeeding. Outside pregnancy and lactation, women with high levels of prolactin may have irregular ovulation cycles and fertility problems.

Assisted conception:

The following methods are currently available for assisted conception.

a. **Intrauterine insemination (IUI)**: At the time of ovulation, a fine catheter is inserted through the cervix into the uterus to place a sperm sample directly into the uterus. The sperm is washed in a fluid and the best specimens are selected. The woman may be given a low dose of ovary stimulating hormones. IUI is more commonly done when the man has a low sperm count, decreased sperm motility, or when infertility does not have an identifiable cause. It can also help if a man has severe erectile dysfunction.

b. **In-vitro fertilization (IVF)**: Sperm are placed with unfertilized eggs in a petri dish, where fertilization can take place. The embryo is then placed in the uterus to begin a pregnancy. Sometimes the embryo is frozen for future use.

c. **Intracytoplasmic sperm injection (ICSI)**: A single sperm is injected into an egg to achieve fertilization during an IVF procedure. The likelihood of fertilization improves significantly for men with low sperm concentrations.

d. **Sperm or egg donation**: If necessary, sperm or eggs can be received from a donor. Fertility treatment with donor eggs is usually done using IVF.

e. **Assisted hatching**: The embryologist opens a small hole in the outer membrane of the embryo, known as the zona pellucid. The opening improves the ability of the embryo to implant into the uterine lining. This improves the chances that the embryo will implant at, or attach to, the wall of the uterus. This may be used if IVF has not been effective, if there has been poor embryo growth rate, and if the woman is older. In some women, and especially with age, the membrane becomes harder. This can make it difficult for the embryo to implant.

f. **Electric or vibratory stimulation to achieve ejaculation**: Ejaculation is achieved with electric or vibratory stimulation. This can help a man who cannot ejaculate normally, for example, because of a spinal cord injury.

g. **Surgical sperm aspiration**: The sperm is removed from part of the male reproductive tract, such as the vas deferens, testicle, or epididymis.

Surgical procedures for women:

If the fallopian tubes are blocked or scarred, surgical repair may make it easier for eggs to pass through. Endometriosis may be treated through laparoscopic surgery. A small incision is made in the abdomen, and a thin, flexible microscope with a light at the end, called a laparoscope, is inserted through it. The surgeon can remove implants and scar tissue, and this may reduce pain and aid fertility.

Probable Questions:

1. Define infertility. What are the types of infertility?
2. Discuss the causes of male sterility.
3. What medicines induce male sterility.
4. Discuss the causes of female sterility.
5. How male sterility can be diagnosed?
6. How female sterility can be diagnosed?
7. What is ectopic pregnancy? Explain.
8. Discuss fertility treatment in females?
9. Discuss fertility treatments in males?

Suggested Readings:

1. Embryology by N. Kumarsen
2. Developmental Biology by Veerbala Rastogi.
3. Embryology by M.P. Arora
4. Developmental Biology by Gilbert.

UNIT-III
Teratogenesis, Stem cells and tissue engineering

Objective: In this unit we ail discus about teratogenesis. We will discuss also about stem cells and their applications in embryonic development and tissue engineering.

Introduction:

Teratogenesis is a prenatal toxicity characterized by structural or functional defects in the developing embryo or fetus. It also includes intrauterine growth retardation, death of the embryo or fetus, and transplacental carcinogenesis (in which chemical exposure of the mother initiates cancer development in the embryo or fetus, resulting in cancer in the progeny after birth). A teratogen is any agent that causes an abnormality following fetal exposure during pregnancy. Teratogens are usually discovered after an increased prevalence of a particular birth defect. For example, in the early 1960's, a drug known as thalidomide was used to treat morning sickness.

Intrauterine human development has three stages: implantation, postimplantation, and fetal development. The first two stages are the embryonic stages and last through the first eight weeks after conception. The fetal stage begins in the ninth week and continues to birth. Depending on the developmental stage, chemical exposure in the mother can result in different degrees of toxicity in the embryo or fetus. In the preimplantation period, a toxic chemical can kill some of the cells in the blastocyst, resulting in the death of the embryo. During the postimplantation period, chemical-induced cell death leads to one of two outcomes. If death is confined to those cells undergoing active cell division at the moment, the corresponding organs are affected, resulting in malformation. If the cell death is generalized without significant replication by the remaining cells to sustain life, the embryo dies. During the third, fetal, period, chemical injury can retard growth or, if severe enough, kill the fetus.

The genesis of a particular organ (organogenesis) occurs at a specific time during gestation and is not repeated. Because organogenesis is a tightly programmed sequence of events, each organ system has a critical period during which it is sensitive to chemical injury. Chemical exposure in a critical period is likely to produce malformations of that organ and not others; however, because there is some overlapping of critical periods of organ development and because chemicals frequently remain in the embryo for a period of time, malformations of more than one organ usually occur. Since organogenesis occurs mostly in the embryonic stages, chemical exposure in the first trimester should be minimized, if possible. Little is known about mechanisms of teratogenesis. It is thought that some

teratogens produce malformations directly by killing the cells in the embryo. Teratogens can also produce malformations indirectly by causing maternal toxicity, resulting in oxygen or nutrient deficiency for the embryo. A few well-known examples are discussed below.

Thalidomide is a drug originally marketed to combat nausea and vomiting in pregnancy. It was discovered in the 1960s in West Germany to cause rare limb defects, among other congenital anomalies. The discoveries about thalidomide triggered legislation requiring teratogenicity testing for drugs. Chronic alcohol ingestion during pregnancy is the most common cause of congenital problems in mental development. Ingestion of more than 30 millilitres (1 ounce) of ethyl alcohol per day during pregnancy can lead to the development of fetal alcohol syndrome, characterized by intrauterine growth retardation and subsequent learning disabilities, such as distractibility, language disorders, and low IQ. Heavier consumption of alcohol, more than 60 millilitres per day, by a pregnant woman can result in malformations of the fetal brain and in spontaneous abortions.

Diethylstilbestrol (DES) is a drug used primarily from the 1940s to the ’50s to prevent miscarriage. The drug is an example of a chemical that can produce transplacental carcinogenesis. It was discovered in the early 1970s that exposures to diethylstilbestrol before the ninth week of gestation could lead to the formation of rare vaginal and cervical cancers in female progenies.

Causes of Teratogenicity:

The toxicants which cause teratogenesis are known as teratogenic agents. A gestating-embryo exhibits great dynamicity of the living cells. The embryonic cells multiply and differentiate at a tremendous rate making the embryo more susceptible to the drugs.

Stage Sensitivity for Teratogenicity:

i. Pre-Differentiation Stage:

During this stage the embryo is not susceptible to teratogenic agents. These agents either cause death to the embryo by killing all or most of the cells, or have no apparent effect on the embryo. Even when some widely harmful effects have been produced, the surviving cells can compensate and form a normal embryo. This resistant stage varies from 5-9 days depending on the species.

ii. Embryonic Stage:

In fact this is the period when the cells undergo intensive differentiation, mobilization and organization. It is during this period that most of the organogenesis takes place. As a result, the embryo becomes most susceptible to the effects of various teratogens.

This period generally ends sometimes from the 10th-14th day in rodents and in the 14th week of the gestation period in humans. All organs are, however, not susceptible in the same period of the pregnancy. Rat embryo is most susceptible between days 8 and 12 for most organs, but the palate and urinogenital organs are more susceptible at a later stage for teratogens.

J. G. Wilson (1965) observed teratogenic treatment on the 10th day of gestation which resulted in the following incidences of malfor-mations in rat:

Brain defects – 35%

Eye defects – 33%

Heart defects – 24%

Skeletal defects – 18%

Urinogential defects – 6%

iii. Fetal Stage:

This stage is characterized by growth and functional maturation. Teratogens are thus unlikely to cause morphological defects during this stage, but they may induce functional abnormalities. Whereas, morphologic defects are, in general, readily detected at birth or shortly thereafter functional abnormalities, viz., CNS impairment, may not be diagnosed for some time even after birth.

Mode of Action of Teratogens:

Various mechanisms are involved in teratogenic effects:

i. Interference with Nucleic Acids:

Various teratogenic agents interfere with nucleic acid replication, transcription, or RNA translation. These include alkylating agents, antimetabolites, intercalating agents and amino acid antagonists.

ii. Inhibition of Enzymes:

Inhibitors of enzymes, e.g. 5-flourouracil, may induce malformation through interference with differentiation or growth by inhibiting thymidylate synthatase. Other examples include 6-aminonicotinamide, which inhibits glucose-6-phosphate dehydrogenase, and folate antagonists which inhibit dihydrofolate reductase.

iii. Deficiency of Energy Supply and Osmolarity:

Certain teratogens can affect the energy supply for the metabolism by restricting the availability of substrates either directly (e.g., dietary deficiencies) or through the presence of analogs for antagonists of vitamins, essential amino acids, and others.

In addition, hypoxia and agents i.e., CO and CO2, can be teratogenic by depriving the metabolic process of the required O2 and probably also by the production of osmolar imbalances. These can induce edema, which, in turn, cause mechanical distortion and tissue ischemia. Physical agents that can cause malfor-mations include radiation, hypothermia, hyperthermia and mechanical trauma.

It shall not be out of place to mention that the mode of action of many teratogens is yet uncertain. Furthermore, a potential teratogen may or may not exert teratogenic effects depending on such factors as bio-activating mechanism, stability and detoxifying capability of the embryonic tissues. Appropriate experimental testing for the teratogenicity of toxicants is, therefore, essential.

Testing Procedures:

Animals:

For teratogenic tests, the animals should be young, mature and healthy. Usually, Prima gravida females are preferred. Rats, rabbits and hamsters are the commonly used animals, because of their ready availability, easy handling, little size and short gestational period.

Pigs, are sometimes also used because they are phylogenetically more similar to humans. WHO (1967) suggested the use of nonhuman primates because of their phylogenetic proximity to humans. Other animals such as dogs and cats have also been used by some investigators.

With rats and rabbits, at least 20 and 12 females, respectively, are placed in each dose group of teratogenic agent. Smaller numbers of large animals such as dogs and nonhuman primates are also used.

Administration of Teratogenic Agent:

Dosage:

At least three dosages are usually used. The lowest dosage should be approximately interspersed between the two extremes.

In addition, two control groups are included. One of these is given the vehicle or physiologic saline and the other receives a substance of known teratogenic activity. These groups provide information on the incidence of spontaneous malformation and the sensitivity of the specific lot of animals under the existing experimental conditions. In addition to these contemporary controls, data from historical controls are also useful.

Route and Timing:

The test compounds should be administered through route that stimulates the human exposure situations. For food additives and contaminants, the chemical is preferably incorporated in the animal feeds. Oral drugs are generally administered by gastric gavage.

The timing of administering the substance is of great importance. For routine teratologic studies, it is customary to administer the substance during the entire period of organogenesis when the embryo is most susceptible. This period varies from one species to another.

Observations:

The Pregnant Animals:

The animals should be examined daily for gross signs of toxicity and many females that show signs of impending abortion or premature delivery (e.g., vaginal bleeding) should be examined.

The Fetuses:

Fetuses are usually surgically removed from the mother about one day prior to the expected delivery. This procedure is intended to avoid cannibalism and permit counting of resorption sites and dead fetuses.

Following observations are to be made and recorded:

i. Number of corpora lutea

ii. Number and position of implantations

iii. Number and position of resorptions

iv. Number and position of dead fetuses

v. Number and position of live fetuses

vi. Sex of each live fetus

vii. Weight of each live fetus

viii. Length of each live fetus, and

ix. Abnormalities of each fetus.

Detailed Examinations:

To determine the different types of abnormalities, each fetus is examined for external defects. In addition, about 2/3rd of random sampled fetuses are closely examined for skeletal abnormalities after staining with Alizarin Red. The remaining one-third of the fetuses are examined for visceral defects after fixations in Bowin's fluid and sectioned by microtome. With larger animals, e.g., dogs, pigs, and non-human primates, the skeletal structure is generally examined with X-ray instead of staining.

Delayed Effects:

With toxicants that are suspected of having effects on the central nervous system or genitourinary system, a sufficient number of pregnant females are allowed to deliver their pups. These pups are nursed either by their biological mothers — thus possibly being exposed to the toxicants via the milk — or by foster mothers. In the latter case, the potential effects of postnatal exposure are eliminated.

Neuromotor and behavioural tests may be applied to detect CNS effects. These include posture, mother activity, coordination, endurance, vision, hearing, learning ability, response to foreign environment, mating behaviour and maternal behaviour.

Evaluation of Teratogenic Effects:

Categories and Relative Significance:

Aberrations:

In addition to functional abnormalities, morphologic defects may involve external/or internal structure. Not all types of aberrations have the same significance. For example, supernumerary ribs decrease, or abnormal sternal ossification might have little or no visible effect on external morphology, functional activity, or survival of the fetus. These have been considered as deviations.

Malformations of doubtful significance include curly tail, straight legs, malrotated limbs and paws, wrist drop, protruding tongue, enlarged atria and/or ventricles, abnormal renal pelvic development, and translucent skin. In general, these have been characterized as minor anomalies. There are, at the other extreme, major malformations that are incompatible with survival, growth, development, fertility, and longevity, e.g., Spina bifda, hydrocephalus.

In practice, the distinction between these categories is not always clear cut. It is then necessary to take other factors into consideration:

(i) Resorption:

This is a manifestation of death of the conceptus. Although the site of resorption can be readily identified with a close examination of the uterus; the number of resorptions is more

reliably determined by subtracting the total near term offsprings from the total implantations, as indicated by the number of corpora lutea.

If there is an appreciable increase in the number of resorptions in the treated group, it may be necessary to alter the testing procedure to differentiated embryotoxicity from teratogenicity, e.g. by lowering the dose used to reduce the toxicity or shorten the exposure period.

(ii) Fetal Toxicity:

This may manifest as reduced body weight on non-viable fetus. This type of data is often useful in assessing the teratogenicity of the toxicant in question. With rabbits, the viability of fetus, if in question, may be determined by incubating it for 24 hours.

Sources of Error:

i. The animals used may exhibit an excessive number of spontaneous malformation or may be resistant to teratogenic effects. These errors can usually be assessed by the response of the animals to the negative and positive control agents.

ii. Poor animal husbandry and mishandling of the animals may also result in an increased incidence of malformations.

iii. The food consumption can.be affected by the toxicant used. This fact may then alter the body weight of the mothers and indirectly affect the fetuses.

iv. Excessively large doses can result in many resorptions but few or no malformations. On the other hand, if the doses are too small, there may not be any evidence of teratogenicity.

Analysis of the Results:

In comparing the treated and control groups, the proper experimental unit is the litter rather than the individual fetuses. The number of litters with malformed fetuses, resorptions, or dead fetuses are the parameters to be used in statistical analysis. However, an increase in the average number of fetuses with defects per litter may provide evidence of teratogenicity.

If the results indicate a relationship between the doses and the response (incidence of malformation), it is generally justifiable to conclude that the agent is teratogenic under the specific experimental conditions.

When the incidence of malformation does not provide a definite conclusion, an analysis of the data from the historical controls may be valuable. Furthermore, a close examination of the data on other parameters of the fetus and on the mother is sometimes useful.

Extrapolation to Humans:

The results obtained in teratogenesis studies in animals cannot be readily extrapolated to humans. The lack of a suitable animal model is evidenced by the fact that the most potent human teratogen, thalidomide, which is effective at a dose of 0.5 -1.0 mg/kg, has no teratogenic effect in rats and mice at 4,000 mg/kg. Only moderate embyropathy is noted in rabbits. On the other hand, acetylsalicylic acid has a long history of safe use in human pregnancy but is a potent teratogen in rat, mice and hamsters.

The mechanism of teratogenesis and the differences in response among various species of animals are poorly understood. The cause of spontaneous congenital malformations in humans are unknown. More basic animal studies and prospective epidemiologic studies are essentially required.

Nevertheless, since all chemicals that are teratogenic in humans were found to be active in certain animals as well, it is, therefore, prudent to carry out appropriate animal tests on all chemicals to which females of child-bearing age are usually exposed.
If positive results are achieved with a substance — especially this is so in more than one species of animal — exposure of females of childbearing age to this substance should be avoided, if possible. In assessing the teratogenic effects of a chemical, not only the incidence but also the severity of the aberrations should be taken into account.

In Vitro Tests:

Actually, these tests are not in routine use as yet, though they may show the mode of actions of teratogens. Some of these tests are — cell culture, organ culture, etc.

Stem cell and tissue engineering:

Stem cells also have the ability to repair damaged cells. These cells have strong healing power. They can evolve into any type of cell.

Research on stem cells is going on, and it is believed that stem cell therapies can cure ailments like paralysis and Alzheimer's as well. Let us have a detailed look at stem cells, their types and their functions.

Types of cells
Stem cells are of the following different types:

Embryonic Stem Cells
Adult Stem Cells
Induced Pluripotent Stem Cells
Mesenchymal stem cells
Embryonic Stem Cells

The fertilized egg begins to divide immediately. All the cells in the young embryo are totipotent cells. These cells form a hollow structure within a few days. Cells in one region group together to form the inner cell mass. This contains pluripotent cells that make up the developing foetus.

The embryonic stem cells can be further classified as:

Totipotent Stem Cells: These can differentiate into all possible types of stem cells.

Pluripotent Stem Cells: These are the cells from an early embryo and can differentiate into any cell type.

Multipotent Stem Cells: These differentiate into a closely related cell type. E.g., the hematopoietic stem cells differentiate into red blood cells and white blood cells.

Oligopotent Stem Cells: Adult lymphoid or myeloid cells are oligopotent. They can differentiate into a few different types of cells.

Unipotent Stem Cells: They can produce cells only of their own type. Since they have the ability to renew themselves, they are known as unipotent stem cells. E.g., Muscle stem cells.

Adult Stem Cells: These stem cells are obtained from developed organs and tissues. They can repair and replace the damaged tissues in the region where they are located. For

eg., hematopoietic stem cells are found in the bone marrow. These stem cells are used in bone marrow transplants to treat specific types of cancers.

Induced Pluripotent Stem Cells:

These cells have been tested and arranged by converting tissue-specific cells into embryonic cells in the lab. These cells are accepted as an important tool to learn about the normal development, onset and progression of the disease and are also helpful in testing various drugs. These stem cells share the same characteristics as embryonic cells do. They also have the potential to give rise to all the different types of cells in the human body.

Mesenchymal Stem Cells:

These cells are mainly formed from the connective tissues surrounding other tissues and organs, known as the stroma. These mesenchymal stem cells are accurately called stromal cells. The first mesenchymal stem cells were found in the bone marrow that is capable of developing bones, fat cells, and cartilage.

There are different mesenchymal stem cells that are used to treat various diseases as they have been developed from different tissues of the human body. The characteristics of mesenchymal stem cells depend on the organ from where they originate.

Applications of Stem Cells:

Following are the important applications of stem cells:

j. Tissue Regeneration

This is the most important application of stem cells. The stem cells can be used to grow a specific type of tissue or organ. This can be helpful in kidney and liver transplants. The doctors have already used the stem cells from beneath the epidermis to develop skin tissue that can repair severe burns or other injuries by tissue grafting.

ii. Treatment of Cardiovascular Disease

A team of researchers have developed blood vessels in mice using human stem cells. Within two weeks of implantation, the blood vessels formed their network and were as efficient as the natural vessels.

iii. Treatment of Brain Diseases

Stem cells can also treat diseases such as Parkinson's disease and Alzheimer's. These can help to replenish the damaged brain cells. Researchers have tried to differentiate embryonic stem cells into these types of cells and make it possible to treat diseases.

iv. Blood Disease Treatment:

The adult hematopoietic stem cells are used to treat cancers, sickle cell anaemia, and other immunodeficiency diseases. These stem cells can be used to produce red blood cells and white blood cells in the body.

Sources of Stem Cells:

Stem Cells originate from different parts of the body. Adult stem cells can be found in specific tissues in the human body. Matured cells are specialized to conduct various functions. Generally, these cells can develop the kind of cells found in tissues where they reside.

Embryonic Stem Cells are derived from 5-day-old blastocysts that develop into embryos and are pluripotent in nature. These cells can develop any type of cell and tissue in the body. These cells have the potential to regenerate all the cells and tissues that have been lost because of any kind of injury or disease.

Stem cell therapy:

Stem-cell therapy is the use of stem cells to cure or prevent a disease or condition. The damaged cells are repaired by the generated stem cells, which can also hasten the healing process in the injured tissue. These cells are essential for the regeneration and transplanting of tissue.

Tissue Engineering:

Tissue engineering is a biomedical engineering discipline that uses a combination of cells, engineering, materials methods, and suitable biochemical and physicochemical factors to restore, maintain, improve, or replace different types of biological tissues.

The four fundamental aspects are: (1) Cell Sources and Culture (2) Cell Orientation (3) Cell Support Materials and (4) Design and Engineering of Tissues.

Tissue engineering (TE) refers to the application of the principles of engineering to cell culture for the construction of functional anatomical units (tissues/organs). The ultimate purpose of TE is to supply various body parts for the repair or replacement of damaged tissues or organs.

Tissue engineering may be regarded as the backbone of reconstructive surgery. It is possible to supply almost all surgical implants (skin, blood vessels, ligaments, heart valves, joint surfaces, nerves) through the developments in tissue engineering.

There are two schools of thought while dealing with tissue engineering techniques:

1. Some workers believe that the living cells possess an innate potential of biological regeneration. This implies that when suitable cells are allowed to grow on an appropriate support matrix, the cells proliferate, and ultimately result in an organized and functional tissue. This tissue resembles the original tissue in structure and function. This approach is very simple, and economical, although the success is limited.

2. According the second school of thought, there are several control processes to produce a new and functional tissue. Thus, tissue regeneration in vivo or tissue production in vitro are very complex. Therefore, tissue engineering is not a simple regeneration of cells, and it requires a comprehensive approach with a thorough understanding of cellular configuration, special arrangement and control process.

Tissue engineering is a complicated process. Some fundamental and basic aspects of TE with special reference to the following aspects are briefly described:

1. Cell sources and culture

2. Cell orientation

3. Cell support materials

4. Design and engineering of tissues.

1. Cell Sources and Culture:

Adequate quantities of cells are required for tissue engineering. There are three types of cell sources-autologous, allogeneic and xenogeneic.

Autologous Cell Sources:

The cell source is said to be autologous when the patient's own cells are used in TE. This is a straight forward approach. A piece of desired tissue is taken by biopsy. It may be enzymatically digested or explant cultured, and the cells are grown to the required number.

The main advantages of autologous cells in TE are:

i. Avoidance of immune complications

ii. Reduction in the possible transfer of inherent infections.

There are certain disadvantages associated with autologous cells.

i. It is not always possible to obtain sufficient biopsy material from the patient.

ii. Disease state and age of the patient will be limiting factors.

Allogeneic Cell Sources:

If the cells are taken from a person other than the patient, the source is said to be allogeneic.

The advantages of allogeneic cell source are listed:

i. Obtained in good quantity from a healthy donor.

ii. Cells can be cultured in a large scale.

iii. Cost-effective with consistent quality.

iv. Available as and when required by a patient.

The major problem of allogeneic cell source is the immunological complications that may ultimately lead to graft rejection. The immune responses however, are variable depending on the type of cells used. For instance, endothelial cells are more immunogenic while fibroblasts and smooth muscle cells are less immunogenic. The age of the donor is another important factor that contributes to immunological complications. Thus, cells from adult donors are highly immunogenic while fetal or neonatal cells elicit little or no immune response.

Xenogeneic Cell Sources:

When the cells are taken from different species (e.g. pig source for humans) the source is said to be xenogeneic. This approach is not in common use due to immulogical complications.

Culture of Cells:

The methods adapted for culturing of cells required for tissue engineering depend on the type and functions of cells. For most of the cells, the conventional monolayer cultures serve the purpose. The major drawback of monolayer cultures is that cells may lose their morphology, functions and proliferative capacity after several generations. Some workers

prefer three dimensional cultures for the cells to be used in tissue engineering. The nutrient and gaseous exchanges are the limiting factors in three dimensional cultures.

Genetic Alterations of Cultured Cells for Use in TE:

Gene therapy can be successfully employed in tissue engineering. This can be achieved by transferring the desired genes to cells in culture. The new genes may increase the production of an existing protein or may synthesize a new protein.

Some success has been achieved in this direction:

i. Genetically altered fibroblasts can produce transferrin, clotting factor VIII and clotting factor IX.

ii. Modified endothelial cells can synthesize tissue plasminogen activator.

iii. Genetically engineered keratinocytes can produce trans-glutaminase-l (This enzyme is lacking in patients suffering from a dermal disorder, lamellar icthyosis). The altered keratinocytes proved successful when transplanted in animal (rat) models of this disease.

2. Cell Orientation:

The orientation of cells with regard to specific shape and spatial arrangement is influenced by the following environmental factors:

1. Substrate or contact guidance.

2. Chemical gradients.

3. Mechanical cues.

Substrate Guidance:

The topographical features of the substrate determine the contact guidance. These features may be in the form of ridges, aligned fibers etc. It is possible to use differential attachment to substrates as a means of producing different alignment of cells. In recent years, synthetic polymer substrate collagen fibrils and fibronectin are used as bioresorbable templates for tissue engineering.

Chemical Gradients:

Development of chemical gradients is required for cellular orientation and for the stimulation of cellular functions. Certain growth factors and extracellular macromolecules are capable of creating chemical gradients e.g. vascular endothelial growth factor (VEGF), oligosaccharide fragments of hyaluronan, fibronectin, and collagen. There are certain

practical difficulties in maintaining effective chemical gradients for the cells in three dimensional cultures. This is particularly the limiting factor when the cells become dense.

Mechanical Cues:

The response of the cells to mechanical signals is complex and this may result in any one or more of the following:

i. Changes in the cell alignment.

ii. Deformation of cytoskeleton.

iii. Altered matrix formation.

iv. Synthesis of regulatory molecules (e.g growth factors, hormones).

There are mainly three mechanical cues governing cell populations:

1. Tensional forces.

2. Compressional forces.

3. Shear forces.

3. Cell Support Materials:

The support materials of cells largely determine the nature of adherent cells or cell types, and consequently tissue engineering.

There are a large number of support materials which may be broadly categorized as follows:

i. Traditional abiotic materials.

ii. Bio-prosthesis materials.

iii. Synthetic material.

iv. Natural polymers.

v. Semi-natural materials.

Traditional Abiotic Materials:

The traditional abiotic support materials include plastics, ceramics and metals. These materials cannot be resorbed or become biologically integrated into the tissues. Therefore, it is preferable to avoid the traditional materials in tissue engineering.

Bio-prosthesis Materials:

The natural materials modified to become biologically inert represent bio-prosthesis materials. They are formed by extensive chemical cross linking of natural tissues. For instance, the natural collagen-based connective tissue (e.g. porcine heart values) can be stabilized by treatment with glutaraldehyde.

The product formed is non-immunogenic that remains unchanged at the site of transplantation for several years. However, growth of some cells or even connective tissue can occur on bio-prosthesis materials. The design and fabrication of these materials is done in such a way that their functions are not affected by the surrounding host tissues.

Synthetic Materials:

A wide range of synthetic bioresorbable polymers are available as support materials. The most commonly used polymers in tissue engineering are poly (glycolic acid) (PGA), poly (lactic acid) (PLA), and copolymer PLGA (poly (lactic-co-glycolic acid). The composition and dimensions of these polymers can be so adjusted to make them stable in vivo, besides supporting in vivo cell growth.

There are certain advantages in using synthetic polymers:

i. Production is easy and relatively cheap.

ii. Composition of polymers is reproducible even in large scale production.

There are however, some disadvantages also:

i. Compatibility with cells is not as good as natural polymers.

ii. On degradation, they may form some products which cause undesirable cellular effects.

Natural Polymers:

The most widely used natural polymer materials are collagen-chondroitin sulfate aggregates. These materials are commercially available with varying composition under the trade name Integra. The other natural polymers for cell support are usually obtained by their aggregation in culture as it occurs in vivo e.g. collagen gels, fibrin glue, Matrigel and some polysaccharides.

Among the polysaccharides, chitosan and hyaluronan are used as hydrated gels. The natural polymers mainly act on the principle of intermolecular interaction within the polymers to promote intimate molecular packing. The so formed molecules can effectively serve as support materials.

Semi-natural Materials:

Semi-natural materials are derived from the natural macromolecular polymers or whole tissues. They are the modified materials to achieve aggregation or stabilization.

Some examples of semi-natural materials are listed below:

i. Chemically cross-linked hyaluronan, stabilized by benzyl esterification.

ii. Collagen cross-linked with agents such as tannic acid or carbodiimide.

4. Design and Engineering of Tissues:

The following surgical criteria are taken into consideration while dealing with tissue engineering:

i. Rapid restoration of the desired function.

ii. Ease of fixing the tissue.

iii. Minimal patient discomfort.

For designing tissue engineering, the source of donor cells is very critical. Use of patients own cells (autologous cells) is favoured to avoid immunological complications. Allogeneic cells are also used, particularly when the TE construct is designed for temporary repair. It is observed that when the cells are cultured and/or preserved (i.e. cryopreservation), the antigenicity of allogeneic cells is reduced.

Another important criteria in TE is the support material, its degradation products, cell adhesion characteristics and mechanical cues. The design and tissue engineering with respect to skin, urothelium and peripheral nerve are briefly described hereunder.

Tissue Engineered Skin:

It was first demonstrated in 1975 that human keratinocytes could be grown in the laboratory in a form suitable for grafting. Many improvements have been made since then. It is now possible to grow epithelial cells to produce a continuous sheet which progresses to form carnified layers.

The major difficulty with TE skin is the dermal layer possessing blood capillaries, nerves, sweat glands and other accessory organs. Some developments have occurred in recent years to produce implantable skin substitutes which may be regarded as tissue engineering skin constructs.

Integra:

This is a bio artificial material composed of collagen-glycosaminoglycan. Integra™ is not a true TE construct. It is mainly used to carry the seeded cells.

Dermagraft:

This is composed of poly (glycolic acid) polymer mesh seeded with human dermal fibroblasts from neonatal foreskins.

Apligraf:

This has human dermal fibroblasts seeded into collagen gel. A layer of human keratinocytes is then placed on the upper surface. The tissue constructs described above have limited shelf-life (about 5 days). However, they can integrate into the surrounding normal tissue and form a good skin cover. Further, there is no evidence of immunological complications with TE constructs.

Tissue Engineered Urothelium:

It is now possible to culture urothelial cells and bladder smooth muscle cells. This raises the hope that the construction of TE urothelium is possible. In fact, some success has been reported in the development of a functional bladder in dogs.

For this purpose, poly (glycolic acid) polymer base was shaped into a bladder and muscle cells were coated on the outer surface. The lumenal surface (i.e. inner surface) coated with pre-cultured urothelial cells. The bladder constructed in this way functioned almost like a normal one, and was maintained for about one year.

Tissue Engineered Peripheral Nerve Implants:

Peripheral nerve injury is a common occurrence of trauma and tumor resection surgery, often leading to irreversible muscle atrophy. Therefore, the repair of injured peripheral nerves assumes significance.

A diagrammatic representation of the basic design of a peripheral nerve implant is depicted in Fig 40.4.

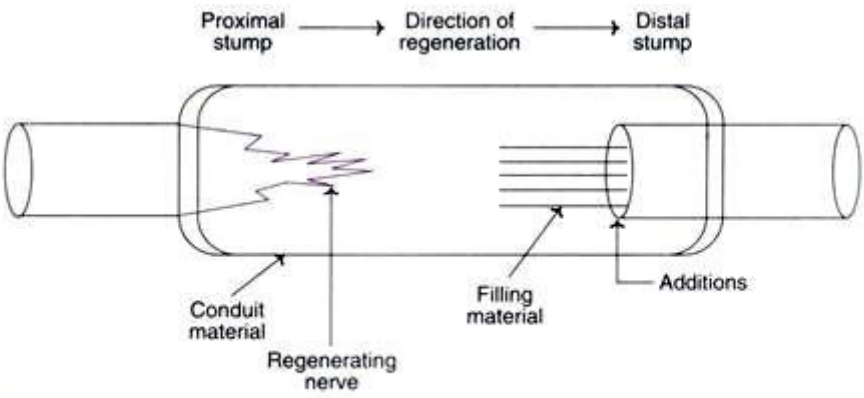


Fig. 40.4 : A diagrammatic representation of the basic design for peripheral nerve implant.

The regeneration of the injured nerve occurs from the proximal stump to rejoin at distal stump. The regeneration is guided by three types of substances.

Conduct material:

This is the outer layer and is the primary source of guidance. Conduct material is composed of collagen-glycosaminoglycan's, PLGA (poly lactic-co-glycolic acid), hyaluronan and fibronectin. All these are bioresorbable materials.

Filling material:

This supports the neural cells for regeneration, besides guiding the process of regeneration. Filling material contains collagen, fibrin, fibronectin and agarose.

Additives:

Additives include a large number of growth factors; neurotrophic factors (in different forms, combinations, ratios) e.g. fibroblast growth factor (FGF), nerve growth factor (NGF). Additions of Schwann cells or transfected fibroblasts promote nerve generation process.

Tissue Modeling:

Research is in progress to create tissue models in the form of artificial organs. Some of the recent development on experimental tissue modeling are briefly outlined.

Artificial liver:

Hepatocytes, cultured as spheroids or hepatocytes and fibroblasts cultured as hetero-spheroids can be used. They are held in the artificial support systems such as porous gelatin sponges, agarose or collagen. Addition of exogenous molecules is useful for the long – term culture of liver cells. Some progress has been reported in creating artificial liver as is evident from the hepatocytes three-dimensional structure and metabolic functions.

Artificial pancreas:

Spheroids of insulin secreting cells have been developed from mouse insulinoma beta cells. Some workers could implant fetal islet-like cell clusters under the kidneys of mice, although the functions were not encouraging due to limitation of oxygen supply.

Other Tissue Models:

Pituitary gland:

Multicellular spheroids could be created to study certain hormonal release e.g. luteinizing hormone (LH), following stimulation by luteinizing hormone releasing hormone (LHRH). Some success has also been achieved to create spheroids for the production of melatonin.

Thyroid gland:

Thyroid cell spheroids can be used for the study of cell adhesion, motility, and thyroid follicle biogenesis.

Brain cell cultures:

Three dimensional brain cell cultures have been used for the study of neural myelination and demyelination, neuronal regeneration, and neurotoxicity of lead. Aggregated brain cells are also used for the study of Alzheimer's disease and Parkinson's disease.

Heart cell cultures:

Aggregated heart cells have been used for the study of cardiac development and physiology.

Probable Questions:

1. Discuss stage sensitivity of teratogenesis.
2. Discuss mode of action of teratogens.
3. How dosage and time and route of administration affect action of teratogens?
4. How teratogenic effects of a drug are evaluated?
5. Discuss different types of stem cells.
6. What are the applications of stem cells?

Suggested Readings:

1. Embryology by N. Kumarsen
2. Developmental Biology by Veerbala Rastogi.
3. Embryology by M.P. Arora
4. Developmental Biology by Gilbert.

UNIT-IV

Structural genomics: Genome sequencing, High resolution genome mapping radiation hybrid mapping

UNIT-V

Physical mapping of genomes, FISH

Objective: In this unit you will learn physical maps, EST, SNPs as physical markers, radiation hybrids, FISH, optical mapping, gene maps, integration of physical and genetic maps; sequencing genomes: high-throughput sequencing, strategies of sequencing, recognition of coding and non-coding regions and annotation of genes, quality of genome-sequence data, base calling and sequence accuracy

Gene mapping describes the methods used to identify the locus of a gene and the distances between genes. The essence of all genome mapping is to place a collection of molecular markers onto their respective positions on the genome. Molecular markers come in all forms. Genes can be viewed as one special type of genetic markers in the construction of genome maps, and mapped the same way as any other markers.

Genetic and Physical Maps:

The convention is to divide genome mapping methods into two categories.

- Genetic mapping is based on the use of genetic techniques to construct maps showing the positions of genes and other sequence features on a genome. Genetic techniques include cross-breeding experiments or, in the case of humans, the examination of family histories (pedigrees).
- Physical mapping uses molecular biology techniques to examine DNA molecules directly in order to construct maps showing the positions of sequence features, including genes.

Genetic Mapping:

As with any type of map, a genetic map must show the positions of distinctive features. In a geographic map these markers are recognizable components of the landscape, such as rivers, roads and buildings. What markers can we use in a genetic landscape?

Genes were the first markers to be used:

The first genetic maps, constructed in the early decades of the 20th century for organisms such as the fruit fly, used genes as markers. This was many years before it was understood that genes are segments of DNA molecules. Instead, genes were looked upon as abstract entities responsible for the transmission of heritable characteristics from parent to offspring. To be useful in genetic analysis, a heritable characteristic has to exist in at least two alternative forms or phenotypes, an example being tall or short stems in the pea plants originally studied by Mendel. Each phenotype is specified by a different allele of the corresponding gene. To begin with, the only genes that could be studied were those specifying phenotypes that were distinguishable by visual examination. So, for example, the first fruit-fly maps showed the positions of genes for body color, eye color, wing shape and suchlike, all of these phenotypes being visible simply by looking at the flies with a low-power microscope or the naked eye. This approach was fine in the early days but geneticists soon realized that there were only a limited number of visual phenotypes whose inheritance could be studied, and in many cases their analysis was complicated because a single phenotype could be affected by more than one gene. For example, by 1922 over 50 genes had been mapped onto the four fruit-fly chromosomes, but nine of these were for eye color; in later research, geneticists studying fruit flies had to learn to distinguish between fly eyes that were colored red, light red, vermilion, garnet, carnation, cinnabar, ruby, sepia, scarlet, pink, cardinal, claret, purple or brown. To make gene maps more comprehensive it would be necessary to find characteristics that were more distinctive and less complex than visual ones.

The answer was to use biochemistry to distinguish phenotypes. This has been particularly important with two types of organisms

- microbes and humans. Microbes, such as bacteria and yeast, have very few visual characteristics so gene mapping with these organisms has to rely on biochemical phenotypes such as those listed in With humans it is possible to use visual characteristics, but since the 1920s studies of human genetic variation have been based largely on biochemical phenotypes that can be scored by blood typing. These phenotypes nclude not only the standard blood groups such as the ABO series (Yamamoto et al., 1990), but also variants of blood serum proteins and of immunological proteins such as the human leukocyte antigens (the HLA system). A big advantage of these markers is that many of the relevant genes have multiple alleles. For example, the gene called *HLA-DRB1* has at least 290 alleles and *HLA- B* has over 400. This is relevant because of the way in which gene mapping is carried out with humans. Rather than setting up many breeding experiments, which is the procedure with experimental organisms such as fruit flies or mice, data on inheritance of human genes have to be gleaned by examining the phenotypes displayed by members of a single family. If all the family members have the same allele for the gene being studied then no useful information can be obtained. It is therefore necessary for the relevant marriages to have occurred, by chance, between individuals with different alleles. This is much more likely if the gene being studied has 290 rather than two alleles.

DNA markers for genetic mapping:

Genes are very useful markers but they are by no means ideal. One problem, especially with larger genomes such as those of vertebrates and flowering plants, is that a map based entirely on genes is not very detailed. This would be true even if every gene could be mapped because, as we saw in Chapter 2, in most eukaryotic genomes the genes are widely spaced out with large gaps between them (see Figure). The problem is made worse by the fact that only a fraction of the total number of genes exists in allelic forms that can be distinguished conveniently. Gene maps are therefore not very comprehensive. We need other types of marker.

Mapped features that are not genes are called DNA markers. As

with gene markers, a DNA marker must have at least two alleles to be useful. There are three Types of DNA sequence feature that satisfy this requirement: restriction fragment length polymorphisms (RFLPs), simple sequence length polymorphisms (SSLPs), and single nucleotide polymorphisms (SNPs). Discussed underneath is the SNPs which is a comparatively modern technique used in current times:

Single nucleotide polymorphisms (SNPs):

These are positions in a genome where some individuals have one nucleotide (e.g. a G) and others have a different nucleotide (e.g. a C) (Figure). There are vast numbers of SNPs in every genome, some of which also give rise to RFLPs, but many of which do not because the sequence in which they lie is not

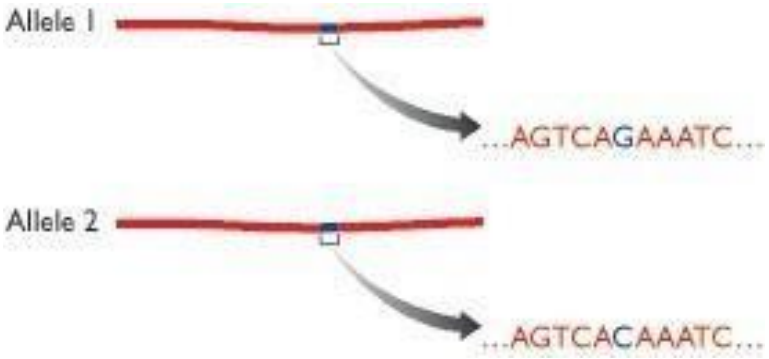
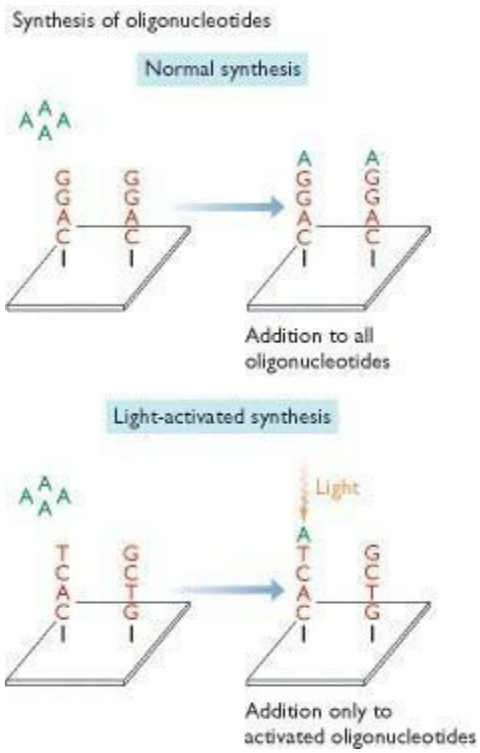


Figure : A single nucleotide polymorphism (SNP)

recognized by any restriction enzyme. In the human genome there are at least 1.42 million SNPs, only 100 000 of which result in an RFLP (SNP Group,2001).

Although each SNP could, potentially, have four alleles (because there are four nucleotides), most exist in just two forms, so these markers suffer from the same drawback as RFLPs with regard to human genetic mapping: there is a high possibility that a SNP does not display any variability in the family that is being studied. The advantages of SNPs are their abundant numbers and the fact that they can be typed by methods that do not involve gel electrophoresis. This is important because gel electrophoresis has proved difficult to automate so any detection method that uses it will be relatively slow and labor-intensive. SNP detection

is more rapid because it is based on oligonucleotide hybridization analysis. An oligonucleotide is a short single-stranded DNA molecule, usually less than 50 nucleotides in length,that is synthesized in the test tube. If the conditions are just right, then an oligonucleotide will hybridize with another DNA molecule only if the oligonucleotide forms a completely base-paired structure with the second molecule. If there is a single mismatch - a single position within the oligonucleotide that does not form a base pair - then hybridization does not occur (Figure below). Oligonucleotide hybridization can therefore discriminate between the two alleles of an SNP. Various screening strategies have been devised (Mir and Southern, 2000), including DNA chip technology and solution hybridization technique



A DNA chip is a wafer of glass or silicon, 2.0 cm^2 or less in area, carrying many different oligonucleotides in a high-density array. The DNA to be tested is labeled with a fluorescent marker and pipetted onto the surface of the chip. Hybridization is detected by examining the chip with a fluorescence microscope, the positions at which the fluorescent signal is emitted indicating which

oligonucleotides have hybridized with the test DNA. Many SNPs can therefore be scored in a single experiment (Wang et al., 1998; Gerhold et al.,1999).

Solution hybridization techniques are carried out in the wells of a microtiter tray, each well containing a different oligonucleotide, and use a detection system that can discriminate between unhybridized single-stranded DNA and the double- stranded product that results when an oligonucleotide hybridizes to the test DNA. Several systems have been developed, one of which makes use of a pair of labels comprising a fluorescent dye and a compound that quenches the fluorescent signal when brought into close proximity with the dye. The dye is attached to one end of an oligonucleotide and the quenching compound to the other end. Normally there is no fluorescence because the oligonucleotide is designed in such a way that the two ends base-pair to one another, placing the quencher next to the dye (Figure 5.9). Hybridization between oligonucleotide and test DNA disrupts this base pairing, moving the quencher away from the dye and enabling the fluorescent signal to be generated (Tyagi et al.,1998).

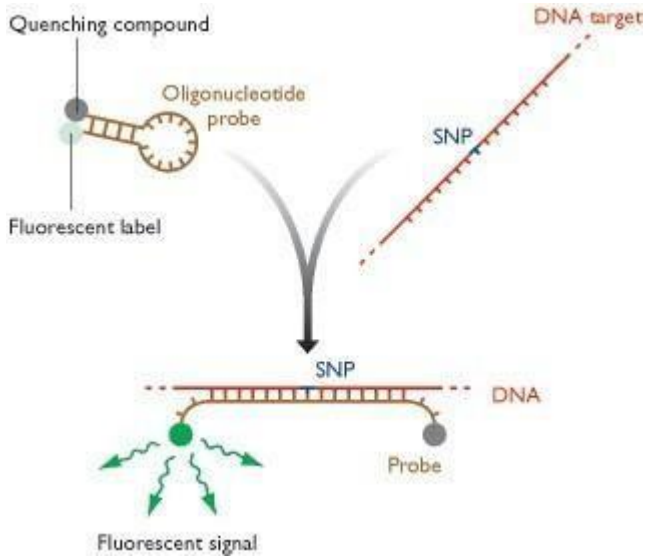


Figure: One way of detecting an SNP by solution hybridization.

Linkage analysis is the basis of genetic mapping:

Now that we have assembled a set of markers with which to construct a genetic map we can move on to look at the mapping techniques themselves. These techniques are all based on genetic

linkage, which in turn derives from the seminal discoveries in genetics made in the mid 19th century by Gregor Mendel.

Working out a genetic map from recombination frequencies:

Once Morgan had understood how partial linkage could be explained by crossing-over during meiosis he was able to devise a way of mapping the relative positions of genes on a chromosome. In fact the most important work was done not by Morgan himself, but by an undergraduate in his laboratory, Arthur Sturtevant (Sturtevant, 1913). Sturtevant assumed that crossing-over was a random event, there being an equalchance of it occurring at any position along a pair of lined-up chromatids. If this assumption is correct then two genes that are close together will be separated by crossovers less frequently than two genes that are more distant from one another. Furthermore, the frequency with which the genes are unlinked by crossovers will be directly proportional to how far apart they are on their chromosome. The recombination frequency is therefore a measure of the distance between two genes. If you work out the

recombination frequencies for different pairs of genes, you can construct a map of their relative positions on the chromosome.

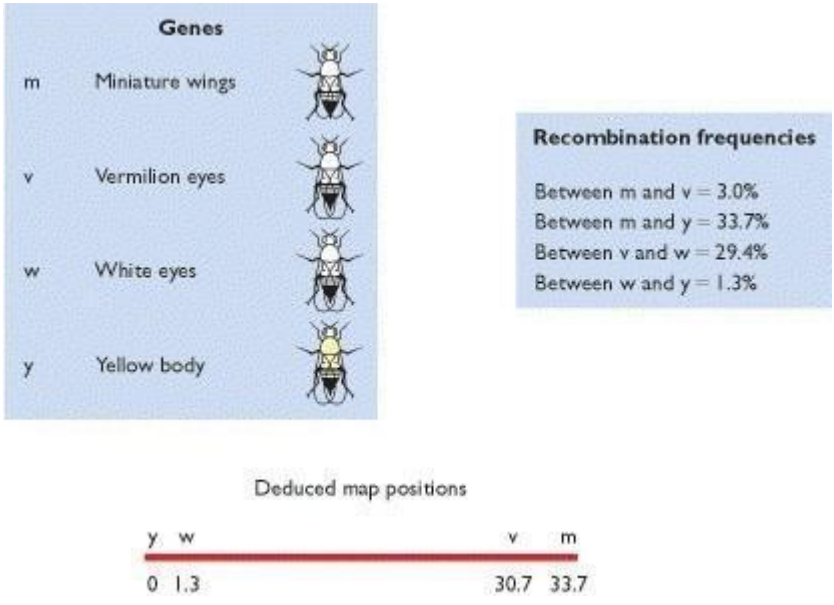


Figure: Working out a genetic map from recombination frequencies. The example is taken from the original

experiments carried out with fruit flies by Arthur Sturtevant. All four genes are on the X chromosome of the fruit fly.

It turns out that Sturtevant's assumption about the randomness of crossovers was not entirely justified. Comparisons between genetic maps and the actual positions of genes on DNA molecules, as revealed by physical mapping and DNA sequencing, have shown that some regions of chromosomes, called recombination hotspots, are more likely to be involved in crossovers than others. This means that a genetic map distance does not necessarily indicate the physical distance between two markers (see Figure). Also, we now realize that a single chromatid can participate in more than one crossover at the same time, but that there are limitations on how close together these crossovers can be, leading to more inaccuracies in the mapping procedure. Despite these qualifications, linkage analysis usually makes correct deductions about gene order, and distance estimates are sufficiently accurate to generate genetic maps that are of value as frameworks for genome sequencing projects.

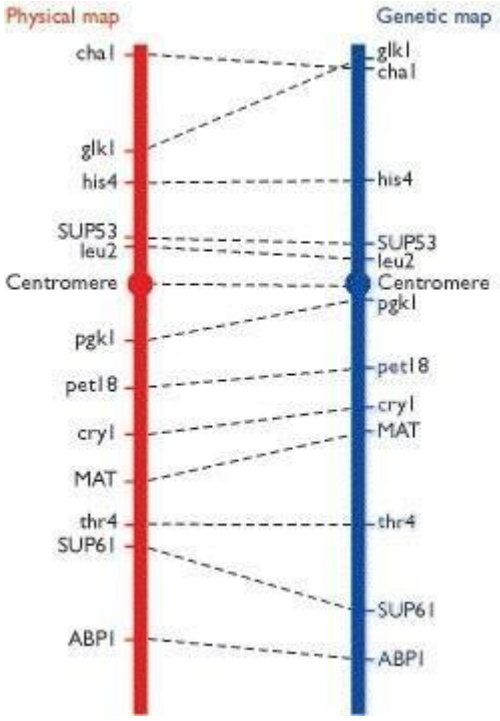


Figure: Comparison between the genetic and physical maps of *Saccharomyces cerevisiae* chromosome III. The comparison shows the discrepancies between the genetic and physical maps, the latter determined by DNA sequencing.

Physical Mapping:

A map generated by genetic techniques is rarely sufficient for directing the sequencing phase of a genome project. This is for two reasons:

- **The resolution of a genetic map depends on the number of crossovers that have been scored .**This is not a major problem for microorganisms because these can be obtained in huge numbers, enabling many crossovers to be studied, resulting in a highly detailed genetic map in which the markers are just a few kb apart. For example, when the *Escherichia coli* genome sequencing project began in 1990, the latest genetic map for this organism comprised over 1400 markers, an average of one per 3.3 kb. This was sufficiently detailed to direct the sequencing program without the need for extensive physical mapping. Similarly, the *Saccharomyces cerevisiae* project was supported

  by a fine-scale genetic map (approximately 1150 genetic markers, on average one per 10 kb). The problem with humans and most other eukaryotes is that it is simply not possible to obtain large numbers of progeny, so relatively few meiosis can be studied and the resolving power of linkage analysis is restricted. This means that genes that are several tens of kb apart may appear at the same position on the genetic map.

- **Genetic maps have limited accuracy .**Sturtevant's assumption that crossovers occur at random along chromosomes is only partly correct because the presence of recombination hotspots means that crossovers are more likely to occur at some points rather than at others. The effect that this can have on the accuracy of a genetic map was illustrated in 1992 when the complete sequence for S*. cerevisiae* chromosome III was published (Oliver et al., 1992), enabling the first direct comparison to be made between a genetic map and the actual positions of markers as shown by DNA sequencing (Figure). There were considerable discrepancies, even to the extent that one pair of genes had been ordered incorrectly by genetic

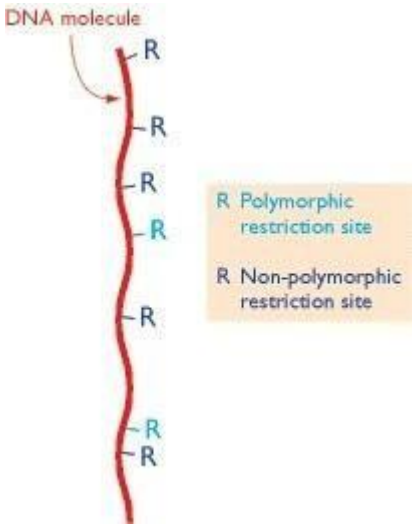
analysis. Bear in mind that *S. cerevisiae* is one of the two eukaryotes (fruit fly is the second) whose genomes have been subjected to intensive genetic mapping. If the yeast genetic map is inaccurate then how precise are the genetic maps of organisms subjected to less detailed analysis?

These two limitations of genetic mapping mean that for most eukaryotes a genetic map must be checked and supplemented by alternative mapping procedures before large-scale DNA sequencing begins. A plethora of physical mapping techniques has been developed to address this problem, the most important being:

- Restriction mapping, which locates the relative positions on a DNA molecule of the recognition sequences for restriction endonucleases;
- **Fluorescent *in situ* hybridization (**FISH**)**, in which marker locations are mapped by hybridizing a probe containing the marker to intact chromosomes;
- Sequence tagged site **(**STS**) mapping**, in which the positions of short sequences are mapped by PCR and/or hybridization analysis of genome fragments.

Restriction mapping:

Genetic mapping using RFLPs as DNA markers can locate the positions of polymorphic restriction sites within a genome, but very few of the restriction sites in a genome are polymorphic, so many sites are not mapped by this technique (Figure below). Could we increase the marker density on a genome map by using an alternative method to locate the positions of some of the non-polymorphic restriction sites? This is what restriction mapping achieves, although in practice the technique has limitations which mean that it is applicable only to relatively small DNA molecules. We will look first at the technique and then consider its relevance to genome mapping.



The basic methodology for restriction mapping:

The simplest way to construct a restriction map is to compare the fragment sizes produced when a DNA molecule is digested with two different restriction enzymes that recognize different target sequences. An example using the restriction enzymes *Eco*RI and *Bam*HI is shown in Figure. First, the DNA molecule is digested with just one of the enzymes and the sizes of the resulting fragments are measured by agarose gel electrophoresis. Next, the molecule is digested with the second enzyme and the resulting fragments again sized in an agarose gel. The results so far enable the number of restriction sites for each enzyme to be worked out, but do not allow their relative positions to be determined. Additional information is therefore obtained by cutting the DNA molecule with both enzymes together. In the example shown in Figure below this double restriction enables three of the sites to be mapped. However, a problem arises with the larger *Eco*RI fragment because this contains two *Bam*HI sites and there are two alternative possibilities for the map location of the outer one of these. The problem is solved by going back to the original DNA molecule and treating it again with *Bam*HI on its own, but this time preventing the digestion from going to completion by, for example, incubating the reaction for only a short time or using a suboptimal incubation temperature. This is called a partial restriction and leads to a more complex set of products, the complete restriction products now being supplemented with partially restricted fragments that still contain one or more uncut

*Bam*HI sites. In the example shown in Figure, the size of one of the partial restriction fragments is diagnostic and the correct map can be identified.

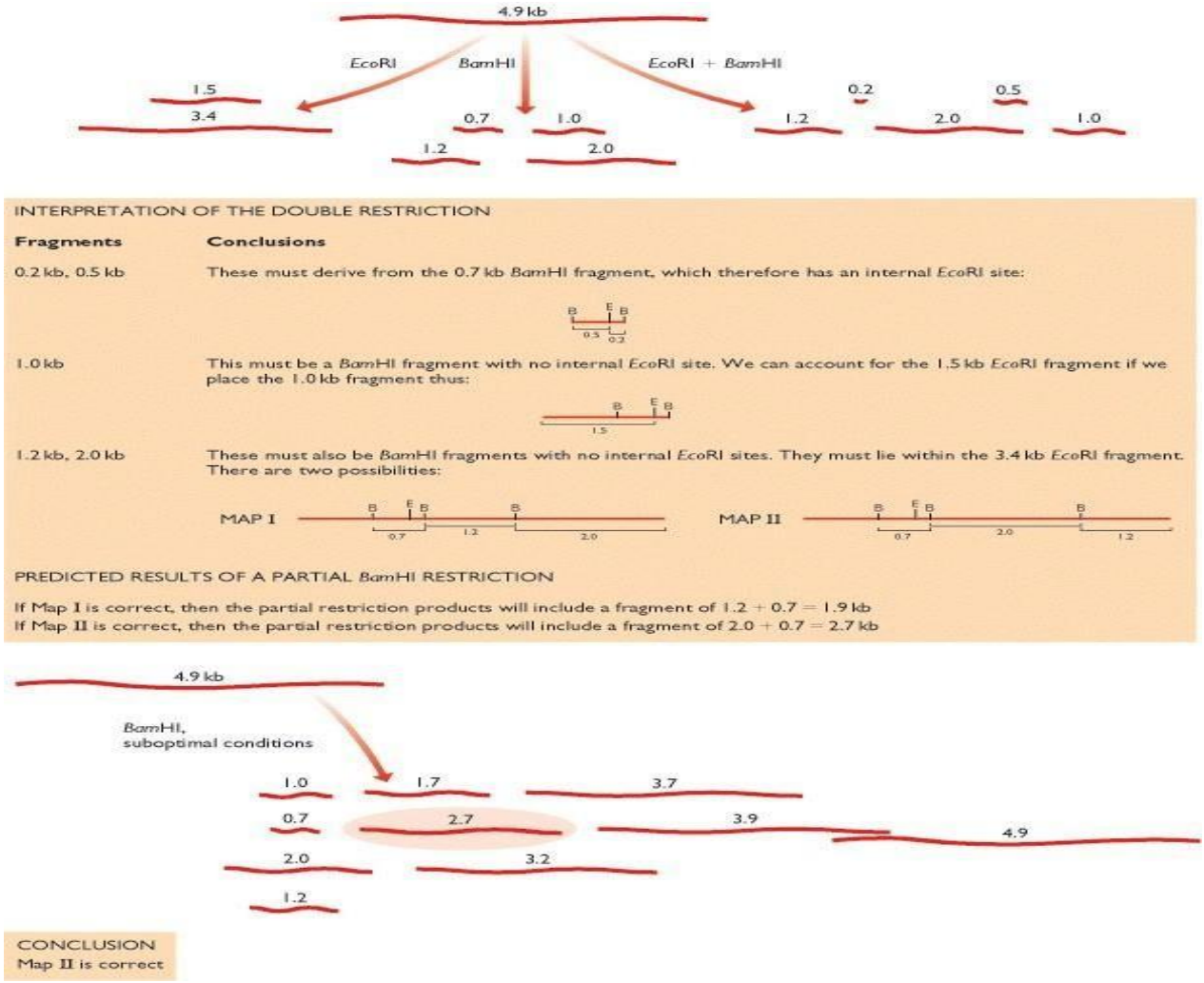


Figure : Restriction mapping. The objective is to map the EcoRI (E) and BamHI (B) sites in a linear DNA molecule of 4.9 kb. The results of single and double restrictions are shown at the top. The sizes of the fragments given after double restriction enable two alternative maps to be constructed, as explained in the central panel, the unresolved issue being the position of one of the three BamHI sites. The two maps are tested by a partial BamHI restriction (bottom), which shows that Map II is the correct one.

A partial restriction usually gives the information needed to complete a map, but if there are many restriction sites then this type of analysis becomes unwieldy, simply because there are so many different fragments to consider. An alternative strategy is simpler because it enables the majority of the fragments to be ignored. This is achieved by attaching a radioactive or other type of marker to each end of the starting DNA molecule before carrying out the partial digestion. The result is that many of the partial restriction products become 'invisible' because they do not contain an end-fragment and so do not show up when the agarose gel is screened for labeled products. The sizes of the partial restriction products that are visible enable unmapped sites to be positioned relative to the ends of the starting molecule.

Restriction maps are easy to generate if there are relatively few cut sites for the enzymes being used. However, as the number of cut sites increases, so also do the numbers of single, double and partial restriction products whose sizes must be determined and compared in order for the map to be constructed. Computer analysis can be brought into play but problems still eventually arise. A stage will be reached when a digest contains so many fragments that individual bands merge on the agarose gel, increasing the chances of one or more fragments being measured incorrectly or missed out entirely. If several fragments have similar sizes then even if they can all be identified, it may not be possible to assemble them into an unambiguous map.

Restriction mapping is therefore more applicable to small rather than large molecules, with the upper limit for the technique depending on the frequency of the restriction sites in the molecule being mapped. In practice, if a DNA molecule is less than 50 kb in length it is usually possible to construct an unambiguous restriction map for a selection of enzymes with six-nucleotide recognition sequences. Fifty kb is of course way below the minimum size for bacterial or eukaryotic chromosomes, although it does cover a few viral and organelle genomes, and whole- genome restriction maps have indeed been important in directing sequencing projects with these small molecules.

Restriction maps are equally useful after bacterial or eukaryotic genomic DNA has been cloned, if the cloned fragments are less than 50 kb, because a detailed restriction map can then be built up as a preliminary to sequencing the cloned region. This is an important application of restriction mapping in sequencing projects with large genomes, but is there any possibility of using restriction analysis for the more general mapping of entire genomes larger than 50 kb?

The answer is a qualified 'yes', because the limitations of restriction mapping can be eased slightly by choosing enzymes expected to have infrequent cut sites in the target DNA molecule. These 'rare cutters' fall into two categories:

- **Enzymes with seven- or eight-nucleotide recognition sequences.** A few restriction enzymes cut at seven- or eight-nucleotide recognition sequences. Examples are SapI (5′- GCTCTTC-3′) and SgfI (5′-GCGATCGC- 3′). The seven-nucleotide enzymes would be expected, on average, to cut a DNA molecule with a GC content of 50% once every 4^7 = 16 384 bp. The eight-nucleotide enzymes should cut once every 4^8 = 65 536 bp. These figures compare with 4^6 = 4096 bp for six-nucleotide enzymes such as BamHI and EcoRI. Seven- and eight-nucleotide cutters are often used in restriction mapping of large molecules but the approach is not as useful as it might be simply because not many of these enzymes are known.

- **Enzymes whose recognition sequences contain motifs that are rare in the target** DNA **.** Genomic DNA molecules do not have random sequences and some are significantly deficient in certain motifs. For example, the sequence 5′-CG-3′ is rare in human DNA because human cells possess an enzyme that adds a methyl group to carbon 5 of the C nucleotide in this sequence. The resulting 5-methylcytosine is unstable and tends to undergo deamination to give thymine (Figure below). The consequence is that during human evolution many of the 5′-CG-3′ sequences that were originally in our genome have become converted to 5′-TG-3′. Restriction enzymes that recognize a site containing 5′-CG-3′ therefore cut

human DNA relatively infrequently. Examples are SmaI (5′- CCCGGG-3′), which cuts human DNA on average once every 78 kb, and BssHII (5′- GCGCGC-3′) which cuts once every 390 kb. Note that NotI, an eight-nucleotide cutter, also targets 5′-CG-3′ sequences (recognition sequence 5′-GCGGCCGC-3′) and cuts human DNA very rarely - approximately once every 10Mb.

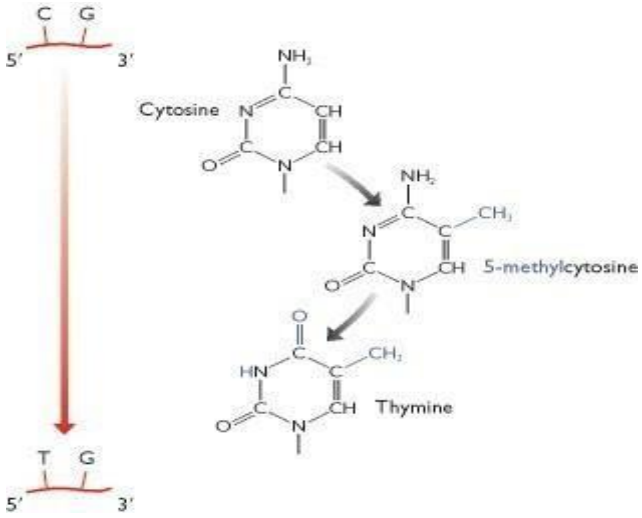


Figure: The sequence 5′-CG-3′ is rare in human DNA because of methylation of the C, followed by deamination to give T.

The potential of restriction mapping is therefore increased by using rare cutters. It is still not possible to construct restriction maps of the genomes of animals and plants, but it is feasible to use the technique with large cloned fragments, and the smaller DNA molecules of prokaryotes and lower eukaryotes such as yeast and fungi.

If a rare cutter is used then it may be necessary to employ a special type of agarose gel electrophoresis to study the resulting restriction fragments. This is because the relationship between the length of a DNA molecule and its migration rate in an electrophoresis gel is not linear, the resolution decreasing as the molecules get longer (Figure A). This means that it is not possible to separate molecules more than about 50 kb in length because all of these longer molecules run as a single slowly migrating band in a standard agarose gel. To separate them it is necessary to replace the linear electric field used in conventional gel electrophoresis with a more complex field. An example is

provided by orthogonal field alternation gel electrophoresis (OFAGE), in which the electric field alternates between two pairs of electrodes, each positioned at an angle of 45° to the length of the gel (Figure B). The DNA molecules still move down through the gel, but each change in the field forces the molecules to realign. Shorter molecules realign more quickly than longer ones and so migrate more rapidly through the gel. The overall result is that molecules much longer than those separated by conventional gel electrophoresis can be resolved. Related techniques include CHEF (contour clamped homogeneous electric fields) and FIGE (field inversion gel electrophoresis).

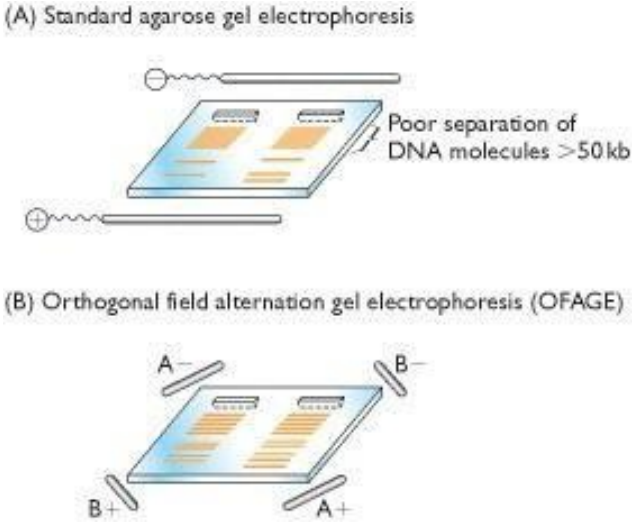
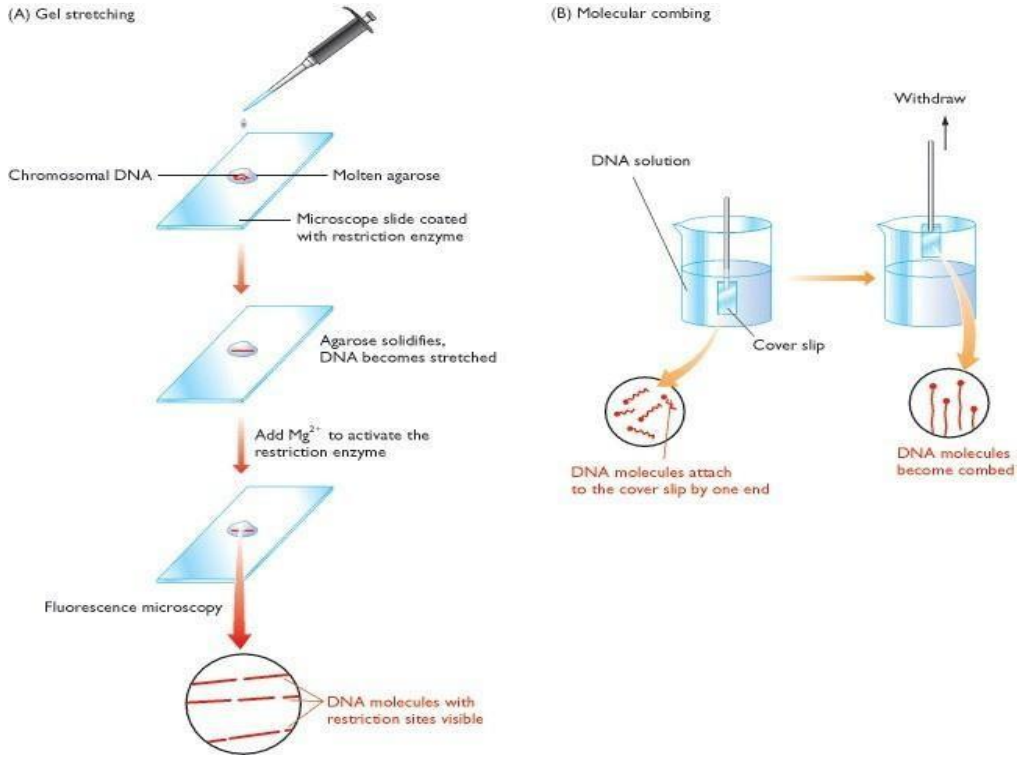


Figure : Conventional and non-conventional agarose gel electrophoresis. (A) In standard agarose gel electrophoresis the electrodes are placed at either end of the gel and the DNA molecules migrate directly towards the positive electrode. Molecules longer than about 50 kb cannot be separated from one another in this way. (B) In OFAGE, the electrodes are placed at the corners of the gel, with the field pulsing between the A pair and the B pair.

Optical mapping:

It is also possible to use methods other than electrophoresis to map restriction sites in DNA molecules. With the technique called optical mapping (Schwartz et al., 1993), restriction sites are directly located by looking at the cut DNA molecules with a

microscope. The DNA must first be attached to a glass slide in such a way that the individual molecules become stretched out, rather than clumped together in a mass. There are two ways of doing this: gel stretching and molecular combing. To prepare gel-stretched DNA fibres (Schwartz et al., 1993), chromosomal DNA is suspended in molten agarose and placed on a microscope slide. As the gel cools and solidifies, the DNA molecules become extended (Figure A). To utilize gel stretching in optical mapping, the microscope slide onto which the molten agarose is placed is first coated with a restriction enzyme. The enzyme is inactive at this stage because there are no magnesium ions, which the enzyme needs in order to function. Once the gel has solidified it is washed with a solution containing magnesium chloride, which activates the restriction enzyme. A fluorescent dye is added, such as DAPI (4,6- diamino-2-phenylindole dihydrochloride), which stains the DNA so that the fibres can be seen when the slide is examined with a high-power fluorescence microscope. The restriction sites in the extended molecules gradually become



gaps as the degree of fibre extension is reduced by the natural springiness of the DNA, enabling the relative positions of the cuts to be recorded.

Figure: Gel stretching and molecular combing. (A) To carry out gel stretching, molten agarose containing chromosomal DNA molecules is pipetted onto a microscope slide coated with a restriction enzyme. As the gel solidifies, the DNA molecules become stretched.

In molecular combing (Michalet et al., 1997), the DNA fibres are prepared by dipping a silicone- coated cover slip into a solution of DNA, leaving it for 5 minutes (during which time the DNA molecules attach to the cover slip by their ends), and then removing the slip at a constant speed of 0.3 mm s^{-1} (Figure B). The force required to pull the DNA molecules through the meniscus causes them to line up. Once in the air, the surface of the cover slip dries, retaining the DNA molecules as an array of parallel fibres.

Optical mapping was first applied to large DNA fragments cloned in YAC and BAC vectors. More recently, the feasibility of using this technique with genomic DNA has been proven with studies of a 1-Mbchromosome of the malaria parasite Plasmodium falciparum (Jing et al., 1999), and the two chromosomes and single mega plasmid of the bacterium *Deinococcus radiodurans* (Lin et al., 1999)

Fluorescent in situ hybridization (FISH):

The optical mapping method described above provides a link to the second type of physical mapping procedure that we will consider - FISH (Heiskanen et al., 1996). As in optical mapping, FISH enables the position of a marker on a chromosome or extended DNA molecule to be directly visualized. In optical mapping the marker is a restriction site and it is visualized as a gap in an extended DNA fibre. In FISH, the marker is a DNA sequence that is visualized by hybridization with a fluorescent probe.

In situ hybridization with radioactive or fluorescent probes:

In situ hybridization is a version of hybridization analysis in which an intact chromosome is examined by probing it with a labeled DNA molecule. The position on the chromosome at which hybridization occurs provides information about the map location of the DNA sequence used as the probe (Figure). For the

method to work, the DNA in the chromosome must be made single stranded ('denatured') by breaking the base pairs that hold the double helix together. Only then will the chromosomal DNA be able to hybridize with the probe. The standard method for denaturing chromosomal DNA without destroying the morphology of the chromosome is to dry the preparation onto a glass microscope slide and then treat with formamide.

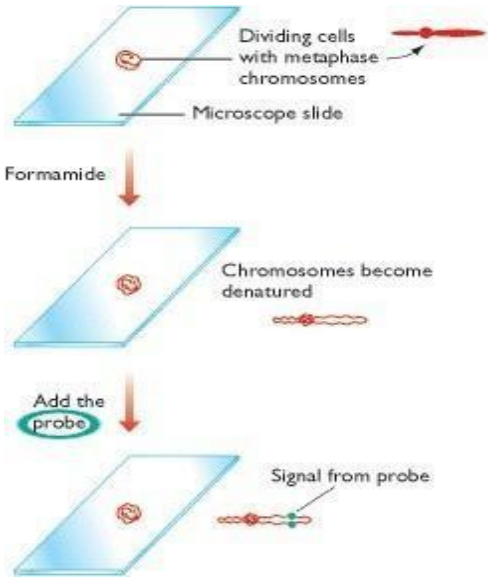


Figure :Fluorescent in situ hybridization. A sample of dividing cells is dried onto a microscope slide and treated with formamide so that the chromosomes become denatured but do not lose their characteristic metaphase morphologies. The position at which the probe hybridizes to the chromosomal DNA is visualized by detecting the fluorescent signal emitted by the labeled DNA.

In the early versions of in situ hybridization the probe was radioactively labeled but this procedure was unsatisfactory because it is difficult to achieve both sensitivity and resolution with a radioactive label, two critical requirements for successful in situ hybridization. Sensitivity requires that the radioactive label has a high emission energy (an example of such a radiolabel is ^{32}P), but if the radiolabel has a high emission energy then it scatters its signal and so gives poor resolution. High resolution is

possible if a radiolabel with low emission energy, such as ^{3}H, is used, but these have such low sensitivity that lengthy exposures are needed, leading to a high background and difficulties in discerning the genuine signal.

These problems were solved in the late 1980s by the development of non- radioactive fluorescent DNA labels. These labels combine high sensitivity with high resolution and are ideal for in situ hybridization. Fluoro-labels with different colored emissions have been designed, making it possible to hybridize a number of different probes to a single chromosome and distinguish their individual hybridization signals, thus enabling the relative positions of the probe sequences to be mapped. To maximize sensitivity, the probes must be labeled as heavily as possible, which in the past has meant that they must be quite lengthy DNA molecules - usually cloned DNA fragments of at least 40 kb. This requirement is less important now that techniques for achieving heavy labeling with shorter molecules have been developed. As far as the construction of a physical map is concerned, a cloned DNA fragment can be looked upon as simply another type of marker, although in practice the use of clones as markers adds a second dimension because the cloned DNA is the material from which the DNA sequence is determined. Mapping the positions of clones therefore provides a direct link between a genome map and its DNA sequence.

If the probe is a long fragment of DNA then one potential problem, at least with higher eukaryotes, is that it is likely to contain examples of repetitive DNA sequences and so may hybridize to many chromosomal positions, not just the specific point to which it is perfectly matched. To reduce this non-specific hybridization, the probe, before use, is mixed with unlabeled DNA from the organism being studied. This DNA can simply be total nuclear DNA (i.e. representing the entire genome) but it is better if a fraction enriched for repeat

sequences is used. The idea is that the unlabeled DNA hybridizes to the repetitive DNA sequences in the probe, blocking these so that the subsequent *in situ* hybridization is driven wholly by the unique sequences (Lichter et al., 1990). Non-specific hybridization is therefore reduced or eliminated entirely (Figure

).

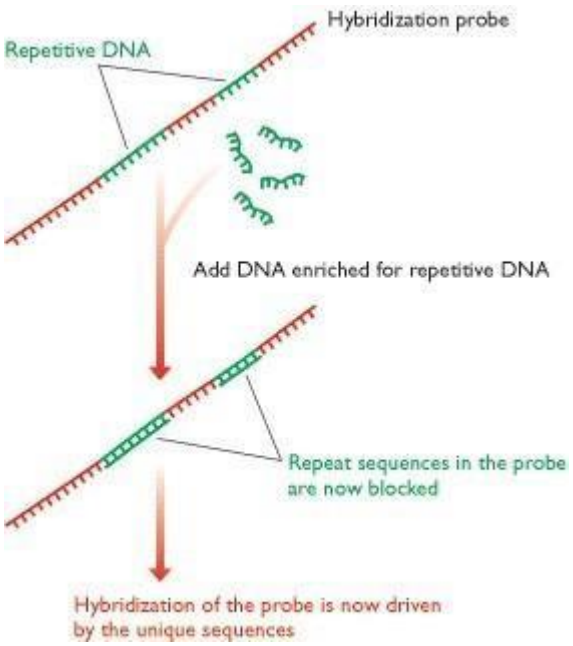


Figure : A method for blocking repetitive DNA sequences in a hybridization probe. In this example the probe molecule contains two genome-wide repeat sequences (shown in green). If these sequences are not blocked then the probe will hybridize to any copies of these genome-wide repeats in the target DNA. To block the repeat sequences, the probe is pre-hybridized with a DNA fraction enriched for repetitive

FISH in action:

FISH was originally used with metaphase chromosomes . These chromosomes, prepared from nuclei that are undergoing division, are highly condensed and each chromosome in a set takes up a recognizable appearance, characterized by the position of its centromere and the banding pattern that emerges after the chromosome preparation is stained. With metaphase chromosomes, a fluorescent signal obtained by FISH is mapped by measuring its position relative to the end of the short arm of the chromosome (the FLpter value). A disadvantage is that the highly condensed nature of metaphase chromosomes means that only low-

resolution mapping is possible, two markers having to be at least 1 Mb apart to be resolved as separate hybridization signals (Trask et al., 1991). This degree of resolution is insufficient for the construction of useful chromosome maps, and the main application of metaphase FISH has been in determining the chromosome on which a new marker is located, and providing a rough idea of its map position, as a preliminary to finer scale mapping by other methods.

For several years these 'other methods' did not involve any form of FISH, but since 1995 a range of higher resolution FISH techniques has been developed. With these techniques, higher resolution is achieved by changing the nature of the chromosomal preparation being studied. If metaphase chromosomes are too condensed for fine-scale mapping then we must use chromosomes that are more extended. There are two ways of doing this (Heiskanen et al., 1996):

- **Mechanically stretched chromosomes** can be obtained by modifying the preparative method used to isolate chromosomes from metaphase nuclei. The inclusion of a centrifugation step generates shear forces which can result in the chromosomes becoming stretched to up to 20 times their normal length. Individual chromosomes are still recognizable and FISH signals can be mapped in the same way as with normal metaphase chromosomes. The resolution is significantly improved and markers that are 200–300kb apart can be distinguished.

- **Non-metaphase chromosomes** can be used because it is only during metaphase that chromosomes are highly condensed: at other stages of the cell cycle the chromosomes are naturally unpacked. Attempts have been made to use prophase nuclei because in these

- the chromosomes are still sufficiently condensed for individual ones to be identified. In practice, however, these preparations provide no advantage over mechanically stretched chromosomes. Interphase chromosomes are more useful because this stage of the cell cycle (between nuclear divisions) is when the chromosomes

  are most unpacked. Resolution down to 25 kb is possible, but chromosome morphology is lost so there are no external reference points against which to map the position of the probe. This technique is therefore used after preliminary map information has been obtained, usually as a means of determining the order of a series of markers in a small region of a chromosome.

Interphase chromosomes contain the most unpacked of all cellular DNA molecules. To improve the resolution of FISH to better than 25 kb it is therefore necessary to abandon intact chromosomes and instead use purified DNA. This approach, called fibre-FISH, makes use of DNA prepared by gel stretching or molecular combing and can distinguish markers that are less than 10 kb apart.

Sequence tagged site (STS)mapping:

To generate a detailed physical map of a large genome we need, ideally, a high-resolution mapping procedure that is rapid and not technically demanding. Neither of the two techniques that we have considered so far - restriction mapping and FISH - meets these requirements. Restriction mapping is rapid, easy, and provides detailed information, but it cannot be applied to large genomes. FISH can be applied to large genomes, and modified versions such as fibre-FISH can give high-resolution data, but FISH is difficult to carry out and data accumulation is slow, map positions for no more than three or four markers being obtained in a single experiment. If detailed physical maps are to become a reality then we need a more powerful technique.

At present the most powerful physical mapping technique, and the one that has been responsible for generation of the most detailed maps of large genomes, is STS mapping. A sequence tagged site or **STS** is simply a short DNA sequence, generally between 100 and 500 bp in length, that is easily recognizable and occurs only once in the chromosome or genome being studied. To map a set of STSs, a collection of overlapping DNA fragments from a single chromosome or from the entire genome is needed. In the example shown in Figure, a fragment collection has been prepared from a single chromosome, with each point along the chromosome represented on average five times in the collection. The data from which the map will be derived are obtained by determining which fragments contain which STSs. This can be done by hybridization analysis but PCR is generally used because it is quicker and has proven to be more amenable to automation. The chances of two STSs being present on the same fragment will, of course, depend on how close together they are in the genome. If they are very close then there is a good chance that they will always be on the same fragment; if they are further apart then sometimes they will be on the same fragment and sometimes they will not (Figure below). The data can therefore be used to calculate the distance between two markers, in a manner analogous to the way in which map distances are determined by linkage analysis.

Remember that in linkage analysis a map distance is calculated from the frequency at which crossovers occur between two markers. STS mapping is essentially the same, except that each map distance is based on the frequency at which *breaks* occur between two markers.

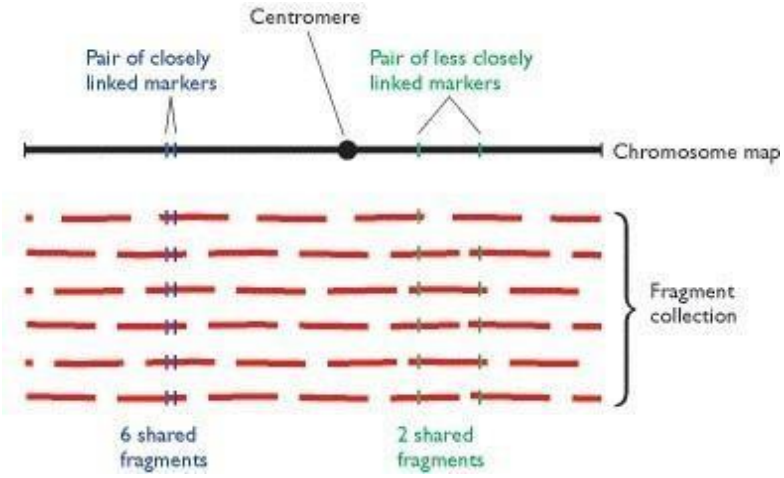


Figure : A fragment collection suitable for STS mapping. The fragments span the entire length of a chromosome, with each point on the chromosome present in an average of five fragments. The two blue markers are close together on the chromosome map and there is a high probability that they will be found on the same fragment. The two green markers are more distant from one another and so are less likely to be found on the same fragment

Any unique DNA sequence can be used as an STS:

To qualify as an STS, a DNA sequence must satisfy two criteria. The first is that its sequence must be known, so that a PCR assay can be set up to test for the presence or absence of the STS on different DNA fragments. The second requirement is that the STS must have a unique location in the chromosome being studied, or in the genome as a whole if the DNA fragment set covers the entire genome. If the STS sequence occurs at more than one position then the mapping data will be ambiguous. Care must therefore be taken to ensure that STSs do not include sequences found in repetitive DNA.

These are easy criteria to satisfy and STSs can be obtained in many ways, the most common sources being **expressed sequence tags (ESTs)**, SSLPs, and **random genomic sequences**.

Expressed sequence tags (ESTs): These are short sequences obtained by analysis of cDNA clones (Marra et al., 1998). Complementary DNA is prepared by converting an mRNA preparation into double-stranded DNA (Figure). Because the mRNA in a cell is derived from protein-coding genes, cDNAs and the ESTs obtained from them represent the genes that were being expressed in the cell from which the mRNA was prepared. ESTs are looked upon

as a rapid means of gaining access to the sequences of important genes, and they are valuable even if their sequences are incomplete. An EST can also be used as an STS, assuming that it comes from a unique gene and not from a member of a gene family in which all the genes have the same or very similar sequences.

SSLPs: In earlier sections, we examined the use of microsatellites and other SSLPs in genetic mapping. SSLPs can also be used as STSs in physical mapping. SSLPs that are polymorphic and have already been mapped by linkage analysis are particularly valuable as they provide a direct connection between the genetic and physical maps.

Random genomic sequences: These are obtained by sequencing random pieces of cloned genomic DNA, or simply by downloading sequences that have been deposited in the databases.

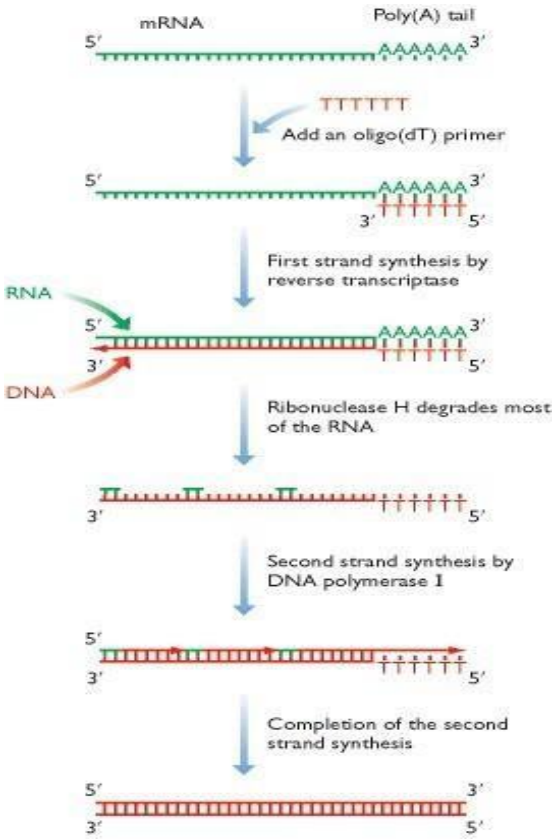


Figure : Most eukaryotic mRNAs have a poly(A) tail at their 3′ end . This series of A nucleotides is used as the priming site for the first stage of cDNA synthesis, carried out by reverse transcriptase a DNA polymerase that copies an RNA template. The primer is a short synthetic DNAoligonucleotide, typically 20 nucleotides in length, made up entirely of Ts (an 'oligo(dT)' primer). When the first strand synthesis has been completed, the preparation is treated with ribonuclease H, which specifically degrades the RNA component of an RNA-DNA hybrid. Under the conditions used, the enzyme does not degrade all of the RNA, instead leaving short segments that prime the second DNA strand synthesis reaction, this one catalyzed by DNA polymerase-I.

Fragments of DNA for STS mapping:

The second component of an STS mapping procedure is the collection of DNA fragments spanning the chromosome orgenome being studied. This collection is sometimes called the mapping reagent and at present there are two ways in which it can be assembled: as a clone library and as a panel of radiation hybrids. We will consider radiation hybrids first.

A radiation hybrid is a rodent cell that contains fragments of chromosomes

from a second organism (McCarthy, 1996). The technology was first developed in the 1970s when it was discovered that exposure of human cells to X-ray doses of 3000–8000 rads causes the chromosomes to break up randomly into fragments, larger X-ray doses producing smaller fragments (Figure below). This treatment is of course lethal for the human cells, but the chromosome fragments can be propagated if the irradiated cells are subsequently fused with non-irradiated hamster or other rodent cells. Fusion is stimulated either chemically with polyethylene glycol or by exposure to Sendai virus. Not all of the hamster cells take up chromosome fragments so a means of identifying the hybrids is needed. The routine selection process is to use a hamster cell line that is unable to make either thymidine kinase (TK) or hypoxanthine phosphoribosyl transferase (HPRT), deficiencies in either of these two enzymes being lethal when the cells are grown in a medium containing a mixture of hypoxanthine, aminopterin and thymidine (HAT medium). After fusion, the cells are placed in HAT medium. Those that grow are hybrid hamster cells that have acquired human DNA fragments that include genes for the human TK and HPRT enzymes, which are synthesized inside the hybrids, enabling these cells to grow in the selective medium. The treatment results in hybrid cells that contain a random selection of human DNA fragments inserted into the hamster chromosomes. Typically the fragments are 5–10 Mb in size, with each cell containing fragments equivalent to 15–35% of the human genome. The collection of cells is called a radiation hybrid panel and can be used as a mapping reagent in STS mapping, provided that the PCR assay used to identify the STS does not amplify the equivalent region of DNA from the hamster genome.

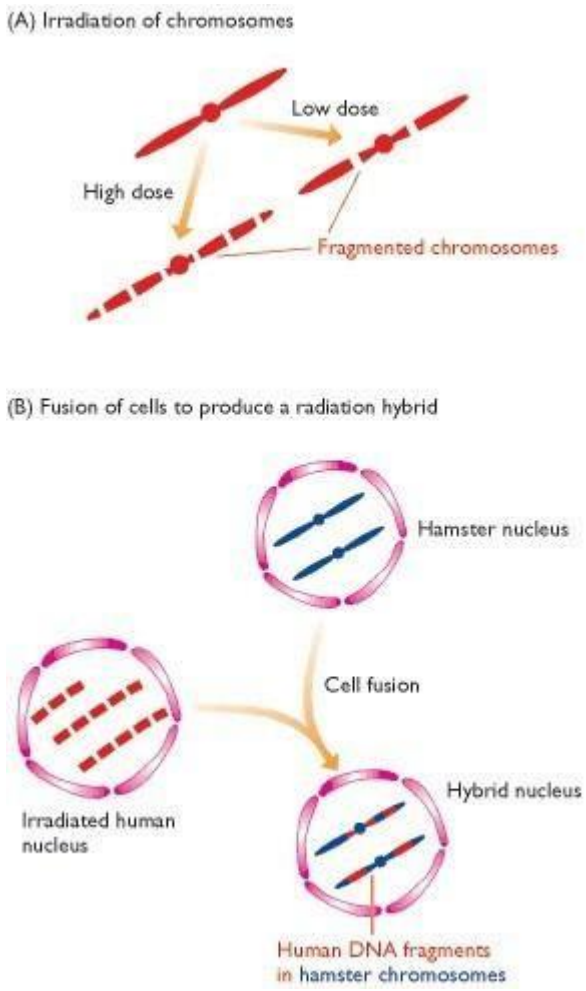


Figure: Radiation hybrids. (A) The result of irradiation of human cells: the chromosomes break into fragments, smaller fragments generated by higher X-ray doses. In (B), a radiation hybrid is produced by fusing an irradiated human cell with an untreated hamster cell. For clarity, only the nuclei are shown.

A second type of radiation hybrid panel, containing DNA from just one human chromosome, can be constructed if the cell line that is irradiated is not a human one but a second type of rodent hybrid. Cytogeneticists have developed a number of rodent cell lines in which a single human chromosome is stably propagated in the rodent nucleus. If a cell line of this type is irradiated and fused with hamster cells, then the hybrid hamster cells obtained after selection will contain either human or mouse chromosome fragments, or a mixture of both. The ones containing human DNA can be identified by probing with a human-specific genome-wide repeat sequence, such as the

short interspersed nuclear element (SINE) called Alu, which has a copy number of just over 1 million and so occurs on average once every 4 kb in the human genome. Only cells containing human DNA will hybridize to Alu probes, enabling the uninteresting mouse hybrids to be discarded and STS mapping to be directed at the cells containing human chromosome fragments.

Radiation hybrid mapping of the human genome was initially carried out with chromosome- specific rather than whole-genome panels because it was thought that fewer hybrids would be needed to map a single chromosome than would be needed to map the entire genome. It turns out that a high-resolution map of a single human chromosome requires a panel of 100– 200 hybrids, which is about the most that can be handled conveniently in a PCR screening program. But whole-genome and single-chromosome panels are constructed differently, the former involving irradiation of just human DNA, and the latter requiring irradiation of a mouse cell containing much mouse DNA and relatively little human DNA. This means that the human DNA content per hybrid is much lower in a single-chromosome panel than in a whole-genome panel. It transpires that detailed mapping of the entire human genome is possible with fewer than 100 whole-genome radiation hybrids, so whole-genome mapping is no more difficult than single- chromosome mapping. Once this was realized, whole-genome radiation hybrids became a central component of the mapping phase of the Human Genome Project. Whole-genome libraries are also being used for STS mapping of other mammalian genomes and for those of the zebra fish and the chicken (McCarthy,1996).

A clone library can also be used as the mapping reagent for STS analysis:

A preliminary to the sequencing phase of a genome project is to break the genome or isolated chromosomes into fragments and to clone each one in a high-capacity vector, one able to handle large fragments of DNA . This results in a clone library, a collection of DNA fragments, which, in this case, have an average size of several hundred kb. As well as supporting the sequencing work, this type of clone library can also be used as a mapping reagent in STS analysis.

As with radiation hybrid panels, a clone library can be prepared from genomic DNA, and so represents the entire genome, or a chromosome-specific library can be made if the starting DNA comes from just one type of chromosome. The latter is possible because individual chromosomes can be separated by flow cytometry. To carry out this technique, dividing cells (ones with condensed chromosomes) are carefully broken open so that a mixture of intact chromosomes is obtained. The chromosomes are then stained with a

fluorescent dye. The amount of dye that a chromosome bind depends on its size, so larger chromosomes bind more dye and fluoresce more brightly than smaller ones. The chromosome preparation is diluted and passed through a fine aperture, producing a stream of droplets, each one containing a single chromosome. The droplets pass through a detector that measures the amount of fluorescence, and hence identifies which droplets contain the particular chromosome being sought. An electric charge is applied to these drops, and no others, enabling the droplets containing the desired chromosome to be deflected and separated

from the rest. What if two different chromosomes have similar sizes, as is the case with human chromosomes 21 and 22? These can usually be separated if the dye that is used is not one that binds non-specifically to DNA, but instead has a preference for AT- or GC-rich regions. Examples of such dyes are Hoechst 33258 and chromomycin A_3, respectively. Two chromosomes that are the same size rarely have identical GC contents, and so can be distinguished by the amounts of AT- or GC-specific dye that they bind.

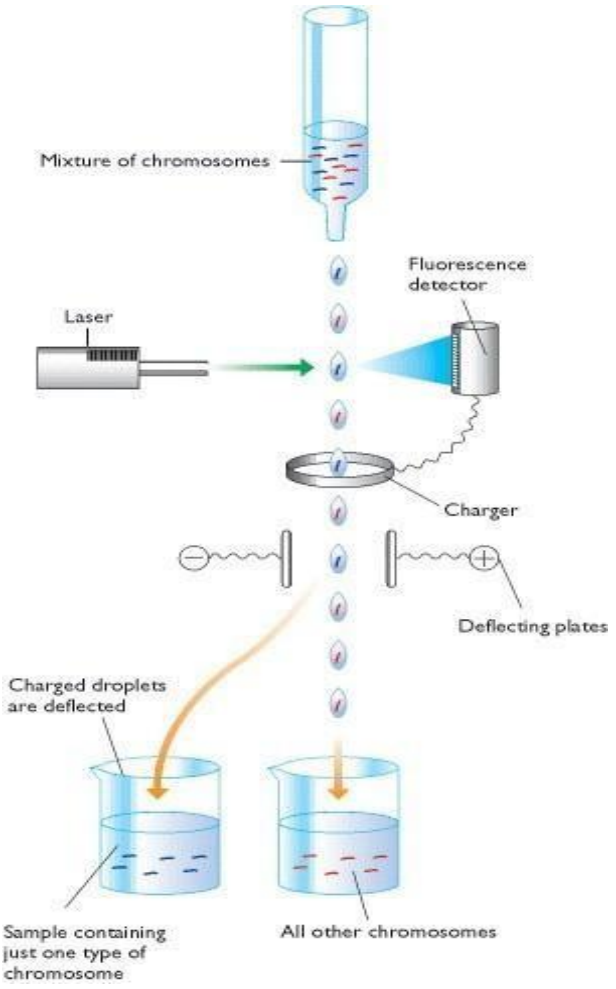


Figure : Separating chromosomes by flow cytometry A mixture of fluorescently stained chromosomes is passed through a small aperture so that each drop that emerges contains just one chromosome. The fluorescence detector identifies the signal from drops containing the correct chromosome and applies an electric charge to these drops. When the drops reach the electric plates, the charged ones are deflected into a separate beaker. All other drops fall straight through the deflecting plates and are collected in the waste beaker.

Compared with radiation hybrid panels, clone libraries have one important advantage for STS mapping. This is the fact that the individual clones can subsequently provide the DNA that is actually sequenced. The data resulting from STS analysis, from which the physical map is generated, can equally well be used to determine which clones contain overlapping DNA fragments, enabling a clone contig to be built up (Figure); for other methods for assembling clone contigs). This assembly of overlapping clones can be used as the base material for a lengthy, continuous DNA sequence, and the STS

data can later be used to anchor this sequence precisely onto the physical map. If the STSs also include SSLPs that have been mapped by genetic linkage analysis then the DNA sequence, physical map and genetic map can all be integrated.

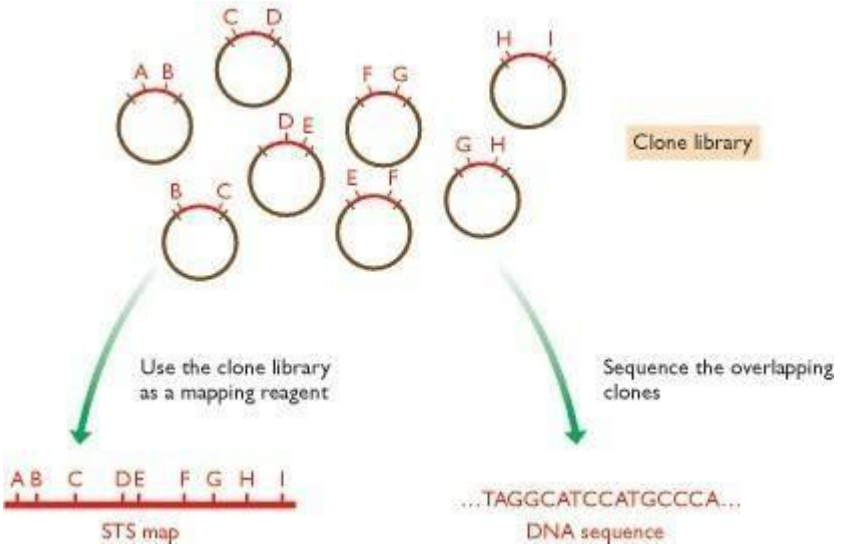


Figure The value of clone libraries in genome projects. The small clone library shown in this example contains sufficient information for an STS map to be constructed, and can also be used as the source of the DNA that will besequenced.

Genome Sequencing:

The objective of a genome project is the complete DNA sequence for the organism being studied, ideally integrated with the genetic and/or physical maps of the genome so that genes and other interesting features can be located within the DNA sequence. This chapter describes the techniques and research strategies that are used during the sequencing phase of a genome project, when this ultimate objective is being directly addressed. Techniques for sequencing DNA are clearly of central importance in this context and we will begin the chapter with a detailed examination of sequencing methodology. This methodology is of little value however, unless the short sequences that result from individual sequencing experiments can be linked together in the correct order to give the master sequences of the chromosomes that make up the genome. The middle part of this chapter

describes the strategies used to ensure that the master sequences are assembled correctly. Finally, we will review the way in which mapping and sequencing were used to produce the two draft human genome sequences that were published in February 2001.

The Methodology for DNA Sequencing:

Rapid and efficient methods for DNA sequencing were first devised in the mid-1970s. Two different procedures were published at almost the same time:

- The chain termination method (Sanger et al., 1977), in which the sequence of a single- stranded DNA molecule is determined by enzymatic synthesis of complementary polynucleotide chains, these chains terminating at specific nucleotide positions;
- The **chemical degradation method** (Maxam and Gilbert, 1977), in which the sequence of a double-stranded DNA molecule is determined by treatment with chemicals that cut the molecule at specific nucleotide positions.

Both methods were equally popular to begin with but the chain termination procedure has gained ascendancy in recent years, particularly for genome sequencing. This is partly because the chemicals used in the chemical degradation method are toxic and therefore hazardous to the health of the researchers doing the sequencing experiments, but mainly because it has been easier to automate chain termination sequencing. As we will see later in this chapter, a genome project involves a huge number of individual sequencing experiments and it would take many years to perform all these by hand. Automated sequencing techniques are therefore essential if the project is to be completed in a reasonable time-span.

Chain termination DNA sequencing:

Chain termination DNA sequencing is based on the principle that single-stranded DNA molecules that differ in length by just a single nucleotide can be separated from one another by polyacrylamide gel electrophoresis. This means that it is possible to resolve a family of molecules, representing all lengths from 10 to 1500 nucleotides, into a series of bands (Figure).

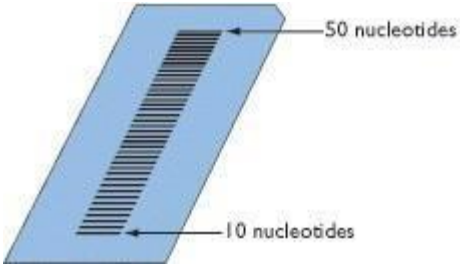


Figure: The banding pattern is produced after separation of single- stranded DNA molecules by denaturing polyacrylamide gel electrophoresis. The molecules are labeled with a radioactive marker and the bands visualized by autoradiography. The bands gradually get closer together towards the top of the ladder. In practice, molecules up to about 1500 nucleotides in length can be separated if the electrophoresis is continued for long enough. Chain termination sequencing in outline

The starting material for a chain termination sequencing experiment is a preparation of identical single-stranded DNA molecules. The first step is to anneal a short oligonucleotide to the same position on each molecule, this oligonucleotide subsequently acting as the primer for synthesis of a new DNA strand that is complementary to the template (Figure). The strand synthesis reaction, which is catalyzed by a DNA polymerase enzyme and requires the four deoxyribonucleotide triphosphates (dNTPs - dATP, dCTP, dGTP and dTTP) as substrates, would normally continue until several thousand nucleotides had been polymerized. This does not occur in a chain termination sequencing experiment because, as well as the four dNTPs, a small amount ofa dideoxynucleotide (e.g.ddATP) is added to the reaction. The polymerase enzyme does not discriminate between dNTPs and ddNTPs, so the dideoxynucleotide can be incorporated into the growing chain, but it then blocks further elongation because it lacks the 3′-hydroxyl group needed to form a connection with the next nucleotide (Figure B).

If ddATP is present, chain termination occurs at positions opposite thymidines in the template DNA. Because dATP is also present the strand synthesis does not always terminate at the first T in the template; in fact it may continue until several hundred nucleotides have been polymerized before a ddATP is eventually incorporated. The result is therefore a set of new chains, all of different lengths, but each ending in ddATP. Now the polyacrylamide gel comes into play. The family of molecules generated in the presence of ddATP is loaded into one lane of the gel, and the families generated with ddCTP, ddGTP and ddTTP loaded into the three adjacent

lanes. After electrophoresis, the DNA sequence can be read directly from the positions of the bands in the gel (Figure D). The band that has moved the furthest represents the smallest piece of DNA, this being the strand that terminated by incorporation of a ddNTP at the first position in the template. In the example shown in

Figure this band lies in the 'G' lane (i.e. the lane containing the molecules terminated with ddGTP), so the first nucleotide in the sequence is 'G'. The next band, corresponding to the molecule that is one nucleotide longer than the first, is in the 'A' lane, so the second nucleotide is 'A' and the sequence so far is 'GA'. Continuing up through the gel we see that the next band also lies in the 'A' lane (sequence GAA), then we move to the 'T' lane (GAAT), and so on. The sequence reading can be continued up to the region of the gel where individual bands are not separated.

Chain termination sequencing requires a single-stranded DNA template:

The template for a chain termination experiment is a single-stranded version of the DNA molecule to be sequenced. There are several ways in which this can be obtained:

- **The** DNA **can be cloned in a plasmid vector :** The resulting DNA will be double stranded so cannot be used directly in sequencing. Instead, it must be converted into single-stranded DNA by denaturation with alkali or by boiling. This is a common method for obtaining template DNA for DNA sequencing, largely because cloning in a plasmid vector is such a routine technique. A shortcoming is that it can be difficult to prepare plasmid DNA that is not contaminated with small quantities of bacterial DNA and RNA, which can act as spurious templates or primers in the DNA sequencing experiment.
- **The** DNA **can be cloned in a bacteriophage M13 vector.** Vectors based on M13 bacteriophage are designed specifically for the production of single-stranded templates for DNA sequencing. M13 bacteriophage has a single-stranded DNA genome which, after infection of Escherichia coli bacteria, is converted into a double-stranded replicative form. The replicative form is copied until over 100 molecules are present in the cell, and when the cell divides the copy number in the new cells is maintained by further replication. At the same time, the infected cells continually secrete new M13 phage particles, approximately 1000 per generation, these phages containing the single-stranded version of the genome. Cloning vectors based on M13 vectors are double-stranded DNA molecules

equivalent to the replicative form of the M13 genome. They can be manipulated in exactly the same way as a plasmid cloning vector. The difference is that cells that have been transfected with a recombinant M13 vector secrete phage particles containing single-stranded DNA, this DNA comprising the vector molecule plus any additional DNA that has been ligated into it. The phages therefore provide the template DNA for chain termination sequencing. The one disadvantage is that DNA fragments longer than about 3 kb suffer deletions and rearrangements when cloned in an M13 vector, so the system can only be used with short pieces of DNA.

- **The** DNA **can be cloned in a phagemid.** This is a plasmid cloning vector that contains, in addition to its plasmid origin of replication, the origin from M13 or another phage with a single-stranded DNA genome. If an E. coli cell contains both a phagemid and the replicative form of a helper phage, the latter carrying genes for the phage replication enzymes and coat proteins, then the phage origin of the phagemid becomes activated, resulting in synthesis of phage particles containing the single-stranded version of the phagemid. The double-stranded plasmid DNA is therefore converted into single-stranded template DNA for DNA sequencing. This system avoids the instabilities of M13 cloning and can be used with fragments up to 10 kb or more.

- PCR **can be used to generate single-stranded** DNA**.** There are various ways of generating single-stranded DNA by PCR, the most effective being to modify one of the two primers so that DNA strands synthesized from this primer are easily purified. One possibility is to attach small metallic beads to the primer and then use a magnetic device to purify the resulting strands.

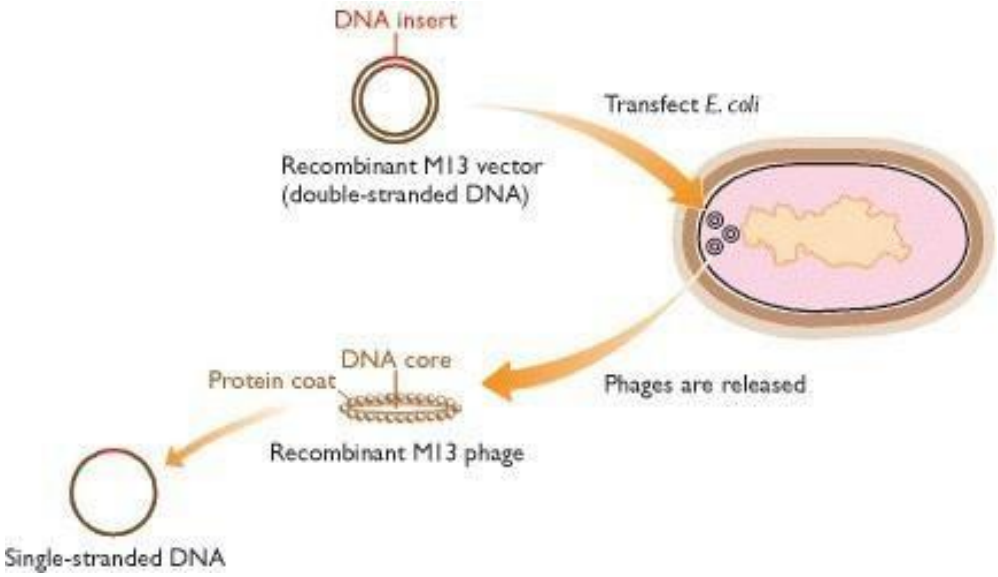


Figure. M13 vectors can be obtained in two forms: the double-stranded replicative molecule and the single-stranded version found in bacteriophage particles. The replicative form can be manipulated in the same way as a plasmid cloning vector (Section 4.2.1) with new DNA inserted by restriction followed by ligation. The recombinant vector is introduced into Escherichia coli cells by transfection. Once inside an E. coli cell, the double-stranded vector replicates and directs synthesis of single-stranded copies, which are packaged into phage particles and secreted from the cell. The phage particles can be collected from the culture medium after centrifuging to pellet the bacteria. The protein coats of the phages are removed by treating with phenol, and the single-stranded version of the recombinant vector is purified for use in DNA sequencing.

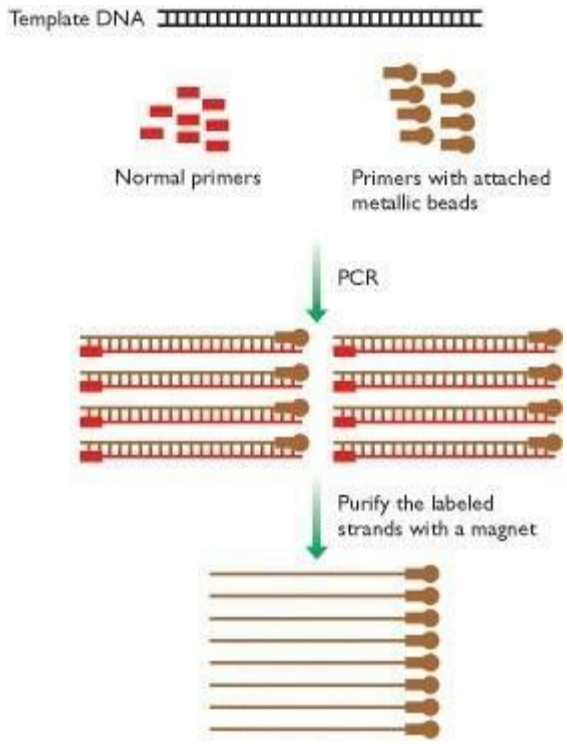


Figure :The PCR is carried out with one normal primer (shown in red), and one primer that is labelled with a metallic bead (shown in brown). After PCR, the labelled strands are purified with a magnetic device. For more details about PCR The primer determines the region of the template DNA that will be sequenced

To begin a chain termination sequencing experiment, an oligonucleotide primer is annealed onto the template DNA. The primer is needed because template-dependent DNA polymerases cannot initiate DNA synthesis on a molecule that is entirely single-stranded: there must be a short double-stranded region to provide a 3′ end onto which the enzyme can add new nucleotides.

The primer also plays the critical role of determining the region of the template molecule that will be sequenced. For most sequencing experiments a 'universal' primer is used, this being one that is complementary to the part of the vector DNA immediately adjacent to the point into which new DNA is ligated (Figure below). The same universal primer can therefore give the sequence of any piece of DNA that has been ligated into the vector. Of course if this inserted DNA is longer than 750 bp or so then only a part of its

sequence will be obtained, but usually this is not a problem because the project as a whole simply requires that a large number of short sequences are generated and subsequently assembled into the contiguous master sequence. It is immaterial whether or not the short

sequences are the complete or only partial sequences of the DNA fragments used as templates. If double-stranded plasmid DNA is being used to provide the template then, if desired, more sequence can be obtained from the other end of the insert. Alternatively, it is possible to extend the sequence in one direction by synthesizing a non- universal primer, designed to anneal at a position within the insert DNA (Figure below). An experiment with this primer will provide a second short sequence that overlaps the previous one.

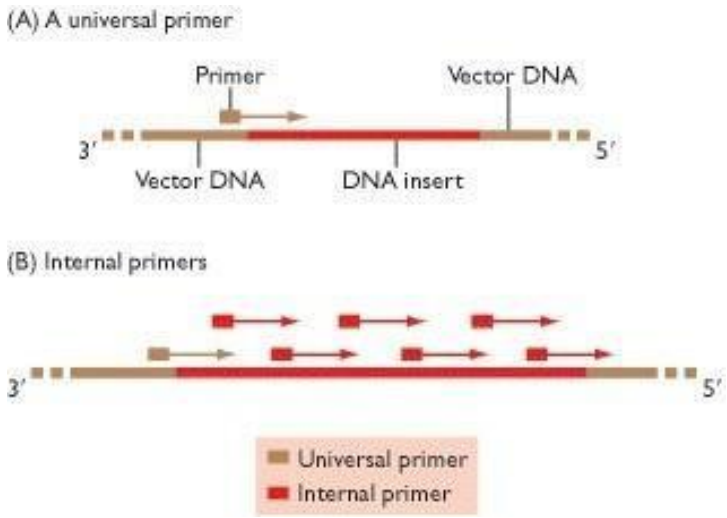


Figure : A universal primer anneals to the vector DNA, adjacent to the position at which new DNA is inserted. A single universal primer can therefore be used to sequence any DNA insert, but only provides the sequence of one end of the insert. (B) One way of obtaining a longer sequence is to carry out a series of chain termination experiments, each with a different internal primer that anneals within the DNA insert.

Thermal cycle sequencing offers an alternative to the traditional methodology:

The discovery of thermostable DNA polymerases, which led to the development of PCR, has also resulted in new methodologies for chain termination sequencing. In particular, the innovation called thermal cycle sequencing (Sears et al., 1992) has two advantages over traditional chain termination sequencing. First, it uses double-stranded rather than single-stranded DNA as the starting material. Second, very little template DNA is needed, so the DNA does not have to be cloned before being sequenced.

Thermal cycle sequencing is carried out in a similar way to PCR but just one primer is used and each reaction mixture includes one of the ddNTPs (Figure below). Because there is only one primer, only one of the strands of the starting molecule is copied, and the product accumulates in a linear fashion, not exponentially as is the case in a real PCR. The presence of the ddNTP in the reaction mixture causes chain termination, as in the standard methodology, and the family of resulting strands can be analyzed and the sequence read in the normal manner by polyacrylamide gel electrophoresis.

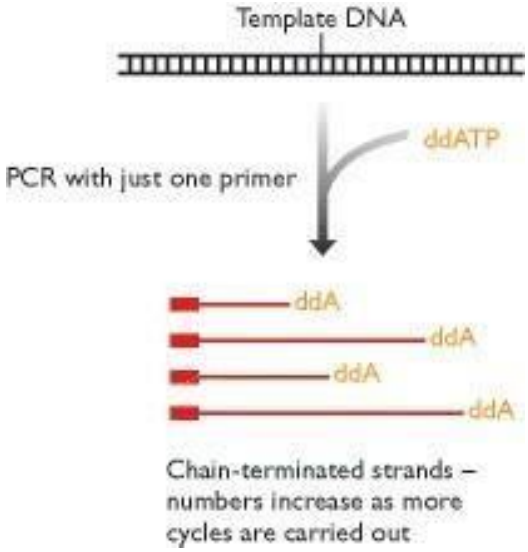


Figure: PCR is carried out with just one primer and with a dideoxynucleotide present in the reaction mixture. The result is a family of chain-terminated strands - the 'A' family in the reaction

shown. These strands, along with the products of the C, G and T reactions, are electrophoresed as in the standard methodology

The standard chain termination sequencing methodology employs radioactive labels, and the banding pattern in the polyacrylamide gel is visualized by autoradiography. Usually one of the nucleotides in the sequencing reaction is labeled so that the newly synthesized strands contain radiolabels along their lengths, giving high detection sensitivity. To ensure good band resolution, ^{33}P or ^{35}S is generally used, as the emission energies of these isotopes are relatively low, in contrast to ^{32}P, which has a higher emission energy and gives poorer resolution because of signal scattering.

Previously we saw how the replacement of radioactive labels by fluorescent ones has given a new dimension to in situ hybridization techniques. Fluorolabeling has been equally important in the development of sequencing methodology, in particular because the detection system for fluorolabels has opened the way to automated sequence reading (Prober et al., 1987). The label is attached to the ddNTPs, with a different fluorolabel used for each one (Figure below). Chains terminated with A are therefore labeled with one fluorophore, chains terminated with C are labeled with a second fluorophore, and so on. Now it is possible to carry out the four sequencing reactions - for A, C, G and T - in a single tube and to load all four families of molecules into just one lane of the polyacrylamide gel, because the fluorescent detector can discriminate between the different labels and hence determine if each band represents an A, C, G or T. The sequence can be read directly as the bands pass in front of the detector and either printed out in a form readable by eye (Figure B) or sent straight to a computer for storage. When combined with robotic devices that prepare the sequencing reactions and load the gel, the fluorescent detection system provides a major increase in throughput and avoids errors that might arise when a sequence is read by eye and then entered manually into a computer. It is only by use of these automated techniques that we can hope to generate sequence data rapidly enough to complete a genome project in a reasonable length of time.

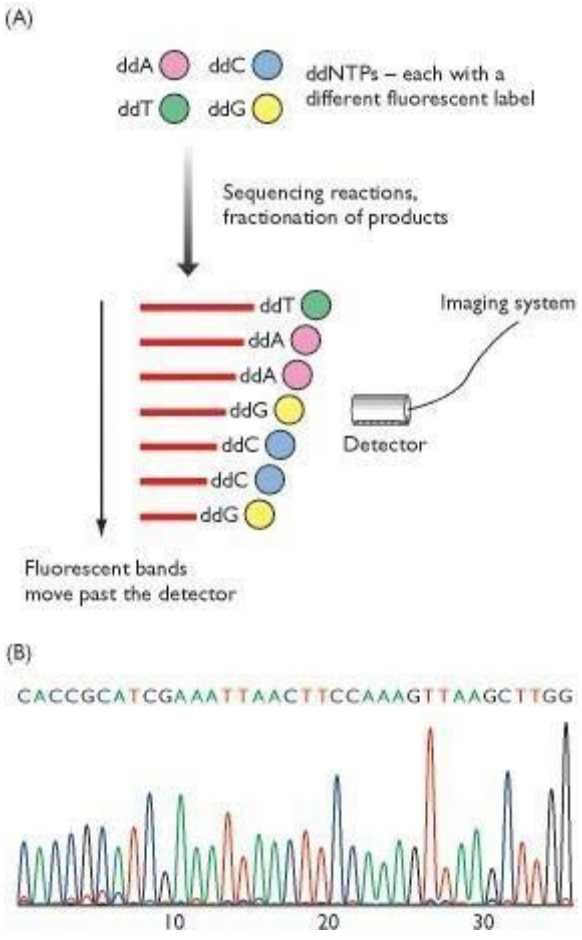


Figure (A) The chain termination reactions are carried out in a single tube, with each dideoxynucleotide labeled with a different fluorophore. In the automated sequencer, the bands in the electrophoresis gel move past a fluorescence detector, which identifies which dideoxynucleotide is present in each band. The information is passed to the imaging system. (B) The printout from an automated sequencer. The sequence is represented by a series of peaks, one for each nucleotide position. In this example, a green peak is an 'A', blue is 'C', black is 'G', and red is 'T'.

High throughput sequencing methods:

In spite of the development of automated techniques, conventional DNA sequencing suffers from the limitation that only a few hundred bp of sequence can be determined in a single experiment. In the context of the

Human Genome Project, this means that each experiment provides only one five-millionth of the total genome sequence. Attempts are continually being made to modify the technology so that sequence acquisition is more rapid, a recent example being the introduction of new automated sequencers that use capillary separation rather than a polyacrylamide gel. These have 96 channels so 96 sequences can be determined in parallel, and each run takes less than 2 hours to complete, enabling up to 1000 sequences to be obtained in a single day (Mullikan and McMurray, 1999). Other systems that are being developed will increase data generation even further by enabling 384 or 1024 sequences to be run at the same time (Rogers, 1999).

There have also been attempts to make sequence acquisition more rapid by devising new sequencing methodologies. One possibility is pyrosequencing, which does not require electrophoresis or any other fragment separation procedure and so is more rapid than chain termination sequencing (Ronaghi et al., 1998). In pyrosequencing, the template is copied in a straightforward manner without added ddNTPs. As the new strand is being made, the order in which the dNTPs are incorporated is detected, so the sequence can be 'read' as the reaction proceeds. The addition of a nucleotide to the end of the growing strand is detectable because it is accompanied by release of a molecule of pyrophosphate, which can be converted by the enzyme sulfurylase into a flash of chemiluminescence. Of course, if all four dNTPs were added at once then flashes of light would be seen all the time and no useful sequence information would be obtained. Each dNTP is therefore added separately, one after the other, with a nucleotidase enzyme also present in the reaction mixture so that if a dNTP is not incorporated into the polynucleotide then it is rapidly degraded before the next dNTP is added. This procedure makes it possible to follow the order in which the dNTPs are incorporated into the growing strand. The technique sounds complicated, but it simply requires that a repetitive series of additions be made to the reaction mixture, precisely the type of procedure that is easily automated, with the possibility of many experiments being carried out in parallel.

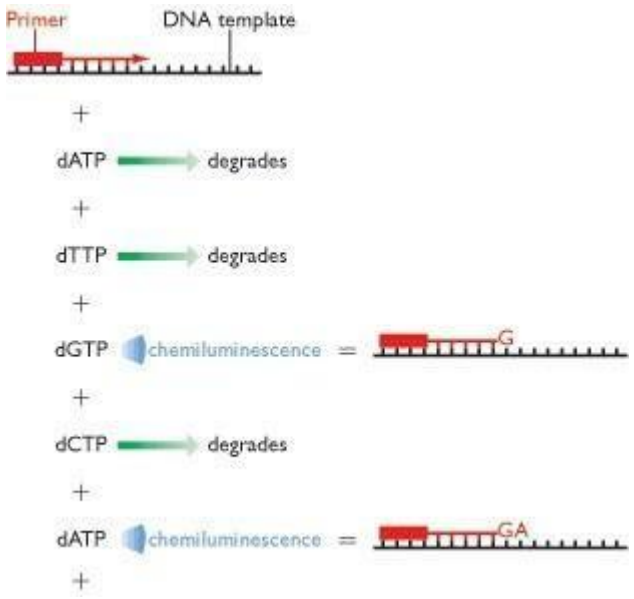


Figure Pyrosequencing. The strand synthesis reaction is carried out in the absence of dideoxynucleotides. Each dNTP is added individually, along with a nucleotidase enzyme that degrades the dNTP if it is not incorporated into the strand being synthesized. Incorporation of a nucleotide is detected by a flash of chemiluminescence induced by the pyrophosphate released from the dNTP. The order in which nucleotides are added to the growing strand can therefore be followed.

A very different approach to DNA sequencing through the use of DNA chips might one day be possible. A chip carrying an array of different oligonucleotides could be used in DNA sequencing by applying the test molecule - the one whose sequence is to be determined - to the array and detecting the positions at which it hybridizes. Hybridization to an individual oligonucleotide would indicate the presence of that particular oligonucleotide sequence in the test molecule, and comparison of all the oligonucleotides to which hybridization occurs would enable the sequence of the test molecule to be deduced (Figure). The problem with this approach is that the maximum length of the molecule that can be sequenced is given by the square root of the number of oligonucleotides in the array, so if every possible 8-mer oligonucleotide (ones containing eight nucleotides) were attached to the chip - all 65 536 of them - then the maximum length of readable sequence would be only 256 bp(Southern, 1996). Even if the chip carried all the 1 048 576 different 10-mer sequences, it could still only be used to sequence a 1 kb molecule. To sequence a 1 Mb molecule (this being the sort of advance in sequence capability that is really needed) the chip

would have to carry all of the 1×10^{12} possible 20-mers. This may sound an outlandish proposition but advances in miniaturization, together with the possibility of electronic rather than visual detection of hybridization, could bring such an array within reach in the future.

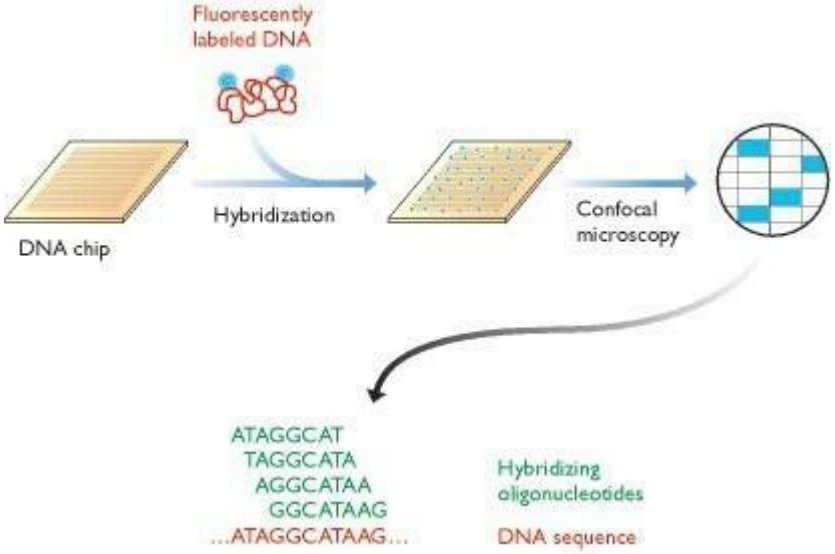


Figure The chip carries an array of every possible 8-mer oligonucleotide. The DNA to be sequenced is labeled with a fluorescent marker and applied to the chip, and the positions of hybridizing oligonucleotides determined by confocal microscopy. Each hybridizing oligonucleotide represents an 8-nucleotide sequence motif that is present in the probe DNA. The sequence of the probe DNA can therefore be deduced from the overlaps between the sequences of these hybridizing oligonucleotides.

DNA Annotation:

DNA annotation or genome annotation is the process of identifying the locations of genes and all of the coding regions in a genome and determining what those genes do. An annotation (irrespective of the context) is a note added by way of explanation or commentary. Once a genome is sequenced, it needs to be annotated to make sense of it.

For DNA annotation, a previously unknown sequence representation of genetic material is enriched with information relating genomic position to intron-exon boundaries, regulatory sequences, repeats, gene names and protein products. This annotation is stored in genomic databases such as

Mouse Genome Informatics, FlyBase, and Worm Base. Educational materials on some aspects of biological annotation from the 2006 Gene Ontology annotation camp and similar events are available at the Gene Ontology website.

The National Center for Biomedical Ontology develops tools for automated annotation of database records based on the textual descriptions of those records.

As a general method, dcGO has an automated procedure for statistically inferring associations between ontology terms and protein domains or combinations of domains from the existing gene/protein-level annotations.

Process:

Genome annotation consists of three main steps:.

1. identifying portions of the genome that do not code for proteins
2. identifying elements on the genome, a process called gene prediction, and
3. attaching biological information to these elements.

Automatic annotation tools try to perform all this by computer analysis, as opposed to manual annotation (a.k.a. curation) which involves human expertise. Ideally, these approaches co-exist and complement each other in the same annotation pipeline.

The simplest way to perform gene annotation relies on homology based search tools, like BLAST, to search for homologous genes in specific databases, the resulting information is then used to annotate genes and genomes. However, nowadays more and more additional information is added to the annotation platform. The additional information allows manual annotators to deconvolute discrepancies between genes that are given the same annotation. Some databases use genome context information, similarity scores, experimental data, and integrations of other resources to provide genome annotations through their Subsystems approach. Other databases (e.g.Ensembl) rely on both curated data sources as well as a range of different software tools in their automated genome annotation pipeline.

Structural annotation consists of the identification of genomic elements.

- ORFs and their localization
- Gene structure
- Coding regions

- location of regulatory motifs

Functional annotation consists of attaching biological information to genomic elements.

- Biochemical function
- Biological function
- involved regulation and interactions
- expression

These steps may involve both biological experiments and in silico analysis. Proteo-genomics based approaches utilize information from expressed proteins, often derived from mass spectrometry, to improve genomics annotations.

A variety of software tools have been developed to permit scientists to view and share genome annotations. Genome annotation remains a major challenge for scientists investigating the human genome, now that the genome sequences of more than a thousand human individuals and several model organisms are largely complete. Identifying the locations of genes and other genetic control elements is often described as defining the biological "parts list" for the assembly and normal operation of an organism. Scientists are still at an early stage in the process of delineating this parts list and in understanding how all the parts "fit together".[

Genome annotation is an active area of investigation and involves a number of different organizations in the life science community which publish the results of their efforts in publicly available biological databases accessible via the web and other electronic means. Here is an alphabetical listing of on-going projects relevant to genome annotation:

- Encyclopedia of DNA elements(ENCODE)
- Entrez Gene
- Ensembl
- GENCODE
- Gene Ontology Consortium
- GeneRIF

- RefSeq
- Uniprot
- Vertebrate and Genome Annotation Project (Vega)

Base calling and sequence accuracy:

Base calling is the process by which an order of nucleotides in a template is inferred during a sequencing reaction. Next generation sequencing platforms that use fluorescently labeled reversible terminators have a unique color for each base. These are incorporated into the complementary strand of the DNA template and captured with a sensitive CCD camera. These images are processed into signals which are used to infer the order of nucleotides, also known as base calling. While sequencing platforms typically have integrated base calling software, the development of high performing base calling algorithms is an area of ongoing research.

Base calling accuracy is typically measured by a Q score (Phred quality score), a common metric to assess the accuracy of a sequencing run. Q scores are defined as logarithmically related to base calling error probability.

$Q = - 10 \log P / \log 10$

If a sequencing run is assigned a Q score of 40, this is equal to the probability of an incorrect base call of 1 in 10,000 times, or 99.99% base calling accuracy.

Q Score 10 - Base calling accuracy 1 in 10 - Probability of incorrect base 90% Q Score 20 - Base calling accuracy 1 in 100 - Probability of incorrect base 99%
Q Score 30 - Base calling accuracy 1 in 1,000 - Probability of incorrect base 99.9%

Q Score 40 - Base calling accuracy 1 in 10,000 - Probability of incorrect base 99.99%

Q Score 50 - Base calling accuracy 1 in 100,000 - Probability of incorrect base 99.999%

A lower Q score of 10 means, there is the probability of an incorrect call in 1 of 10 bases. Lower Q scores can lead to increases in false positive variant calls and reduces the overall confidence an investigator has in their sequencing data

Probable Questions:

1. How DNA markers can be used for genetic mapping?
2. Write down the importance of SNP markers in genetic mapping.
3. Describe basic methodology of restriction mapping.
4. What is optical mapping. Describe the technique.
5. Describe FISH method. What is its significance?
6. Describe STTS mapping.
7. Describe the methodology of DNA sequencing?
8. What is high throughput DNA sequencing?
9. What do you mean by base calling?
10. What is DNA annotation?

Suggested Readings:

1. Jing JP, Lai ZW, Aston C. et al. Optical mapping of Plasmodium falciparum chromosome 2. Genome Res (1999);9:175–181.
2. Lichter P, Tang CJ, Call K. et al. High resolution mapping of human chromosome 11 by in situ hybridization with cosmid clones. Science. (1990);247:64–69.[PubMed]
3. Lin J, Qi R, Aston C. et al. Whole-genome shotgun optical mapping of Deinococcusradiodurans. Science. (1999);285:1558–1562.[PubMed]
4. Marra MA, Hillier L, Waterston RH. Expressed sequence tags - ESTablishingbridges between genomes. Trends Genet. (1998);14:4–7.[PubMed]
5. McCarthy L. Whole genome radiation hybrid mapping. Trends Genet.(1996);12:491–

   493. [PubMed]
6. Oliver SG, van der Aart QJM, Agostoni-Carbone ML. et al. The complete DNA sequence of yeast chromosome III. Nature. (1992);357:38–46.[PubMed]
7. Schwartz DC, Li X, Hernandez LI, Ramnarain SP, Huff EJ, Wang Y-

K. Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. Science.(1993);262:110–114.[PubMed]

8. SNP Group (The International SNP Map Working Group). A map of human genome sequence variation containing 1.42 million singlenucleotide polymorphisms. Nature. (2001);409:928–933. [PubMed]
9. Sturtevant AH. The linear arrangement of six sex-linked factors in Drosophila asshown by mode of association. J. Exp. Zool.(1913);14:39–45.
10. Yamamoto F, Clausen H, White T, Marken J, Hakamori S. Molecular genetic basis of the histo-blood group ABO system. Nature. (1990);345:229–233.[PubMed]
11. Maxam AM, Gilbert W. A new method for sequencing DNA. Proc. Natl Acad. Sci. USA. (1977);74:560–564.[PMC free article][PubMed]
12. Mullikan JC, McMurray AA. Sequencing the genome, fast. Science. (1999);283:1867– 1868.[PubMed]
13. Murray JC, Buetow KH, Weber JL. et al. A comprehensive human linkage map with centimorgan density. Science. (1994);265:2049–2054.[PubMed]
14. Prober JM, Trainor GL, Dam RJ. et al. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. Science.(1987);238:336–341. [PubMed]
15. Rogers J. Gels and genomes. Science. (1999);286:429.[PubMed]
16. Ronaghi M, Ehleen M, Nyrn P. A sequencing method based on real-time pyrophosphate. Science. (1998);281:363–365. [PubMed]
17. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain terminating inhibitors. Proc. Natl Acad. Sci. USA. (1977);74:5463–5467. [PMC freearticle]

UNIT-VI

Functional genomics: Study of gene interaction by the yeast two-hybrid system; Protein-DNA interaction, ChIP Assay

Objective: In this unit you will know about functional genomics and gene interaction study by yeast two hybrid system and microarray technique using DNA chips.

Functional Genomics

Genomics involves the mapping of an entire genome and, eventually, the determination of a species' complete DNA sequence. The amount of information found within a species' genome is enormous. The goal of functional genomics is to elucidate the roles of genetic sequences-DNA, RNA, and amino acid sequences-in a given species. In most cases, functional genomics is aimed at understanding gene function. At the genomic level, researchers can study genes as groups. For example, the information gained from a genome-sequencing project can help researchers study entire metabolic pathways. This provides a description of the ways in which gene products interact to carry out cellular processes. In addition, a study of genetic sequences can help to identify regions that play particular functional roles. For example, an analysis of certain species of bacteria helped to identify DNA sequences that promote the uptake of DNA during bacterial transformation. Because most genes encode proteins, a goal of many molecular biologists is to understand the functional roles of all the proteins a species produces. The entire collection of proteins a given cell or organism can make is called its proteome, and the study of the function and interactions of these proteins is termed proteomics. An objective of researchers in the field of proteomics is to understand the interplay among many proteins as they function to create cells and, ultimately, the traits of a given species.

Methods to Study Protein- Protein Interactions:

The operation of the genome can be evaluated by the study of proteome. Thus, by studying the functions of proteins, it is possible to understand how the genome operates and how a dysfunctional genome activity can result in disease states such as cancer. Proteomics broadly involves the methodology for characterizing the protein content of the cell. This can be done by protein electrophoresis, mass spectrometry etc.

Identification of protein-protein interaction is a recent approach to study proteome. The protein interaction maps can be constructed to understand the relation between the proteome and cellular biochemistry. Phage display and yeast two-hybrid system are commonly used to study protein- protein interactions.

Phage Display:

Phage display is a novel technique to evaluate genome activity with particular reference to identify proteins that interact with one another. It basically involves insertion of a foreign DNA into phage genome, and its expression as fusion product with a phage coat protein (Fig. 5.14A). This is followed by screening of test protein by phage display library (Fig. 5.14B). The technique is briefly described below.

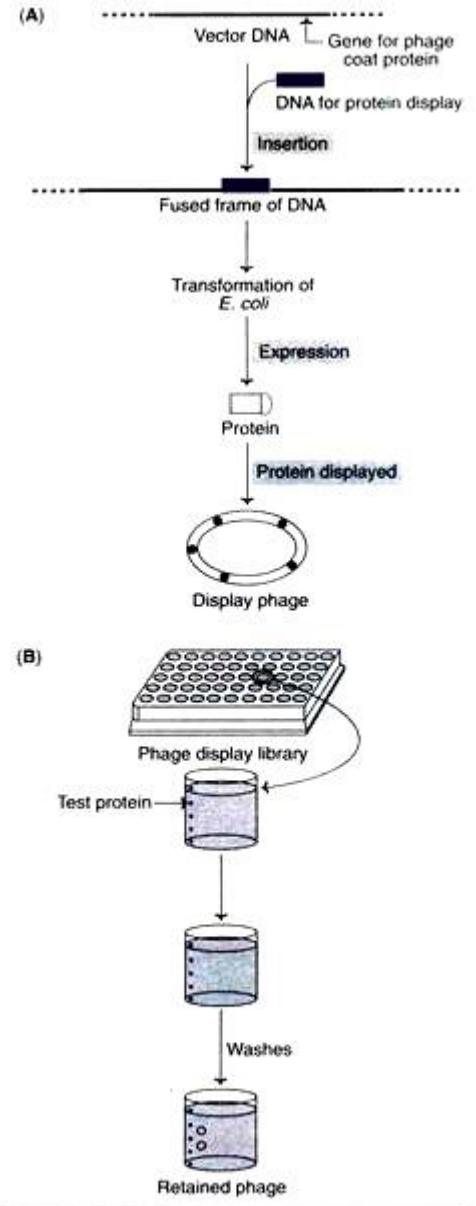


Fig. 5.14 : Elucidation of protein–protein interaction by phage display (A) Production of fusion protein displayed on phage (B) Screening of test protein by phage display library.

A special type of cloning vector such as a bacteriophage or filamentous bacteriophage (e.g. M13) are used for phage display. A fragment of DNA coding for the test protein is

inserted into the vector DNA (adjacent to phage coat protein gene). After transformation of E. coli, this recombinant gene (fused frame of DNA) results in the synthesis of hybrid protein. The new protein is made up of the test protein fused with the phage coat protein. The phage particles produced in the transformed E. coli display the test protein in their coats.

The test protein interaction can be identified by using a phage display library. For this purpose, the test protein is immobilized within a well of a micro-titer tray, and the phage display library added. After several washes, the phages that are retained in the well are those displaying a protein that interacts with the test protein. Phage-displaying peptides can be isolated, based on their antibody-binding properties, by employing affinity chromatography. Several rounds of affinity chromatography and phage propagation can be used to enrich phages with desired proteins.

Phagemid display:

Phagemid in place of plasmid can also be used for the display of proteins. In fact, special types of phagemid display vectors have been developed for this purpose. Phage and phagemid display can be successfully used for selecting and engineering polypeptides with novel functions.

Yeast Two-Hybrid System:

When two proteins interact with each other, their corresponding genes are known as interacting genes. The yeast two-hybrid system uses a reporter gene to detect the physical interaction of a pair of proteins inside a yeast nucleus.

The two-hybrid method is based on the observation that most of the transcriptional proteins (i.e. the proteins involved in promoting transcription of a gene) contain two distinct domains—DNA binding domain and transcriptional activation domain. When these two domains are physically separated, the protein loses its activity. However, the same protein can be reactivated when the domains are brought together. These proteins can bind to DNA and activate transcription.

The target protein is fused to a DNA-binding domain to form a bait. When this target protein binds to another specifically designed protein namely the prey in the nucleus, they interact, which in turn switches on the expression of the reporter gene (Fig. 5.15). The reporter genes can be detected by growing the yeast on a selective medium.

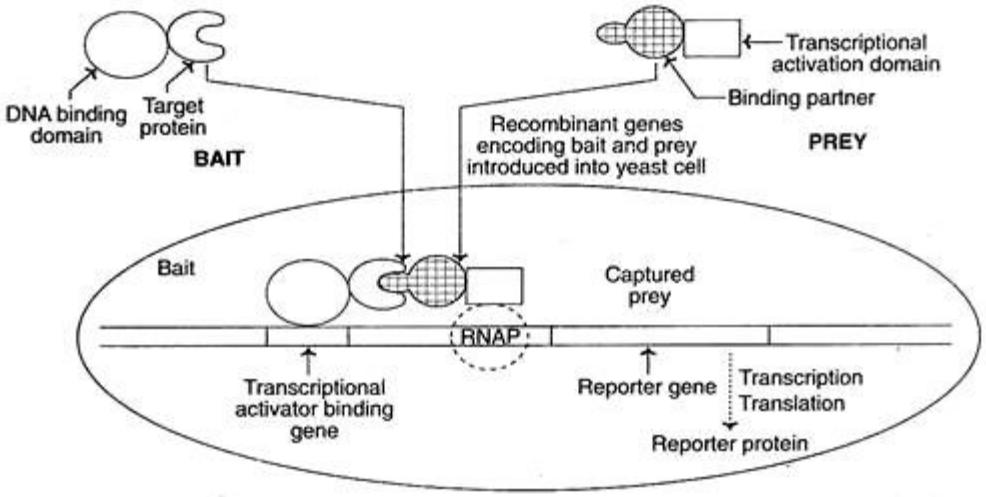


Fig. 5.15 : Elucidation of protein-protein interaction by yeast two-hybrid system (RNAP-RNA polymerase)

It is possible to generate the bait and prey fusion proteins by standard recombinant DNA techniques. A single bait protein is frequently used to fish out interacting partners among the collection of prey proteins. A large number of prey proteins can be produced by ligating DNA encoding the activation domain of a transcriptional activator to a mixture of DNA-fragments from a cDNA library.

The yeast two-hybrid system originally created by Fields and Song is a genetic system wherein the interaction between two proteins of interest is detected via the reconstitution of a transcription factor and the subsequent activation of reporter genes under the control of this transcription factor. In yeast, Galactose is imported into the cell and converted to galactose-6-phosphate by six enzymes (GAL1, GAL2, PGM2, GAL7, GAL10, MEL1) which are transcriptionally regulated by the proteins Gal80, Gal3, and Gal4, the latter of which plays the central role of DNA-binding transactivator. Gal80 binds Gal4 and inhibits its transcriptional ability. Gal3, in the presence of galactose, binds and causes a conformational change in Gal80, which then allows Gal4 to function as a transcriptional activator. Gal4, like other transcriptional activators, is a modular protein that requires both DNA-binding (BD) and activation domains (AD). The "two hybrid" technique exploits the fact that Gal4 cannot function as a transcriptional activator unless physically bound to an activation domain. Furthermore, it has been demonstrated that this interaction does not need to be covalent: an experiment was performed where the negative regulatory protein, Gal80, was fused with an activation domain to produce Gal80-AD, and was able to act as a transcriptional activator through its natural binding interaction with a Gal4 protein that was missing its own activation domain.

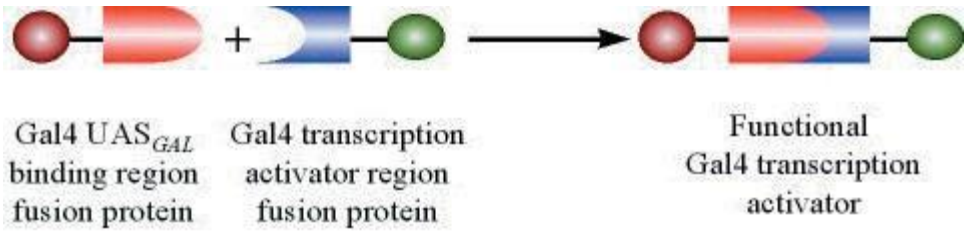


Fig 1: Gal 4 transcriptional activator.

In a classical assay system, a protein X is expressed as a fusion to a DNA binding domain (DBD). The DBD–X fusion is commonly termed the "bait." Because of the affinity of the DBD for its operator sequences the bait is bound to a promoter element upstream of a reporter gene but does not activate it because it lacks an activation domain. A second protein Y is expressed as a fusion to an activation domain (AD) and is commonly termed the "prey." The prey is capable of activating transcription but usually does not do so because it has no affinity for the promoter elements upstream of the reporter gene. If bait and prey are co-expressed and the two proteins X and Y interact, then a functional transcription factor is reconstituted at the promoter site upstream of the reporter gene. Consequently, transcription of the reporter gene is activated. Thus, in a yeast two-hybrid assay a protein-protein interaction is measured through the activation of one or several reporter genes in response to the assembly of a transcription factor by the said protein-protein interaction. In common yeast two-hybrid screening schemes the prey is usually replaced by a collection of unknown preys expressed from a cDNA or genomic library. Screening of entire libraries against a defined bait may then lead to the discovery of novel interaction For large-scale screenings, two approaches are commonly used: the library screening approach, in which multiple baits are screened against a library, and the matrix approach, in which an array of defined preys is substituted for the library.

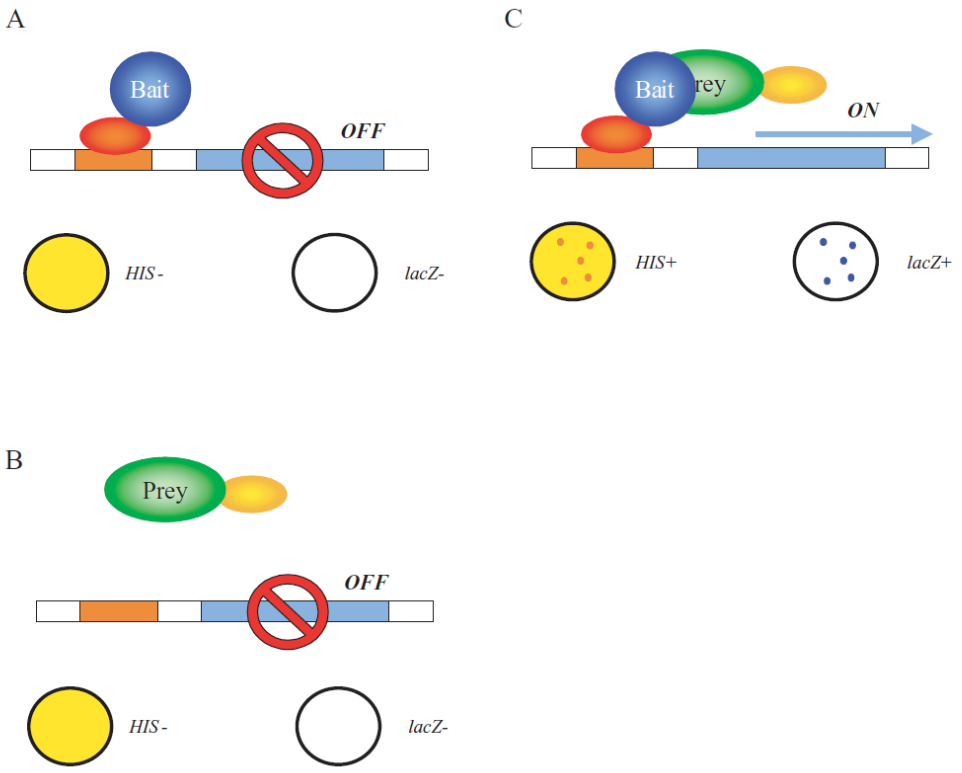


Fig 2: Yeast-Two-Hybrid system. The hybrid transcription factor is bound to the promoter upstream of the reporter gene and therefore activates transcription. The readout of the activated reporter gene is measured either as growth on selective medium (auxotrophic selection markers, such as *HIS3*, *URA3*, or *ADE2)* or in a colour reaction *(lacZ*). Yeast expressing only the DBD–bait or the AD–prey on its own do not grow on selective medium (*HIS*–) and do not display blue staining in a colour assay (*lacZ*-), whereas yeast harbouring an interacting DBD–bait and AD–prey display growth (*HIS*+) and blue colour (*lacZ*+).

Library Screening approach: In high throughput library approach a particular bait is expressed in a yeast reporter strain of the mating type a, whereas a collection of preys (the library) is transformed into a yeast reporter strain of the mating type α. The bait bearing strain is then mated with the mixture of library strains, and clones expressing an interaction pair are isolated on selective media. To determine the identity of the interacting prey, the library plasmid encoding it has to be isolated from the yeast strain and amplified in *Escherichia coli*. The region encoding the prey is then sequenced.

The Matrix screening Approach: In the matrix approach a collection of defined preys is used instead of a random collection of open reading frames (ORFs) or ORF fragments. Each prey is separately introduced into yeast and the transformants are arrayed on plates using a robot. A bait-bearing strain of the opposite mating type is then mated with every prey-bearing strain and the resulting diploid strains are replicated onto selective medium. If a particular diploid within the array grows on selection medium, its prey must interact with the bait under investigation. As opposed to the

library screen, no plasmid isolation or sequencing is necessary since the position of the growing diploid on the array identifies the prey it expresses. In essence, a matrix screen consists of a series of defined interactions between a bait and a numbers of prey, rather than a screen of a bait against a collection of unknown preys.

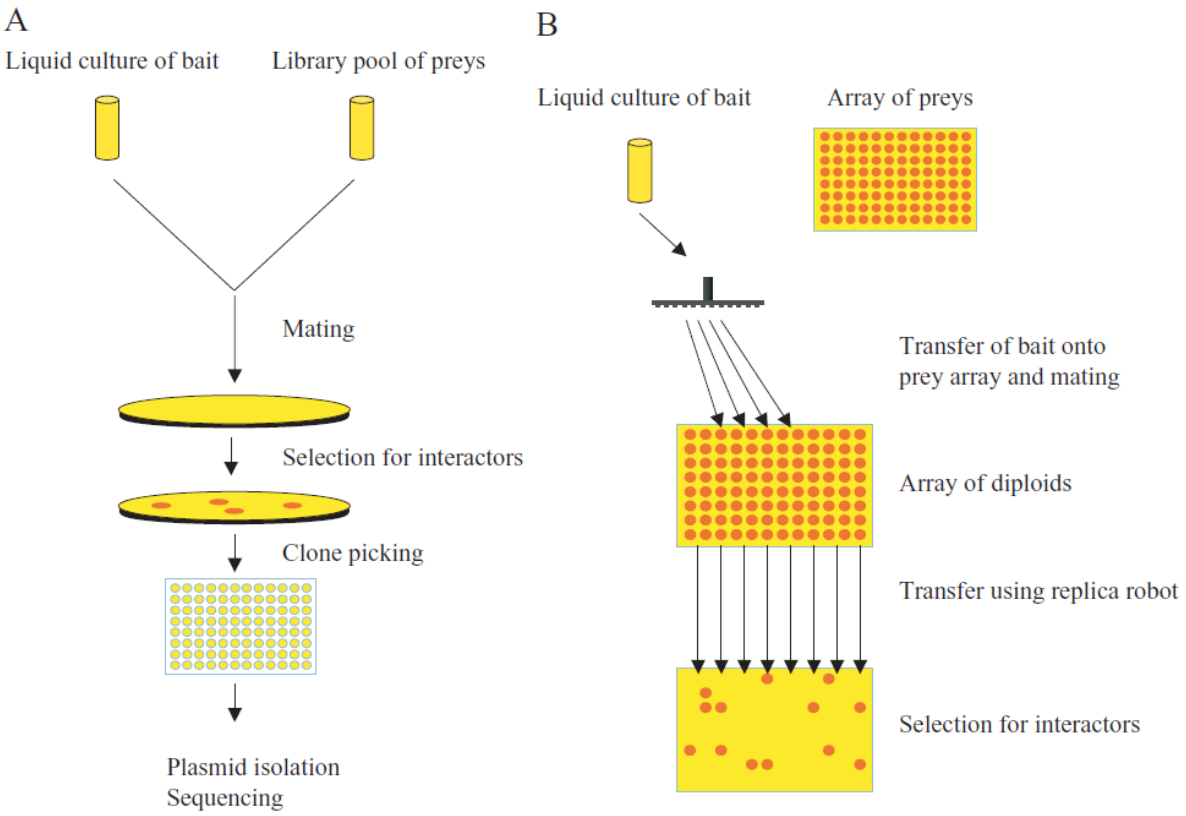


Fig 3: High throughput screening. (A) Library approach (B) Matrix approach

Yeast Three-Hybrid System:

The interactions between protein and RNA molecules can be investigated by using a technique known as yeast three-hybrid system.

DNA Microarray:

The DNA Microarray technology is used to determine the level of expression of many thousands of genes simultaneously. This new approach is used not for individual genetic loci, rather, for the analysis of genome-wide patterns of gene expression. Using DNA microarrays, it is possible to estimate the relative level of gene expression of each gene in the genome.

The DNA microarray or chip is a high density grid system, consisting of a flat solid substrate about the size of a postage stamp that can be used to detect hybridisation of target DNA under appropriate conditions. The chip contains 10,000 to 100,000 distinct spots, from 75 to 150 μm in diameter.

The spacing between spots on an array is usually 100 to 200 μm. Each spot contains a different immobilised DNA sequence that can be hybridised with DNA (or RNA) from a large number of different cells. Two types of chips are currently available: one, in which oligonucleotides have been synthesised directly on the chip, one nucleotide at a time, by automated procedures.

These chips have hundreds of thousands of spots per array; second, chips in which double-stranded DNA sequences of 500 to 5000 base pairs have been deposited through drops by capillary action from miniaturized devices mounted on the movable head of a robotic workstation. These chips have tens of thousands of spots per array. The surface onto which DNA is spotted is critically important. The ideal surface immobilizes the target DNAs, and is compatible with stringent probe hybridisation conditions.

The procedure shown (Fig. 24.1) depicts only 6 spots in a chip, each of which contains a DNA sequence that serves as a probe for a different gene. Experimental cells are used for the extraction of cellular mRNA, and a control sample of mRNA from another source. The samples are subjected to reverse transcription to obtain DNA strands. In the experimental material, the primer for reverse transcription is tagged with a green fluorescent label, while primers of the control material receive red fluorescent label. After the DNA strands have been obtained in sufficient quantity, the fluorescent samples are mixed and hybridised with the DNA in the spots in the chip. The hybridisation is competitive because the two samples were mixed.

Therefore, the density of red and green strands bound to the chip is proportional to the concentration of red or green molecules in the mixture. Genes that are over-expressed in the experimental sample relative to the control will have more green strands hybridised to the spot, whereas those that are under-expressed in the experimental sample relative to the control will have more of red strands hybridised to the spot.

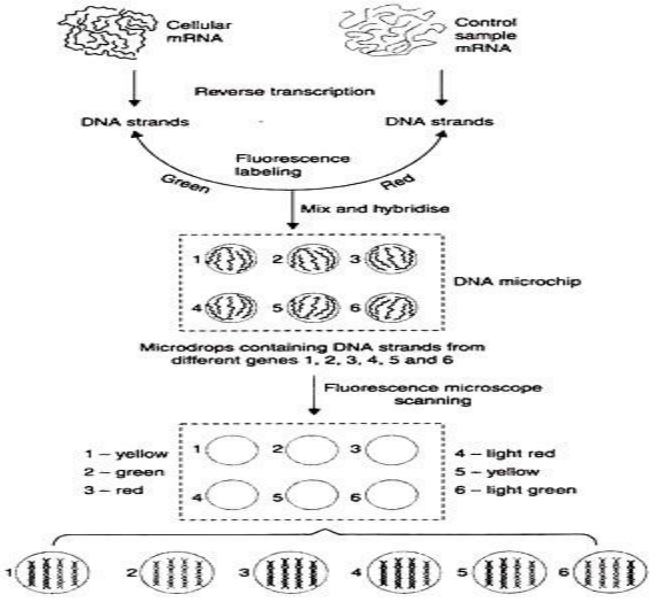


Fig. 24.1 Procedure for DNA microarrays. Six dried microdrops are introduced into a DNA microchip. Each drop contains immobilised DNA strands from a different gene numbered 1 to 6. These are hybridised with fluorescence-labelled DNA samples obtained by reverse transcription of cellular mRNA (green) and red labelled control DNA sample. Competitive hybridisation of green (experimental) and red (control) label is proportional to the *relative* abundance of each mRNA in the sample. The intensity of red and green fluorescence is analysed by microscopy and interpreted as overexpression, underexpression and equal expression of the gene depending on the intensity of red, green, orange, yellow-green and yellow fluorescence.

The intensity of fluorescence is viewed by placing the chip under a laser scanning microscope or a fluorescence microscope that scans each pixel, which is the smallest discrete unit in a visual image. The intensity of fluorescent label is recorded. The signals are synthesised to produce a signal value for each spot in the microarray.
The signals indicate the relative levels of gene expression through colour. Green or yellow green indicate over-expression in experimental sample, while red or orange indicates under-expression in experimental sample. Yellow indicates equal expression in both experimental and control samples.
DNA microarray technology is useful for study of large number of cells growing under different conditions, at different developmental stages, or at different stages of a disease. Besides detection of gene expression, this technology can be used to detect mutations and polymorphisms, to map genomic DNA clones, and to compare the gene expression pattern in normal and diseased tissues.

Fabrication:

Microarrays can be fabricated using a variety of technologies, including printing with fine-pointed pins onto glass slides, photolithography using pre-made masks, photolithography using dynamic micro-mirror devices, ink-jet printing, or electrochemistry on microelectrode arrays.
DNA microarrays can be used to detect RNAs that may or may not be translated into active proteins. Scientists refer to this kind of analysis as “expression analysis” or expression profiling. Since there can be tens of thousands of distinct probes on an array, each microarray experiment can accomplish the equivalent number of genetic tests in parallel. Arrays have, therefore, dramatically accelerated many types of investigations. The use of microarrays for gene expression profiling was first published in 1995 (Science) and the first complete eukaryotic genome {Saccharomyces cerevisiae) on a microarray was published in 1997 (Science).

1. Spotted Microarrays*:*

In spotted microarrays (or two-channel or two-colour microarrays), the probes are oligonucleotides, cDNA or small fragments of PCR products that correspond to mRNAs and are spotted onto the microarray surface. This type of array is typically hybridized with cDNA from two samples to be compared (e.g., diseased tissue versus healthy tissue) that are labelled with two different fluorophores (e.g., Rhodamine (Cyanine 5, red) and Fluorescein (Cyanine 3, green)). The two samples are mixed and hybridized to a single microarray that is then scanned in a microarray scanner to visualize fluorescence of the two fluorophores. Relative intensities of each fluorophore are then used to identify up-regulated and down-regulated genes in ratio-based analysis. Absolute levels of gene expression cannot be determined in the two-colour array, but relative differences in expression among different spots (= genes) can be estimated with some oligonucleotide arrays.

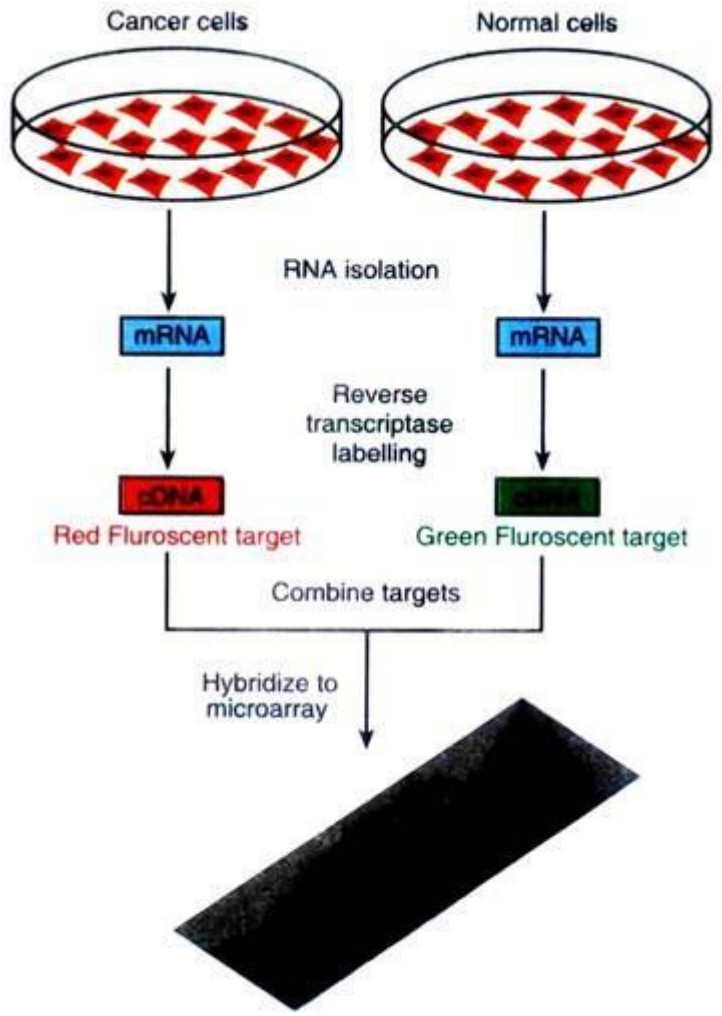


Fig. 16.2: Diagram of typical dual-colour microarray experiment

2. Oligonucleotide Microarrays:

In oligonucleotide microarrays (or single-channel microarrays), the probes are designed to match parts of the sequence of known or predicted mRNAs. There are commercially available designs that cover complete genomes from companies such as GE Healthcare, Affymetrix, Ocimum Bio-solutions, or Agilent. These microarrays give estimations of the absolute value of gene expression and, therefore, the comparison of two conditions requires the use of two separate micro- arrays.

Fig. 16.3: Two Affymetrix chips

Oligonucleotide Arrays can be either produced by piezoelectric deposition with full length oligonucleotides or in situ synthesis. Long Oligonucleotide Arrays are composed of 60-mers, or 50-mers and are produced by ink-jet printing on a silica substrate. Short Oligonucleotide Arrays are composed of 25-mer or 30-mer and are produced by photolithographic synthesis (Affymetrix) on a silica substrate or piezoelectric deposition (GE Healthcare) on an acrylamide matrix. More recently, Maskless Array Synthesis from NimbleGen Systems has combined flexibility with large numbers of probes. Arrays can contain up to 390,000 spots, from a custom array design. New array formats are being developed to study specific pathways or disease states for a systems biology approach.

Oligonucleotide microarrays often contain control probes designed to hybridize with RNA spike-ins. The degree of hybridization between the spike-ins and the control probes is used to normalize the hybridization measurements for the target probes.

Genotyping Microarrays:

DNA microarrays can also be used to read the sequence of a genome in particular positions. SNP microarrays are a particular type of DNA microarrays that are used to identify genetic variation in individuals and across populations.

Short oligonucleotide arrays can be used to identify the single nucleotide polymorphisms (SNPs) that are thought to be responsible for genetic variation and the source of susceptibility to genetically caused diseases. Generally termed genotyping applications, DNA microarrays may be used in this fashion for forensic applications, rapidly discovering or measuring genetic predisposition to disease, or identifying DNA-based drug candidates. These SNP microarrays are also being used to profile somatic mutations in cancer, specifically loss of heterozygosity events and amplifications and deletions of regions of DNA. Amplifications and deletions can also be detected using

comparative genomic hybridization, or aCGH, in conjunction with microarrays, but may be limited in detecting novel Copy Number Polymorphisms, or CNPs, by probe coverage. Re-sequencing arrays have also been developed to sequence portions of the genome in individuals. These arrays may be used to evaluate germ line mutations in individuals, or somatic mutations in cancers. Genome tiling arrays include overlapping oligonucleotides designed to blanket an entire genomic region of interest. Many companies have successfully designed tiling arrays that cover whole human chromosomes.

Microarrays and Bioinformatics:

1. Experimental Design:

Due to the biological complexity of gene expression, the considerations of experimental design that are discussed in the expression profiling article are of critical importance if statistically and biologically valid conclusions are to be drawn from the data.

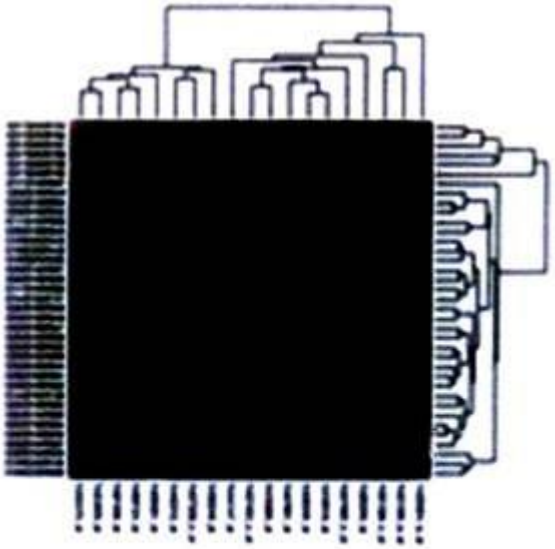


Fig. 16.4: Gene expression values from microarray experiments can be represented as heat maps to visualize the result of data analysis

There are three main elements to consider when designing a microarray experiment.

First, replication of the biological samples is essential for drawing conclusions from the experiment.

Second, technical replicates (two RNA samples obtained from each experimental unit) help to ensure precision and allow for testing differences within treatment groups. The technical replicates may be two independent RNA extractions or two aliquots of the same extraction.

Third, spots of each cDNA clone or oligonucleotide are present at least as duplicates on the microarray slide, to provide a measure of technical precision in each hybridization. It is critical that information about the sample preparation and handling is discussed in order to help identify the independent units in the experiment as well as to avoid inflated estimates of significance.

2. Standardization:

The lack of standardization in arrays presents an interoperability problem in bioinformatics, which hinders the exchange of array data. Various grass-roots open-source projects are attempting to facilitate the exchange and analysis of data produced with non-proprietary chips.

a. The "Minimum Information about a Microarray Experiment" (MIAME) checklist helps define the level of detail that should exist and is being adopted by many journals as a requirement for the submission of papers incorporating microarray results. MIAME describes the minimum required information for complying experiments, but not its format. Thus, as of 2007, whilst many formats can support the MIAME requirements there is no format which permits verification of complete semantic compliance.
b. The "MicroArray Quality Control (MAQC) Project" is being conducted by the FDA to develop standards and quality control metrics which will eventually allow the use of MicroArray data in drug discovery, clinical practice and regulatory decision-making.
3. The MicroArray and Gene Expression (MAGE) group is working on the standardization of the representation of gene expression data and relevant annotations.

3. Statistical Analysis:

The analysis of DNA microarrays poses a large number of statistical problems, including the normalization of the data. There are dozens of proposed normalization methods in the published literature; as in many other cases where authorities disagree, a sound conservative approach is to try a number of popular normalization methods and compare the conclusions reached; how sensitive are the main conclusions to the method chosen?
From a hypothesis-testing perspective, the large number of genes present on a single array means that the experimenter must take into account a multiple testing problem; even if the statistical P-value assigned to a given gene indicates that it is extremely unlikely that differential expression of this gene was due to random rather than treatment effects, the very high number of genes on an array makes it likely that differential expression of some genes represents false positives or false negatives.
Statistical methods tailored to microarray analyses have recently become available that assess statistical power based on the variation present in the data and the number of experimental replicates, and can help minimize type I and type II errors in the analyses.
A basic difference between microarray data analysis and much traditional biomedical research is the dimensionality of the data. A large clinical study might collect 100 data items per patient for thousands of patients. A medium-size microarray study will obtain

many thousands of numbers per sample for perhaps a hundred samples. Many analysis techniques treat each sample as a single point in a space with thousands of dimensions, then attempt by various techniques to reduce the dimensionality of the data to something humans can visualize.

4. Relation between Probe and Gene:

The relation between a probe and the mRNA that it is expected to detect is problematic. On the one hand, some mRNAs may cross-hybridize probes in the array that are supposed to detect another mRNA. On the other hand, probes that are designed to detect the mRNA of a particular gene may be relying on genomic EST information that is incorrectly associated with that gene.

Public Databases of Microarray Data:

Database	Microarray Experiment Sets	Sample Profiles	As of Date
Gene Expression Omnibus - NCBI	5366	134669	April 1, 2007
Stanford Microarray database	12742	?	April 1, 2007
UNC Microarray database	~31	2093	April 1, 2007
MUSC database	~45	555	April 1, 2007
ArrayExpress at EBI	1643	136	April 1, 2007
caArray at NCI	41	1741	November 15, 2006

Online Microarray Data Analysis Programs and Tools:

Several Open Directory Project categories list online microarray data analysis programs and tools:

i. Bioinformatics: Online Services:

Gene Expression and Regulation at the Open Directory Project

ii. Gene Expression:

Databases at the Open Directory Project

iii. Gene Expression:

Software at the Open Directory Project

iv. Data Mining:

Tool Vendors at the Open Directory Project

v. Bio-conductor:

Open source and open development software project for the analysis and comprehension of genomic data

vi. Genevestigator:

Web-based database and analysis tool to study gene expression across large sets of tissues, developmental stages, drugs, stimuli, and genetic modifications.

DNA Chip and Development study:

The mechanistic basis of metazoan development represents one of the unsolved mysteries of biology: how does a single fertilized egg, through successive cell divisions and differentiation events, mature into an adult organism? The fruitfly *Drosophila melanogaster* has been a pioneering model organism for geneticists and developmental biologists for many decades. Drosophilogists have been quick to exploit the power of genome-wide expression profiling using DNA microarrays. One notable example is the study of the expression of 4028 genes analysed in wild-type flies throughout *Drosophila* development during 66 sequential time periods. These included sampling RNA at fertilization, embryonic, larval and pupal periods as well as the first 30 days of adulthood. Each experimental sample was compared with a common reference sample, allowing the relative abundance of any transcript to be determined at every developmental stage. The analysis of such a huge amount of data conventionally proceeds by the use of algorithms that group or cluster genes according to similarity in their expression profiles.

The analysis of the *Drosophila* dataset revealed that, despite the use of whole animals, it was possible to discern expression profiles in specific organs, as well as those associated with particular biological processes. For example, one cluster of 23 genes included eight known to be expressed in terminally differentiated muscle. The profile of this cluster has two peaks of expression, one coinciding with the larval stage and a second with adult muscle development. Initiation of larval muscle development is regulated by the basic helix–loop–helix (bHLH) transcription factor Twist, which induces expression of *dMef 2*, which itself encodes a MADS box transcription factor regulating the transcription of muscle differentiation genes. Crucially, this muscle-specific regulatory hierarchy was recapitulated in the microarray data: the peak of *twist* expression preceded that of *dMef 2*, which preceded transcription of genes in a muscle differentiation cluster. Moreover, 15 of the 23 genes in this latter cluster contained pairs of predicted dMEF2-binding sites. Similar clusters were identified revealing coordinate expression profiles associated with particular biochemical and cellular functions, including mitochondrial proteins, components of the 26S proteasome complex and cytoskeletal/neuronal factors.

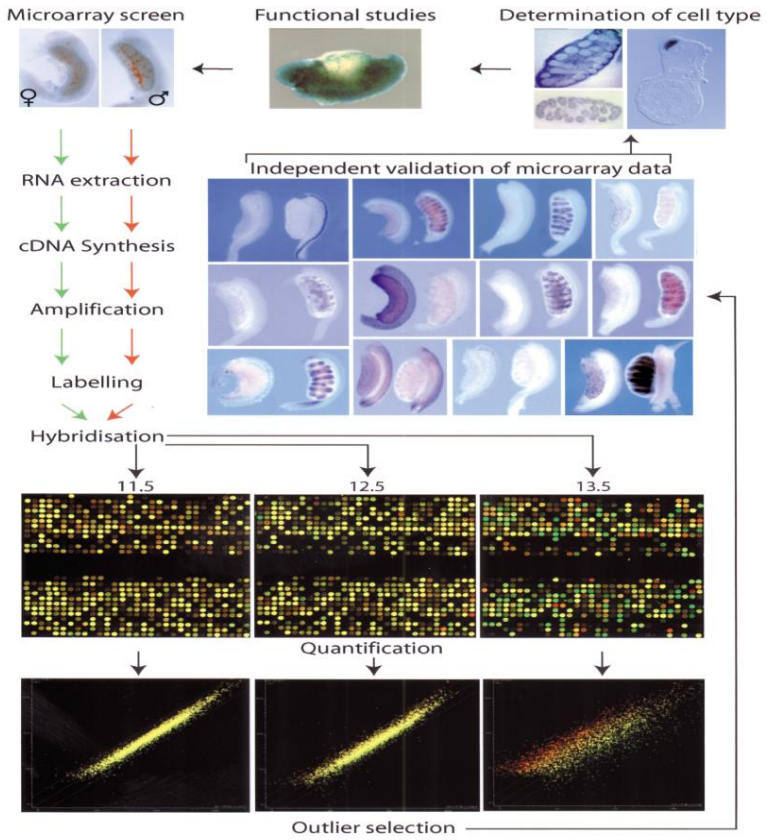


Fig 5: A screen for genes expressed in a sexually dimorphic fashion during mouse gonad and mesonephros development using DNA microarrays.

Global transcriptional information during morphogenesis was also readily available: the vast majority of genes (>88%) that exhibit transcriptional modulation during the stages analysed are expressed during the first 20 h of development, before the end of embryogenesis. A total of 2103 changed during embryogenesis, 445 changed during larval life, 646 during the pupal stage and 118 during adult life. The transcript levels of only 16 genes changed significantly during the adult time period sampled. These data suggest a strong association between modulation of transcriptional activity and morphogenesis.

The pioneering experiments in invertebrates suggest that the notion that gene expression profiles alone do not reveal biological function needs to be re-examined. Surveying gene expression under a wide range of conditions and tissues, in wild-type and mutant animals, seems to transform the significance of data that on a smaller scale would be considered descriptive. Of course, for any individual gene residing within a 'functionally loaded' cluster the task remains to determine the phenotypic consequences of its mutation. Yet perhaps such experiments should be seen as complementing our understanding of that gene's function developed by other means, rather than being exclusively definitive thereof: particularly given the high frequency with which no clear phenotype is observed after mutagenesis. To infer function from

gene expression profiles is not mere speculation if the design of the experiment and the complexity of the dataset permit otherwise

Studies were done in the mammalian embryo at a genome-wide level throughout its development, in a manner reminiscent of those discussed in invertebrates. Given the widespread accessibility of microarray technology today, the observations and analysis are very complex, involving references to both technical and 'cultural' issues. The most common technical problem concerns the small amounts of RNA available from standard dissections of mammalian embryos. By 'cultural', we mean the familiarity that developmental biologists have with in situ hybridization (ISH), their relative lack of familiarity with microarrays and the common attitude that descriptions of gene expression patterns support only speculation about function. However, these remarks are equally applicable to developmental biologists using flies and worms as a model. The use of DNA microarrays to examine mammalian development is a small but rapidly growing field of study. It is currently dominated by the exploitation of arrays to perform screens for molecules involved in particular developmental processes.

Systematic genome-wide studies of mammalian development using microarrays stand out due to their rarity. Studies in mouse and the analysis of the expression of 18 816 mouse genes in 49 different embryonic and adult tissues, permitted some clustering of genes pertinent to the development of specific tissues, such as the central nervous system. However, the limited number of embryonic samples, totalling 11, means that this study falls short of providing a transcriptional profile of mouse development. Perhaps due to the relative complexity of the mammalian embryo, more familiar are studies aimed at profiling expression at specific embryonic stages or in specific embryonic tissues, including (without attempting to be comprehensive): 12.5 days post coitum (dpc) mouse placenta, mouse retina, mouse lung, mouse mammary gland, preimplantation mouse embryos, mouse hippocampus, and mouse B cells. Developmental biologists have also been quick to adapt familiar techniques for the purposes of exploiting microarray technology, including the use of cell line models and organ cultures too.

Applications of these Arrays include:

1. mRNA or gene expression profiling:

Monitoring expression levels for thousands of genes simultaneously is relevant to many areas of biology and medicine, such as studying treatments, disease, and developmental stages. For example, microarrays can be used to identify disease genes by comparing gene expression in diseased and normal cells.

2. Comparative genomic hybridization (Array CGH):

Assessing large genomic rearrangements

3. SNP detection arrays:

Looking for single nucleotide polymorphism in the genome of populations

4. Chromatin immunoprecipitation (ChIP) studies:

Determining protein binding site occupancy throughout the genome, employing ChlP-on-chip technology

Protein DNA Interaction study:

DNA footprinting is a method of investigating the sequence specificity of DNA-binding proteins in vitro. This technique can be used to study protein-DNA interactions both outside and within cells.

The regulation of transcription has been studied extensively, and yet there is still much that is unknown. Transcription factors and associated proteins that bind promoters, enhancers, or silencers to drive or repress transcription are fundamental to understanding the unique regulation of individual genes within the genome. Techniques like DNA footprinting help elucidate which proteins bind to these associated regions of DNA and unravel the complexities of transcriptional control.

History:

In 1978, David J. Galas and Albert Schmitz developed the DNA footprinting technique to study the binding specificity of the lac repressor protein. It was originally a modification of the Maxam-Gilbert chemical sequencing technique.

Method:

The simplest application of this technique is to assess whether a given protein binds to a region of interest within a DNA molecule.[2] Polymerase chain reaction (PCR) amplify and label region of interest that contains a potential protein-binding site, ideally amplicon is between 50 and 200 base pairs in length. Add protein of interest to a portion of the labeled template DNA; a portion should remain separate without protein, for later comparison. Add a cleavage agent to both portions of DNA template. The cleavage agent is a chemical or enzyme that will cut at random locations in a sequence independent manner. The reaction should occur just long enough to cut each DNA molecule in only one location. A protein that specifically binds a region within the DNA template will protect the DNA it is bound to from the cleavage agent. Run both samples side by side on a polyacrylamide gel electrophoresis. The portion of DNA template without protein will be cut at random locations, and thus when it is run on a gel, will produce a ladder-like distribution. The DNA template with the protein will result in

ladder distribution with a break in it, the "footprint", where the DNA has been protected from the cleavage agent. Note: Maxam-Gilbert chemical DNA sequencing can be run alongside the samples on the polyacrylamide gel to allow the prediction of the exact location of ligand binding site.

Labeling:

The DNA template labeled at the 3' or 5' end, depending on the location of the binding site(s). Labels that can be used are: radioactivity and fluorescence. Radioactivity has been traditionally used to label DNA fragments for footprinting analysis, as the method was originally developed from the Maxam-Gilbert chemical sequencing technique. Radioactive labeling is very sensitive and is optimal for visualizing small amounts of DNA. Fluorescence is a desirable advancement due to the hazards of using radio-chemicals. However, it has been more difficult to optimize because it is not always sensitive enough to detect the low concentrations of the target DNA strands used in DNA footprinting experiments. Electrophoretic sequencing gels or capillary electrophoresis have been successful in analyzing footprinting of fluorescent tagged fragments.

Cleavage agent:

A variety of cleavage agents can be chosen. a desirable agent is one that is sequence neutral, easy to use, and is easy to control. Unfortunately no available agents meet all of these standards, so an appropriate agent can be chosen, depending on your DNA sequence and ligand of interest. The following cleavage agents are described in detail: DNase I is a large protein that functions as a double-strand endonuclease. It binds the minor groove of DNA and cleaves the phosphodiester backbone. It is a good cleavage agent for footprinting because its size makes it easily physically hindered. Thus is more likely to have its action blocked by a bound protein on a DNA sequence. In addition, the DNase I enzyme is easily controlled by adding EDTA to stop the reaction. There are however some limitations in using DNase I. The enzyme does not cut DNA randomly; its activity is affected by local DNA structure and sequence and therefore results in an uneven ladder. This can limit the precision of predicting a protein's binding site on the DNA molecule. Hydroxyl radicals are created from the Fenton reaction, which involves reducing $Fe2+$ with $H2O2$ to form free hydroxyl molecules. These hydroxyl molecules react with the DNA backbone, resulting in a break. Due to their small size, the resulting DNA footprint has high resolution. Unlike DNase I they have no sequence dependence and result in a much more evenly distributed ladder. The negative aspect of using hydroxyl radicals is that they are more time consuming to use, due to a slower reaction and digestion time.Ultraviolet irradiation can be used to excite nucleic acids and create photoreactions, which results in damaged bases in the DNA strand. Photoreactions can

include: single strand breaks, interactions between or within DNA strands, reactions with solvents, or crosslinks with proteins. The workflow for this method has an additional step, once both your protected and unprotected DNA have been treated, there is subsequent primer extension of the cleaved products. The extension will terminate upon reaching a damaged base, and thus when the PCR products are run side-by-side on a gel; the protected sample will show an additional band where the DNA was crosslinked with a bound protein. Advantages of using UV are that it reacts very quickly and can therefore capture interactions that are only momentary. Additionally it can be applied to in vivo experiments, because UV can penetrate cell membranes. A disadvantage is that the gel can be difficult to interpret, as the bound protein does not protect the DNA, it merely alters the photoreactions in the vicinity.

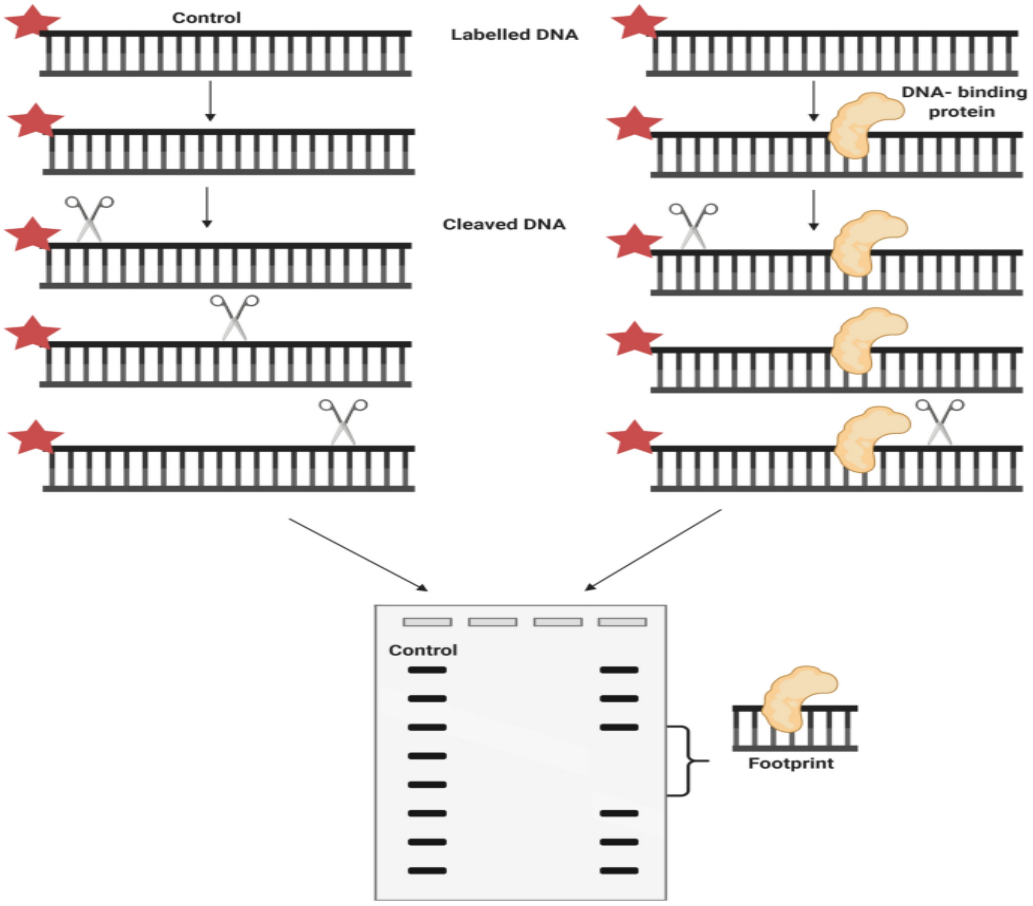


Figure: DNA Footprinting assay

Advanced applications

In vivo footprinting:

In vivo footprinting is a technique used to analyze the protein-DNA interactions that are occurring in a cell at a given time point. DNase I can be used as a cleavage agent if the cellular membrane has been permeabilized. However the most common cleavage agent used is UV irradiation because it penetrates the cell membrane without disrupting cell state and can thus capture interactions that are sensitive to cellular changes. Once the DNA has been cleaved or damaged by UV, the cells can be lysed and DNA purified for analysis of a region of interest. Ligation-mediated PCR is an alternative method to footprint in vivo. Once a cleavage agent has been used on the genomic DNA, resulting in single strand breaks, and the DNA is isolated, a linker is added onto the break points. A region of interest is amplified between the linker and a gene-specific primer, and when run on a polyacrylamide gel, will have a footprint where a protein was bound. In vivo footprinting combined with immunoprecipitation can be used to assess protein specificity at many locations throughout the genome. The DNA bound to a protein of interest can be immunoprecipitated with an antibody to that protein, and then specific region binding can be assessed using the DNA footprinting technique.

Quantitative footprinting:

The DNA footprinting technique can be modified to assess the binding strength of a protein to a region of DNA. Using varying concentrations of the protein for the footprinting experiment, the appearance of the footprint can be observed as the concentrations increase and the proteins binding affinity can then be estimated.

Detection by capillary electrophoresis:

To adapt the footprinting technique to updated detection methods, the labelled DNA fragments are detected by a capillary electrophoresis device instead of being run on a polyacrylamide gel. If the DNA fragment to be analyzed is produced by polymerase chain reaction (PCR), it is straightforward to couple a fluorescent molecule such as carboxyfluorescein (FAM) to the primers. This way, the fragments produced by DNaseI digestion will contain FAM, and will be detectable by the capillary electrophoresis machine. Typically, carboxytetramethyl-rhodamine (ROX)-labelled size standards are also added to the mixture of fragments to be analyzed. Binding sites of transcription factors have been successfully identified this way.

Genome-wide assays:

Next-generation sequencing has enabled a genome-wide approach to identify DNA footprints. Open chromatin assays such as DNase-Seq and FAIRE-Seq have proven to

provide a robust regulatory landscape for many cell types. However, these assays require some downstream bioinformatics analyses in order to provide genome-wide DNA footprints. The computational tools proposed can be categorized in two classes: segmentation-based and site-centric approaches.

Segmentation-based methods are based on the application of Hidden Markov models or sliding window methods to segment the genome into open/closed chromatin region. Examples of such methods are: HINT, Boyle method[18] and Neph method.[19] Site-centric methods, on the other hand, find footprints given the open chromatin profile around motif-predicted binding sites, i.e., regulatory regions predicted using DNA-protein sequence information (encoded in structures such as position weight matrix). Examples of these methods are CENTIPEDE and Cuellar-Partida method.

DNase footprinting assay:

A DNase footprinting assay is a DNA footprinting technique from molecular biology/biochemistry that detects DNA-protein interaction using the fact that a protein bound to DNA will often protect that DNA from enzymatic cleavage. This makes it possible to locate a protein binding site on a particular DNA molecule. The method uses an enzyme, deoxyribonuclease (DNase, for short), to cut the radioactively end-labeled DNA, followed by gel electrophoresis to detect the resulting cleavage pattern.

For example, the DNA fragment of interest may be PCR amplified using a 32P 5' labeled primer, with the result being many DNA molecules with a radioactive label on one end of one strand of each double stranded molecule. Cleavage by DNase will produce fragments. The fragments which are smaller with respect to the 32P-labelled end will appear further on the gel than the longer fragments. The gel is then used to expose a special photographic film.

The cleavage pattern of the DNA in the absence of a DNA binding protein, typically referred to as free DNA, is compared to the cleavage pattern of DNA in the presence of a DNA binding protein. If the protein binds DNA, the binding site is protected from enzymatic cleavage. This protection will result in a clear area on the gel which is referred to as the "footprint".

By varying the concentration of the DNA-binding protein, the binding affinity of the protein can be estimated according to the minimum concentration of protein at which a footprint is observed. This technique was developed by David J. Galas and Albert Schmitz at Geneva in 1977.

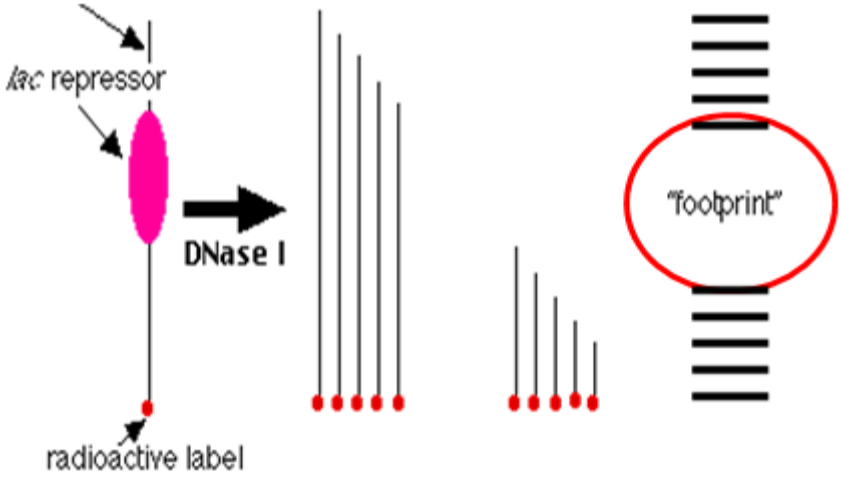


Figure: DNase Footprinting asay

Toeprinting assay:

The toeprinting assay, also known as the primer extension inhibition assay, is a method used in molecular biology that allows one to examine the interactions between messenger RNA and ribosomes or RNA-binding proteins. It is different from the more commonly used DNA footprinting assay. The toeprinting assay has been utilized to examine the formation of the translation initiation complex.

To do a toeprint assay, one needs the mRNA of interest, ribosomes, a DNA primer, free nucleotides, and reverse transcriptase (RT), among other reagents. The assay involves letting the RT generate cDNA until it gets blocked by any bound ribosomes, resulting in shorter fragments called toeprints when the results are observed on a sequencing gel.

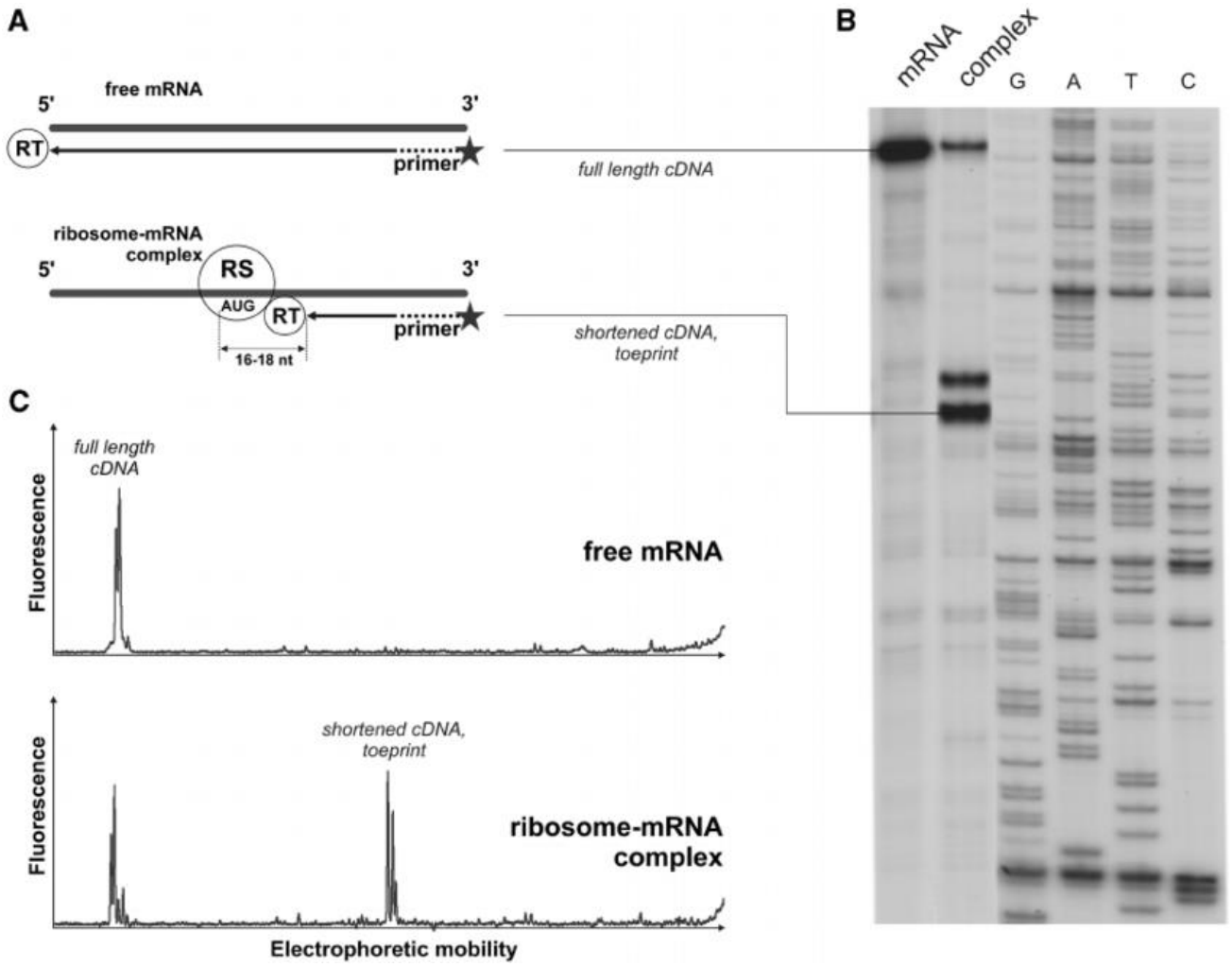


Figure: Toeprinting assay

Probable Questions:

1.How phage display technique can be used for protein protein interaction assay?
2. What is yeast two hybrid system? How it is used to determine protein protein interaction?
3. What is DNA chip?
4. State the basic principle of DNA microarray analysis.
5. What is spotted microarray? Explain.
6. What is oligonucleotide microarray? Explain .
7. How DNA chip can be used in developmental study?
8. What are the applications of DNA microarray?
9. Describe principle of DNA footprinting assay.
10. What is toeprinting assay? When it is applied?
11. Describe DNase footprinting assay.

Suggested readings:

1. Molecular Cell Biology by Lodish, Fourth Edition.
2. The Cell – A Molecular Approach by Cooper and Hausman, Fourth Edition
3. Principles of Genetics by Snustad and Simmons, Sixth Edition.
4. Molecular Biology of the Cell – by Bruce Alberts
5. Cell and Molecular Biology by Gerald Karp, 7th Edition.
6. Gene cloning and DNA Analysis by T. A. Brown, Sixth Edition.
7. Genetics . Verma and Agarwal.
8. Brown TA. (2010) Gene Cloning and DNA Analysis. 6th edition. Blackwell Publishing, Oxford, U.K.
9. Primrose SB and Twyman RM. (2006) Principles of Gene Manipulation and Genomics, 7th edition. Blackwell Publishing, Oxford, U.K.

UNIT-VII

Site Directed Mutagenesis, RNAi, Knockdown / knockout model

Objective: In this unit we will discuss about mutagenesis, RNA interference and also will discuss about knock down / knockout model for evaluating gene function.

RNA Interference:

Cells use dicer to trim double stranded RNA to form small interfering RNA or microRNA.

An exogenous dsRNA or endogenous pre-miRNA can be processed by dicer and incorporated into the RNA-induced silencing complex (RISC), which targets single-stranded messenger RNA molecules and triggers translational repression; incorporation into the RNA-induced transcriptional silencing complex (RITS) induces genome maintenance activities such as histone methylation and chromatin reorganization. RNA interference (also called "RNA-mediated interference", abbreviated RNAi) is a mechanism for RNA-guided regulation of gene expression in which double-stranded ribonucleic acid inhibits the expression of genes with complementary nucleotide sequences. Conserved in most eukaryotic organisms, the RNAi pathway is thought to have evolved as a form of innate immunity against viruses and also plays a major role in regulating development and genome maintenance.

The RNAi pathway is initiated by the enzyme dicer, which cleaves double-stranded RNA (dsRNA) to short double-stranded fragments of 20-25 base pairs. One of the two strands of each fragment, known as the guide strand, is then incorporated into the RNA-induced silencing complex (RISC) and base-pairs with complementary sequences.

The well-studied outcome of this recognition event is a form of post-transcriptional gene silencing. This occurs when the guide strand base pairs with a messenger RNA (mRNA) molecule and induces degradation of the mRNA by argonaute, the catalytic component of the RISC complex. The short RNA fragments are known as small interfering RNA (siRNA), when they derive from exogenous sources and microRNA (miRNA), when they are produced from RNA-coding genes in the cell's own genome. The RNAi pathway has been particularly well-studied in certain model organisms such as the nematode worm Caenorhabditis elegans, the fruit fly Drosophila melanogaster, and the flowering plant Arabidopsis thaliana.

The selective and robust effect of RNAi on gene expression makes it a valuable research tool, both in cell culture and in living organisms; synthetic dsRNA introduced into cells can induce suppression of specific genes of interest. RNAi may also be used for large-scale screens that systematically shut down each gene in the cell, which can help identify the components necessary for a particular cellular process or an event such as cell division. Exploitation of the pathway is also a promising tool in biotechnology and medicine. Historically, RNA interference was known by other names, including post transcriptional gene silencing, transgene silencing, and quelling. Only after these apparently-unrelated processes were fully understood it became clear that they all described the RNAi phenomenon.

RNAi has also been confused with antisense suppression of gene expression, which does not act catalytically to degrade mRNA but instead involves single-stranded RNA fragments physically binding to mRNA and blocking translation. In 2006, Andrew Fire and Craig C. Mello shared the Nobel Prize in Physiology or Medicine for their work on RNA interference in the nematode worm C. elegans, which they published in 1998.

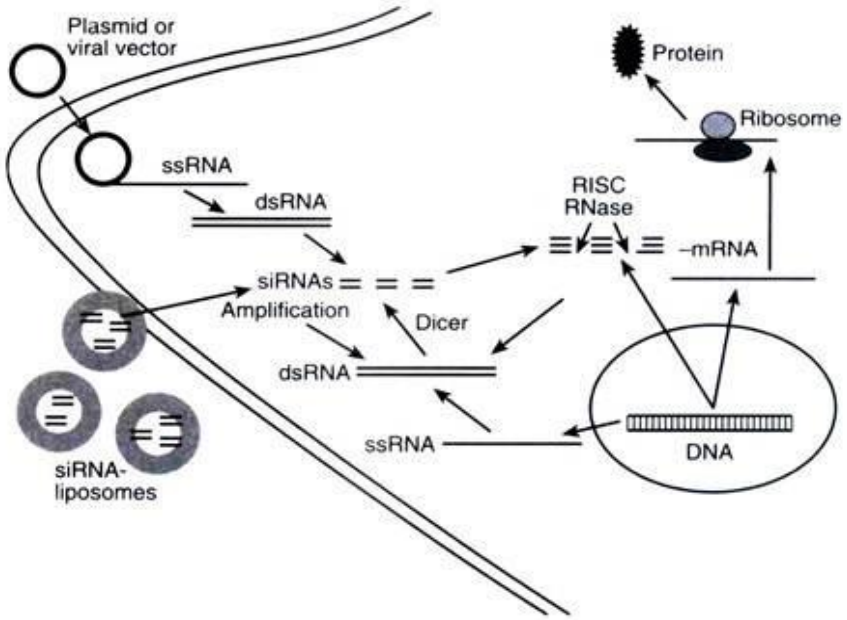


Fig. 17.1: Pathways of RNAi. Cells produce single-stranded RNA (ssRNA), which provides a template for the formation of dsRNA, which involves the activity of RNA-dependent RNA polymerases. The dsRNA is then cleaved by a protein called Dicer to form small 21-23 nucleotide siRNAs. The siRNAs (blue) then associate with the specific mRNA targeted by their nucleotide sequence (red) in a nucleic acid-protein complex called RISC, which includes RNase activity that degrades the mRNA at sites not bound by the siRNAs. The synthesis of the protein encoded by the mRNA targeted by the siRNAs is prevented, and that protein is selectively depleted from the cell. RNAi-mediated silencing can be induced experimentally by introducing synthetic siRNAs into cells using various transfection methods including liposomes (bottom left). Viral vectors can also be used to express dsRNAs against a specific gene, which are then acted upon by Dicer.

Functional Genomics:

Most functional genomics applications of RNAi in animals have used C. elegans and D. melanogaster, as these are the common model organisms in which RNAi is most effective. C. elegans is particularly useful for RNAi research for two reasons: firstly, the effects of the gene silencing are generally heritable, and secondly, because delivery of the dsRNA is extremely simple. Through a mechanism whose details are poorly understood, bacteria such as E. coli that carry the desired dsRNA can be fed to the worms and will transfer their RNA payload to the worm via the intestinal tract.

This "delivery by feeding" is just as effective at inducing gene silencing as more costly and time-consuming delivery methods, such as soaking the worms in dsRNA solution and injecting dsRNA into the gonads. Although delivery is more difficult in most other organisms, efforts are also underway to undertake large-scale genomic screening applications in cell culture with mammalian cells.

Approaches to the design of genome-wide RNAi libraries can require more sophistication than the design of a single siRNA for a defined set of experimental conditions. Artificial neural networks are frequently used to design siRNA libraries and to predict their likely efficiency at gene knockdown.

Mass genomic screening is widely seen as a promising method for genome annotation and has triggered the development of high-throughput screening methods based on microarrays. However, the utility of these screens and the ability of techniques developed on model organisms to generalize to even closely-related species has been questioned; for example, from C. elegans to related parasitic nematodes.

Functional genomics using RNAi is a particularly attractive technique for genomic mapping and annotation in plants because many plants are polyploid, which presents substantial challenges for more traditional genetic engineering methods. For example, RNAi has been successfully used for functional genomics studies in the hexaploid wheat Triticum aestivum, as well as more common plant model systems Arabidopsis thaliana and Zea mays.

History and Discovery:

The discovery of RNAi was preceded first by observations of transcriptional inhibition by antisense RNA expressed in transgenic plants and more directly by reports of unexpected outcomes in experiments performed by plant scientists in the USA and The Netherlands in the early 1990s.

In an attempt to alter flower colours in petunias, researchers introduced additional copies of a gene encoding chalcone synthase, a key enzyme for flower pigmentation into petunia plants of normally pink or violet flower colour. The overexpressed gene was expected to result in darker flowers, but instead produced less pigmented, fully or

partially white flowers, indicating that the activity of chalcone synthase had been substantially decreased; in fact, both the endogenous genes and the transgenes were down regulated in the white flowers.

Soon after, a related event termed quelling was noted in the fungus Neurospora crassa, although it was not immediately recognized as related. Further investigation of the phenomenon in plants indicated that the down regulation was due to post-transcriptional inhibition of gene expression via an increased rate of mRNA degradation. This phenomenon was called co-suppression of gene expression, but the molecular mechanism remained unknown.

Not long after, plant virologists working on improving plant resistance to viral diseases observed a similar unexpected phenomenon. While it was known that plants expressing virus-specific proteins showed enhanced tolerance or resistance to viral infection, it was not expected that plants carrying only short, non-coding regions of viral RNA sequences would show similar levels of protection.

Researchers believed that viral RNA produced by transgenes could also inhibit viral replication. The reverse experiment, in which short sequences of plant genes were introduced into viruses, showed that the targeted gene was suppressed in an infected plant. This phenomenon was labelled “virus-induced gene silencing” (VIGS), and the set of such phenomena were collectively called post-transcriptional gene silencing.

After these initial observations in plants, many laboratories around the world searched for the occurrence of this phenomenon in other organisms. Craig C. Mello and Andrew Fire’s 1998 Nature paper reported a potent gene silencing effect after injecting double stranded RNA into *C. elegans*. In investigating the regulation of muscle protein production, they observed that neither mRNA nor antisense RNA injections had an effect on protein production, but double-stranded RNA successfully silenced the targeted gene.

As a result of this work, they coined the term RNAi. Fire and Mello’s discovery was particularly notable because it represented the first identification of the causative agent of a previously inexplicable phenomenon. Fire and Mello were awarded the Nobel Prize in Physiology or Medicine in 2006 for their work.

Cellular Mechanisms:

RNAi is an RNA-dependent gene silencing process that is mediated by the RNA-induced silencing complex (RISC) and is initiated by short double- stranded RNA molecules in the cytoplasm, where they interact with the catalytic RISC component argonaute. When the dsRNA is exogenous, coming from infection by a virus with an RNA genome or laboratory manipulations, the RNA is imported directly into the cytoplasm and cleaved to short fragments by the enzyme dicer.

The initiating dsRNA can also be endogenous, as in pre-microRNAs expressed from RNA-coding genes in the genome. The primary transcripts from such genes are first processed to the characteristic stem-loop structure of pre-miRNA in the nucleus, and then exported to the cytoplasm to be cleaved by dicer. Thus the two pathways for exogenous and endogenous dsRNA converge at the RISC complex, which mediates gene silencing effects.

Fig. 17.2: The dicer protein from Giardia intestinalis, which catalyzes the cleavage of dsRNA to siRNAs. The RNase domains are coloured green, the PAZ domain yellow, the platform domain red, and the connector helix blue. The distance between the RNase and PAZ domains, determined by the length and angle of the connector helix, may determine the length of siRNA molecules produced by dicer variants.

dsRNA Cleavage:

Exogenous dsRNA initiates RNAi by activating the ribonuclease protein dicer, which binds and cleaves double-stranded RNAs (dsRNAs) to produce double-stranded fragments of 20-25 base pairs with a few unpaired overhang bases on each end. Bioinformatics studies on the genomes of multiple organisms suggest this length maximizes target-gene specificity and minimizes non-specific effects.

These short double-stranded fragments are called small interfering RNAs (siRNAs). These siRNAs are then separated into single strands and integrated into an active RISC complex. After integration into the RISC, siRNAs base-pair to their target mRNA and induce cleavage of the mRNA, thereby preventing it from being used as a translation template. Exogenous dsRNA is detected and bound by an effector protein known as

RDE-4 in C. elegans and R2D2 in Drosophila that stimulates dicer activity. This protein only binds long dsRNAs, but the mechanism producing this length specificity is unknown. These RNA-binding proteins then facilitate transfer of cleaved siRNAs to the RISC complex.

This initiation pathway may be amplified by the cell through the synthesis of a population of 'secondary' siRNAs using the dicer-produced initiating or 'primary' siRNAs as templates. These siRNAs are structurally distinct from dicer-produced siRNAs and appear to be produced by an RNA-dependent RNA polymerase (RdRP).

Fig. 17.3: *Left:* A full-length argonaute protein from the archaea species *Pyrococcus furiosus*. *Right:* The PIWI domain of an argonaute protein in complex with double-stranded RNA. The base-stacking interaction between the 5′ base on the guide strand and a conserved tyrosine residue (light blue) is highlighted; the stabilizing divalent cation (magnesium) is shown as a gray sphere.

microRNA:

MicroRNAs (miRNAs) are genomically encoded non-coding RNAs that help regulate gene expression, particularly during development. The phenomenon of RNA interference, broadly defined, includes the endogenously induced gene silencing effects of miRNAs as well as silencing triggered by foreign dsRNA.

Mature miRNAs are structurally similar to siRNAs produced from exogenous dsRNA, but miRNAs must first undergo extensive post-transcriptional modification. An miRNA is expressed from a much longer RNA-coding gene as a primary transcript known as a pri-miRNA, which is processed in the cell nucleus to a 70-nucleotide stem-loop structure called a pre-miRNA by the microprocessor complex.

This complex consists of an RNase III enzyme called Drosha and a dsRNA-binding protein Pasha. The dsRNA portion of this pre-miRNA is bound and cleaved by dicer to produce the mature miRNA molecule that can be integrated into the RISC complex; thus, miRNA and siRNA share the same cellular machinery downstream of their initial processing.

The siRNAs derived from long dsRNA precursors differ from miRNAs; in that miRNAs, especially those in animals, typically have incomplete base pairing to a target and

inhibit the translation of many different mRNAs with similar sequences. In contrast, siRNAs typically base-pair perfectly and induce mRNA cleavage only in a single, specific target. In Drosophila and C. elegans, miRNA and siRNA are processed by distinct argonaute proteins and dicer enzymes.

RISC Activation and Catalysis:

The catalytically-active components of the RISC complex are endonucleases called argonaute proteins, which cleave the target mRNA strand complementary to their bound siRNA. As the fragments produced by dicer are double-stranded, they could each in theory produce a functional siRNA.

However, only one of the two strands, which is known as the guide strand, binds the argonaute protein and directs gene silencing. The other anti-guide strand or passenger strand is degraded during RISC activation. Although it was first believed that an ATP-dependent helicase separated these two strands, the process is actually ATP-independent and performed directly by the protein components of RISC.

The strand selected as the guide tends to be that with a more stable 5′-end, but strand selection is unaffected by the direction in which dicer cleaves the dsRNA before RISC incorporation. Instead, the R2D2 protein may serve as the differentiating factor by binding the less- stable 5′-end of the passenger strand.

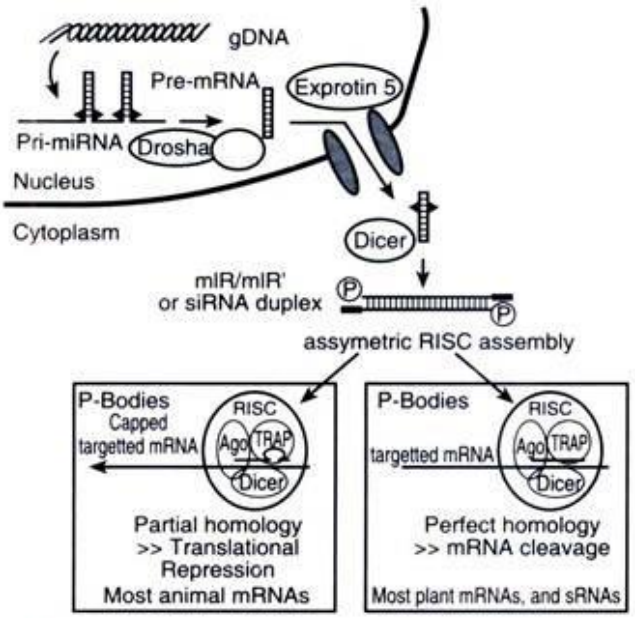


Fig. 17.4: Illustration of the major differences between plant and animal gene silencing. Natively expressed microRNA or exogenous small interfering RNA is processed by dicer and integrated into the RISC complex, which mediates gene silencing. In general, miRNAs in plants match perfectly to their gene targets and induce direct messenger RNA cleavage, while miRNAs in animals often have less than perfect base pairing to a number of targets, and induce translational repression.

The structural basis for binding of RNA to the argonaute protein was examined by X-ray crystallography of the binding domain of an RNA-bound argonaute protein. Here, the phosphorylated 5′-end of the RNA strand enters a conserved basic surface pocket and makes contacts through a divalent cation such as magnesium and by aromatic stacking between the 5′ nucleotide in the siRNA and a conserved tyrosine residue. This site is thought to form a nucleation site for the binding of the siRNA to its mRNA target.

It is not understood how the activated RISC complex locates complementary mRNAs within the cell. Although the cleavage process has been proposed to be linked to translation, translation of the mRNA target is not essential for RNAi-mediated degradation. Indeed, RNAi may be more effective against mRNA targets that are not translated.

Argonaute proteins, the catalytic components of RISC, are localized to specific regions in the cytoplasm called P-bodies (also cytoplasmic bodies or GW bodies), which are regions with high rates of mRNA decay; miRNA activity is also clustered in P-bodies. Disruption of P bodies in cells decreases the efficiency of RNA interference, suggesting that they are the site of a critical step in the RNAi process.

Variation among Organisms:

Organisms vary in their ability to take up foreign dsRNA and use it in the RNAi pathway. The effects of RNA interference can be both systemic and heritable in plants and C. elegans, although not in Drosophila or mammals. In plants, RNAi is thought to propagate by the transfer of siRNAs between cells through plasmodesmata.

A broad general distinction between plants and animals lies in the targeting of endogenously produced miRNAs; in plants, miRNAs are usually perfectly or nearly perfectly complementary to their target genes and induce direct mRNA cleavage by RISC, while animals' miRNAs tend to be more divergent in sequence and induce translational repression. This translational effect may be produced by inhibiting the interactions of translation initiation factors with the messenger RNA's polyadenine tail.

Some eukaryotic protozoa such as Leishmania major and Trypanosoma cruzi lack the RNAi pathway entirely. Most or all of the components are also missing in some fungi, most notably the model organism *Saccharomyces cerevisiae*. Certain ascomycetes and basidiomycetes are also missing RNA interference pathways; this observation indicates that proteins required for RNA silencing have been lost independently from many fungal lineages, possibly due to the evolution of a novel pathway with similar function, or to the lack of selective advantage in certain niches.

Biological Functions:

Immunity:

RNA interference is a vital part of the immune response to viruses and other foreign genetic material, especially in plants where it may also prevent self-propagation by transposons. Plants such as Arabidopsis thaliana express multiple dicer homologues that are specialized to react differently when the plant is exposed to different types of viruses.

Even before the RNAi pathway was fully understood, it was known that induced gene silencing in plants could spread throughout the plant in a systemic effect, and could be transferred from stock to scion plants via grafting. This phenomenon has since been recognized as a feature of the plant innate immune system, and allows the entire plant to respond to a virus after an initial localized encounter.

In response, many plant viruses have evolved elaborate mechanisms that suppress the RNAi response in plant cells. These include viral proteins that bind short double-stranded RNA fragments with single-stranded overhang ends, such as those produced by the action of dicer. Some plant genomes also express endogenous siRNAs in response to infection by specific types of bacteria. These effects may be part of a generalized response to pathogens that down regulates any metabolic processes in the host that aid the infection process.

Although animals generally express fewer variants of the dicer enzyme than plants, RNAi in some animals has also been shown to produce an antiviral response. In both juvenile and adult Drosophila, RNA interference is important in antiviral innate immunity and is active against pathogens such as Drosophila X virus. A similar role in immunity may operate in C. elegans, as argonaute proteins are up-regulated in response to viruses, and worms that overexpress components of the RNAi pathway are resistant to viral infection. The role of RNA interference in mammalian innate immunity is poorly understood and relatively little data is available.

However, the existence of viruses that encode genes able to suppress the RNAi response in mammalian cells may be evidence in favour of an RNAi-dependent mammalian immune response. However, this hypothesis of RNAi-mediated immunity in mammals has been challenged as poorly substantiated. Alternative functions for RNAi in mammalian viruses also exist, such as miRNAs expressed by the herpes virus that may act as heterochromatin organization triggers to mediate viral latency.

Genome Maintenance:

Components of the RNA interference pathway are used in many eukaryotes in the maintenance of the organisation and structure of their genomes. Modification of histones and associated induction of heterochromatin formation serve to down regulate

genes pre-transcriptionally; this process is referred to as RNA-induced transcriptional silencing (RITS), and is carried out by a complex of proteins called the RITS complex.

In fission yeast this complex contains argonaute, a chromo domain protein Chp1, and a protein called Task of unknown function. As a consequence, the induction and spread of heterochromatic regions requires the argonaute and RdRP proteins. Indeed, deletion of these genes in the fission yeast S. pombe disrupts histone methylation and centromere formation, causing slow or stalled anaphase during cell division. In some cases, similar processes associated with histone modification have been observed to transcriptionally up-regulate genes.

Fig. 17.5: The stem-loop secondary structure of a pre-microRNA *Brassica oleracea*

The mechanism by which the RITS complex induces heterochromatin formation and organization is not well understood and most studies have focused on the mating-type region in fission yeast, which may not be representative of activities in other genomic regions or organisms.

In maintenance of existing heterochromatin regions, RITS forms a complex with siRNAs complementary to the local genes and stably binds local methylated histones, acting co-transcriptionally to degrade any nascent pre-mRNA transcripts that are initiated by RNA polymerase. The formation of such a heterochromatin region, though not its maintenance, is dicer-dependent, presumably because dicer is required to generate the initial complement of siRNAs that target subsequent transcripts. Heterochromatin maintenance has been suggested to function as a self-reinforcing feedback loop, as new siRNAs are formed from the occasional nascent transcripts by RdRP for incorporation into local RITS complexes.

The relevance of observations from fission yeast mating-type regions and centromeres to mammals is not clear, as heterochromatin maintenance in mammalian cells may be independent of the components of the RNAi pathway.

miRNAs and Gene Regulation:

Endogenously expressed miRNAs, including both intronic and intergenic miRNAs, are most important in translational repression and in the regulation of development, especially the timing of morphogenesis and the maintenance of undifferentiated or incompletely differentiated cell types such as stem cells.

The role of endogenously expressed miRNA in down regulating gene expression was first described in C. elegans in 1993. In plants this function was discovered when the "JAW microRNA" of Arabidopsis was shown to be involved in the regulation of several genes that control plant shape.

In plants, the majority of genes regulated by miRNAs are transcription factors; thus miRNA activity is particularly wide-ranging and regulated entire gene networks during development by modulating the expression of key regulatory genes, including transcription factors as well as F- box proteins. In many organisms, including humans, miRNAs have also been linked to the formation of tumours and dysregulation of the cell cycle. Here, miRNAs can function as both oncogenes and tumour suppressors.

Crosstalk with RNA Editing:

The type of RNA editing that is most prevalent in higher eukaryotes convert's adenosine nucleotides into inosine in dsRNAs via the enzyme adenosine deaminase (ADAR). It was originally proposed in 2000 that the RNAi and A→I RNA editing pathways might compete for a common dsRNA substrate. Indeed, some pre-miRNAs do undergo A→I RNA editing, and this mechanism may regulate the processing and expression of mature miRNAs.

Furthermore, at least one mammalian ADAR can sequester siRNAs from RNAi pathway components. Further support for this model comes from studies on ADAR-null C. elegans strains indicating that A→1 RNA editing may counteract RNAi silencing of endogenous genes and transgenes.

Related Prokaryotic Systems:

Gene expression in prokaryotes is influenced by an RNA-based system similar in some respects to RNAi. Here, RNA-encoding genes control mRNA abundance or translation by producing a complementary RNA that binds to an mRNA by base pairing. However, these regulatory RNAs are not generally considered to be analogous to miRNAs because the dicer enzyme is not involved. It has been suggested that CRISPR systems in prokaryotes are analogous to eukaryotic RNA interference systems; although none of the protein components are orthologous.

Evolution:

Based on parsimony-based phylogenetic analysis, the most recent common ancestor of all eukaryotes most likely already possessed an early RNA interference pathway; the

absence of the pathway in certain eukaryotes is thought to be a derived characteristic. The ancestral RNAi system probably contained at least one dicer-like protein, one argonaute, one PIWI protein, and an RNA dependent RNA polymerase that may have also played other cellular roles.

A large-scale comparative genomics study likewise indicates that the eukaryotic crown group already possessed these components, which may then have had closer functional associations with generalized RNA degradation systems such as the exosome. This study also suggests that the RNA-binding argonaute protein family, which is shared among eukaryotes, most archaea, and at least some bacteria (such as Aquifex aeolicus), is homologous to and originally evolved from components of the translation initiation system. The ancestral function of the RNAi system is generally agreed to have been immune defense against exogenous genetic elements such as transposons and viral genomes. Related functions such as histone modification may have already been present in the ancestor of modern eukaryotes, although other functions such as regulation of development by miRNA are thought to have evolved later.

RNA interference genes, as components of the antiviral innate immune system in many eukaryotes, are involved in an evolutionary arms race with viral genes. Some viruses have evolved mechanisms for suppressing the RNAi response in their host cells, an effect that has been noted particularly for plant viruses. Studies of evolutionary rates in Drosophila have shown that genes in the RNAi pathway are subject to strong directional selection and are among the fastest-evolving genes in the Drosophila genome.

Gene Knockdown:

A wild-type adult *Caenorhabditis elegans* nematode worm, grown under RNAi suppression of a nuclear hormone receptor involved in desaturase regulation. These worms have abnormal fatty acid metabolism but are viable and fertile. The RNA interference pathway is often exploited in experimental biology to study the function of genes in cell culture and in vivo in model organisms.

Double-stranded RNA is synthesized with a sequence complementary to a gene of interest and introduced into a cell or organism, where it is recognized as exogenous genetic material and activates the RNAi pathway. Using this mechanism, researchers can cause a drastic decrease in the expression of a targeted gene.

Studying the effects of this decrease can show the physiological role of the gene product. Since RNAi may not totally abolish expression of the gene, this technique is sometimes referred to as a "knockdown", to distinguish it from "knockout" procedures in which expression of a gene is entirely eliminated.

Extensive efforts in computational biology-have been directed toward the design of successful dsRNA reagents that maximize gene knockdown but minimize "off-target"

effects. Off-target effects arise when an introduced RNA has a base sequence that can pair with and thus reduce the expression of multiple genes at a time. Such problems occur more frequently when the dsRNA contains repetitive sequences. It has been estimated from studying the genomes of H. sapiens, C. elegans, and S. pombe that about 10% of possible siRNAs will have substantial off-target effects. A multitude of software tools have been developed implementing algorithms for the design of general, mammal-specific, and virus-specific siRNAs that are automatically checked for possible cross-reactivity.

Depending on the organism and experimental system, the exogenous RNA may be a long strand designed to be cleaved by dicer, or short RNAs designed to serve as siRNA substrates. In most mammalian cells, shorter RNAs are used because long double-stranded RNA molecules induce the mammalian interferon response, a form of innate immunity that reacts nonspecifically to foreign genetic material.

Mouse oocytes and cells from early mouse embryos lack this reaction to exogenous dsRNA and are, therefore, a common model system for studying gene-knockdown effects in mammals. Specialized laboratory techniques have also been developed to improve the utility of RNAi in mammalian systems by avoiding the direct introduction of siRNA, for example, by stable transfection with a plasmid encoding the appropriate sequence from which siRNAs can be transcribed, or by more elaborate lentiviral vector systems allowing the inducible activation or deactivation of transcription, known as conditional RNAi.

Mutagenesis:

Site-directed mutagenesis is the production of either random or specific mutations in a piece of cloned DNA. Typically, the DNA will then be reintroduced into a cell or an organism to assess the results of the mutagenesis.

It is a technique conceived in the 70s used to mutate specific DNA sequences *in vitro*. It relies on synthetic short single-stranded DNA fragments, or *oligonucleotides*, that contain designated mutations to act as templates in the presence of *DNA polymerase* enzyme.

In the late 80s, *Polymerase Chain Reaction (PCR)*, another laboratory technique, was incorporated into site-directed mutagenesis workflow. The modified site-directed mutagenesis is a simple and efficient technique that has rapidly become the cornerstone of gene and protein function studies. As a result, Michael Smith, the inventor of site-directed mutagenesis, and Kary B. Mullis, the inventor of PCR, were awarded the equally shared 1993 Nobel Prize in Chemistry.

Background and Conception:

Originally termed oligonucleotide-directed mutagenesis, the technique was primarily developed in a single-stranded DNA *Escherichia coli* phage ΦX174. The core idea of the technique stemmed from the following discoveries in the bacteriophage:

1. Short oligonucleotides could form a stable double-stranded DNA complex despite a few mismatches in the nucleotides;
2. A small number of bacteriophages, which possess point mutations, could be reverted into wild-type when a wild-type DNA was used as a complementarystrand during DNA annealing before transfection;
3. Short oligonucleotides could act as a primer on a circular single-stranded DNA template for DNA replication by *DNA polymerase I* in *E. coli*; and
4. The resulting double-stranded, open circular DNA produced in (3) can be converted into closed double-stranded circular DNA using enzymatic *ligation*.

These discoveries inspired Smith to develop an oligonucleotide-based mutagenic method. In this **method**, a short strand of synthetic oligonucleotides contains the predefined mutation and acts as a mutagen that alters a specific nucleobase on the specified DNA position. The first site-directed mutagenesis reported by Smith and his

team used *one* strand of oligonucleotides as mutagens to prime with a single-stranded circular DNA of the phage ΦX174. Their nucleotide sequences were complementary to the template, except for one nucleobase designated to introduce a nonsense mutation (Figure 1).

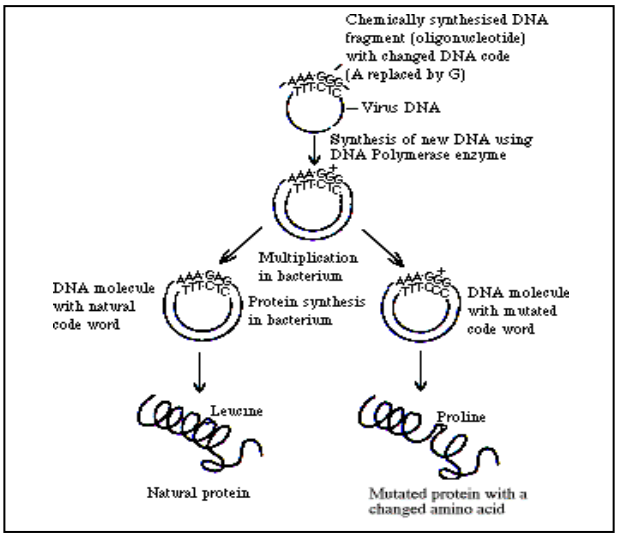


Fig 1: The principle of site-directed mutagenesis

The mutagenic oligonucleotides annealed the DNA template and served as a primer for DNA replication in the presence of the enzyme *DNA polymerase*.

The resulting product from DNA replication was a double-stranded DNA molecule that contained the point mutation derived from the mutagenic oligonucleotides.

The double-stranded DNA was circularized using the enzyme DNA ligase, separated from the residual template, and introduced into a suitable host organism for multiplication, selection, and phenotypic observation.

Reportedly, the method efficiency depended on removing non-mutagenized molecules and incomplete DNA duplexes before introducing them into the host organisms for multiplication.

Polymerase Chain Reactions (PCR)-Based Site-Directed Mutagenesis

The principle used in the mutagenesis of a circular single-stranded DNA template for introducing a predefined point mutation can be applied, with some modifications, to introduce deletion, insertion, and point mutations to circular and linear double- stranded DNA templates.

Nonetheless, the earlier site-directed mutagenesis methods were low in efficiency, laborious, and time-consuming. To increase the method efficiency, researchers focused

on the strategy used in selecting the mutated DNA molecules and removing those that were not mutated.

The invention of polymerase chain reaction in the late 80s has enabled researchers to address the efficiency of this technique using another approach. Instead of focusing on the post-mutagenesis process, researchers can shift the focus directly to the mutagenesis process.

Can PCR Create Multiple Copies of a Specific DNA Fragment *In Vitro*?

PCR is an *in vitro* technique that creates multiple copies of a designated region on the DNA template. It uses *a pair* of oligonucleotides as primers to bind with the template on opposite DNA strands and *deoxynucleic triphosphates* as building blocks to synthesize new DNA strands in the presence of a thermostable DNA polymerase enzyme. The newly synthesized DNA strands serve as templates in the following PCR rounds, thus the term *chain reactions*.

PCR consists of three steps: *denaturation*, *annealing,* and *extension*. Each step

is performed only for a few seconds at different temperatures after the other in a cycle.

Theoretically, the number of the targeted DNA fragments increases twofold after the three steps are complete. After several PCR cycles, $2^{number\ of\ cycles}$ PCR products are produced from the PCR reaction.

Reactions of PCR-Based Site-Directed Mutagenesis Process

PCR is integrated into site-directed mutagenesis via the use of oligonucleotides containing the designated mutations. Only one or both PCR primers act as mutagens for site-directed mutagenesis.

The following reactions take place during PCR-based site-directed mutagenesis:

1. In the *denaturation step*, high temperature breaks the hydrogen bonds between the double-stranded DNA template, separating the strands.
2. During the *annealing step*, the mutagenic oligonucleotides bind to complementary nucleobases on one strand. The other oligonucleotides bind to the complementary nucleobases on the opposite strand.
3. In the *extension step*, new daughter strands are synthesized by the enzyme DNA polymerase, extending the primers.
4. At the end of the extension step, the resulting two complementary daughter strands possess the mutations derived from the mutagenic primers and serve as templates in the subsequent PCR cycle, in addition to the original template.

In vitro mutagenesis

Another use of cloned DNA is *in vitro* mutagenesis in which a mutation is produced in a segment of cloned DNA. The DNA is then inserted into a cell or organism, and the effects of the mutation are studied. Mutations are useful to geneticists in enabling them to investigate the components of any biological process. However, traditional mutational analysis relied on the occurrence of random spontaneous mutations—a hit-or-miss method in which it was impossible to predict the precise type or position of the mutations obtained. In vitro mutagenesis, however, allows specific mutations to be tailored for type and for position within the gene. A cloned gene is treated in the test tube (in vitro) to obtain the specific mutation desired, and then this fragment is reintroduced into the living cell, where it replaces the resident gene.

One method of in vitro mutagenesis is oligonucleotide-directed mutagenesis. A specific point in a sequenced gene is pinpointed for mutation. An

oligonucleotide, a short stretch of synthetic DNA of the desired sequence, is made chemically. For example, the oligonucleotide might have adenine in one specific location instead of guanine. This oligonucleotide is hybridized to the complementary strand of the cloned gene; it will hybridize despite the one base pair mismatch. Various enzymes are added to allow the oligonucleotide to prime the synthesis of a complete strand within the vector. When the vector is introduced into a bacterial cell and replicates, the mutated strand will act as a template for a complementary strand that will also be mutant, and thus a fully mutant molecule is obtained. This fully mutant cloned molecule is then reintroduced into the donor organism, and the mutant DNA replaces the resident gene.

Another version of *in vitro* mutagenesis is gene disruption, or gene knockout. Here the resident functional gene is replaced by a completely nonfunctional copy. The advantage of this technique over random mutagenesis is that specific genes can be knocked out at will, leaving all other genes untouched by the mutagenic procedure.

Gene knockout:

A knockout refers to the use of genetic engineering to inactivate or remove one or more specific genes from an organism. Scientists create knockout organisms to study the impact of removing a gene from an organism, which often allows them to then learn something about that gene's function.

Gene knockout is the complete elimination of genes from an organism. Gene knockdownis the reduction of the expression of a gene in an organism. It can happen only by genetic engineering techniques**.**

Process of gene knockout:

A typical process of KO involves steps like– gene selection, Vector design and construction, transformation, Homologous recombination, selection, identification and validation, insert validation, animal model generation and phenotypic assays. We can consider gene KO as one kind of genetic engineering experiment.

Gene selection:

The very first step in KO is selecting a gene we want to knock out. A gene is selected only if its function is known and should have been involved in any biological activity or biological process we wish to study.

Some dry labwork has been carried out to understand the structure of a gene. It is mapped on a chromosome and other parameters such as gene length, No of exons and introns, promoter length and sequence, are closely examined.Dry labwork helps construct an effective vector for the experiment

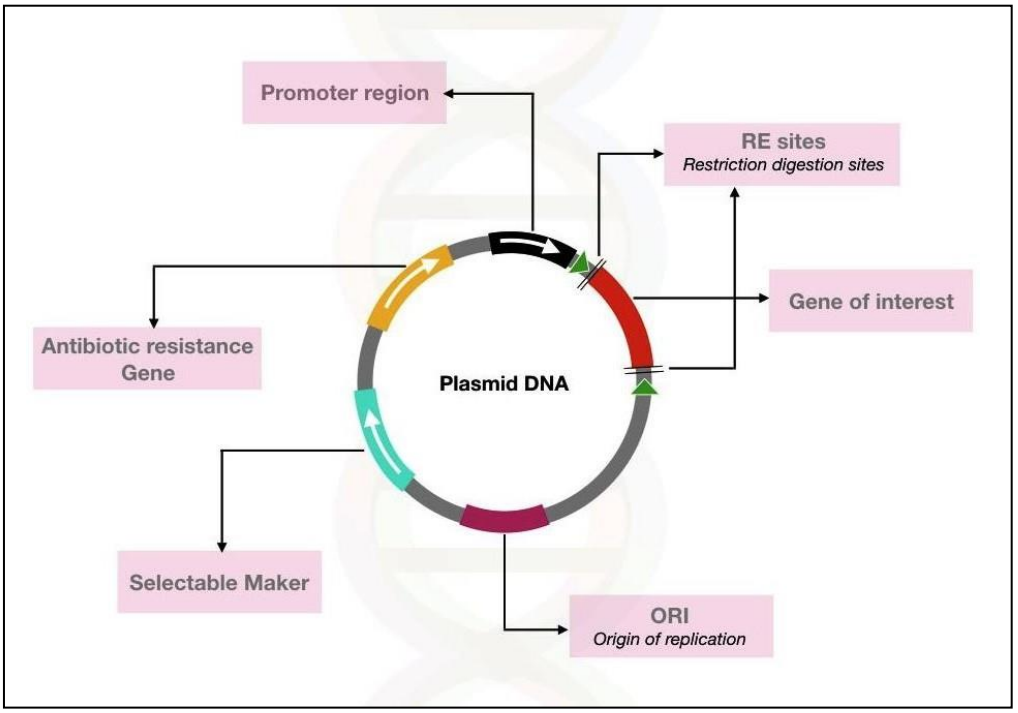


Fig 2: A general structure of a plasmid used in genetic engineering experiments

Vector design and construction:

A vector is a vehicle that carries our genetic material and transfers it to another location. Viral vectors, plasmids, cosmids and bacterial artificial chromosomes are a few common types of vectors used in gene therapy experiments.

Among all these, a plasmid is the most popular and common vector used in KO. The plasmid is bacterial extrachromosomal and circular DNA that can carry the target mutation or sequence for KO.

A vector consists of the selected marker, a gene-specific homologous region, and adisrupting element- a mutation or sequence. After vector construction and validation, our plasmid vector is ready for transformation.

Transformation:

Now, our plasmid is ready to insert into a target cell. Scientists usually select EmbryonicStem Cells (ESCs) for transformation. ESCs can divide at a faster rate into different cell types and form various tissues.

The vector once finds the gene-specific target location, it recombines and inserts our target 'change' with the marker gene in the target cells. Transformed cells are then cultured using an appropriate media.

Electroporation, Sonication and micro-injection are some common techniques for ESCs transformation. Note that cells will only grow if transformation occurs.

Homologous recombination:

After successful transformation, and after reaching the complementary region, homologous recombination will occur which results in the disruption of a gene's function later on. HR introduces the 'change' that eventually knocks out the target gene.

Selection and validation:

Now, it's important to validate whether the transformation fully occurred or not. Meaning, the experiment is successfully placed or not. Techniques like polymerase chainreaction and DNA sequencing can validate KO.

A few transformed cells are taken for DNA isolation. DNA sequencing can validate KO at the sequence level. PCR and qPCR can also be used for validation using sequence-specific primers.

qPCR can determine the concentration of the target gene from the transformed cells and determines if knockout occurred or not. Besides, restriction digestion can also be employed for KO verification but is less recommended.

After the completion of validation, the successfully transformed cells are selected for thegeneration of genetically modified model organisms.

Introduction into the animal model:

Now the transformed cells are employed to generate a KO animal using breeding and other techniques.

Phenotypic investigation:

Phenotypic analysis includes physical, biochemical and functional investigations. *In vitro*, experiments are carried out to study the effect of gene loss. Various chemical assays are also available for protein studies as well.

The entire process of gene knockout is represented in the figure below,

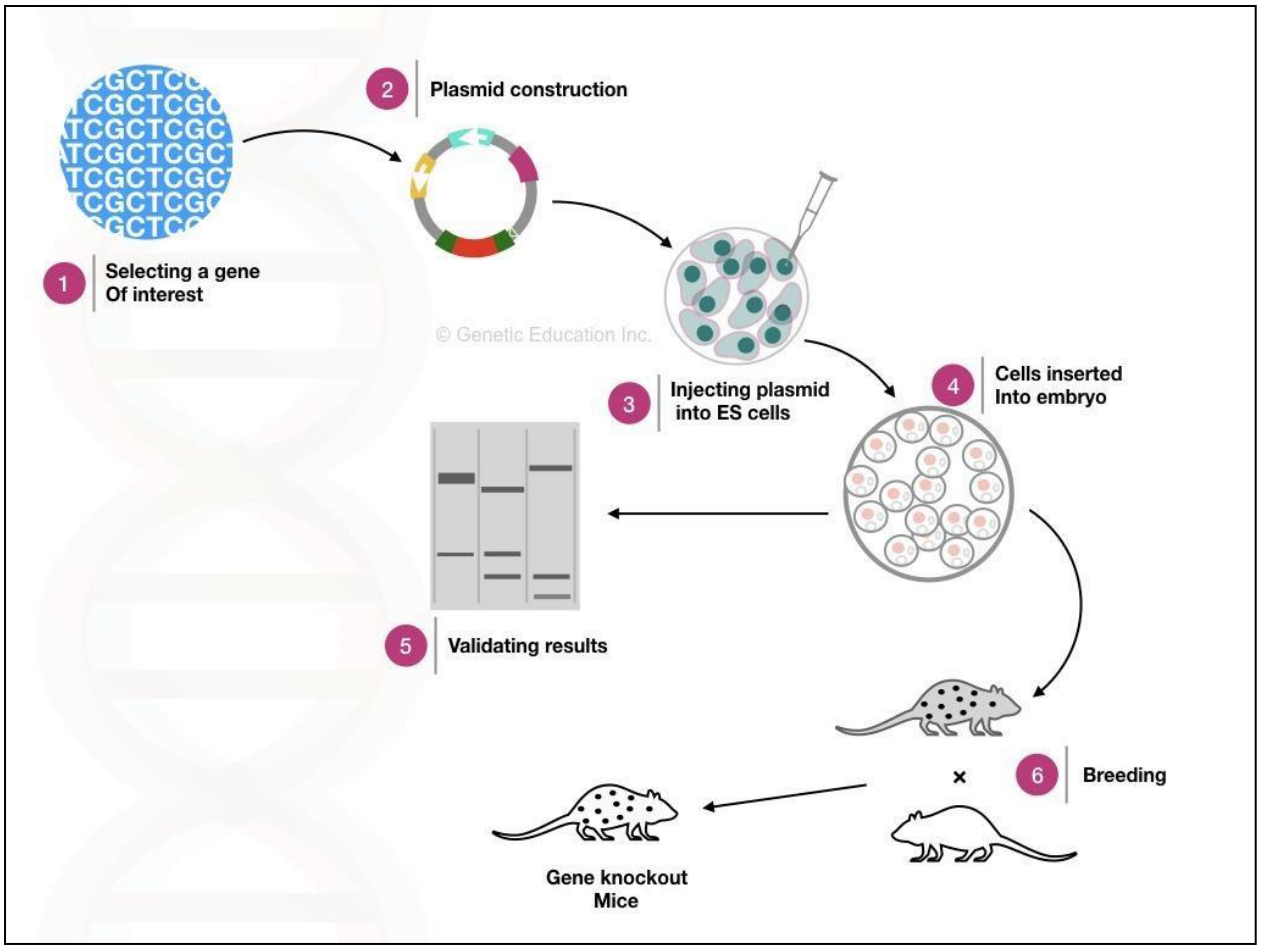


Fig 3: Entire gene knockout method

Gene knockout Methods:

So far we have discussed the KO experiment in context with the homologous recombination, however, there are other methods as well those scientists can use. Gene silencing, conditional knockout, gene editing and knockout by mutation are a few common methods for KO. We will discuss each technique one after another in this section. But remember, in any method, the purpose is to inactivate or make silence the target gene *viz* knockout.

Conditional knockout:

Conditional KO is 'specific.' It allows the inactivation of a target gene in a specific tissue or specific developmental stage. Or gene inactivation has been carried out to study a specific function in a specific cell population, tissue or developmental stage.

Conditional KO is a great tool to study tissue-specific cancer. The present technique uses homologous recombination and tissue-specific promoter for

conditional KO where

knockout only occurs in selected tissues or cells. CKO is the best way to study a gene's function at a tissue or cellular level.

Interestingly, it is not only performed in the embryonic stage but also on adult animals as well. Cre-LoxP is one such technique in which the "Cre" recombinase works on LoxP tissue-specific inactivation. Recent studies showed that CKO is widely used in breast cancer and ovarian cancer research.

Conditional knockout is very popular and by far the most successful technique for gene inactivation. We will cover the whole topic in some other article.

Gene knockout by mutation:

KO by mutation is yet another important technique. In this technique, a mutation is incorporated in a target gene or sequence in a way that deactivates or alters the target gene's function.

Such a mutation is known as loss of function mutation which can be either insertion, deletion, duplication or translocation. Artificial mutagenesis techniques like chemical alteration, radiation-induced mutagenesis or physical mutagenesis have been used to incorporate a mutation. Notedly, artificial mutagenesis induces mutation in a gene's coding, promoter or any regulatory region to knock it out.

Gene knockout by gene therapy:

Knockout by gene therapy is a novel approach, which involves the delivery of a therapeutic agent, engineered vector or system to suppress a gene's function.

CRISPR-CAS9 is a popular and modern gene therapy technique that can introduce a mutation or remove a sequence from the target gene. Any target element, for example, the sgRNA in the case of the CRISPR technique, is engineered in a way that introduces a desired change in the DNA sequence.

Such techniques are very accurate and precise for gene knockout. Gene therapy approaches hold promise for potential applications in treating genetic diseases. If you would like to learn more, you can read these articles.

Homologous Recombination:

Homologous recombination involves the exchange of genetic material between two homologous sequences or genes. Homologous recombination has a higher success rate. Thus, Scientists use this phenomenon for KO.

Here, scientists prepare homologous sequences complementary to the flanking region of the target gene and introduce them in the vector. The vector transfers it to the target location, where the recombinase enzyme performs recombination and exchanges the native sequence with our target sequence.

Interestingly, this process either disrupts a gene's function, eliminates it or suppresses it completely. Notedy, as aforesaid, a marker gene is also incorporated for validation.

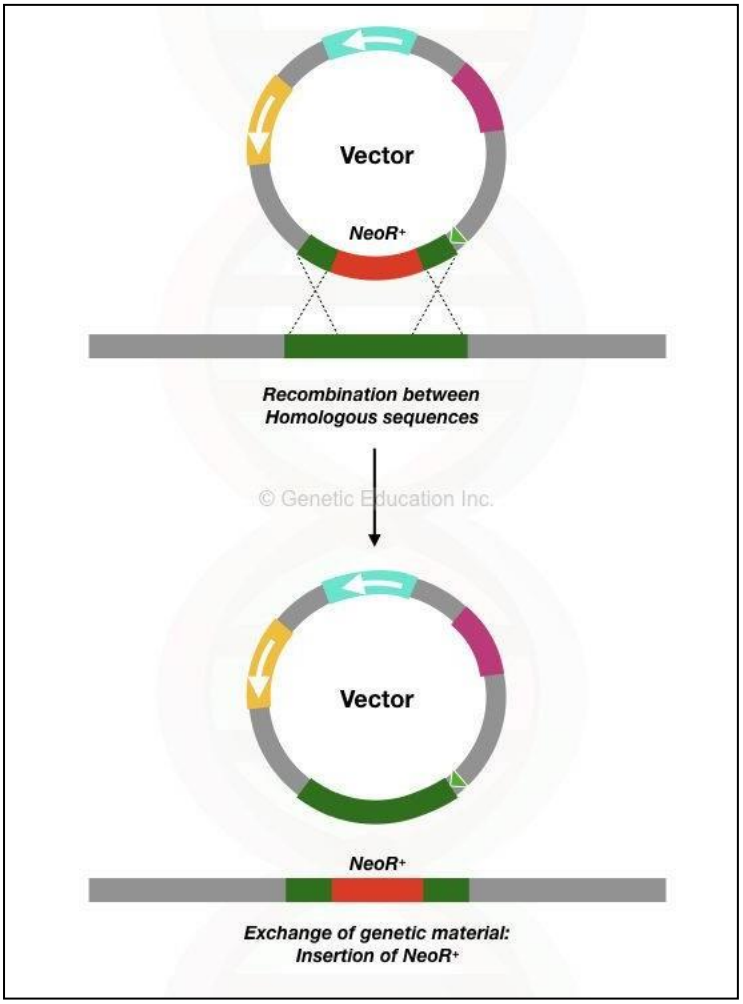


Fig 4: Hypothetical representation of recombination event between the plasmid and target gene.

Applications of Gene Knockout:

The key function of KO is to eliminate or inactivate a gene. Thus, making it functionally inactive. The present revolutionary approach has opened new doors for researchers to study diseases, genes and traits. Here are some of the amazing and important applications of gene knockout.

1. Understanding a gene's function:

Discussion so far demonstrates that the pivotal function of KO is to understand a gene's function. Scientists can study alterations in phenotypes or traits by selective knockout or target gene inactivation. KO experiments on model organisms help study genes' role in development, behavior, biological activities, physiology and disease progression. Eventually, such studies contribute to develop new strategies and therapies for diseases.

2. Genotype and phenotype relation:

Gene knockout is a great tool to establish the relationship between a gene and a phenotype. Through selective gene inactivation, its effect on a phenotype or group of phenotypes can be studied. This way, scientists understand the relationship between a gene and a phenotype.

3. Functional genomics:

Functional genomics is largely defined as 'to understand the overall function of the genome.' KO is a crucial tool in functional genomic studies. Through selective gene targeting, scientists can systematically investigate the impact of individual genes on related functions, pathways, cellular activities, signaling pathways, etc.

Scientists can understand the function as well as the evolutionary importance of homologous genes and gene families too.

4. Evolutionary studies:

Gene knockout is also an amazing tool for studying Evolution and the role of a gene in evolutionary events. By selectively inactivating a gene across different organisms, scientists can understand a gene's effect and its evolutionary importance. This helps to understand the underlying mechanism of evolution and biodiversity.

5. Disease studies:

KO also helps understand the molecular mechanism behind diseases through gene-level analysis. Scientists prepare animal models to perform KO for a particular disease- governing gene and understand its effect.

These studies help us understand how genetic diseases happen and how certain genes or groups of genes play a role in causing disease. This understanding helps in creating new treatments for the disease.

6. Drug studies and development:

Assessing a disease at the molecular level and using gene knockout techniques also aids in understanding drug responses. This means that scientists can evaluate the effectiveness, efficiency, and response of a drug when targeting specific genes or when those genes are not targeted.

Such valuable information provides value in drug development, studies and discovery. Italso strengthens the therapeutic approach against a disease.

7. Developing gene therapy:

Gene knockout recently has given more strength to the field of gene therapy.

Newtherapeutic approaches and gene therapy treatments can be developed against particular or many gene defects.

Various CRISPR-cas9 therapies mediated by KO are now under clinical trials and ready to rule the world. Other therapies are also under research.

8. Agriculture and livestock improvement:

Not only in disease have studies and human genetics, but gene knockout techniques also had great value in the agriculture industry and livestock improvement. Through knockout studies, researchers can understand the function of an economically important trait and thereby can enhance it.

Desirable and economically important traits like disease-resistant, yield improvement, nutrient values in plants and growth, development and milking capacity in animals can be improved.

Limitations of Gene Knockout:

Gene knockout is a highly useful and precise tool in the field of genetic engineering. However, like other techniques, it also has certain important limitations. Here are some limitations of gene knockout (KO):

1. Off-targeting effects:

Any gene knockout experiment is designed to specifically inactivate a target gene, however, often times, it can suppress or affect other gene's activity, unintentionally. Or in other cases, it inactivates entirely a different gene. These two scenarios are known as an off-targeting effect.

Research shows substantial off-targeting gene knockout effects, which may directly or indirectly impact the function of other gene(s) or regulatory elements.

2. Gene compensatory effects:

Compensatory mechanisms may exist for several genes that compensate or fill the function of an inactivated gene. This particularly occurs during the developmental process in which KO triggers the developmental compensatory mechanism.

Resultantly, even after successful gene knockout, the function will remain intact. Such events pose challenges to understanding the full potential effect of a gene and its governing traits.

3. Compensatory action:

In some cases, redundant genes can compensate for the impact of a target gene. Therefore, simply knocking out a single gene may not result in a loss of

function. Other genes, known as redundant genes, can step in and continue to produce the expected traits. These compensatory effects of genes can restrict the effectiveness of gene knockout.

4. Gene knockout for essential genes:

There are a set of genes in organisms that are essential for their survival. Alteration in such genes can cause lethal effects or genetic abnormalities. Scientists can't manipulate and study these genes by gene knockout.

5. Tissue specificity:

Gene knockout has less tissue specificity as it can impact a broad range of cells and tissues. So while targeting a specific gene for specific tissues, adequate results cannot bœobtained.

Probable questions:

1. What do you mean by mutagenesis?
2. Describe the principle of site-directed mutagenesis with diagram.
3. Can PCR Create Multiple Copies of a Specific DNA Fragment *In Vitro*?
4. Describe the steps of Reactions of PCR-Based Site-Directed Mutagenesis.
5. Write short notes on *In vitro* mutagenesis.
6. What is Gene knockout? Name the processes by which gene knockout is done.
7. Discuss the process of gene knockout with diagram.
8. Define Homologous recombination.
9. Discuss the method of gene knockout.
10. Discuss the Applications of Gene Knockout.
11. What are the limitations of gene knockout process?
12. Discuss mechanism of RNAi technology.
13. What are the functions of RNAi technology?
14. How genome editing can be performed by RNAi?
15. Differentiate miRNA and siRNA?
16. Discuss about RISC activation.

Suggested readings*:*

1. Snustad D P, Simmons MJ. 2009. Principles of Genetics. V Edition. John Wiley andSons Inc
2. Strickberger M. W – Genetics; Macmillam
3. Tamarin R. H. – Principles of Genetics; McGraw Hill

UNIT-VIII

Comparative genomics: Homologous genes-Orthologous, paralogous; Sequence homology; Evolutionary relationships

Objective: In this unit we will discuss about Comparative genomics: Homologous genes-Orthologous, paralogous; Sequence homology; Evolutionary relationships

Introduction:

Orthologous and paralogous genes are two types of homologous genes, that is, genes that arise from a common DNA ancestral sequence. Orthologous genes diverged after a speciation event, while paralogous genes diverge from one another within a species. Put another way, the terms orthologous and paralogous describe the relationships between genetic sequence divergence and gene products associated with speciation or genetic duplication.

Understanding Homologous Genes

Orthologous and paralogous genes are different types of homologous genes. Homologous genes are two or more genes that descend from a common ancestral DNA sequence. An example of homologous genes are the genetic codes underlying a bat wing and a bear arm. Both retain similar features and are utilized in similar manners. These traits, which were passed down from their last common ancestor, have adaptive pressures that may lead to variations within the gene. The point or event in evolutionary history that accounts for the DNA sequence variation within the gene determines whether the homologous genes are considered ‘ortho’ or ‘para’.

Orthologous Genes

Orthologous genes are homologous genes that diverged after evolution gives rise to different species, an event known as speciation. The genes generally maintain a similar function to that of the ancestral gene that they evolved from. In this type of homologous gene, the ancestral gene and its function is maintained through a speciation event, though variations may arise within the gene after the point in which the species diverged.

Paralogous Genes

Paralogous genes are homologous genes that have diverged within one species. Unlike orthologous genes, a paralogous gene is a new gene that holds a new function. These genes arise during gene duplication where one copy of the gene receives a mutation that gives rise to a new gene with a new function, though the function is often related to the role of the ancestral gene.

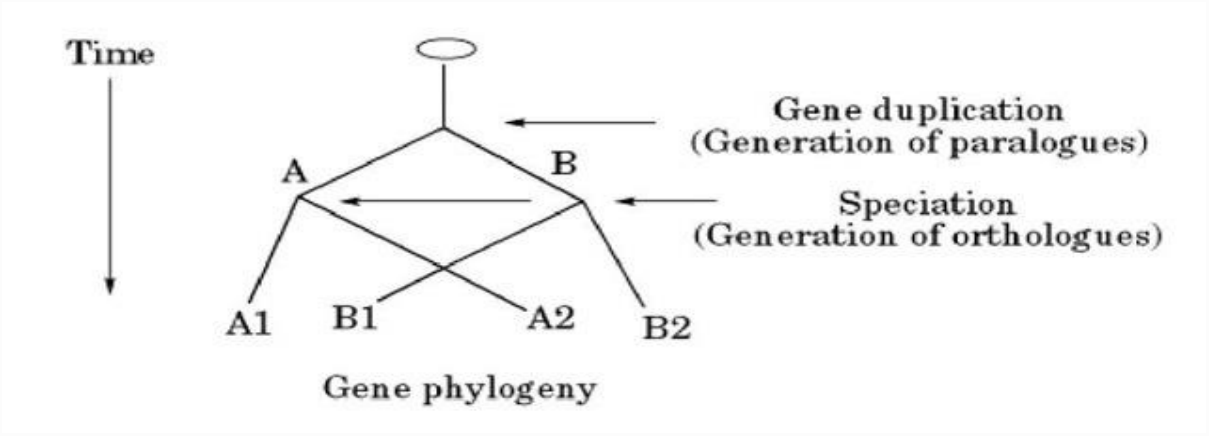


Figure 1. Generation of orthologous and paralogous genes

Examples of Paralogous and Orthlogous Genes

The genes that produce the haemoglobin and myoglobin proteins are homologous genes that have both orthologous and paralogous relationships. Both humans and dogs hold the genes for both haemoglobin and myoglobin proteins, indicating that the haemoglobin and myoglobin genes evolved before human's and dog's last common ancestor. Myoglobin arose in this ancestral species as a paralogous gene to haemoglobin; a mutation in the haemoglobin gene during a duplication event resulted in a separate myoglobin gene that carries out a new, yet similar, function. Since divergence in human and dog haemoglobin did not occur until after speciation, these genes are orthologous. Human myoglobin and dog haemoglobin, however, are homologous genes that are neither paralogous or orthologous.

Gene displacement

Comparative genomics has revealed many examples in which the same function is performed by unrelated or distantly related proteins in different cellular lineages. In some cases, this has been explained by the replacement of the original gene by a paralogue or non-homologue, a phenomenon known as non-orthologous gene

displacement. Such gene displacement probably occurred early on in the history of proteins involved in DNA replication, repair, recombination and transcription (DNA informational proteins), i.e. just after the divergence of archaea, bacteria and eukarya from the last universal cellular ancestor (LUCA). This would explain why many DNA informational proteins are not orthologues between the three domains of life. However, in many cases, the origin of the displacing genes is obscure, as they do not even have detectable homologues in another domain. I suggest here that the original cellular DNA informational proteins have often been replaced by proteins of viral or plasmid origin. As viral and plasmid-encoded proteins are usually very divergent from their cellular counterparts, this would explain the puzzling phylogenies and distribution of many DNA informational proteins between the three domains of life.

Non-orthologous gene displacement in the evolution of DNA informational proteins

Most DNA replication proteins that are functionally analogous in bacteria and eukarya have either no sequence similarities between these two domains or only limited similarities restricted to a few amino acids involved in active sites. For example, the polymerization subunit of the bacterial replicase, DNA polymerase III, has no sequence similarity at all to its eukaryotic counterparts, the catalytic subunits of DNA polymerases a, d and e. In fact, bacterial and eukaryotic replicases belong to two distinct DNA polymerase families, C and B, respectively (Ito and Braithwaite, 1991). Similarly, although the bacterial and eukaryotic initiator proteins (DnaA and ORC/CDC6 respectively) belong to the same superfamily (the AAA‡ superfamily described recently by Neuwald et al., 1999), they are very distantly related, only sharing a common ATP binding site. The situation is similar for primases, helicases and ligases. Most bacterial replicative proteins thus have no clear-cut orthologues in eukarya and vice versa (Mushegian and Koonin, 1996; Koonin and Galperin, 1997; Galperin et al., 1998). To a lesser extent, a similar situation can be observed in repair/recombination and transcription systems, as some bacterial proteins involved in these processes have no homologues in eukarya, and vice versa (Aravind et al., 1999). In contrast, most functionally analogous proteins involved in translation (e.g. ribosomal proteins, elongation factors, tRNA synthetases) are clearly orthologues in bacteria and eukarya. From comparative genomics (Brown and Doolittle, 1997; Olsen and Woese, 1997; Aravind et al., 1999), it turns out that all bacterial replicative proteins and many bacterial proteins involved in repair, recombination and transcription also have no detectable homologues in archaea. In contrast, some eukaryotic proteins involved in these mechanisms have readily detectable archaeal homologues, which are most probably their orthologues. To explain why DNA replicative proteins were so different in bacteria on one side and in eukarya/archaea on the other, Koonin and colleagues suggested that LUCA was a member of the RNA world and that DNA informational proteins originated independently in the bacterial and eukaryal (archaeal) lineages from unrelated RNA informational proteins (Mushegian and Koonin, 1996; Aravind et al., 1999). However, not all DNA informational proteins are unrelated between bacteria and the two other domains. Some of them are indeed clearly homologues in all extant cellular organisms (e.g. large subunits of DNA-dependent RNA polymerases,

ribonucleotide reductases, Topo IA, recombinases of the RecA family and a few other DNA recombination/repair proteins), suggesting to several authors that these proteins dealing with DNA were already present in a LUCA with a DNA genome. If one realizes that DNA is a specific form of modified RNA (thymine-dRNA), I think that it is indeed difficult to imagine that DNA and DNA-processing enzymes have been independently invented twice in two lineages, as many other types of nucleic acids could have been produced from RNA modification. An alternative to the RNA-LUCA hypothesis would be that DNA informational proteins have diverged to such an extent from their ancestors in LUCA that their orthology can no longer be recognized between bacteria and eukarya/archaea. However, it is not clear why DNA informational proteins should have evolved much more rapidly than proteins involved in the translation machinery. Moreover, one should explain why, despite their involvement in common mechanisms, some of them have evolved very rapidly to become apparently unrelated between bacteria and the other two domains, whereas others have evolved slowly, with their orthology still recognizable in all three domains. It has been argued that the replicative mechanism was very primitive in LUCA and was only re®ned after the divergence of the three domains (Olsen and Woese, 1997). This could indeed have produced a core of proteins present in the primitive system and still homologues in the three domains and a set of `re®nement' proteins, added later independently in each lineage and, thus, unrelated from one domain to another. However, a primitive replicative system should contain at least a replicase, a primase and a helicase, three proteins that are clearly non-orthologues between bacteria and eukarya/archaea, whereas putative `refinement' proteins, such as clamp loading factors, are homologues and possibly orthologues in the three domains! It has been suggested recently that informational proteins were so different between archaea and bacteria because archaeal proteins evolved to escape antibiotics produced by Gram-positive relatives of archaeal ancestors (Gupta, 1998). But this does not explain why these differences are especially high in the case of replicative proteins, as proteins involved in translation are, in fact, more often the targets of antibiotics. Furthermore, resistance to antibiotics usually involves only minor modifications that are restricted to a limited part of the protein. For example, archaeal DNA polymerases of the B family can be either sensitive or resistant to the drug aphidicolin in the same genus, despite a high degree of sequence identity (Forterre et al., 1994). Finally, although most archaeal and eukaryal replicative proteins are orthologues, a few archaeal replicative proteins are also unrelated or very distantly related to their eukaryal functional analogues, and vice versa. For example, the archaeal Topo VI, which is probably involved in chromosome segregation, and its functional counterparts in eukaryotes, Topo II, belong to different DNA topoisomerase II families (Bergerat et al., 1997). Similarly, eukaryal Topo IB, which is involved in the relaxation of positive superturn at the eukaryotic replication fork, has no homologues in archaeal genomes. Such an erratic pattern of relationships between archaeal and eukaryal replicative proteins cannot be explained by the LUCA-RNA theory because, as many archaeal and eukaryal replicative proteins are most probably orthologues, the common ancestor of archaea and eukarya was certainly a DNA-based organism. It cannot be

explained either by differences in evolutionary rates, as, for example, eukaryotic Topo IB (a type I DNA topoisomerase) is not even structurally and mechanistically related to its bacterial and archaeal proteins that perform the same function in DNA replication (which are both type II DNA topoisomerases) (Forterre et al., 1994). In that case, it is clear that the puzzling distribution of DNA topoisomerase between archaea and eukarya can only be explained by non-orthologous gene displacement.

The term non-orthologous gene displacement has been coined recently by Koonin to describe the presence of nonorthologous proteins (unrelated or paralogues) for the same function in different organisms (Mushegian and Koonin, 1996; Koonin and Galperin, 1997; Koonin et al., 1996; 1997) (Fig. 1A). Now that several completely sequenced archaeal and bacterial genomes are available, it is clear that the displacement of proteins responsible for essential functions by evolutionary unrelated or distantly related proteins has been extensive in the archaeal and eukaryal domains, and even more so between domains (Koonin et al., 1997; Doolittle, 1998a). A priori, one would have thought that non-orthologous gene displacement should be limited to proteins that do not physically interact with other proteins, because it is difficult to envisage the replacement of a protein that physically interacts with several partners by a phylogenetically distantly related or unrelated protein (Jain et al., 1999). However, although DNA informational proteins are often part of macromolecular complexes, many well-documented cases of non-orthologous displacement between various lineages of a single domain have now been reported for genes encoding such proteins. For example, most of the primosome components are non-homologous between E. coli and B. subtilis, whereas one of them (PriA) and other parts of the replication apparatus (DNA helicase, primase, replicase) are clearly orthologues (Bruand et al., 1995; Kunst et al., 1997). E. coli and B. subtilis also use analogous but non-homologous systems to produce single-stranded DNA for genetic recombination, the RecBCD and the AddAB/RexAB helicase/ exonuclease respectively (El Karoui et al., 1998), whereas they both use orthologous RecA to complete the recombination pathway. This example is highly significant as some Gram-positive bacteria possess the RecBCD system, whereas others have the AddAB/RexAB system, indicating that non-orthologous displacement of DNA informational proteins even occurred during the diversification of Gram-positive bacteria. The absence of homologues of the eukaryotic DNA replication initiator proteins ORC1/ CDC6 in the genome of the archaeon *Methanococcus jannaschii* also suggests recent non-orthologous displacement. Indeed, this protein, which is present in all other completely sequenced archaeal genomes, probably plays an essential role in the initiation of archaeal DNA replication (Lopez et al., 1999a). Accordingly, its function should have been taken over by another unknown protein in *M. jannaschii* (Bernander, 1998). In all these cases, one cannot explain the puzzling pattern of protein distribution observed inside one domain by divergent rates of evolution or by the RNA-LUCA hypothesis! If non-orthologous gene displacement of DNA informational proteins thus clearly occurred between both archaea and eukarya (as in the case of type II DNA topoisomerases) and after the diversification of each domain in multiple lineages it most likely that non-orthologous gene displacement also occurred between bacteria and

archaea/eukarya in the replicative, repair, recombination and transcription apparatus, explaining why some DNA informational proteins are unrelated between bacteria and archaea/eukarya. This hypothesis is more parsimonious than either the RNA-LUCA hypothesis or the replicative protein fast evolution hypothesis, as it gives the same explanation (nonorthologous gene displacement) for the presence of phylogenetically unrelated functional analogues between domains and between different lineages of a domain, whereas competitive hypotheses involved two unrelated explanations. The idea that non-orthologous gene displacement in general has played a major role in the history of DNA informational proteins is supported by many examples of functional complementation that have been observed experimentally in studying DNA replication. For example, while the removal of RNA primers from Okazaki fragments is normally performed in E. coli by the 58 to 38 exonuclease activity of DNA polymerase I, it can be done by RNase H in mutants lacking this exonuclease activity (Ogawa and Okazaki, 1984). Similarly, E. coli DNA polymerase II can be used as replicase in some E. coli DNA polymerase III mutants (Rangarajan et al., 1997). The participation of DNA polymerase II at the E. coli replication fork is in line with the observation that, despite belonging to different DNA polymerase families (B and C respectively), both E. coli DNA polymerases II and III can interact in vitro with the bacterial processivity factor b-clamp and clamp-loading factors (Bonner et al., 1992). Finally, a well-known example of functional complementation is the displacement of a thermosensitive E. coli DnaA protein by the initiator protein of another replicon at non-permissive temperature (Tresguerres et al., 1975). In that case, the E. coli mutant is rescued using the replication origin of an integrated replicon as a new oriC.

Sequence homology is the biological homology between DNA, RNA, or protein sequences, defined in terms of shared ancestry in the evolutionary history of life. Two segments of DNA can have shared ancestry because of three phenomena: either a speciation event (orthologs), or a duplication event (paralogs), or else a horizontal (or lateral) gene transfer event (xenologs).

Homology among DNA, RNA, or proteins is typically inferred from their nucleotide or amino acid sequence similarity. Significant similarity is strong evidence that two sequences are related by evolutionary changes from a common ancestral sequence. Alignments of multiple sequences are used to indicate which regions of each sequence are homologous.

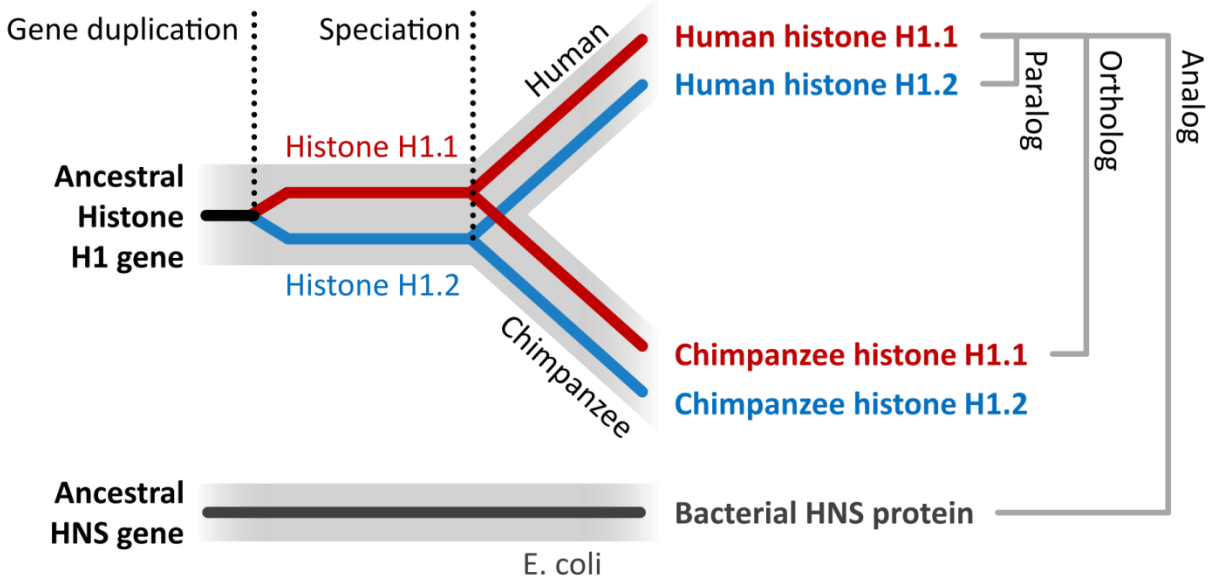


Figure: Gene phylogeny as red and blue branches within grey species phylogeny. Top: An ancestral gene duplication produces two paralogs (histone H1.1 and 1.2). A speciation event produces orthologs in the two daughter species (human and chimpanzee). Bottom: in a separate species (E. coli), a gene has a similar function (histone-like nucleoid-structuring protein) but has a separate evolutionary origin and so is an analog.

Identity, similarity, and conservation:

The term "percent homology" is often used to mean "sequence similarity", that is the percentage of identical residues (*percent identity*), or the percentage of residues conserved with similar physicochemical properties (*percent similarity*), e.g. leucine and isoleucine, is usually used to "quantify the homology." Based on the definition of homology specified above this terminology is incorrect since sequence similarity is the observation, homology is the conclusion. Sequences are either homologous or not. This involves that the term "percent homology" is a misnomer.[

As with morphological and anatomical structures, sequence similarity might occur because of convergent evolution, or, as with shorter sequences, by chance, meaning that they are not homologous. Homologous sequence regions are also called conserved. This is not to be confused with conservation in amino acid sequences, where the amino acid at a specific position has been substituted with a different one that has functionally equivalent physicochemical properties.

Partial homology can occur where a segment of the compared sequences has a shared origin, while the rest does not. Such partial homology may result from a gene fusion event

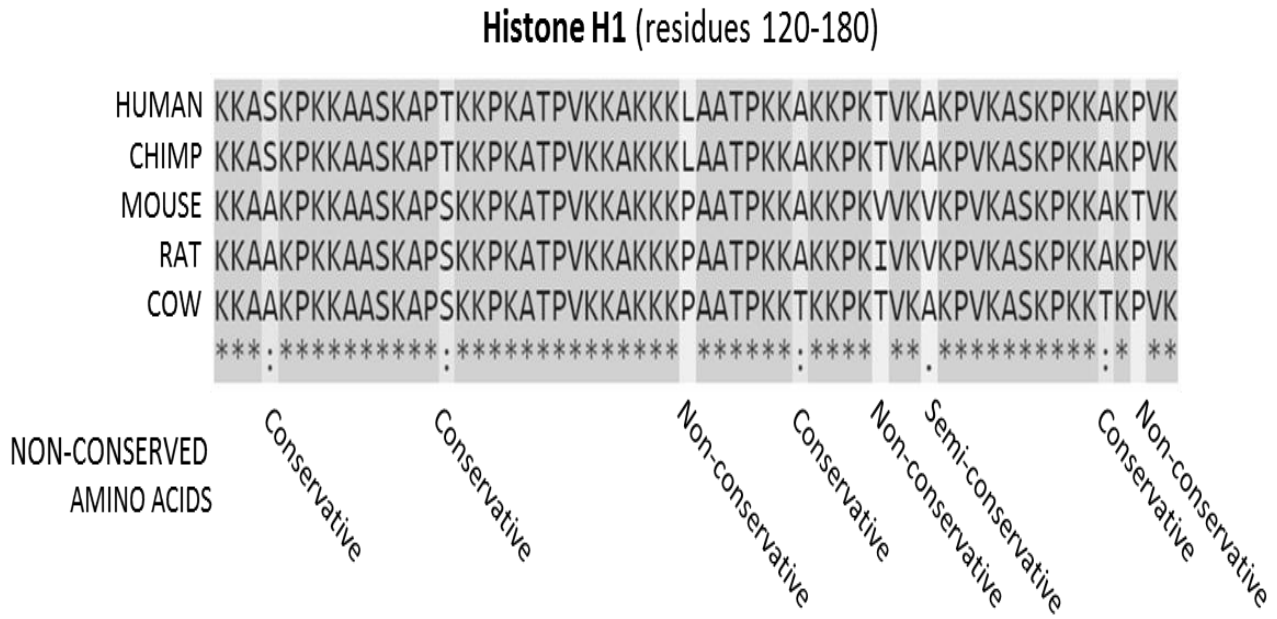


Figure: A sequence alignment of mammalian histone proteins. Sequences are the middle 120-180 amino acid residues of the proteins. Residues that are conserved across all sequences are highlighted in grey. The key below denotes conserved sequence (*), conservative mutations (:), semi-conservative mutations (.), and non-conservative mutations

Orthology:

Homologous sequences are orthologous if they are inferred to be descended from the same ancestral sequence separated by a speciation event: when a species diverges into two separate species, the copies of a single gene in the two resulting species are said to be orthologous. Orthologs, or orthologous genes, are genes in different species that originated by vertical descent from a single gene of the last common ancestor. The term "ortholog" was coined in 1970 by the molecular evolutionist Walter Fitch.

For instance, the plant Flu regulatory protein is present both in *Arabidopsis* (multicellular higher plant) and *Chlamydomonas* (single cell green algae). The *Chlamydomonas* version is more complex: it crosses the membrane twice rather than once, contains additional domains and undergoes alternative splicing. However it can fully substitute the much simpler *Arabidopsis* protein, if transferred from algae to plant genome by means of genetic engineering. Significant sequence similarity and shared functional domains indicate that these two genes are orthologous genes, inherited from the shared ancestor.

Orthology is strictly defined in terms of ancestry. Given that the exact ancestry of genes in different organisms is difficult to ascertain due to gene duplication and genome rearrangement events, the strongest evidence that two similar genes are orthologous is

usually found by carrying out phylogenetic analysis of the gene lineage. Orthologs often, but not always, have the same function.

Orthologous sequences provide useful information in taxonomic classification and phylogenetic studies of organisms. The pattern of genetic divergence can be used to trace the relatedness of organisms. Two organisms that are very closely related are likely to display very similar DNA sequences between two orthologs. Conversely, an organism that is further removed evolutionarily from another organism is likely to display a greater divergence in the sequence of the orthologs being studied.

Databases of orthologous genes:

Given their tremendous importance for biology and bioinformatics, orthologous genes have been organized in several specialized databases that provide tools to identify and analyze orthologous gene sequences. These resources employ approaches that can be generally classified into those that use heuristic analysis of all pairwise sequence comparisons, and those that use phylogenetic methods. Sequence comparison methods were first pioneered in the COGs database in 1997. These methods have been extended and automated in twelve different databases the most advanced being AYbRAH Analyzing Yeasts by Reconstructing Ancestry of Homologs as well as these following databases right now.

- eggNOG
- GreenPhylDB for plants
- InParanoid focuses on pairwise ortholog relationships
- OHNOLOGS is a repository of the genes retained from whole genome duplications in the vertebrate genomes including human and mouse.
- OMA
- OrthoDB appreciates that the orthology concept is relative to different speciation points by providing a hierarchy of orthologs along the species tree.
- OrthoInspector is a repository of orthologous genes for 4753 organisms covering the three domains of life
- OrthologID
- OrthoMaM for mammals
- OrthoMCL
- Roundup

Paralogy:

Paralogous genes are genes that are related via duplication events in the last common ancestor (LCA) of the species being compared. They result from the mutation of duplicated genes during separate speciation events. When descendants from the LCA share mutated homologs of the original duplicated genes then those genes are considered paralogs.

As an example, in the LCA, one gene (gene A) may get duplicated to make a separate similar gene (gene B), those two genes will continue to get passed to subsequent generations. During speciation, one environment will favor a mutation in gene A (gene A1), producing a new species with genes A1 and B. Then in a separate speciation event, one environment will favor a mutation in gene B (gene B1) giving rise to a new species with genes A and B1. The descendants' genes A1 and B1 are paralogous to each other because they are homologs that are related via a duplication event in the last common ancestor of the two species.

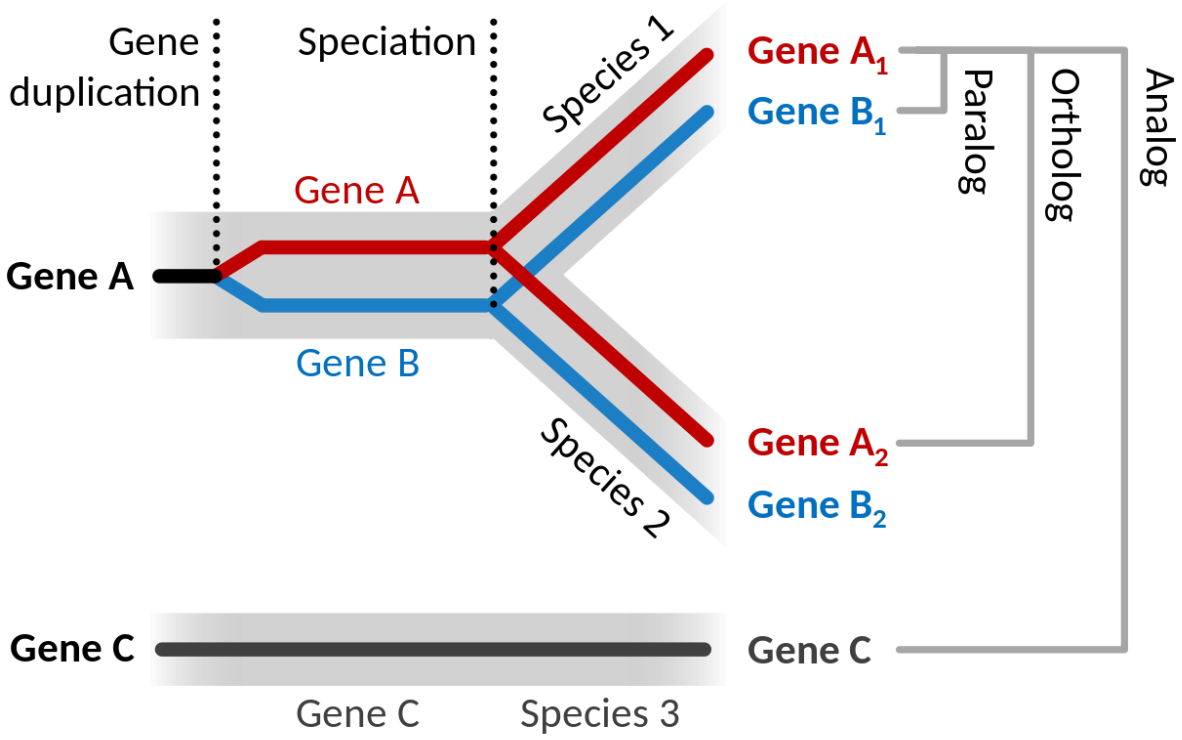


Figure: An ancestral gene duplicates to produce two paralogs (Genes A and B). A speciation event produces orthologs in the two daughter species. Bottom: in a separate species, an unrelated gene has a similar function (Gene C) but has a separate evolutionary origin and so is an analog.

Additional classifications of paralogs include alloparalogs (out-paralogs) and symparalogs (in-paralogs). Alloparalogs are paralogs that evolved from gene duplications that preceded the given speciation event. In other words, alloparalogs are paralogs that evolved from duplication events that happened in the LCA of the organisms being compared. The example above is an example alloparalogy. Symparalogs are paralogs that evolved from gene duplication of paralogous genes in subsequent speciation events. From the example above, if the descendant with genes A1

and B underwent another speciation event where gene A1 duplicated, the new species would have genes B, A1a, and A1b. In this example, genes A1a and A1b are symparalogs.

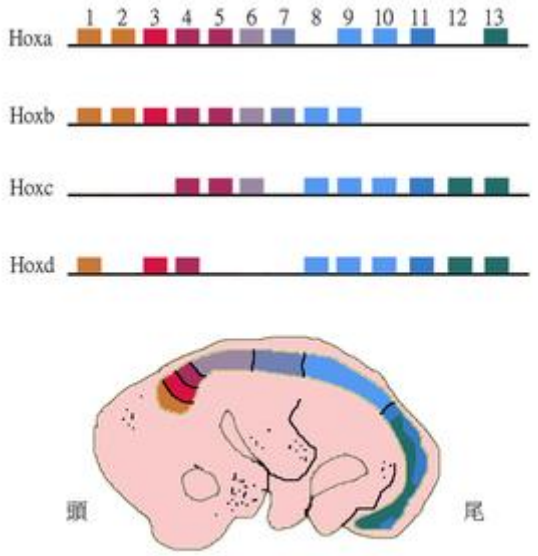


Figure: Vertebrate Hox genes are organized in sets of paralogs. Each Hox cluster (HoxA, HoxB, etc.) is on a different chromosome. For instance, the human HoxA cluster is on chromosome 7. The mouse HoxA cluster shown here has 11 paralogous genes (2 are missing).

Paralogous genes can shape the structure of whole genomes and thus explain genome evolution to a large extent. Examples include the Homeobox (Hox) genes in animals. These genes not only underwent gene duplications within chromosomes but also whole genome duplications. As a result, Hox genes in most vertebrates are clustered across multiple chromosomes with the HoxA-D clusters being the best studied.

Another example are the globin genes which encode myoglobin and hemoglobin and are considered to be ancient paralogs. Similarly, the four known classes of hemoglobins (hemoglobin A, hemoglobin A2, hemoglobin B, and hemoglobin F) are paralogs of each other. While each of these proteins serves the same basic function of oxygen transport, they have already diverged slightly in function: fetal hemoglobin (hemoglobin F) has a higher affinity for oxygen than adult hemoglobin. Function is not always conserved, however. Human angiogenin diverged from ribonuclease, for example, and while the two paralogs remain similar in tertiary structure, their functions within the cell are now quite different. It is often asserted that orthologs are more functionally similar than paralogs of similar divergence, but several papers have challenged this notion.

Regulation:

Paralogs are often regulated differently, e.g. by having different tissue-specific expression patterns (see Hox genes). However, they can also be regulated differently on

the protein level. For instance, *Bacillus subtilis* encodes two paralogues of glutamate dehydrogenase: GudB is constitutively transcribed whereas RocG is tightly regulated. In their active, oligomeric states, both enzymes show similar enzymatic rates. However, swaps of enzymes and promoters cause severe fitness losses, thus indicating promoter–enzyme coevolution. Characterization of the proteins shows that, compared to RocG, GudB's enzymatic activity is highly dependent on glutamate and pH.

Paralogous chromosomal regions:

Sometimes, large regions of chromosomes share gene content similar to other chromosomal regions within the same genome. They are well characterised in the human genome, where they have been used as evidence to support the 2R hypothesis. Sets of duplicated, triplicated and quadruplicated genes, with the related genes on different chromosomes, are deduced to be remnants from genome or chromosomal duplications. A set of paralogy regions is together called a paralogon. Well-studied sets of paralogy regions include regions of human chromosome 2, 7, 12 and 17 containing Hox gene clusters, collagen genes, keratin genes and other duplicated genes,[44] regions of human chromosomes 4, 5, 8 and 10 containing neuropeptide receptor genes, NK class homeobox genes and many more gene families, and parts of human chromosomes 13, 4, 5 and X containing the ParaHox genes and their neighbors. The Major histocompatibility complex (MHC) on human chromosome 6 has paralogy regions on chromosomes 1, 9 and 19. Much of the human genome seems to be assignable to paralogy regions.

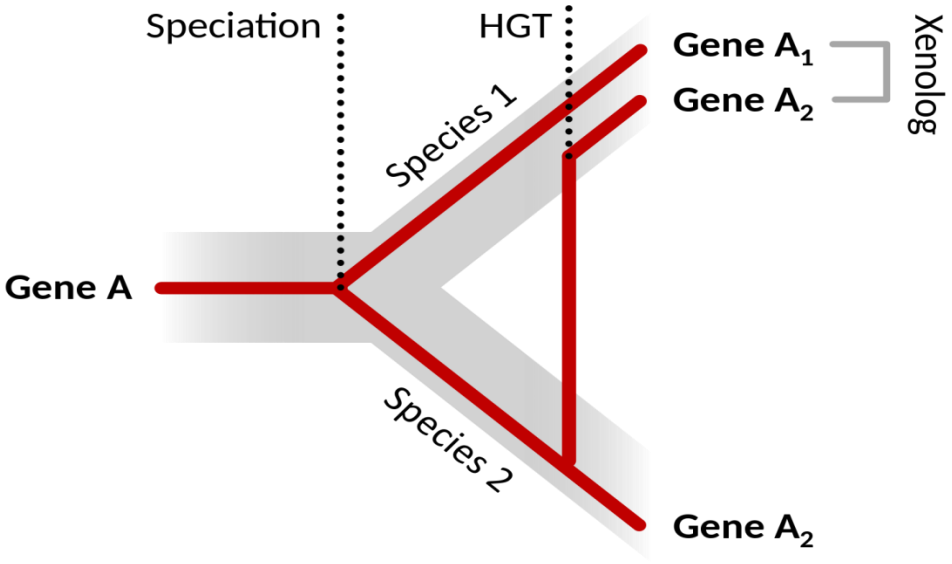


Figure: A speciation event produces orthologs of a gene in the two daughter species. A horizontal gene transfer event from one species to another adds a xenolog of the gene to its genome.

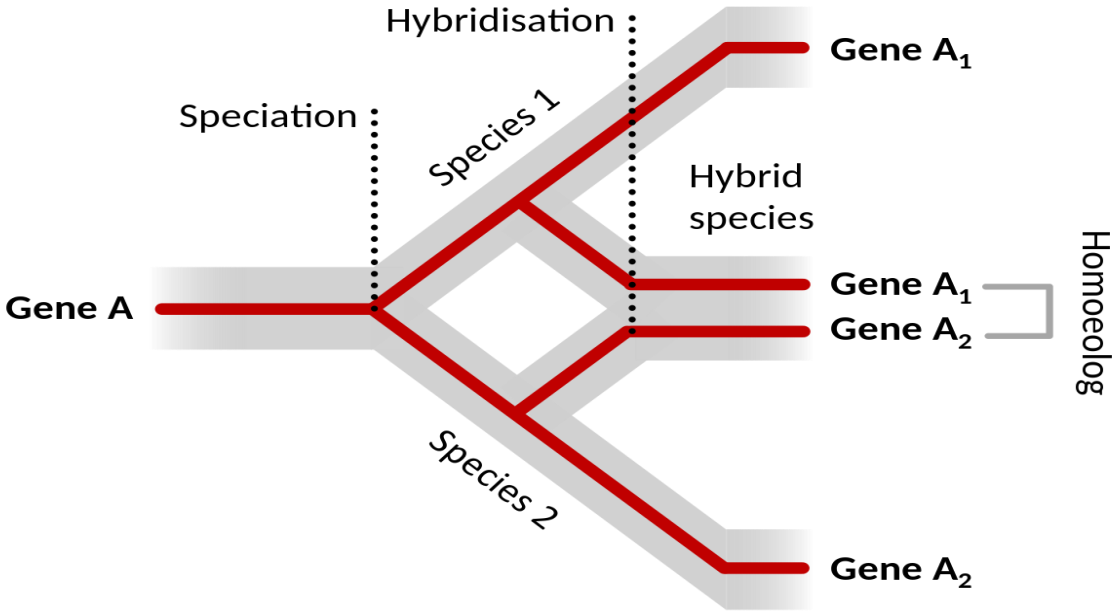


Figure: A speciation event produces orthologs of a gene in the two daughter species. Subsequent hybridisation of those species generates a hybrid genome with a copy of each gene from both species.

Ohnology: Ohnologous genes are paralogous genes that have originated by a process of whole-genome duplication. The name was first given in honour of Susumu Ohno by Ken Wolfe. Ohnologues are useful for evolutionary analysis because all ohnologues in a genome have been diverging for the same length of time (since their common origin in the whole genome duplication). Ohnologues are also known to show greater association with cancers, dominant genetic disorders, and pathogenic copy number variations

Xenology: Homologs resulting from horizontal gene transfer between two organisms are termed xenologs. Xenologs can have different functions if the new environment is vastly different for the horizontally moving gene. In general, though, xenologs typically have similar function in both organisms. The term was coined by Walter Fitch.

Homoeology:

Homoeologous (also spelled homeologous) chromosomes or parts of chromosomes are those brought together following inter-species hybridization and allopolyploidization to form a hybrid genome, and whose relationship was completely homologous in an ancestral species. In allopolyploids, the homologous chromosomes within each parental sub-genome should pair faithfully during meiosis, leading to disomic inheritance; however in some allopolyploids, the homoeologous chromosomes of the parental genomes may be nearly as similar to one another as the homologous chromosomes,

leading to tetrasomic inheritance (four chromosomes pairing at meiosis), intergenomic recombination, and reduced fertility.[

Gametology:

Gametology denotes the relationship between homologous genes on non-recombining, opposite sex chromosomes. The term was coined by García-Moreno and Mindell. 2000. Gametologs result from the origination of genetic sex determination and barriers to recombination between sex chromosomes. Examples of gametologs include CHDW and CHDZ in birds.[

Probable Questions:

1. what is paralogous gene? Give examples.
2. What is orthologous gene? Give example?
3. What is homologous gene? Give example?
4. Describe the basic steps of DNA Fingerprinting?
5. What is the significance of DNA Fingerprinting?
6. What is the significance of mt DNA analysis?
7. What are the significance of Y chromosome analysis?

Suggested Readings:

1. Molecular Cell Biology by Lodish, Fourth Edition.
2. The Cell – A Molecular Approach by Cooper and Hausman, Fourth Edition
3. Principles of Genetics by Snustad and Simmons, Sixth Edition.
4. Molecular Biology of the Cell – by Bruce Alberts
5. Cell and Molecular Biology by Gerald Karp, 7th Edition.
6. Gene cloning and DNA Analysis by T. A. Brown, Sixth Edition.
7. Genetics . Verma and Agarwal.
8. Brown TA. (2010) Gene Cloning and DNA Analysis. 6th edition. Blackwell Publishing, Oxford, U.K.
9. Primrose SB and Twyman RM. (2006) Principles of Gene Manipulation and Genomics, 7th edition. Blackwell Publishing, Oxford, U.K.
10. Sambrook J and Russell D. (2001) Molecular Cloning-A Laboratory Manual. 3rd edition. Cold Spring Harbor Laboratory Pres

UNIT-IX

Allele frequencies and genotype frequencies: Hardy-Weinberg relationship

Objective: In this unit we will discuss about Hardy-Weinberg principle and how allele and genotype frequency changes causes evolution.

Gene Pool: The gene pool is all the different alleles that are present in a population. For example; you have a population of bunnies of different colours. The colours are caused by different variations of the same gene. All the different variations of the gene make up the gene pool. The gene pool does not bother with frequencies; a variant is either present or not. You can compare the size of the gene pool between populations or in the course of time. When there are more gene variants present in a population, the gene pool is big. It can get smaller because of stochasticity (e.g. a bottleneck in population size accidentally losing all the individuals with a certain allele), selective breeding or natural selection. It can get bigger because of mutations or introduction of individuals of the same species from a different population.

Measurement of Gene pool : If you want to measure the gene pool, you need toknow how many variant of a gene are present in a given population. To do so you need to sample the DNA of a certain amount of individuals in a population. Then you can use DNA sequencing or gel electrophoresis to determine how many variants of the gene are present. You can really only look at one gene or a set of genes at one time. Usually scientists look at genes that are not selected for (neutral genes) or genes that are very important for survival chances of the population (like variation in immune defence genes). The size of the gene pool is a direct measure of the amount of genetic variation.

What happens if the gene pool gets smaller?

If the gene pool of a population gets smaller you are stating that the amount of genetic variation of a population goes down. In itself this is not bad and it is happening all the time, as natural selection filters out some genetic variation. The problem comes from a too small gene pool. Low genetic variation causes the population to be vulnerable to changes in the environment and stochasticity. When a population has a low variation, there are fewer genes to select on if the environment changes. E.g. if you have a population of only black rabbits they will be more vulnerable to predation when there is snow. A bigger gene pool, with black, white and grey rabbits, gives natural selection a hand hold to select on.

Population Genetics:

A population consisting of sexually interbreeding organisms carrying one or more particular genes, which follow the Mendelian Principles of Inheritance, is called 'Mendelian Population'. Gene pool and gene frequencies are considered to be two important attributes of a population.

A gene pool is the sum total of genes in reproductive gametes of a population. The nature of gene pool depends on random mating of gametes to form zygotes in the next generation.

Gene frequency can be defined as proportions of different alleles of a gene in a population, and in a particular generation these frequencies depend on their frequencies in the previous generation and also on the proportion of various genotypes in total population.

In any population, if a character is controlled by two alleles, then the frequency of these alleles or genes can be calculated very easily by phenotypic observation of that character under homozygous and heterozygous conditions. The frequency of an allele in a population is the number of occurrence of that allele divided by the total number of alleles of that gene locus.

Regarding the genetic structure of the population, the following two hypotheses have been proposed:- **1. Classical Hypothesis 2. Balance Hypothesis.**

a. Classical Hypothesis:

It was developed by T.H. Morgan (1932) and supported by H.J. Muller and Kaplan (1966). The classical hypothesis proposes that the gene pool of a population consists at each gene locus of a wild-type allele with a frequency approaching one.

Mutant alleles in very low frequencies may also exist at each locus. A typical individual would be homozygous for the wild-type allele at most gene loci; at a very small proportion of its loci, the individual would be heterozygous for a wild and a mutant allele.

Except in the progenies of consanguineous mating, individuals homozygous for a mutant allele would be extremely rare. The "normal" ideal genotype would be an individual homozygous for the wild-type allele.

According to classical hypothesis, mutant alleles are continuously introduced in the population by mutation pressure, but are generally deleterious and, thus, are more or less gradually removed from the population by natural selection. Periodically, a beneficial mutant allele might arise, conferring higher fitness upon its carriers than the pre-existing wild-type allele.

This beneficial allele would gradually increase in frequency by natural selection to become the new wild-type allele, while the former wild-type allele would be eliminated. Evolution, thus, consists of the replacement at an occasional locus of the pre-existing wild-type allele by a new wild-type allele.

b. Balance Hypothesis:

This hypothesis was proposed by Dobzhansky (1970) and E.B. Ford (1971). This hypothesis was derived by direct study of natural populations. According to the balance model, there is generally no single wild-type or 'normal' allele. Rather, the gene pool of a population is envisioned as consisting at most loci of an array of alleles in moderate frequencies.

A typical individual is heterozygous at a large proportion of its gene loci. There is no 'normal' or ideal genotype, only an adaptive norm consisting of an array of genotypes that yield a satisfactory fitness in most environments encountered by the population.

The proponents of the balance hypothesis argue that the common allelic polymorphisms are maintained in populations by various forms of balancing natural selection. The fitness granted on its carriers by an allele depends on what other alleles exist in the genotype at that and other gene loci. It also depends, of course, on the environment.

Gene pools are co-adapted systems; the sets of alleles favoured at one locus depend on the sets of alleles that exist at other loci. Evolution occurs by gradual change in the frequencies and kinds of alleles at many gene loci. As the configuration of the set of alleles changes at one locus, it also changes at many other loci.

The balance model of genetic structure of populations has now become definitely established, although some controversy remains regarding the process maintaining the common polymorphisms.

Hardy Weinberg's Law:

In 1908, the mathematician G. H. Hardy in England and the physician W. Weinberg in Germany independently developed a quantitative theory for defining the genetic structure of populations. The Hardy-Weinberg Law provides a basic algebraic formula for describing the expected frequencies of various genotypes in a population.

The similarity of their work however, remained unnoticed until Stern (1943) drew attention to both papers and recommended that names of both discoverers be attached to the population formula. The Law states that gene frequencies in a population remain constant from generation to generation if no evolutionary processes like migration, mutation, selection and drift are operating.Thus if matings are random, and no other factors disturb the reproductive abilities of any genotype, the equilibrium genotypic frequencies are given by the square of the allelic frequencies.

If there are only two alleles A and a with frequencies p and q respectively, the frequencies of the three possible genotypes are:

$(p + q)^2 = p^2 + 2pq + q^2$

If there are 3 alleles say A1, A2 and A3 with frequencies p, q and r, the genotypic frequencies would be; $(p + q + r)^2 = p^2 + q^2 + r^2 + 2pq + 2pr + 2qr$

This square expansion can be used to obtain the equilibrium genotypic frequencies for any number of alleles. It must also be noted that the sum of all the allelic frequencies,

and of all the genotypic frequencies must always be 1. If there are only two alleles p and q, then $p + q = 1$, and therefore $p^2 + 2pq + q^2 = (p + q)^2 = 1$. If there are 3 alleles with frequencies p, q, and r, then $p + q + r = 1$, as well as $(p + q + r)^2 = 1$.

The time required for attaining equilibrium frequencies has been determined. If a certain population of individuals with one set of allele frequencies mixes with another set and complete panmixis occurs (that is, random mating), then the genotypes of the next generation will be found in the proportion $p^2 + 2pq + q^2$ where p and q are allele frequencies in the new mixed populations. Thus it takes only one generation to reach Hardy-Weinberg equilibrium provided the allelic frequencies are the same in males and females. If the allelic frequencies are different in the two sexes, then they will become the same in one generation in the case of alleles on autosomes, and genotypic frequencies will reach equilibrium in two generations.

In general equilibrium is arrived at within one or at the most a few generations. Once equilibrium is attained it will be repeated in each subsequent generation with the same frequencies of alleles and of genotypes.

The Hardy-Weinberg law is applicable when there is random mating. Random mating occurs in a population when the probability of mating between individuals is independent of their genetic constitution. Such a population is said to be panmictic or to undergo panmixis. The matings between the genotypes occur according to the proportions in which the genotypes are present.

The probability of a given type of mating can be found out by multiplying the frequencies of the two genotypes that are involved in the mating. Mating are not random for instance when a population consists of different races such as blacks and whites in the U.S., or different communities as in India as there are preferred mating between members of the same racial or communal group.

Assumptions of Hardy-Weinberg Equilibrium:

We will consider a population of diploid, sexually reproducing organisms with a single autosomal locus segregating two alleles (i.e., every individual is one of three genotypes – MM, MN and AW).

The following major assumptions are necessary for the Hardy-Weinberg equilibrium to hold:

1. Random Mating,
2. Large Population Size,
3. No Mutation or Migration, and
4. No Natural Selection.

1. Random Mating:

The first assumption of Hardy-Weinberg equilibrium is random mating which means that the probability that two genotypes will mate is the product of the frequencies (or probabilities) of the genotypes in the population.

If an MM genotypes makes up 90% of a population, then any individual has a 90% chance (probability = 0.9) of mating with a person with an MM genotype. The probability of an MM by MM mating is (0.9) (0.9), or 0.81.
Any deviation from random mating comes about for two reasons: choice or circumstance. If members of a population choose individuals of a particular phenotype as mates more or less often than at random, the population is engaged in assortative mating.
If individuals with similar phenotypes are mating more often than at random, positive assortative mating is in force; if mating occur between individuals with dissimilar phenotypes more often than at random, negative assortative mating or disassortative mating is at work.
Further, deviation from random mating also arise when mating individuals are either more closely related genetically or more distantly related than individuals chosen at random from the population.

Inbreeding is the mating of related individuals, and outbreeding is the mating of genetically unrelated individuals. Inbreeding is a consequence of pedigree relatedness (e.g., cousins) and small population size.
One of the first distinct observations of population genetics is that deviation from random mating alter genotypic frequencies but not allelic frequencies. Imagine a population in which every individual is the parent of two children on the average, each individual will pass on one copy of each of his or her alleles.
Assortative mating and inbreeding will change the zygotic (genotypic) combinations from one generation to the next, but will not change which alleles are passed into the next generation. Thus, genotypic, but not allelic frequencies change under non-random mating.

2. Large Population Size:

Although an extremely large number of gametes are produced in each generation, each successive generation is the result of a sampling of a relatively small portion of the gametes of the previous generation. A sample may not be an accurate representation of a population, especially if the sample is small.
Thus, the second assumption of the Hardy-Weinberg equilibrium is that the population is infinitely large. A large population produces a large sample of successful gametes. The larger the sample of successful gametes, the greater the probability that the allelic frequencies of the offspring will accurately represent the allelic frequencies in the parental population.
When populations are small or when alleles are rare, changes in allelic frequencies take place due to chance alone. These changes are referred to as random genetic drift or just genetic drift.

3. No Mutation or Migration:

Allelic and genotypic frequencies may change through the loss or addition of alleles through mutation or migration (immigration or emigration) of individuals from or into a population, ne third and fourth assumptions of the Hardy-Weinberg equilibrium are that neither mutation nor migration causes such allelic loss or addition in the population.

4. No Natural Selection:

The final assumption necessary to the Hardy-Weinberg equilibrium is that no individual will have a reproductive advantage over another individual because of its genotype. In other words, no natural selection in occurring. (Note. Artificial selection, as practised by animal and plant breeders, will also perturb the Hardy-Weinberg equilibrium of captive population).

The significance of Hardy-Weinberg equilibrium was not immediately appreciated. A rebirth of biometrical genetics was later brought about with the classical papers of R.A. Fisher, beginning in 1918 and those of Sewall Wright, beginning in 1920.

Under the leadership of these mathematicians, emphasis was placed on the population rather than on the individual or family group, which had previously occupied the attention of most Mendelian geneticists. In about 1935, T. Dobzhansky and others started to interpret and to popularize the mathematical approach for studies of genetics and evolution.

Genetic Equilibrium:

As shown by Hardy and Weinberg, alleles segregating in a population tend to establish an equilibrium with reference to each other. Thus, if two alleles should occur in equal proportion in a large, isolated breeding population and neither had a selective or mutational advantage over the other, they would be expected to remain in equal proportion generation after generation. This would be a special case because alleles in natural populations seldom if ever, occur in equal frequency. They may, however, be expected to maintain their relative frequency, whatever it is, subject only to such factors as chance, natural selection, differential mutation rates or mutation pressure, meiotic drive and migration pressure, all of which alter the level of the allele frequencies. A genetic equilibrium is maintained through random mating.

Applications of the Hardy-Weinberg Law:

(a) Complete Dominance:

When Hardy-Weinberg equilibrium exists, allele frequencies can even be found out in presence of complete dominance where two genotypes cannot be distinguished. If two genotypes AA and Aa have the same phenotype due to complete dominance of A over a the allele frequencies can be determined from the frequencies of individuals showing the recessive phenotype aa.

The frequency of aa individuals must be equal to the square of the frequency of the recessive allele q. Let us suppose q = 0.5, then $q^2 – (0.5)^2 = 0.25$. In other words when aa phenotype is 0.25 in the population, then it follows that the frequency of the recessive allele a is $\sqrt{0.25} – 0.5$. The frequency of the dominant allele A would be 1 – q or 1 – 0.25 = 0.75.

(b) Frequencies of Harmful Recessive Alleles:

The Hardy-Weinberg Law can also be used to calculate the frequency of heterozygous carriers of harmful recessive genes. If there are two alleles A and a at an autosomal locus with frequencies p and q in the population and p + q = 1, then the frequency of AA, Aa, and aa genotypes would be $p^2 + 2pq + p^2$.

If the aa genotype expresses a harmful phenotype such as cystic fibrosis, then the proportion of affected individuals in the population would be q^2, and the frequency of the heterozygous carriers of the recessive allele would be 2pq.

To illustrate with figures, suppose one out of 1,000 children is affected with cystic fibrosis, then the frequency $q^2 = 0.001$, so that $q = \sqrt{0.001}$ which is about 0.032, then 2pq = 2 x 0.032 x 0.968 = 0.062. This means that about 62 individuals out of 1000 or one out of 16 is a carrier of the allele for cystic fibrosis. As already mentioned the number of individuals (aa) who are actually affected is one out of 1000. This implies that the frequency of heterozygous carriers is much higher than that of affected homozygotes. Similar calculation shows that when an allele is very rare in the population the proportion of carriers is still much higher and of affected homozygotes much lower. Thus, lower the frequency of an allele, greater the proportion of that allele that exists in the heterozygotes.

(c) Multiple Alleles:

The Hardy-Weinberg Law permits calculation of genotypic frequencies at loci with more than two alleles, such as the ABO blood groups. There are 3 alleles I^A, I^B and I° with frequencies p, q and r. Here p + q + r = 1. The genotypes of a population with random mating would be $(p + q + r)^2$.

(d) Sex-Linked Loci:

It is possible to apply Hardy-Weinberg Law for calculating gene frequencies in case of sex-linked loci in males and females. Red green colour blindness is a sex- linked recessive trait. Let r denote the recessive allele which produces affected individuals, and R the normal allele. The frequency of R is p and of r is q where p + q = 1. The frequencies of females having RR, Rr, rr genotypes would be p^2, 2pq, q^2 respectively.

Males are different as they are hemizygous, have only one X chromosome derived from the mother with a single allele either R or r. The frequency of affected r males would be the same as the frequency of the r allele among the eggs that is q. The frequency of normal R males would be p. Suppose the frequency of r alleles is 0.08, then the incidence of affected males would be 0.08 or about 8%. The frequency of affected rr females would

be $(0.08)^2 = 0.0064$ or 0.64%. Thus the Hardy-Weinberg Law explains that males would be affected a hundred times more frequently than females. This is actually what is observed. Males are more affected by sex-linked recessive traits than females. The difference between the sexes is even more pronounced if the recessive allele is still more rare. The incidence of a common form of haemophilia is one in a thousand males; thus q = 0.001. However, only one in 1000,000 females will be affected. Thus males could have haemophilia one thousand times more often than females.

(e) Linkage Disequilibrium:

Consider two or more alleles at one locus and another locus on the same chromosome with two or more alleles. Due to genetic exchange by recombination occurring regularly over a period of time, the frequencies of the allelic combinations at the two syntenic loci will reach equilibrium. If equilibrium is not reached, the alleles are said to be in linkage disequilibrium. The effect is due to tendency of two or more linked alleles to be inherited together more often than expected. Such groups of genes have also been referred to as supergenes.

Measurement of Gene Frequency:

In a diploid species, a population having N individuals has 2N alleles for each gene locus. If there are two alleles 'A' and 'a' of a particular gene in this population, the number of A alleles is twice the number of AA homozygotes plus the number of Aa heterozygotes, as each homozygote has two 'A' alleles, and each heterozygote has one 'A' allele. So the frequency of 'A' is the number of 'A' alleles divided by total number of alleles, i.e., 2N.

If the number is denoted by 'n' then the equation can be written as:

$$n_A = 2n_{AA} + n_{Aa}$$

$$n_a = 2n_{aa} + n_{Aa}$$

If the **frequency of allele A** is denoted by 'p' then,

$$p = \frac{n_A}{2N} = \frac{2n_{AA} + n_{Aa}}{2N}$$

Similarly, if the **frequency of allele a** is denoted by 'q' then,

$$q = \frac{n_a}{2N} = \frac{2n_{aa} + n_{Aa}}{2N}$$

It must be remembered that for all alleles the total frequency will be always 1, i.e., p + q = 1 and $n_A + n_a = 2N$.

Example 1:

In human population, a sample of 100 individuals for MN blood group character shows 50MM, 20MN and 30 NN individuals, then the frequency of M and N allele can be calculated using the above formula.

The frequency of 'M' will be $= \frac{2 \times 50 + 20}{200} = \frac{120}{200} = 0.6$

where $2n_{MM} = 2 \times 50$ $n_{MN} = 20.$ $2N = 200.$

The frequency of 'N' will be $= \frac{2 \times 30 + 20}{200} = \frac{80}{200} = 0.4$

where $2n_{NN} = 2 \times 30.$ $n_{MN} = 20.$ $2N = 200.$

The frequency can be calculated by another formula:

Frequency of a gene = frequency of homozygotes of that gene

$+ \frac{1}{2}$ frequency of heterozygotes

In the above example.

Frequency of M = 0.5 MM + $\frac{1}{2}$ (0.2 MN) = 0.6

Frequency of N = 0.3 NN + $\frac{1}{2}$ (0.2 MN) = 0.4.

Genotype Frequency and Hardy-Weinberg Equilibrium:

In a randomly mating Mendelian population the gene and genotype frequencies reach to equilibrium in a single generation. The Hardy-Weinberg law states that the gene and genotype frequencies in a Mendelian population remain constant generation after generation if there is no selection, mutation, migration or genetic drift.

Homozygotes (AA & aa) are produced by the union of gametes carrying similar alleles. The frequency with which a male gamete carrying allele A (frequency p) fuses with a female gamete carrying 'A' (also frequency p) will be p x p = p^2. Similarly the frequency with which a male gamete carrying allele 'a' (frequency q)"fuses with a female gamete also carrying 'a' will be q x q = q^2. Heterozygotes (Aa) are produced by the fusion of gametes carrying different alleles, the frequency with which a male gamete carrying allele 'A' fuses with a female gamete 'a' will be p x q = pq. Similarly in opposite way the frequency with which a female gamete carrying allele 'A' fuses with a male gamete carrying 'a' will be p x q = pq. So, the total frequency of heterozygotes will be 2 pq.

♂ \ ♀	A (p)	a (q)
A(p)	AA (p^2)	Aa (pq)
a (q)	Aa (pq)	aa (q^2)

Fig. 14.1: The frequencies of three genotypes produced due to random mating in a population having 'A' allele with frequency 'p' and allele 'a' with frequency 'q'

According to the theory of probability, due to random mating, the genotype frequency can easily be calculated from the following formula:

$(p + q) \times (p + q) = p^2 + 2pq + q^2$

(p + q) = Total frequency of two alleles 'A' and 'a'.

This equation is called Hardy-Weinberg equation which is named after British Mathematician G.H. Hardy and German Physician W. Weinberg.

As according to theory of probability the total frequency of p + q is always 1. So,

$(p + q)^2 = p^2 + 2pq + q^2 = 1$

In case of genes showing co-dominance, this equation can be used very easily to calculate the gene frequency by observing the phenotypic characters of homozygotes and heterozygotes.

Example 2:

In the following example it is shown how the observed value help to calculate the frequency:

MN blood group of Human

Phenotype	Genotype	No. of Individuals	Frequency Calculation
M	L_ML_M	1787	$L_M = \frac{2\times1787 + 3039}{2\times6129} = 0.5395$
MN	L_ML_N	3039	
N	L_NL_N	1303	$L_N = \frac{2\times1303 + 3039}{2\times6129} = 0.4605$
	Total	6129	

If we put these two values in a frequency table we can get the expected number of phenotypes due to random mating of each allele as follows:

	L_M p = 0.5395	L_N q = 0.4605
L_M p = 0.5395	$p^2 = 0.2911$	pq = 0.2484
L_N q = 0.4605	pq = 0.2484	$q^2 = 0.2121$

So, the expected values in total population can be calculated

$p^2 = 0.2911 \times 6129 = 1784.2$

$2pq = 2 \times 0.2484 \times 6129 = 3044.8$

$q^2 = 0.2121 \times 6129 = 1300$

The above calculation shows that the observed number of individuals and expected number of phenotypes are very close to each other which follows the Hardy-Weinberg Equilibrium Frequencies. In case of completely dominant genes of a population, the

heterozygotes cannot be phenotypically distinguished from homozygotes for the dominant character.

In this case, by observing the recessive phenotype, the recessive allele frequency can be measured very easily using Hardy-Weinberg equation. The frequency of homozygotes of recessive allele is q^2. So the frequency of recessive allele in a population is $\sqrt{q^2} = q$.

The frequency of dominant allele in that population is $(1 - q) = p$. This concept can also be applied to determine the number of individuals affected or carrying a particular disease, vis-a-vis, the frequency of the gene causing the disease.

PKU (Phenylketonuria) is a serious disease of human being which is a genetically controlled metabolic disorder. If in any population 0.04% is PKU affected then the frequency of PKU allele in that population is 7.0004 = 0.02 and the frequency of dominant allele is (1 – 0.02) = 0.98. If in a population the frequency of deleterious allele is 0.01, assuming random mating the frequency of homozygous population will be $(0.01)^2 = .0001$. Now the heterozygous frequency will be 2 x .01 x (1 – 0.01) = 2 x .01 x 0.99 = 0.0198.

So, about 1.98% population are carrier of this allele in heterozygous condition which is not expressed. If the frequency of recessive allele is too small then it is very difficult to measure it correctly, thus it is desirable to draw a large sample which will give more error free estimation.

1. Migration:

Migration occurs when a large influx of people moves into another population and interbreeds with the latter. The phenomenon called gene flow takes place if one population contributes an allele to the other population. Let us suppose that a migrating population m interbreeds with members of another population.

Then the descendants of the next generation will have m genes from the migrants and 1 – m genes from members of the original population. Consider an allele A occurring with frequency p in migrant population.

In the original population this allele has frequency q. In the next generation the frequency of A in the new population would be:

$$r = (1 - m)q + mp$$

$$= q - m (q - p)$$

That means the frequency of allele A in the new population now would be the original allelic frequency q multiplied by the genes (1 – m) present in the original population plus the product of reproducing migrant individuals and their gene frequency imp). Thus there will be a new gene frequency in the next generation.

Migration is a complex phenomenon in humans, influenced by many factors. It seems however, that it leads to make populations genetically more similar than they would be otherwise.

2. Mutation:

The ultimate source of all genetic variation is mutation. Both chromosomal rearrangements and point mutations are implied here as they follow the same rules of population dynamics. However, mutations occur with an extremely low frequency. In humans where there may be from 30-50 successive mitotic divisions in the germ cells in each generation, only one gene in a million or 10 million roughly undergoes a mutation.

Consider an allele A that is homozygous in many individuals in a population. Assume that in every generation one A allele in a million mutates to a. This will reduce the frequency of A allele over many generations, while a allele will gradually accumulate in the population. The change in frequencies of A and a occurs at an unimaginably slow rate.

The a allele however, can also back mutate to A; this event will take place as the frequency of a alleles increases. After a very long time the number of A alleles lost by forward mutation would be balanced by the number of A alleles arising from back mutation of a to A. When this happens gene frequencies of A and a are said to be in mutation equilibrium. Thereafter, no further change in

can be affected by environmental factors like radiation and chemicals. In some instances mutations are under genetic control. There is a recessive mutator gene on the second chromosome of Drosophila melanogaster. In stocks of flies homozygous for the mutator gene, sex-linked recessive lethals occur spontaneously with high frequency. In maize the recessive mutator gene frequency of A and a will occur in subsequent generations. This applies however, only when the other evolutionary forces such as migration, selection and genetic drift are not operating to affect gene frequencies in the population.

Experimental work has shown that the rate of mutation Dt acts on an unlinked locus a which controls synthesis of the purple anthocyanin pigment. The mutator gene Dt causes mutation of recessive a alleles to the dominant form A which leads to synthesis of anthocyanin appearing as purple spots on the stem, leaves and kernels of the maize plant. Mutator genes also occur in micro-organisms including E. coli. As the action of mutator genes is directed at a particular locus, they are said to produce directed mutations in contrast to random mutations which are not specific.

Normally all genes could mutate. In the case of a rare allele its mutation to other alleles is difficult to detect due to low frequency and slow rate of mutation. But when an allele occurs more frequently in the population it leads to higher mutation rate and increases the population's potential for evolutionary change. If a rare deleterious allele accumulates in the population, it is a disadvantage and constitutes what is referred to as the mutational load.

3. Selection:

Selection is one of the forces that change gene frequencies in the population and a fundamental process of evolutionary change. The idea was first conceived by Charles Darwin in his Origin of Species published in 1859 and by Alfred Russel Wallace.

Selection is defined as differential survival or fertility of different genotypes. If individuals carrying gene A are more successful in reproduction than individuals

carrying its allele a, then the frequency of gene A will tend to be greater than that of gene a.

The wide variety of mechanisms responsible for modifying the reproductive success of a genotype is collectively included under selection. It is the process that determines the contribution that people of different genotypes will make as parents of the next generation. Selection does not act on individual genes, but rather on the organism bearing the genes.

The reproductive efficiency of a genotype is measured in terms of the average number of offspring born to the bearers of the genotype and is called Darwinian fitness or relative fitness. It is also referred to as the organism's adaptive value. The fitness value of 1 is usually assigned to the genotype with highest reproductive efficiency.

However, fitness does not have an absolute value, and is expressed in relative terms as a ratio. Relative fitness (w) is obtained by dividing the fitness of all the genotypes by the fitness of any one genotype. Fitness simply describes the average number of progeny that survive and reproduce. The related term selection coefficient s = 1 – w. Some aspects in the individual's life are likely to affect the survival, growth and reproduction; consequently they affect the fitness of genotypes and are referred to as fitness components.

Basically fitness depends upon survival and fertility. Persons affected with Huntington's chorea, a dominant condition may have 25% reproductive efficiency as compared to normal human beings. On the other hand children with Tay Sachs disease usually die before reproductive age. Thus the fitness of a person with Huntington's chorea is 0.25 and of Tay Sachs patient is zero.

When fitness of two alleles at a locus differs then selection favours survival of alleles with greater fitness and elimination of the other alleles. Thus frequency of one allele increases and of the other will decrease in the subsequent generations. However, if a rare allele occurs with low frequency, then selection is not able to cause much change in gene frequency.

Specifically selection occurs against a recessive allele, or a dominant allele, resulting in its elimination; it could occur in favour of a heterozygote or against a heterozygote leading to polymorphism in a given trait. When selection occurs in favour of a heterozygote over both homozygotes it is called over dominance or heterosis. It occurs when the fitness of the heterozygous genotype is greater than the fitness of both homozygotes. Assume that the relative fitness of the genotypes AA, Aa and aa are 0.9,1 and 0.8 respectively.

The greater fitness of the Aa- genotype will not allow either A or a alleles from homozygotes to become fixed. Ultimately equilibrium gene frequencies would be attained. In humans over dominance has led to polymorphism in sickle cell trait, thalassaemia and G6PD. The effect of selection is also counterbalanced by mutation. While selection is eliminating some genes from the population, mutation is creating new ones. The two forces selection and mutation operate in opposite directions, and tend to compensate each other. After a long time, gene frequencies will reach equilibrium.

There could be partial selection against recessives. This is a less complete form of selection against homozygous recessive individuals. In this case selection coefficient s is less than one, and the relative fitness w of the homozygous recessive individual is 1 – s, having value greater than zero. A popular example of selection is industrial melanism as exhibited by the pepper moth **Biston betularia**. In the mid19th century the light coloured forms of the moth were abundant on the pale barks of trees growing in unpolluted, non-industrialised regions of England. The dark form of Biston was extremely rare. As industry developed in the area, the environment became polluted and the barks of trees turned dark grey with smoke and dust.

The light moths on the dark coloured bark were easily noticed by the predators and were preyed upon. Their number began to decrease. In the following decades, the population of dark moths was observed to gradually increase to more than 95%; the light moths were hardly seen. Industrial melanism is thus a clear cut example of selection disturbing gene frequencies in the population.

Artificial and Natural Selection:

Selection was being practiced by humans since antiquity. Plant and animal breeders have been attempting to modify hereditary transmission of traits by selecting most desirable individuals to serve as parents for the next generation. This is called artificial selection. By contrast, when organisms are selected by natural forces instead of by human choice, they are said to be subject to natural selection.

4. Random Genetic Drift:

These are unexpected random changes that occur in gene frequencies from generation to generation in all populations. They are particularly noticeable as sampling variation in small populations. In some generations the frequency of a certain allele will by chance increase, in others it will decrease, in still others it may remain the same. These fluctuations in gene frequency occur at random. In small samples there is greater variation as compared to big samples. Drift however, does not depend upon the total size of the population, rather on the number of breeding individuals who would produce the next generation. It is unlikely that random drift alone will affect allelic frequencies at a gene locus over long periods of time. It is more likely that selection, mutation or migration would also take place at one time or another.

Measurement of Genotype Frequency:

One way of measuring genotype frequency is from phenotype frequency. Consider the case of 3 blood groups A, AB and B determined by two alleles I^A and I^B at a single locus. In a random sample of 1000 humans, the A group occurred in 210, AB in 450 and B in 340 individuals.

The frequencies of the blood group phenotypes and their respective genotypes are obtained by dividing the number of individuals for each blood group by the total. Thus, the frequency of blood group B for instance would be 340/1000 = 0.34.

Another way of estimating genotype frequency is to first calculate gene frequency of genes A and B in the population. Assume that the above sample contains 210 AA, 450 AB and 340 BB individuals.

The gene frequency of A in the population is represented by the probability to find A allele at the AB locus and is exactly equivalent to the proportion of A alleles among all alleles at this locus in the sample or in the population (that is because we cannot determine the frequency of A in the whole population).

As each individual carries two alleles at the AB locus, the total number of alleles in the sample is 1000 x 2 = 2000. Out of these 210 + 210 + 450 = 870 are A. Therefore the frequency of the A allele is 870/2000 = 0.435. The number of B alleles is 450 + 340 + 340 = 1130, and the frequency of B allele is 1130/2000 = 0.565. If we represent the gene frequency of A by p, then p represents a value between 0 and 1 (because the proportion of allele A must lie between 0 and 100 per cent). In our example p = 0.435. Similarly, if we symbolize the frequency of B by q, then q = 0.565. It may be noted that q = 1-p or 1-0.435. Similarly p = 1 – q or 0.565. Thus p + q = 1.

For predicting genotype frequencies some assumptions have to be made, such as random mating in the population. That is to say, with respect to the trait of blood groups, an individual will mate with another without regard to whether the blood group of the mate is AA, AB or BB. Random mating implies random union of eggs and sperm, which in the example cited is the frequency of A and B alleles among the eggs and sperm (the gametes). Now the probability for the allele A at the AB locus in the population is p, and this is also the probability for a randomly chosen gamete to carry allele A. Similarly the probability for a randomly chosen gamete to carry B is q. For producing individuals with genotype AA, an A sperm must fertilize an A egg.

This occurs with probability p x p $=p^2$. An AB genotype results from fertilisation of A sperm with B egg or vice versa, and the probability is p x q or pq + qp = 2pq (that is pq for AB genotype and qp for BA genotype). The frequency of BB genotypes depends upon the chance fertilisation of a B sperm and B egg; this has probability q x q = q^2. Thus the frequencies of AA, AB and BB genotypes in the population should be expected to be p^2, 2pq, and q^2 respectively. If we substitute the values of p = 0.435 and q = 0.565 we can know that the frequency of AA = $(0.435)^2$ = 0.189, AB = 2 x 0.435 x 0.565 = 0.246, and of BB = $(0.565)^2$ = 0.319.

Significance of Population Genetics:

1. Knowledge of gene and genotype frequency in a population is useful for a plant breeder in the assessment of competitive ability of various genotypes in varietal mixtures. Such studies help in identification of genotypes with high adaptive value. If such studies are conducted over multiplications, the varietal flexibility or stability can also be assessed in varietal blends. Hardy-Weinberg Law operates in random mating or panmictic species.

2. Study of gene frequency in a population also reveals significance of various factors in natural evolution. In cross pollinated crops, development of composite and synthetic varieties is based on Hardy-Weinberg principle.

Probable questions:

1. What is gene pool? How it is measured?
2. What happens if the gene pool gets smaller?
3. Explain classical hypothesis regarding population structure.
4.. Explain balance hypothesis regarding population structure.
5. State Hardy Weinberg's Law.
6. Define gene frequency. How it is calculated ?
7. Define genotype frequency. How it is calculated ?
8. What are assumptions of Hardy Weinberg's law. Explain.
9. What are the applications of Hardy Weinberg's law ?
10. How mutation disrupt Hardy Weinberg's law ?
11. How natural selection disrupts Hardy Weinberg's law ?
12. How migration disrupts Hardy Weinberg's law ?
13. What are the significance of population genetics ?

Suggested Readings:

1. Molecular Cell Biology by Lodish, Fourth Edition.
2. The Cell – A Molecular Approach by Cooper and Hausman, Fourth Edition
3. Principles of Genetics by Snustad and Simmons, Sixth Edition.
4. Molecular Biology of the Cell – by Bruce Alberts
5. Cell and Molecular Biology by Gerald Karp, 7th Edition.
6. Gene cloning and DNA Analysis by T. A. Brown, Sixth Edition.
7. Genetics . Verma and Agarwal.
8. Brown TA. (2010) Gene Cloning and DNA Analysis. 6th edition. Blackwell Publishing, Oxford, U.K.

UNIT-X

Haplotype frequencies and linkage disequilibrium, changing allele frequencies

Objective: In this unit we will discuss about haplotype frequency and linkage disequilibrium and their role in changing allele frequencies.

Linkage disequilibrium — the nonrandom association of alleles at different loci — this is a sensitive indicator of the population genetic forces that structure a genome. Because of the explosive growth of methods for assessing genetic variation at a fine scale, evolutionary biologists and human geneticists are increasingly exploiting linkage disequilibrium in order to understand past evolutionary and demographic events, to map genes that are associated with quantitative characters and inherited diseases, and to understand the joint evolution of linked sets of genes. At present, linkage disequilibrium is used much more extensively in the study of humans than in non-humans, but that is changing as technological advances make extensive genomic studies feasible in other species.

Linkage disequilibrium (LD) is one of those unfortunate terms that does not reveal its meaning. As every instructor of population genetics knows, the term is a barrier not an aid to understanding. LD means simply a nonrandom association of alleles at two or more loci, and detecting LD does not ensure either linkage or a lack of equilibrium. The term was first used in 1960 by Lewontin and Kojima and it persists because LD was initially the concern of population geneticists who were not picky about terminology as long as the mathematical definition was clear. At first, there were few data with which to study LD, and its importance to evolutionary biology and human genetics was unrecognized outside of population genetics. However, interest in LD grew rapidly in the 1980s once the usefulness of LD for gene mapping became evident and large-scale surveys of closely linked loci became feasible. By then, the term was too well established to be replaced.

LD is of importance in evolutionary biology and human genetics because so many factors affect it and are affected by it. LD provides information about past events and it constrains the potential response to both natural and artificial selection. LD throughout the genome reflects the population history, the breeding system and the pattern of geographic subdivision, whereas LD in each genomic region reflects the history of natural selection, gene conversion, mutation and other forces that cause gene-frequency evolution. How these factors affect LD between a particular pair of loci or in a genomic

region depends on local recombination rates. The population genetics theory of LD is well developed and is being widely used to provide insight into evolutionary history and as the basis for mapping genes in humans and in other species.

Definitions of LD:

Different definitions of linkage disequilibrium (LD) have been proposed because they capture different features of nonrandom association. All of them are related to D, which is defined in equation 1 in the text. Although D completely characterizes the extent to which two alleles, A and B, are nonrandomly associated, it is often not the best statistic to use when comparing LD at different pairs of loci because the range of possible values of D for each pair is constrained by the allele frequencies. The smallest possible value, $D_{\min}$, is the less negative value of $-p_A p_B$ and $-(1 - p_A)(1 - p_B)$, where p is the frequency of the allele. The largest possible value, $D_{\max}$, is the smaller of $p_A(1 - p_B)$ and $p_B(1 - p_A)$. Lewontin[1] defined D' to be the ratio of D to its maximum possible absolute value, given the allele frequencies. This definition has the convenient property that when $D' = 1$ it indicates that at least one of the four possible haplotypes is absent, regardless of the allele frequencies (BOX 2), a situation commonly described as a `perfect' disequilibrium.

Another commonly used way to quantify LD is with r^2:

$$r2=D2pA(1-pA)pB(1-pB)$$

which is a correlation coefficient of 1/0 (all or none) indicator variables indicating the presence of A and B. In general, r^2 is similar to D' in that it can be nearly one even if one or both alleles are in low frequency.

Still another measure is δ_A, defined to be $p_A + D/p_B$. It is the conditional probability that a chromosome carries an A allele, given that it carries a B allele. It is useful for characterizing the extent to which a particular allele is associated with a genetic disease.

One pair of loci:

LD between alleles at two loci has been defined in many ways (BOX 1), but all definitions depend on the quantity:

$$D_{AB} = p_{AB} - p_A p_B \qquad 1$$

which is the difference between the frequency of gametes carrying the pair of alleles A and B at two loci (p_{AB}) and the product of the frequencies of those alleles (p_A and p_B). Originally, the definition was in terms of gamete frequencies because that allows for the possibility that the loci are on different chromosomes. The usual application now is to loci on the same chromosome, in which case the allele pair AB is called a haplotype and p_{AB} is the haplotype frequency. As defined, D_{AB} characterizes a population; in

practice, D_{AB} is estimated from allele and haplotype frequencies in a sample. Standard sampling theory has to be applied to find the confidence intervals of estimated values·

The quantity D_{AB} is the coefficient of linkage disequilibrium. It is defined for a specific pair of alleles, A and B, and does not depend on how many other alleles are at the two loci — each pair of alleles has its own D. The values for different pairs of alleles are constrained by the fact that the allele frequencies at both loci and the haplotype frequencies have to add up to 1. If both loci are diallelic, as is the case with virtually all SNPs, the constraint is strong enough that only one value of D is needed to characterize LD between those loci. In fact, $D_{AB} = -D_{Ab} = -D_{aB} = D_{ab}$, where a and b are the other alleles. In this case, the D is used without a subscript. The sign of D is arbitrary and depends on which pair of alleles one starts with.

If either locus has more than two alleles, no single statistic quantifies the overall LD between them. Although several have been suggested, none has gained wide acceptance. Such a statistic is needed when both loci have numerous alleles, as is the case for many loci in the major histocompatibility complex in vertebrates, which have dozens or even hundreds of alleles, or for microsatellite loci, which often have 10 to 20 alleles. If there is no one pair of alleles of particular interest, the question is often whether there is more LD between one pair of loci than another pair, or more LD between a pair of loci in one species than in another.

Linkage equilibrium:

If $D = 0$ there is linkage equilibrium (LE), which has similarities to the Hardy–Weinberg equilibrium (HWE) in implying statistical independence. When genotypes at a single locus are at HWE, whether an allele is present on one chromosome is independent of whether it is present on the homologue. Consequently, the frequency of the AA homozygote is the square of the frequency of A (pAA=p2A) and the frequency of the Aa heterozygote is twice the product of p_A and p_a, the two being necessary to allow for both Aa and aA. The essential feature of HWE is that, regardless of the initial genotype frequencies, HWE is established in one generation of random mating. Any initial deviation from HWE disappears immediately. Significant departures from HWE indicate something interesting is going on, for example, extensive inbreeding, strong selection or genotyping error.

LE is similar to HWE because it implies that alleles at different loci are randomly associated. The frequency of the AB haplotype is the product of the allele frequencies ($p_A p_B$). LE differs from HWE, however, because it is not established in one generation of random mating. Instead, D decreases at a rate that depends on the recombination frequency, c, between the two loci:

$$D_{AB}\,(t+1) = (1-c)\,D_{AB}\,(t) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 2$$

where t is time in generations. Even for unlinked loci ($c = 0.5$), D decreases only by a factor of a half each generation, something proved by Weinberg in 1909. The general formula was obtained first by Jennings.

Although LE will eventually be reached, it will occur slowly for closely linked loci. That is the basis for the uses of LD discussed in later sections. Other population genetic forces, including selection, gene flow, genetic drift and mutation, all affect D, so substantial LD will persist under many conditions. Now that very large numbers of polymorphic loci can be surveyed, the extent of LD in a genome can be quantified with great precision, allowing a fine-scale analysis of forces governing genomic variation.

The coefficient of LD and related quantities are descriptive statistics. Their magnitude does not indicate whether or not there is a statistically significant association between alleles in haplotypes. Standard statistical tests, including the chi squared and Fisher's exact test, are commonly used to test for significance.

Haplotype phase:

D and related statistics implicitly assume that haploid individuals or gametes can be typed. But often, only diploid genotypes and not haplotypes can be determined. That is the case with all SNP surveys, other than those of the X chromosome in males (assuming males are the heterogametic sex) or when haploids can be typed. The problem is sketched in BOX 2. The extent of LD in genotypic data can be quantified, but the lack of information about the haplotype phase weakens the signal of nonrandom association sufficiently that this approach is not often taken. It is more common to use a statistical method based on population genetics theory (BOX 2) to infer haplotype phase from genotypic data and then to treat the inferred haplotypes as if they were data. Although this procedure is intuitively appealing and usually leads to reasonable results, especially for common haplotypes, it ignores the uncertainty that is inherent in the inference step and that might be important in some cases. Often, genotypes can be resolved into several possible haplotypes, and inferred frequencies of rare haplotypes can be quite wrong. It is preferable, although sometimes difficult, to use methods that account for the uncertainly in the inferred haplotype frequencies, as is done in likelihood analysis with Metropolis algorithm using random coalescence (LAMARC) and some other computer program packages.

LD at more than two loci:

When more than two loci are considered together, a common practice is to distinguish graphically those pairs that have high levels of LD from those that do not. The result is a graph of the type introduced by Miyashita and Langley to describe patterns of LD in *Drosophila melanogaster*. This figure indicates that a 216 kb segment in the class II region of the major histocompatibility complex in humans is made up of non-overlapping sets of loci in strong LD with each other. Each group is called a `haplotype block' and boundaries were shown to be associated with hot spots of recombination. Similar patterns were found in other genomic regions in humans, leading to the hypothesis that most of the human genome had a block-like pattern of LD. Haplotype blocks in humans vary in size from a few kb to more than 100 kb

Haplotype blocks were a surprising discovery that was of great practical importance for the mapping of inherited diseases. Before their discovery, the prevailing view of LD in humans was represented by results from the simulation study of Kruglyak, which showed that, under assumptions that were intended to approximate the history of modern humans, little LD would be expected beyond 3 kb. The discovery of haplotype blocks showed that LD usually extended over much longer chromosomal distances and suggested that testing one SNP within each block for significant association with a disease might be sufficient to indicate association with every SNP in that block, thus reducing the number of SNPs that need to be tested in case–control studies of disease association. The situation turned out to be more complex both because some genomic regions were found to not have a block-like structure and because different ways of defining haplotype blocks resulted in different block boundaries. Nevertheless, the observation that LD in humans extended over relatively large chromosomal distances provided a major part of the impetus for the International HapMap Project, which in its first generation identified over 1 million SNPs in humans and characterized the LD in 269 individuals in four ethnically different populations (European, Han Chinese, Japanese and Yoruban). The second generation HapMap published recently characterized 3.1 million SNPs in the same group of individuals.

Haplotype blocks vary somewhat among human populations — they tend to be somewhat shorter in African populations. Haplotype blocks have been studied in other species as well, both model organisms, including the mouse and rat, and domesticated species, including cows and dogs. The isolation of strains and breeds in these species results in much longer block lengths than are found in humans.

Variance in heterozygosity:

A simple and often useful statistic describing the overall extent of LD in a genomic region is the variance in heterozygosity across loci, which increases as a linear function of D^{-2} . This statistic is useful when the density of polymorphic loci is low and the goal is to obtain a general idea of the importance of recombination. Maynard Smith *et al.* used this statistic to assess the overall degree of clonality of various pathogenic bacteria.

Higher-order disequilibria:

When considering more than two loci, equation 1 can be generalized to define higher-order coefficients of LD. For alleles at three loci (A, B, and C) the third-order coefficient is:

$$D_{ABC} = p_{ABC} - p_A D_{BC} - p_B D_{AC} - p_C D_{AB} - p_A p_B p_C \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3$$

where D_{AB}, D_{BC} and D_{AC} are the pairwise disequilibrium coefficients. D_{ABC} is analogous to the three-way interaction term in an analysis of variance and can be interpreted as the non-independence among these alleles that is not accounted for by the pairwise coefficients. The decay of these higher-order coefficients under random mating was studied by Geiringer and has been worked out in some detail by later authors. Little practical use of these higher-order coefficients has been made, other than in the analysis of variation of human leukocyte antigen loci in humans, which suggested that two loci that are closely linked to a selected locus would display unusual patterns of LD. It is worth considering whether higher-order disequilibrium coefficients can help to understand the patterns found in the HapMap and other large data sets.

LD within and between populations:

When data for more than one population are available, LD between a pair of loci can be partitioned into contributions within and between populations. This partitioning, first suggested by Ohta, is similar to Wright's partitioning of deviations from HWE frequencies into F_{IS}, the average deviation within populations, and F_{ST}, the average deviation that is attributable to differences in allele frequency among populations. Ohta partitioned D_T, the total disequilibrium in a subdivided population, into D_{IS}, the average disequilibrium within subpopulations, and D_{ST}, the contribution to the overall disequilibrium caused by differences in allele frequencies among subpopulations. Computer programs such as Genepop are available to calculate D_{IS} and D_{ST}.

These statistics are used widely in the analysis of data from non-human populations but only rarely for human populations, probably because the focus in humans is on each population whereas the focus in other species is often on the overall pattern of LD. Natural selection favouring adaptations to local conditions will increase D_{ST} whenever alleles at different loci are favoured. Partitioning overall LD is an appropriate first step when trying to determine whether differences in LD result only from differences in allele frequency or from other factors that vary among populations.

Population genetics of LD:

a. Natural selection:

Initial interest in LD arose from questions about the operation of natural selection. If alleles at two loci are in LD and they both affect reproductive fitness, the response to selection on one locus might be accelerated or impeded by selection affecting the other.

One line of research in this area concerns the effect of LD on long-term trends in evolution. Kimura, Nagylaki and others showed that unless interacting loci are very closely linked or selection is very strong, recombination dominates and, to a good approximation, LD can be ignored. This theory supports Fisher's depiction of natural selection steadily increasing the average fitness of a population. This theory also shows that when selection is strong and fitness interactions among loci are complex, average fitness might not increase every generation because LD constrains the way in which haplotype frequencies respond to selection. In that case, linkage must be accounted for explicitly before even qualitative predictions can be made.

In some cases, selection alone can increase LD. This occurs when fitnesses are more than multiplicative, meaning that the average fitness of an individual carrying the AB haplotype exceeds the product of the average fitnesses of individuals carrying A alone or B alone. The pattern is easiest to see with diallelic loci in haploid organisms. If the relative fitness (w) of the ab, Ab, and aB haplotypes are 1, w_{Ab} and w_{aB}, then selection will increase LD if $w_{AB} > w_{Ab}w_{aB}$.

If both A and B are maintained by balancing selection, then LD can persist indefinitely. Furthermore, when more than two loci interact in this way, large blocks of LD can be maintained by selection, leading to the suggestion that an individual locus is not the appropriate unit of selection. Interest in this kind of theory diminished in the 1970s when it was discovered that LD could not be detected between alleles that are distinguishable by protein electrophoresis. This theory will become popular again or perhaps be reinvented as studies find increasing evidence of intragenic interactions that can create strong epistasis in fitness .

b. Genetic drift:

Genetic drift alone can create LD between closely linked loci — the effect is similar to taking a small sample from a large population. Even if two loci are in LE, sampling only a few individuals will create some LD. Results first obtained in the late 1960s suggested that genetic drift balanced by mutation and recombination would maintain only low levels of LD, and the expectation of D^2 is small even if there is no recombination because the flux of mutations at both loci tends to eliminate most LD. For that reason, drift was largely ignored as a cause of LD. However, the expectation of D^2 does not tell the whole story because it includes cases in which one or both loci are monomorphic (when D is necessarily 0). The expectation of D^2 when both loci are polymorphic cannot be calculated analytically, but simulations show that much larger values are seen·

Genetic drift interacts with selection in a surprising way. Selection affecting closely linked loci becomes slightly weakened because drift creates small amounts of LD that, on average, reduces the response to selection. This effect, called the Hill–Robertson effect, is relatively weak when only two loci are considered but is much stronger per locus when many selected loci are closely linked .

Felsenstein showed that the Hill–Robertson effect might have a crucial role in the evolution of recombination and sexual reproduction. The basic idea is that the Hill–Robertson effect causes selection to be inefficient in purging deleterious mutations in a species with a low recombination rate. Hence, natural selection will favour any mutation that increases recombination rates. This early result has been confirmed and extended by many others. As interactions among intragenic SNPs become better understood, the Hill–Robertson effect will have to be taken into account when considering the evolution of gene function, especially in the first few generations of a new selective regime.

c. Population subdivision and population bottlenecks:

Natural selection affects only one or a small number of loci. By contrast, population subdivision, changes in population size and the exchange of individuals among populations all affect LD throughout the genome. Consequently, genome-wide patterns of LD can help us understand the history of changes in population size and the patterns of gene exchange.

The intentional or unintentional mixing of individuals from subpopulations that have different allele frequencies creates LD. The effect is obvious in an extreme case. Suppose that one subpopulation is fixed for A and B whereas another is fixed for a and b. Any mixture of individuals from the two subpopulations would contain only the AB and ab haplotypes, implying that there is perfect LD ($D' = 1$; D' is the ratio of D to its maximum

possible absolute value, given the allele frequencies), when in fact there is no LD in either subpopulation. This effect is similar to the Wahlund effect — the inbreeding coefficient at a locus when subpopulations with different allele frequencies are mixed. The reason is the same: the inbreeding coefficient measures the covariance between alleles at a locus just as *D* measures the covariance between alleles at different loci. Differences in allele frequencies among subpopulations create additional covariance in both cases.

The movement of individuals or gametes among subpopulations causes gene flow, which increases LD in each subpopulation whenever allele frequencies differ among subpopulations. The decay of LD under recombination alone can be greatly retarded. If selection maintains differences in allele frequencies at two or more loci among subpopulations, LD in each subpopulation will persist.

Changes in population size, particularly an extreme reduction in size (a population bottleneck), can increase LD. Colonizing species undergo repeated bottlenecks in size, and many models of the history of hominids assume a bottleneck occurred when modern humans first left Africa. After a bottleneck, some haplotypes will be lost, generally resulting in increased LD. A subsequent period of small population size will augment LD by increasing the effect of genetic drift. Several studies of humans have argued that long-distance LD in humans is the results of a bottleneck early in human history. Detecting higher levels of genome-wide LD in one population than in another can then indicate a past bottleneck.

d. Inbreeding, inversions and gene conversion:

Other forces create LD as well. Inbreeding creates LD for the same reason as population subdivision. Because of recent common ancestry, inbreeding augments the covariance between alleles at different loci. Theory predicts that this effect is largest in selfing species, but the expected pattern is not evident in the most thoroughly studied selfing species, *Arabidopsis thaliana*.

Genomic inversions greatly reduce recombination between the inverted and non-inverted segments because recombination produces aneuploid gametes. Consequently, the inverted and original segments become equivalent to almost completely isolated subpopulations between which LD accumulates. This fact has long been appreciated by *Drosophila* geneticists.

Gene conversion affects LD at a pair of loci in the same way that reciprocal recombination does. The equivalence can be seen by considering a pair of diallelic loci A/a and B/b. Gene conversion at the B/b locus will result in an individual with

haplotype phase AB/ab who will produce Ab or aB gametes depending on whether B converts b or the reverse. However, gene conversion differs from recombination when more than two loci are considered together. Reciprocal crossing over affects LD between all pairs of loci on opposite sides of where the crossing over took place. By contrast, gene conversion affects loci only within the conversion track, which is generally quite short. Loci that are not within the track are unaffected. For example, if three loci, A/a, B/b and C/c, are on a chromosome in that order and only B/b is within a conversion track, LD between A/a and B/b and between B/b and C/c is affected by conversion but the LD between A/a and C/c is not. Several methods for inferring the relative rates of gene conversion and recombination have been based on this idea

Applications of LD:

a. Mutation and gene mapping:

Mutation has a unique role in creating LD. When a mutant allele, M, first appears on a chromosome, it is in low frequency, $p_M = 1/(2N)$ (N is the population size) and is in perfect LD with the alleles at other loci that are on the chromosome carrying the first copy of M; perfect LD means that $D' = 1$ (BOX 1). If $D' = 1$, only three of the four possible haplotypes are present in the population (BOX 3). Perfect LD will persist until recombination involving an M-bearing chromosome creates a non-ancestral haplotype. Consequently, loci that are closely linked to M will remain in perfect LD for a long time and in strong LD for even longer.

The persistence of strong LD between a mutant allele and the loci closely linked to it has many practical implications. Rare marker alleles in strong LD with a monogenic disease locus have to be closely linked to the causative locus. Relatively simple mathematical theory indicates just how close. The resulting method, called LD mapping, has been successfully used with several diseases (BOX 4).

The same idea underlies association mapping of complex diseases. Closely linked polymorphic SNPs tend to be in strong LD with one another. The fine-scale pattern of LD in humans confirms that the human genome is comprised of haplotype blocks within which most or all SNPs are in high LD. These high levels of LD among SNPs are assumed to be true for alleles that increase the risk of complex inherited diseases. This idea, combined with the development of efficient methods for surveying large numbers of SNPs, has led to the many recent genome-wide association (GWA) studies that have detected SNPs that are significantly associated with breast cancer, colorectal cancer, type 2 diabetes and heart disease, among other diseases. However, one potential problem in GWA studies is that, as mentioned above, LD can be created by unrecognized population subdivision. Several methods have been proposed to account for such LD.

Although GWA studies have been successful in finding new causative alleles, the overall proportion of risk that is accounted for is often low. For example, Easton *et al*. found five new variants associated with familial breast cancer, but only 3.6% of familial breast cancer is accounted for by those alleles. Seventy percent of the genetic basis of familial breast cancer remains unaccounted for. Alleles accounting for a greater proportion of risk might be found in even larger studies, but it is unclear whether most causative variants will ultimately be found this way. The reason is that GWA studies are more effective in finding causative alleles that are in relative high frequency. Other methods might be needed for low-frequency causative alleles.

b. Detecting natural selection:

Strong positive selection quickly increases the frequency of an advantageous allele, with the result that linked loci remain in unusually strong LD with that allele. This idea originated with Maynard Smith and Haigh, who called it genetic hitch-hiking. Their paper focused on an advantageous allele that goes to fixation and causes a substantial reduction of heterozygosity at closely linked neutral loci. Recently, methods have been developed for detecting regions of unusually low heterozygosity that are indications of past hitch-hiking events.

If an advantageous allele has not gone to fixation, variability at linked markers will be lower on chromosomes bearing that allele than on other chromosomes. Several tests of neutrality have been based on this idea. One class of methods assumes that a potentially advantageous allele at a locus has been identified and tests whether there is significantly more LD with that allele than with other alleles at the same locus. A second class of methods assumes only that the potentially selected locus has been identified and tests whether patterns of haplotype variation at that locus are consistent with neutrality. Recently, Sabeti *et al*. and Voight *et al*. developed computationally efficient methods for detecting evidence of selection in whole genomes and have applied those methods to the HapMap populations. These studies found several regions in the human genome that were previously not suspected to harbour selected variants.

c. Estimating allele age:

Strong LD with an allele in a relatively large region indicates that not much time has passed since the allele arose by mutation. If the mutant allele has reached a relatively high frequency in a short time, it is likely to have done so under the effect of positive selection. This tendency provides the basis for the various tests of selection mentioned in the previous section. In addition to testing for selection, LD can indicate the point in time that the allele arose by mutation, that is, the allele age. The idea is based on equation 2 above. From an observed level of LD, allele age is estimate by solving

for t . This approach is straightforward and leads to reasonable estimates of allele age, but it does not take account of the stochastic nature of recombination and genetic drift, and hence exaggerates the accuracy of the resulting estimates. Various statistical methods have been developed that provide more realistic confidence intervals of estimated ages.

Genotype data and haplotype phase:

When the genotype of a dipoid individual is determined, the result is a list of genotypes for each locus surveyed. If three diallelic loci are surveyed, the genotypes of four individuals might be AA bb CC, Aa BB cc, aa Bb Cc and Aa Bb Cc. The haplotypes of the first two individuals are immediately apparent. Individual 1 has two copies of AbC and individual 2 has ABc and aBc. There is no uncertainty if no more than one locus is heterozygous. Otherwise, haplotypes cannot be determined without further information. Individual 3 could have haplotypes aBC/abc or aBc/abC. The number of possible resolutions increases exponentially with the number of heterozygous loci. Individual 4 could have haplotypes ABC/abc, ABc/abC, aBc/AbC or aBC/Abc.

There are several ways to determine haplotyes from genotypes; this is commonly referred to as resolving haplotype phase. If the parental genotypes are known, the haplotype phase of the offspring can usually, but not always, be determined. If the parents of individual 3 have genotypes Aa BB Cc and Aa Bb cc, then the individual's haplotype phase has to be aBC/abc. However, if instead the parents' genotypes are Aa Bb Cc and Aa Bb cc, then the haplotype phase still cannot be resolved.

Another way to resolve haplotype phase is to use a biochemical method that separately amplifies each chromosome, allowing direct determination of haplotype phase. Although such methods exist, they are currently too slow and costly to be used in large genomic surveys.
It is much more common to use a statistical method based on the assumption that haplotypes are randomly joined into genotypes. The basic idea is that individuals that are homozygous at all loci or all but one locus provide some information about haplotype frequencies that can then be used to infer the haplotype phase of the other individuals. Various methods — including those based on maximum likelihood, parsimony, combinatorial theory and *a priori* distribution derived from coalescent theory — have been developed. The last method is the basis for the program PHASE, which has performed the best in extensive simulation studies. The emerging view of this problem is that inferring haplotype phase is similar to other cases in which missing data (in this case the haplotype phase of a diploid genotype) has to be imputed.

The future of LD studies:

In human population genetics, the future of LD is now. Very large-scale GWA studies have already been carried out and many more are in progress. The technological problem of efficiently genotyping 500,000 or more SNPs has been solved, and costs of genotyping will continue to decline. And soon new technologies will allow large resequencing studies, including the 1000 Genomes Project, to take place. The limiting factor will be the availability of people who are willing to participate in GWA studies and the resources needed for accurate clinical assessment.

The methods currently used for association mapping will be used even more extensively in the future to study the history of human populations. At present, most analysis is done on the four HapMap populations, but large-scale surveys of SNPs and resequencing studies will be complete for a much broader range of populations. Advances in understanding human history will be increasingly limited by people's willingness to participate in genetic studies, something that is influenced by political and ethical concerns in addition to scientific ones.

With the increased resolution of LD patterns, the study of human history will shift in focus from understanding the average history of populations to understanding the history of different genomic regions. Unusually large variation in LD within the human genome already suggests that ancient human populations were subdivided and that some genomic regions of modern humans were brought by introgression from an extinct ancestor, possibly Neanderthals.

In model organisms, large SNP and resequencing studies on the scale of human studies are now being done. Other species will have to wait a technological generation or two before such large-scale surveys will be possible, in part because of the lower levels of funding available for studying non-model species. The range of possible selective regimes and population histories is vastly greater for non-humans than for humans, and new theoretical methods will no doubt be needed. Extensive studies of variation and examination of LD patterns will probably reveal levels of complexity not seen in humans. In some groups, widespread trans-species polymorphism and evidence of inter-specific gene transfer will be so apparent that some higher organisms might come to resemble bacteria in their genetic promiscuity.

At present, the emphasis is on LD between SNPs, which are diallelic and which mutate at such low rates that current patterns of LD are nearly unaffected by recent mutation. Less attention has been paid to studying LD between other kinds of genetic variants, including microsatellites, insertions, deletions and inversions. The relatively high rate of mutation in microsatellites makes possible the assessment of LD that was created

recently. The potential selective effect of indels and inversions, combined with more efficient means of their detection, will provide additional rich sources of LD in humans and other species

Probable Questions:

1. What is linkage disequilibrium?
2. What is linkage equilibrium?
3. Discuss haplotype phase.
4. Discuss population bottleneck with suitable examples.
5. Discuss natural selection with suitable examples.
6. Discuss genetic drift with suitable examples.
7. Discuss applications of linkage disequilibrium?

Suggested Readings:

1. Molecular Cell Biology by Lodish, Fourth Edition.
2. The Cell – A Molecular Approach by Cooper and Hausman, Fourth Edition
3. Principles of Genetics by Snustad and Simmons, Sixth Edition.
4. Molecular Biology of the Cell – by Bruce Alberts
5. Cell and Molecular Biology by Gerald Karp, 7th Edition.
6. Gene cloning and DNA Analysis by T. A. Brown, Sixth Edition.
7. Genetics . Verma and Agarwal.

UNIT-XI
Population structure and inbreeding

Objective: In this unit we will discuss about different aspects of inbreeding and heterosis. We will discuss about inbreeding depression, its effect and theories of heterosis. Comparison between hybrid vigour and heterosis will also be discussed in this unit.

Inbreeding:

The process of mating of individuals which are more closely related than the average of the population to which they belong, is called inbreeding. For example, parthenogenesis in animals and apomixes and self-fertilization in plants are the most extreme types of inbreeding.

Inbreeding in self-fertilizing pea plants was a real advantage to Mendel in his studies which provided pure lines of pea plants for his hybridization experiments. The term 'pure line' was coined by W. Johannsen in 1903 for the true breeding, self-fertilized plants.

Methods of Inbreeding:

In plants ova fertilized by the pollen of either the same plants (in case of bisexual plants) or of the other plant of the same genotype (in case of unisexual as well as bisexual plants), is called self-fertilization. However, in bisexual plants numerous structural and functional adaptations have been recorded which help plants with bisexual or hermaphrodite flowers avoid self-fertilization. Normally, inbreeding is affected by restrictions in population size or area which brings about the mating between relatives. Since close relatives have similar genes because of common heritage, inbreeding increases the frequency of homozygotes, but does not bring about a change in overall gene frequencies.

Thus, a mating between two heterozygotes as regards two alleles, A and a will result in half of the population, homozygous for either gene A or a and half of the population heterozygous like the parent but the overall frequencies of A and a remain unchanged:

$$Aa \times Aa$$

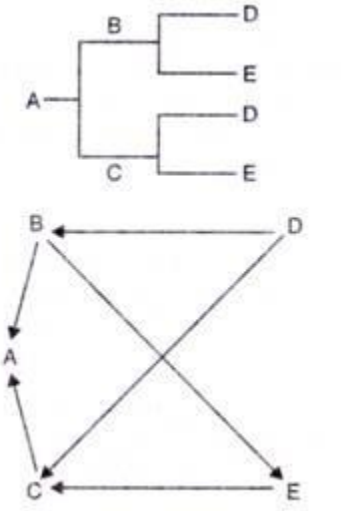
$$1AA : 1Aa$$

Thus, inbreeding brings about the recessive gene to appear in a homozygous stale (aa). Once a recessive allele is in a homozygous state, natural selection can operate upon the rare recessives. Artificial selection is also possible as the homozygous recessives are phenotypically differentiated from the dominant population.

The inbred pedigrees can be depicted as follows:

Here, B and C are full sibs, i.e., have common parents.

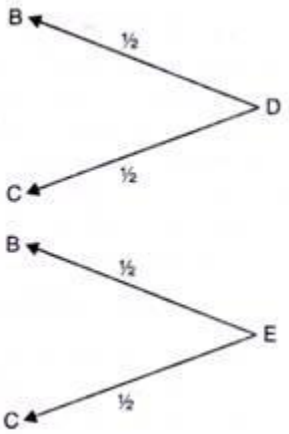
This pedigree can also be represented by the following arrow diagram:



1. Coefficient of Relationship (R):

Coefficient is expression of the amount or degree of any quality possessed by a substance. It is also the degree of physical or chemical change normally occurring in that substance under stated conditions. The coefficient of relationship (R) characterises the percentage of genes held in common by two individuals due to their common ancestry.

Each individual gets only a half of his genotype from one of his parent, each arrow in the above arrow diagram represents a probability of half. The sum (Σ) of all pathways between two individuals through common ancestors is the coefficient of relationship and is represented by R:



(i) R_{BC} = the coefficient of relationship between the full sibs B and C and is calculated as follows:

i.e., individuals B and C contain ½ × ½ = ¼ of their genes in common through ancestor D.

(ii) i.e., individuals B and C contain ½ × ½ = ¼ of their genes in common through ancestor E.

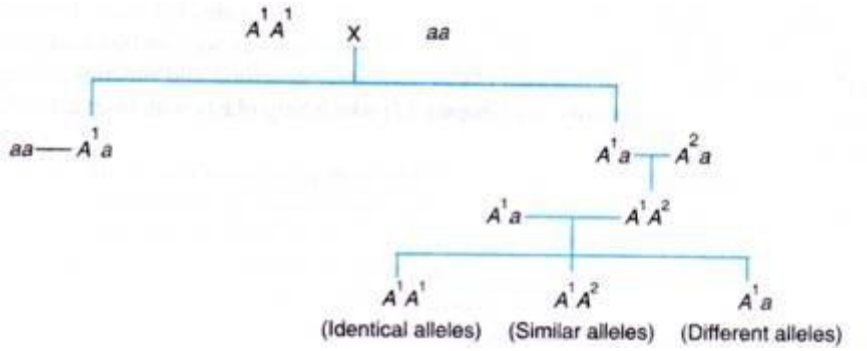
(iii) The sum of these two pathways, the coefficient of relationship, between the full sibs B and C = ¼ + ¼ = ½ = 50 per cent.

2. Inbreeding Coefficient:

In a diploid organism, each gene has two alleles occupying the same locus. They are called identical genes if they have descended from the same gene; such genes are homozygous at the locus.

Such a homozygosity is also caused when two alleles in a diploid organism have not descended from the common gene but the alleles of identical origin are brought together through mating between first cousins. Such alleles are called similar alleles.

The fine difference between these two types of alleles becomes clear by the following chart:



The probability that the two alleles in a zygote are identical by descent, i.e., are the replication product of the same gene of an ancestor is measured by the inbreeding coefficient (F) and is calculated as follows:

1. If the parents B and C are full sibs, i.e., B and C parents are 50 per cent related, the inbreeding coefficient of individual (A) can be calculated by the equation F_A = ½ R_{BC}, where R_{BC} is the coefficient of relationship between the full sib parents (B and C) of A.

2. If the common ancestors are not inbred, the inbreeding coefficient is calculated by the equation:

$$F = \Sigma (½)^{n1+n2+1}$$

where n_1, is the number of generations (arrows) from one parent back to the common ancestor and n_2 is the number of generations from the other parent back to the same ancestor.

3. In case the common ancestors are inbred, the inbreeding coefficient is calculated as follows:

$$F = \Sigma\ (½)^{n1+n2+1(1+Ancestor)}$$

4. The coefficient of inbreeding is also calculated by counting the number of arrows connecting the individual through one parent back to the common ancestor and back again to his other parent by the following equation:

$$F = \Sigma\ (½)^{n}\ (1+F_A)$$

n = number of arrows which connect the individual through one parent back to the common ancestor and back again to his other parent. F_A is the inbreeding coefficient of the common ancestor. For example, the inbreeding coefficient for A in the following arrow diagram can be calculated by following method:

B and C are the parents of A. There is only one pathway from B and C and that goes through ancestor E. Ancestor E is inbred, because its parents (G and H) are full sibs and are 50 per cent related.

The inbreeding coefficient can be calculated as follows:

$F_E = ½\ R_{GH}$ (R = the coefficient of relationship between the full sibs G and H)

or $F_E = ½\ (0.5) = 0.25$

$F_A = \Sigma\ (½)^{n}\ (1 + F_{E\ (ancestor)})$

or $F_A = (½)^3\ (1 + 0.25) = 0.156$

3. Panmixis (Random Mating):

If the breeder assigns no mating restraints upon the selected individuals, their gametes are likely to randomly unite by chance alone. This is commonly the case with outcrossing (non-self-fertilizing) plants. Wind or insect carry pollen from one plant to another in essentially a random manner.

Even livestock such as sheep and range cattle are usually bred panmicticly. The males locate females as they come into heat, copulate with ("cover") and inseminate them without any artificial restrictions as they forage for food over large tracts of grazing land. This mating method is most likely to generate the greatest genetic diversity among the progeny.

4. Assortative and Disassortative Mating:

In sexually reproducing organisms, the most rapid inbreeding system is that between brothers and sisters who share both parents in common. This type of mating is called full-sib mating and produces inbreeding coefficient of 25 per cent in the first generation of inbreeding (F_2 of Mendel).

This rate is reduced in succeeding generations since some of the alleles are now already identical. Within 10 generations, full-sib mating can produce an inbreeding coefficient of

90 per cent. The other inbreeding systems are half-sib mating, parent-offspring mating, third-cousin mating and so on.

All these inbreeding systems are called genetic assortative mating since the parents of each mating type are sorted and mated together on the basis of their genetic relationship. Such a breeding method tends to increase the inbreeding coefficient.

The assortative mating is also of the phenotypic type, i.e., the mating between two like phenotypes, two like dominant phenotypes or between two like recessive phenotypes. If assortative selective mating is continued for many generations, the heterozygotes are eliminated and the resulting population consists of homozygous dominants and homozygous recessives. If more than one locus is considered at a time, the rate of homozygosity achievement will be slower than for one locus. This is so because now the kind of heterozygotes produced will be more combinations of different loci, e.g., Aa BB, AA Bb, …) and eliminating these will need more number of generations. Disassortative mating refers to the mating of unlike phenotypes and genotypes and tends to maintain heterozygosity, as in the case of mating between unlike sexes. This preserves the dissimilarities both genetic as well as phenotypic.

In primitive organism, sexual differences arose at a single gene locus, i.e., one sex was homozygous and the other heterozygous for that locus, and the disassortative mating were the matings between an homozygous and an heterozygous individual for sex locus.

Disassortative mating also results from dichogamy, (Dichogamy = producing mature male and female reproductive structures at different times); self-sterility in plants in which the mating of like phenotypes (inbreeding) is not possible and fertilization between plants with different genotype is favoured. This maintains heterozygosity within a diploid breeding population.

5. Line Breeding:

It is a special form of inbreeding Utilized for the purpose of maintaining a high genetic relationship to a desirable ancestor. D possesses 50 per cent of B's genes and transmits 25 per cent to C. B also contributes 50 per cent of his genes to C. Hence, C contains 50 per cent +25 per cent= 75 per cent B genes and transmits half of them (37.5 per cent) to A. B also contributes 50 per cent of his genes to A. Therefore, A has 50 per cent + 37.5 per cent = 87.5 per cent of B's genes.

Genetic Effects of Inbreeding:

The continuous inbreeding results, genetically, in homozygosity. It produces homozygous stocks of dominant or recessive genes and eliminate heterozygosity from the inbred population.

For example, if we start with a population containing 100 heterozygous individuals (Aa) as shown in figure, the expected number of homozygous genotype increasing by 50% due to selfing or inbreeding in each generation. Thus, due to inbreeding in each generation the heterozygosity is reduced by 50% and after 10 generations we can expect the total elimination of heterozygosity from the inbred line and production of two homozygous or pure lines.

But, because a heterozygous individual possesses several heterozygous allelic pairs, we can conclude that inbreeding will operate on all genes loci to produce totally pure or homozygous offspring's. In human beings if inbreeding continued over a number of generations, it would results in increasing homozygosity, but somewhat slowly.

Inbreeding Depression:

In a heterozygote, the inbreeding increases the probability of homozygosity of deleterious recessive alleles in an inbred population. In other words, one of the consequence of inbreeding is a loss in vigour (i.e., less productive vegetatively and reproductively) which commonly accompanies an increase in homozygosity. This is called Inbreeding depression.

Inbreeding depression is found to occur due to following four features of inbreeding:

(1) Increase in frequency of homozygotes,
(2) Increase in variability between different inbred families,
(3) Reduction in value of quantitative character in the direction of recessive values, and
(4) The dependence of this reduction in value upon dominance.
If this inbreeding effect is multiplied for many genes at many loci, there may be a large reduction in value for many traits, including those that affect fitness and survival. In com (maize) for example, E.M. East (1908) and G. H. Shull (1909) studied the effects of inbreeding for 30 generations of inbreeding and found independently, that the yielding ability in these lines finally reduced to about one-third of the open-pollinated variety from which these samples were derived.

Both of these authors draw the following important conclusions:

(1) A number of lethal and sub-vital types appear in early generations of selfing.
(2) The material rapidly separates into distinct lines, which become increasingly uniform for differences in various morphological and functional characteristics.
(3) Many of the lines decrease in vigour and fecundity until they cannot be maintained even under the most favourable culture conditions.
(4) The lines that survive show a general decline in size and vigour.

Figure 52.4 shows the decline in size and vigour due to inbreeding in maize; here, the inheritance of two quantitative traits namely plant height and grain yield of three lines are shown for 30 generations of inbreeding. It can be noticed that fixation for plant height occurred after five generations of inbreeding. However, yield continued to decline for at least 20 generations until it reached one-third that of open-pollinated variety from which they were derived.

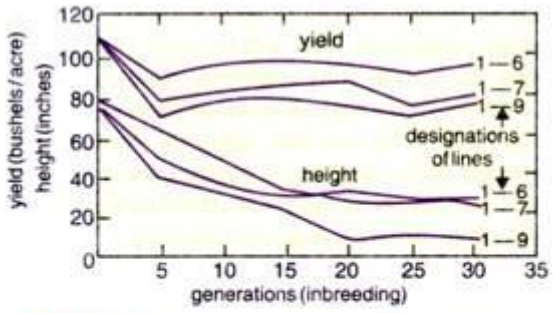


Fig. 52.4. A comparison of three lines of maize, derived from a variety, self-fertilized for 30 generations. Intitially, there were four lines, but it became impossible to maintain one of them beyond 20 generations of inbreeding.

Despite this conspicuous decline, maize was found more tolerant to inbreeding than some organisms where few strains survive two or three generations of inbreeding, e.g., alfalfa and onions. In alfalfa, upon selfing many sub-vital and lethal types appear and the rate of decline of general vigour and productivity is alarming. The very small number of lines which survive give a greatly reduced forage yield.

But onions (a normally cross- pollinated species) are quite tolerant to inbreeding, i.e., they show much less depression in vigour due to inbreeding than alfalfa and maize. Carrot is another cultivated species in which inbreeding leads to loss in vigour and production.

The following cross- pollinated plants are found to be fairly tolerant to inbreeding: sunflowers, rye, timothy, smooth broom-grass and orchard grass. In certain self-pollinated species and normally cross-fertilizing species such as cucurbits, inbreeding is found to be continued indefinitely without any ill effect.

In most animals, inbreeding is found to have less remarkable effects on vigour. For example, in rats continuous brother-sister mating were performed for 25 generations, but no drastic deterioration was detected. In Drosophila, inbreeding usually results in a rapid loss of vigour, but some strains compare favourably with outbreed populations after long continued inbreeding. However, in certain breeds of cattle, intensive inbreeding has led to an unfortunate condition; for example, exhaustive inbreeding and selection of beef cattle breed (Hereford) produced dwarf calves of low economic value.

These calves show characteristic head and body features of the brachycephalic dwarfism (i.e., the characteristic short, broad head, extra long lower jaw, bulging forehead, out of proportion abdomen and short legs). Breeding data indicate that a basic recessive gene is necessary for dwarfing, but additional modifier genes have been postulated to account for the different types of dwarfs.

Practical Applications of Inbreeding:

The correlation of inbreeding and homozygosity exhibits how inbreeding may cause deleterious effects. As we already know that in a heterozygous individual, the harmful recessive alleles remain masked by their normal dominant alleles.

If a heterozygous individual undergoes inbreeding for various generations, there will be equal chances of homozygosity for dominant as well as recessive alleles. In homozygous condition, recessive alleles will be able to express their deleterious phenotypic effects on an individual. On the other hand, the homozygosity for dominant alleles have equal opportunity to express their beneficial phenotypic effects on inbred races.

The practical applications of inbreeding are following:

1. Because inbreeding cause homozygosity of deleterious recessive genes which may result in defective phenotype, therefore, in human society, the religious ethics unknowingly and modem social norms consciously have condemned and banned the marriage of brothers and sisters. Further, the plant breeders and animal breeders too avoid inbreeding's in the individuals due to this reason.

2. The inbreeding because, results in the homozygosity of dominant allele, therefore, it is a best means of mating among hermaphrodites and self-pollinating plant species of several families. The animal breeder have employed the inbreeding to produce best races of horses, dogs, bulls, catties, etc.

The modern race horses, far example, are all descendants of three Arabian stallions imported into England between 1689 and 1730 and mated with several local mares of the slow, heavy type that had carried the medieval knights in heavy armour. The fast runners of F_1 were selected and inbred and stallions of the F_2 appear as beginning points in the pedigrees of almost all modem race horses. This sort of inbreeding in also called line breeding which has been defined as the mating of animals in such a way that their descendants will be kept closely related to an unusually desirable individual.

Similarly, merino sheep are widely known as fine wool producers. They are the result of about 200 years of inbreeding. This strain was being developed in Spain in the 17th century by stock raisers. They observed that the ancestors of the present day merino sheep had two coats of wool, one composed of long, coarse fibres arising from primary follicles, and a second coat composed of short fine wool arising from clusters of secondary follicles.
Intensive artificial selection was maintained for animals with more uniform production of fine wool and a lesser amount of coarse wool. For a time, Spain had a monopoly on the valuable merino sheep. When France invaded Spain, merino sheep were moved to France where they were maintained and eventually distributed to other parts of the world. Merino sheep were taken to South Africa and in 1796 they were introduced into Australia which has since become the world's largest producer of fine wool.

Heterosis:

When two homozygous inbreeds (a true breeding line obtained by continuous inbreeding) of genetically unlike constituents are crossed together, the resulting hybrids

obtained from the crossed seeds are usually robust, vigorous, productive and taller than the either parents.

This increased productivity or superiority over the parents is known as heterosis or hybrid vigour. Heterosis can be defined as the superiority of F_1 hybrid over both the parents in terms of yield or some other character.

History of Heterosis:

Heterosis has been known since the art of hybridization came into existence. Koelreuter (1763) was the first to report hybrid vigour in the hybrids of tobacco, Datura etc. Mendel (1865) observed this in pea crosses.

Darwin (1876) also reported that inbreeding in plants results in deterioration of vigour and the crossing in hybrid vigour. On the basis of his experiments Beal (1877-1882) concluded that F_1 hybrids yield as much as 40 percent more of the parental varieties. From subsequent studies on inter-varietal crosses in maize, it was observed that some of the hybrids show heterosis. While discussing the work on maize during a lecture at Gottingen (West Germany), Dr. G.H. ShuII (1914) proposed the term heterosis (Gr. heteros different and osis = condition). Poweri (1944, 45) reported that the crossing, however, may result in either weak or vigorous hybrids as compared to parental inbreeds.

Hybrid vigour is used as synonym of heterosis. It is generally agreed that hybrid vigour describes only superiority of the hybrid over the parents while heterosis describes the other situation as well i.e., crossing over may result in weak hybrids e.g., many hybrids in tomato are earlier (vegetative phase is replaced by reproductive phase). Earliness in many crops is agriculturally desirable so, it is argued that F, shows faster development in which vegetative phase is replaced by the reproductive phase more quickly than in the parents. On the basis of this explanation it was justified to use the term hybrid vigour as synonym of heterosis.

However, Whaley (1944) was of the opinion that it would be more appropriate to term the developed superiority of the hybrids as hybrid vigour and to refer to the mechanism by which the superiority is developed as heterosis. Smith (1955) opined that the use of heterosis and hybrid vigour as synonyms is highly desirable on the basis of their long usage.

Types of Heterosis:

Heterosis is of two types:

True heterosis (euheterosis) and pseudo-heterosis.

1. True heterosis:

It is inherited.

It can be further divided into two types:

(a) Mutational true heterosis:

It is the sheltering or shadowing of the deleterious, un-favourable, often lethal, recessive mutant genes by their adaptively superior dominant alleles.

(b) Balanced true heterosis:

It arises out of balanced gene combinations with better adaptive value and agricultural usefulness.

2. Pseudo-heterosis:

Crossing of the two parental forms brings in an accidental, excessive and un-adaptable expression of temporary vigour and vegetative overgrowth. It is also called luxuriance.

Manifestation of Heterosis:

Performance or expression of any character or trait is influenced by many genetic factors — some are positive (stimulating) and others are negative (decreasing). Expressivity of the genes or the degree of manifestation of a character is the result of genetic balance in the action of differently directed factors.

The various manifestations of heterosis may be summarised as follows:

1. Increased Yield:

Increase in yield which may be measured in terms of grain, fruit, seed, leaf, tuber or the whole plant is one of the most important manifestations of heterosis.

2. Increase in Size and General Vigour:

Heterosis results in more vigorous growth which ultimately leads to healthier and faster growing plants with increase in size than the parents.

3. Better Quality:

In many cases heterosis yields better quality which may be accompanied with higher yield.

4. Greater Adaptability:

Hybrids are generally more adapted to environmental changes than the inbred lines due to heterozygosity.

5. More Disease Resistant:

Heterosis sometimes results into development of more disease resistant character in the hybrids.

6. Increased Reproductive Ability:

Hybrids exhibit heterosis by expressing high fertility rate or reproductive ability, which is ultimately expressed in yield character.

7. Increase in Growth Rate:

In many cases the hybrids show faster growth rate than the parents, but that does not always produce larger plant size than the parents.

8. Early Flowering and Maturity:

In many cases the hybrids may show early-ness in flowering and maturity than the parents, for some crops these are the desirable characters for crop improvement. All these manifestations of heterosis can be traced at all levels of hybrid plant organisation.

Heterosis can be observes at different level such as :

a. Molecular Level:

Heterosis is manifested in increased rate of DNA reduplication, transcription and translation influencing the formation of genetic information, enzymatic activity, other regulatory mechanisms and also hybrid protein molecule formation.

b. Functional Level:

Heterosis is expressed as an effective regulation in metabolic processes and morphogenesis in hybrid organism.

c. Cellular Level:

Due to change in electro-kinetic properties of hybrid cell nuclei, the heterosis is manifested by increased mitosis.

d. Organism Level:

Heterosis is expressed as increased growth and differentiation of vegetative organs, synthesis and accumulation of nutritional substances and utilisation of metabolic process for yield formation.

Genetic Basis of Heterosis:

There are two main theories to explain the genetic cause of heterosis.

(A) Dominance Hypothesis:

This hypothesis was proposed by Davenport and further expanded by others. This hypothesis suggests that at each locus dominant allele has the favourable character, whereas the recessive allele has the unfavourable character.

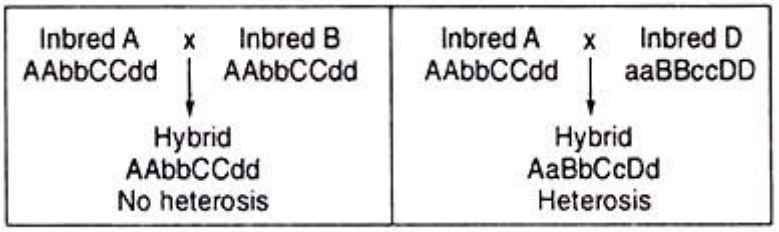
When they are combined together; i.e., in heterozygous condition in the hybrids, the favourable characters get expressed whereas the unfavourable characters are masked. So the heterosis results from the masking of harmful effects of recessive alleles by their dominant alleles.

Dominance Hypothesis has Assumptions:

(a) Dominant genes are beneficial and recessive genes are deleterious.
(b) The loci show addition effects, non-allelic interactions are absent.
(c) No recombination barrier between the genes.

With the help of following example heterosis can be explained:

In a cross between Inbred A (AAbbCCdd) with Inbred B (AAbbCCdd), there will be no heterosis in F_1 hybrid, there is no masking of recessive gene in hybrid. But in another cross, Inbred A (AAbbCCdd) is crossed with Inbred D (aaBBccDD), where the F_1 hybrid is (AaBbCcDd) with all the genes having dominant allele.
As a result the harmful effects of a, b, c, d are hidden by the dominant alleles A, B, C and D. Thus some parents produce heterotic progeny while others do not. Generally parents of diverse or different origin are more likely to produce heterotic progeny than those of similar origin.



Objection:

1. Failure in Isolation of Inbreds as Vigorous as Hybrids:

According to dominance hypothesis it should be possible to get the inbred line with all the dominant genes. Such inbreds should be as vigorous as the F_1 hybrids, but such inbreds have not been isolated.

2. Symmetrical Distribution in F_2:

According to dominance hypothesis, the quantitative characters should not show symmetrical distribution as because dominant and recessive alleles should segregate in the proportion of 3: 1, but generally the F_2 shows symmetrical distribution.

Above two objections can be explained by linked genes. Many of the quantitative characters are governed by linked genes together, so to get the inbred line with all dominant genes require several precisely placed crossovers. In another explanation it

can be showed that if the number of genes governing the quantitative characters is large, symmetrical distribution would be obtained even without linkage.

(B) Over-dominance Hypothesis:

This hypothesis was independently proposed by East and Shull. This is sometimes known as single gene heterosis, super-dominance, cumulative action of divergent alleles and stimulation of divergent alleles. According to this hypothesis, heterozygotes are superior to both the homozygotes.

So the heterozygote Aa would be superior to both the homozygotes AA and aa. Consequently, heterozygosity is essential for the cause of heterosis. In case of maize, the gene ma affects maturity. The heterozygote Ma/ma is more vigorous with late maturity than the homozygotes Ma/Ma or ma/ma.

Another proposal by East was that there are several alleles, e.g., a_1, a_2, a_3, a_4…….. etc. with increasingly different functions. Heterozygotes between more divergent alleles would be more heterotic than those involving less divergent genes, e.g., a_1a_4 is more heterotic than a_1a_2, a_2a_3, a_3a_4, etc. In these cases due to presence of divergent alleles the hybrids have the capacity to perform different functions which is not possible by any of the heterozygotes.

Objection:

1. There are many examples where the superiority is due to the epistatic affect of several non-allelic genes, not due to over-dominance (which is the interaction between allelic genes).

2. There is another objection against over-dominance hypothesis that there are many examples where the homozygotes are superior to the heterozygotes.

Physiological Basis of Heterosis:

Hybrid vigour, the product of heterotic mechanism, is essentially a physiological manifestation.

This better physiological efficiency of hybrids is derived chiefly from:

1. Better initial growth.
2. Greater uptake followed by better utilisation of nutrients by hybrids.

The initial growth activities include the different physiological processes during germination:

(a) Efficient water absorption,
(b) Better activity of enzymes,
(c) Rapid mobilization and utilization of stored food matter,
(d) Transformation and building up of active protoplasmic synthesis.

To explain all these processes different hypotheses have been put forwarded:

1. Initial Capital and Physiological Stimulus:

Large embryo and seed size in hybrids provide initial advantage to the hybrid during germination and early growth of seedlings. This hypothesis is debatable due to two reasons: the greater seedling vigour always not associated with maturity and also hybrid seeds are need not to be always with large size to attain hybrid vigour

2. Balanced Metabolism and Heterosis at Molecular Level:

The hybrids are endowed with a more balanced metabolism than their inbred parents. Many of the enzymes of heterotic plants exhibit greater efficiency over those of their better parents. The hybrids show better and rapid unfolding of balanced metabolic processes.

3. Mitochondrial Complementation and Heterosis:

ATPase activity of the mixture of mitochondria from different inbred lines of maize sometimes exceed that of the mitochondria of individual lines. This heterotic effect is called as mitochondrial complementation.

The mitochondria of heterotic hybrids absorb more O_2 and have high P/O index, i.e., phosphorylation/oxidation ratio than those of inbred lines and non-heterotic hybrids. This suggests that high level of oxidising phosphorylation and synthesis of high energy ATP bonds create favourable conditions for biosynthetic processes and important requisite for heterotic development.

4. Greater Ability for uptake and Utilisation of Nutrients:

Heterosis in post germination seedling growth is associated with better absorption and assimilation of several specific substances essential to the fundamental growth processes of the organism; such as nutritional factors, water absorption and other factors.

Efficient uptake and assimilation of nutrients by heterotic hybrid seedlings confer the following advantages:

1. Larger number of leaf primordia.
2. High carboxylase and photophosphorylation activity.
3. Greater leaf area and larger number of leaves.
4. . More branches per panicle and more grains per branch.
5. High grain weight, etc.

Effects or Manifestations of Heterosis:

Whatever may be the cause (genetical or physiological), heterosis is a well known phenomenon.

It is basically the result of the increased metabolic activity of the heterozygote Its effects are well established or manifested in the following three ways:

1. Quantitative Effects:

(a) Increase in size and genetic vigour:

Hybrids are generally more vigorous I;e larger, healthier and faster growing than the parents e.g., head size in cabbage jowar cob size in maize, fruit size in tomato etc.

(b) Increase in yield:

Yield may be measured in terms of grain, fruit, seed, leaf tuber or the whole plant. Hybrids usually have increased yield.

(c) Better quality:

Hybrids show improved quality e.g., hybrids in onion show better keeping quality.

2. Physiological Effects:

(a) Greater resistance to diseases and pests:

Some hybrids show greater resistance to insects or diseases than parents.

(b) Greater flowering and maturity:

Earliness is highly desirable in vegetables In many cases, hybrids are earlier in flowering and maturity than the parents, e.g. tomato hybrids are earlier than their parents.

(c) Greater Adaptability:

Hybrids are usually less susceptible to adverse environmental conditions.

3. Biological Effects:

Hybrids exhibiting heterosis show an increase in biological efficiency i.e., an increase in fertility (reproduction ability) and survival ability.

Heterosis in animals:

(i) Mule is a hybrid from a cross between Jack (Equus hemicus) and Mare (Earns equus) which has been known since ancient times for its well-known qualities of strength and stubbornness.

(ii) Cross between red Sindhi breed of Indian Cattle and Jersey breed of America contams 30% more butter fat in milk.

(iii) Increased pork yield in pigs, more egg laying hens, silk production in silk worms etc.

Crop varieties developed and released by Division of Genetics, IARI during 1991-2001.

Crops	Number of varieties	Names of the varieties/hybrids
Wheat	19	Sonali, Vaishali, HI8381, Kanchan, HP1731, Ganga, HW2004, HP 1761, HP1744, Vidisha, HS365, HI8498, HW1085, Shresth, HD4672, HW2044, HD2733, HI1454, HI1418
Triticale	1	DT46
Rice	9	PNR381, Pusa 44, PNR 162, Pusa 839, Pusa 677, PNR 519, RH-10, Pusa Sugandh-2, Pusa Sugandh-3
Maize	5	PEHM-1, PEHM-2, PEHM-3, Pusa Comp. 3, Pusa Comp. 4
Pearl Millet	6	Pusa 322, Pusa 444, Pusa Bajra 266, Pusa 605, Pusa 415, Pusa 334.
Sorghum	2	RusaChari 121, Pusa chari hybrid 106
Chick pea	7	Pusa 329, Pusa 372, Pusa 362, Pusa 311, Pusa 1003, BGD72, Pusa1053
Pigeon pea	2	Pusa 855, Pusa 9
Mungbean	3	Pusa 9072, Psua 9531, Pusa vishal
Field pea	4	DMR 7, DDR 13, P1542, DDR 23
Lentil	2	Shivalik, Pusa vaibhav
Cow pea	3	Rambha, Rusa safed, Pusa sampada
Mustard	3	Pusa Bahar, Pusa Agrani, Pusa gaurav
Cotton	2	Pusa 8-6, Pusa 31

Comparison between inbreeding depression and hybrid vigour.

1. Increase in Homozygosity vs. Development of Heterozygosity:

Due to inbreeding each line becomes increasingly homozygous, as a consequence the variation within a line decrease rapidly. After 7-8 generations of selfing, the lines becomes almost uniform (99% homozygosity) which are called inbred lines.

Hybridization always favours heterozygosity; the species which reproduce by cross-fertilisation are heterozygous. Due to heterozygosity the effects of many recessive alleles are not expressed in heterosis, only the dominant effects or the multiple effects are expressed.

2. Appearance vs. No/Less Expression of Some Lethal and Sub-Lethal Alleles:

Inbreeding may result in appearance of many harmful characters due to accumulation of harmful recessive alleles after selfing, e.g., chlorophyll deficiency (albina, chlorina), rootless seedlings, defective floral parts, etc. This type of effects is not found in case of heterosis as most of the lethal characters are expressed in homozygous condition. Heterosis or hybrid vigour prevails heterozygosity, so appearance of such characters does not happen.

3. Reduction vs. Increase in Vigour, Yield and Reproductive Ability:

Due to inbreeding there is a general reduction in vigour of the population, plants become shorter and weaker. The hybrids are generally more vigorous, healthier and increased in size. The reproductive ability also decreases in the population rapidly due

to inbreeding, many lines reproduce so poorly that these cannot be maintained. The hybrids exhibiting heterosis show an increase in fertility or reproductive ability.
Inbreeding generally leads to loss in yield; the inbred lines yield much less than the open pollinated varieties from which they are derived. Heterosis is generally expressed as an increase in yield of the hybrid. Commercially this phenomenon is of great importance as an objective of plant breeding.

4. More Susceptible to Disease vs. Increase in Disease Resistance Property:

Due to inbreeding as homozygosity increases, there may be rapid loss of vigour as well as disease resistance property. Whereas due to heterosis the hybrids are known to exhibit a greater resistance to insects or diseases than the parents.

5. Less Adaptable to Changed Environment vs. Greater Adaptability in Hybrids:

The inbred lines are homozygous, so they are less adaptable to changed environment as the modification of characters are not possible according to need. Whereas the hybrids are generally more adaptable to environmental changes than the inbreeds. Variation due to heterozygosity offers the hybrids more adapted to environmental variations

Probable Questions:

1. Define inbreeding? How it affect genetic diversity?
2. Define inbreeding coefficient? How it can be calculated?
3. Define random mating.
4. Define Assortative and Disassortative Mating.
5. What is inbreeding depression? Why it occurs?
6. What are practical applications of inbreeding.
7. What is heterosis? How it differs from inbreeding depression?
8. Describe types of heterosis?
9. Describe various manifestations of heterosis.
10. Explain two main theories to explain the genetic cause of heterosis.

Suggested Readings:

1. Molecular Cell Biology by Lodish, Fourth Edition.
2. The Cell – A Molecular Approach by Cooper and Hausman, Fourth Edition
3. Principles of Genetics by Snustad and Simmons, Sixth Edition.
4. Molecular Biology of the Cell – by Bruce Alberts
5. Cell and Molecular Biology by Gerald Karp, 7th Edition.
6. Gene cloning and DNA Analysis by T. A. Brown, Sixth Edition.
7. Genetics . Verma and Agarwal.

UNIT-XII

Evolutionary genetics: Origin of species

Objective: In this unit we will discuss abut origin of species and different methods of speciation.

Introduction to Genetic Evolution of Species:

The concept of 'organic evolution' envisages that all the living forms of today developed from a common ancestor. That is, the various life forms are related by descent, which accounts for the similarities among them. The idea of organic evolution was not widely accepted until 1859 when Darwin published his classic work 'The Origin of Species'.

This work contained a large body of evidence in favour of the idea that evolution continuous and it provided an attractive hypothesis to explain the mode of evolution.

Subsequently, various concepts regarding the mechanism of evolution were developed Haldane, Fischer, Wright and several others, Information's from diverse areas of study, such as, geology, palaeontology, taxonomy, population genetics, biochemistry, molecular genetics and others have been collated and resynthesized to understand evolution.

Present Status of Genetic Evolution of Species:

The modality of evolution of species in the plant kingdom involves a combination of processes and phenomena in nature. The processes cover all the changes inherent in the concepts of Drawin, de Vries, and lately by Stebbins.

The basic materials bringing about changes in the individual of a population, are the genes and their alterations. In fact, the random gene changes provide with basic raw materials in the evolutionary process.

Such changes may be major or minor, involving alterations in structure and numbers of genes as well as of chromosomes and chromosome segments. In short, genie and chromosomal alterations occurring at random in the individuals of a population, provide the basic materials for the evolution.

The next step in the evolutionary process at the population level, is the recombination of genes between different individuals. The random hybridization between different individuals containing different genetic changes leads to the origin of new individuals with newer gene combinations. At this step, the population may represent a heterogeneous mass of individuals containing different gene combinations.

The next step in evolution is the operation of natural selection in the struggle for existence among the heterogeneous recombination's, for optimum utilization of the

resources in their specific environments. Ultimately through natural selection, certain individuals with altered gene complements occupy the environmental niche with the gradual exclusion of others. Through cross breeding amongst themselves, such a population ultimately becomes stable with specific altered gene combinations and becomes a stable genotype.

The stable population characterized by a particular gene combination, stands apart from the parental species to which the population initially belonged. Such a stabilized population, characterizing a genotype differing in phenotype from its predecessors, is often considered as attaining an incipient species level. Such an incipient species can even undergo intercrossing with individuals of the parental population and may lose identity.

Allopatric Speciation:

As such, the attainment of a species status from the level of incipient species, would require a compatibility barrier between the new and the old populations. Without this barrier, despite phenotypic differences, the identity of the new population cannot be maintained.

There is every possibility of its merger with parental species through breeding in absence of barrier leading to the origin of a series of graded phenotypes. The barrier to compatibility, essential for attaining species status, can be achieved through different means.

The method without involving any genie changes leading to compatibility barrier is migration. The migration of the new population to new environment, far removed from the original, leads to geographical isolation. Such geographical isolation enables a population to develop its own phenotypic characteristic adapted to the changed environment, far removed from the original. Such species are also termed allopatric species.

Sympatric speciation:

The common method, other than the migration and consequent geographical isolation, is the genie changes or mutations leading to a barrier to fertilization.

Such barrier to fertilization between species-occupying the same geographical area, otherwise termed as sympatric species, can be achieved through seasonal isolation, i.e., blooming at different seasons caused by genie changes in the individual.

Not necessarily seasonal, but the barrier may be present even between two species maintaining their individuality, occupying the same habitat and blooming in the same season. The compatible barrier between the two species, original and derived, can also be due to incompatibility of germinal line, the pollens and ovule.

Such genie sterility may be manifested either in the absence of fertilization or barrier to post-fertilization embryonic development. Such sterility barrier at the genie level is the principal factor in stabilization and as such evolution of species.

Meaning of Isolation:

The first step in the development of a new species is isolation. A population should split up into two or more separate demes each with its own gene pool. These demes must be isolated from one another. If genes are exchanged between them they will effectively behave as one population. If they are isolated, mutation and selection can operate independently in the two populations and each can develop into a distinct species.

Mechanisms of Isolation:

There are three isolating mechanisms important that can lead to the origin of new species:

1. Geographical Isolation:

Frequently isolation is geographic. The population may be widely separated geographically or divided by impenetrable barriers such as mountain ranges and rivers. Even if they occupy the same locality they may be separated by having a preference for slightly different habitats. This is also called as ecological isolation.

2. Reproductive Isolation:

Even though two populations may not be ecologically separated, they may be effectively isolated by reproductive isolation i.e. they cannot interbreed. This might be caused by lack of attraction between males and females or by physical non-correspondence of genetalia.

In case of flowering plants it may be due to the fact that pollination is impossible between the two populations. In animals with elaborate behaviour patterns, it may be because the courtship behaviour of one fails to stimulate the other. This is also called as behavioural isolation.

3. Genetic Isolation:

Some times mating may be possible, but reproduction is not possible because of fundamental differences in genetic constitution. Thus, the gametes may be prevented from fusing. For example, the pollen grain of one population of plants, may fail to germinate on the stigmas of the other.

Even if fertilization does occur, the zygote may be inferior in some way and fail to develop properly. Sometimes offsprings are produced but the hybrids may be adaptively inferior, living only for a short time.

A reproductive isolating mechanism is a structural, functional, or behavioural characteristic that prevents successful reproduction from occurring between different species. This helps in accumulating genetic variations in species. If reproductive isolation does not exist, variant forms freely interbreed with the normal forms and this would lead to intermixing of their genotypes.

These mechanisms are of two types:

1. Pre-Mating Mechanisms:

Premating isolating mechanisms are anatomical or behavioural differences between two species that prevent the possibility of mating

i. Habitat isolation occurs when two species occupy different habitats, even within the same geographic range, so that they are less likely to meet and attempt to reproduce.

ii. Temporal isolation occurs when two species live in the same location, but each reproduces at a different time of the year, preventing a successful mating.

iii. Behavioural isolation occurs when there are differences in mating behaviour between two species.

iv. Mechanical isolation is the result of differences between two species in reproductive structures or other body parts, so that mating is prevented.

2. Post-Mating Mechanisms:

Post-mating isolating mechanisms are the result of developmental or physiological differences between the members of two species after mating.

i. Gamete isolation is the physical or chemical incompatibility of gametes of two different species. If the gametes lack receptors to facilitate fusion, they cannot form a zygote. An egg may have receptors only for the sperm of its own species.

ii. Zygote mortality is a mechanism when the zygote dies soon after its formation.

iii. Hybrid in viability – The offspring of parents of two different species is known as a hybrid. It dies before reaching sexual maturity.

iv. Hybrid sterility – The hybrid fails to reproduce sexually. For example, the mule is a sterile hybrid between a male donkey and a mare, a hinny is a sterile hybrid between a stallion and a female donkey.

Plants are bisexual and can establish a reproductively isolated species very rapidly. Polyploidy and hybridisation are important speciation mechanisms in plants.

Mechanism of Origin of Species:

There are two distinct ways in which new species arise from the preexisting one:

1. Splitting of species

2. Transformation of species

1. Splitting of species:

Suppose species A is ancestral. During the course of evolution, it will give rise to species B and specie C.

2. Transformation of species:

In this type of evolution only one species exists at a time. As for example, species A evolves into species B and B into C and so on.

According to Simpson, there are two types of transformations:

(a) Phyletic evolution:

This involves the sustained directional changes in the average characters of a population. This is caused either due to adaptations to shifting environment or due to increasing specializations for a particular environment or improved adaptation in a constant environment This results in the origin of new genera and families.

(b) Quantum evolution:

It involves rapid shift or sudden changes in the organization of a population to a new equilibrium, distinctly different from the ancestral forms and adapted to occupy new conditions. This results in the origin of higher taxonomic groups such as orders and classes. That is, quantum evolution is macro and mega evolution operating above species level.

Genetic Evolution of Species:

The concept of 'organic evolution' envisages that all the living forms of today developed from a common ancestor. That is, the various life forms are related by descent, which accounts for the similarities among them. The idea of organic evolution was not widely accepted until 1859 when Darwin published his classic work 'The Origin of Species'.

This work contained a large body of evidence in favour of the idea that evolution continuous and it provided an attractive hypothesis to explain the mode of evolution.

Subsequently, various concepts regarding the mechanism of evolution were developed Haldane, Fischer, Wright and several others, Information's from diverse areas of study, such as, geology, palaeontology, taxonomy, population genetics, biochemistry, molecular genetics and others have been collated and resynthesized to understand evolution.

Present Status of Genetic Evolution of Species:

The modality of evolution of species in the plant kingdom involves a combination of processes and phenomena in nature. The processes cover all the changes inherent in the concepts of Drawin, de Vries, and lately by Stebbins.

The basic materials bringing about changes in the individual of a population, are the genes and their alterations. In fact, the random gene changes provide with basic raw materials in the evolutionary process. Such changes may be major or minor, involving alterations in structure and numbers of genes as well as of chromosomes and chromosome segments. In short, genie and chromosomal alterations occurring at random in the individuals of a population, provide the basic materials for the evolution.

The next step in the evolutionary process at the population level, is the recombination of genes between different individuals. The random hybridization between different individuals containing different genetic changes leads to the origin of new individuals with newer gene combinations. At this step, the population may represent a heterogeneous mass of individuals containing different gene combinations. The next step in evolution is the operation of natural selection in the struggle for existence among the heterogeneous recombination's, for optimum utilization of the resources in their specific environments. Ultimately through natural selection, certain individuals with altered gene complements occupy the environmental niche with the gradual exclusion of others.

Through cross breeding amongst themselves, such a population ultimately becomes stable with specific altered gene combinations and becomes a stable genotype. The stable population characterized by a particular gene combination, stands apart from the parental species to which the population initially belonged. Such a stabilized population, characterizing a genotype differing in phenotype from its predecessors, is often considered as attaining an incipient species level. Such an incipient species can even undergo intercrossing with individuals of the parental population and may lose identity.

Allopatric Speciation:

Allopatric speciation occurs when the new species evolves in geographic isolation from the parent species. The species range, becomes subdivided by a barrier such as a new mountain range or the change in the course of a river.

Gene flow between the two subpopulations becomes impossible allowing evolution to proceed independently in each. Natural selection may favour different genotypes on either side of the barrier and random genetic drift and mutation could contribute to divergence. Over time, divergence may proceed to the point that were the two populations to meet again, they would not be able to interbreed and speciation would be

complete. This form of speciation may take place most readily in small populations at the extreme edge of a species range. The peripheral population could become isolated, for example, during contraction of the main species range in response to changing climate.

The isolated population would be subject to the founded effect and could be genetically different from the parent population. The combined effect of a small atypical population and extreme environmental conditions can cause rapid and extensive genetic reorganization through random genetic drift and strong natural selection, or, in other words a genetic revolution.

Allopatric speciation (1) occurs when a species separates into two separate groups that are isolated from one another. A physical barrier, such as a mountain range or a waterway, makes it impossible for them to breed with one another. Each species develops differently based on the demands of their unique habitat or the genetic characteristics of the group that are passed to their offspring.

When Arizona's Grand Canyon formed, squirrels and other small mammals that had once been part of a single population could no longer contact and reproduce with each other across this new geographic barrier. They could no longer interbreed. The squirrel population underwent allopatric speciation. Today, two separate squirrel species inhabit the north and south rims of the canyon. On the other hand, birds and other species that could easily cross this barrier continued to interbreed and were not divided into separate populations.

When small groups of individuals break off from the larger group and form a new species, this is called peripatric speciation (2). As in allopatric speciation, physical barriers make it impossible for members of the groups to interbreed with one another. The main difference between allopatric speciation and peripatric speciation is that in peripatric speciation, one group is much smaller than the other. Unique characteristics of the smaller groups are passed to future generations of the group, making those traits more common among that group and distinguishing it from the others.

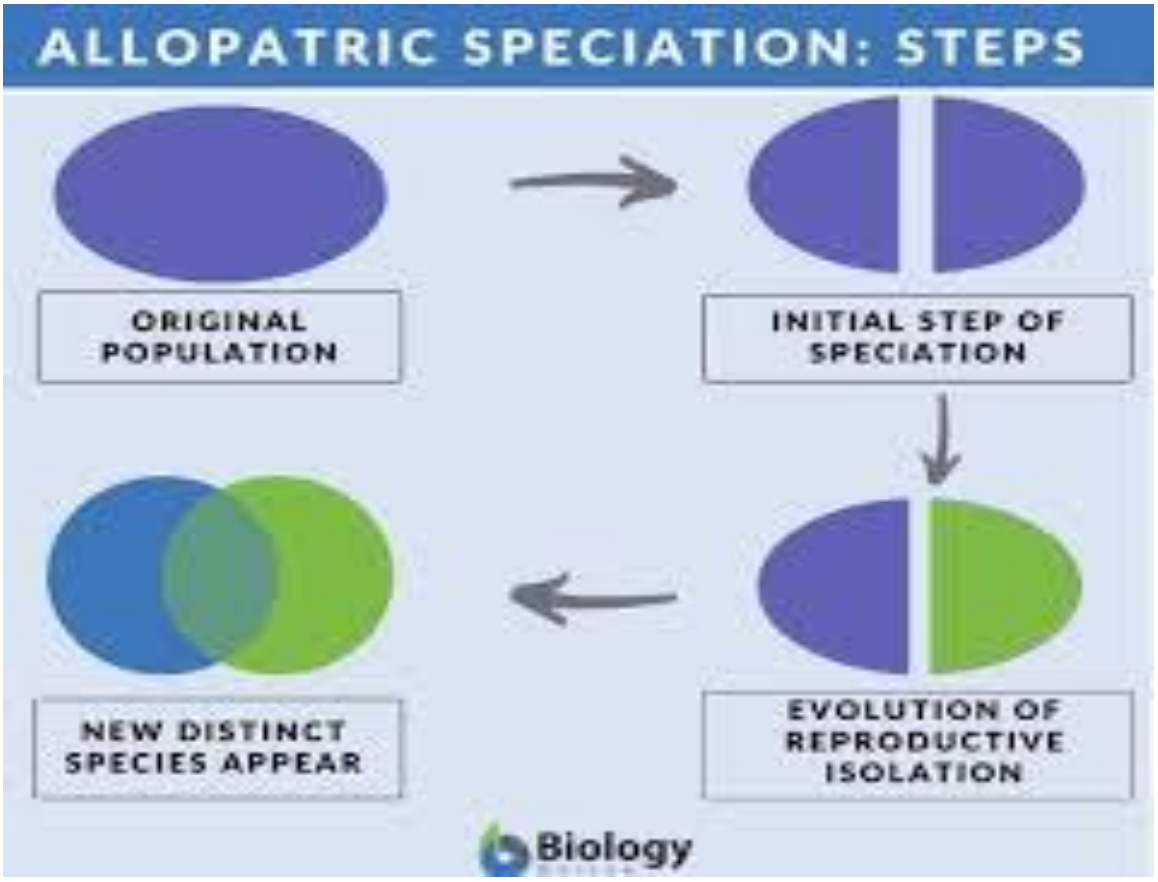


Figure: Steps of allopatric speciation

Parapatric Speciation:

This form of speciation occurs where the speciating populations are contiguous and hence only partially geographically isolated. They are able to across a common boundary during the speciation process. Where a species occupies a large geographical range it may become adapted to different environmental (e.g. climatic) conditions in different parts of that range.

Intermediate or hybrids, will be found but the large distances involved prevent the two types from merging completely. For example, the herring gull Larus argentatus is a ring species whose distribution covers a large geographical area. Westwards from Britain toward North America its appearance changes gradually, but it is still recognizable herring gull. Further west in Siberia it begins to look more like the lesser black-backed gull Larus fuscus.

From Siberia to Russia and into northern Europe it becomes progressively more like the lesser black-backed gull. The ends of the ring meet in Europe and the two geographical extremes appear to be two good biological species.

In parapatric speciation (3), a species is spread out over a large geographic area. Although it is possible for any member of the species to mate with another member,

individuals only mate with those in their own geographic region. Like allopatric and peripatric speciation, different habitats influence the development of different species in parapatric speciation. Instead of being separated by a physical barrier, the species are separated by differences in the same environment.

Parapatric speciation sometimes happens when part of an environment has been polluted. Mining activities leave waste with high amounts of metals like lead and zinc. These metals are absorbed into the soil, preventing most plants from growing. Some grasses, such as buffalo grass (Bouteloua dactyloides), can tolerate the metals. Buffalo grass, also known as vanilla grass, is native to Europe and Asia, but is now found throughout North and South America, too. Buffalo grass has become a unique species from the grasses that grow in areas not polluted by metals. Long distances can make it impractical to travel to reproduce with other members of the species. Buffalo grass seeds pass on the characteristics of the members in that region to offspring. Sometimes a species that is formed by parapatric speciation is especially suited to survive in a different kind of environment than the original species.

Sympatric Speciation:

Sympatric speciation describes a situation where there is no geo- graphical separation between the speciating populations. All individuals are, in theory, able to meet each other during the speciation process. This model usually requires a change in host preference, food preference or habitat preference in order to prevent the new species being swamped by gene flow.

Whether sympatric speciation happens at all is a contentious issue. In theory it can occur where there is a polymorphism in the population conferring adaptation to two different habitats or niches. Reproductive isolation could then arise if the two morphs had a preference for 'their' habitat. There is some evidence for this in natural populations. For example, caterpillars of the ermine moth, *Yponomeuta padellus*, feed on apple and hawthorn trees. Females prefer to lay their eggs on the species on which they were raised.

Caterpillars also prefer to feed on the plant on which their mothers were raised and adult moths prefer to mate with individuals from the same plant. The apple and hawthorn types are not completely isolated, but may represent an intermediate point in on-going sympatric speciation.

An un-contentious example of sympatric speciation occurs in plants through polyploidy. Polyploidy is the spontaneous duplication of the entire genome resulting in an individual with a multiple of the original chromosome number. Polyploidy is common in plants, where it often results in larger, more vigorous forms.

It is usually fatal in animals, although some amphibians are polyploids. The polyploid plant is no longer sexually compatible with the parent population but is able to establish a distinct population which may occupy a different habitat. The sand dune grass, Spartina townsendii, is a polyploid derived from the original S. anglica. It is more vigorous than the parent and has colonized large areas of sand dune in Britain.

A possible example of sympatric speciation is the apple maggot (Rhagoletis pomonella), an insect that lays its eggs inside the fruit of an apple, causing it to rot. As the apple falls from the tree, the maggots dig in the ground before emerging as flies several months later. The apple maggot originally laid its eggs in the fruit of a relative of the apple—a fruit called a hawthorn. After apples were introduced to North America in the 19th century, a type of maggot developed that only lays its eggs in apples. The original hawthorn species still only lays its eggs in hawthorns. The two types of maggots are not different species yet, but many scientists believe they are undergoing the process of sympatric speciation.

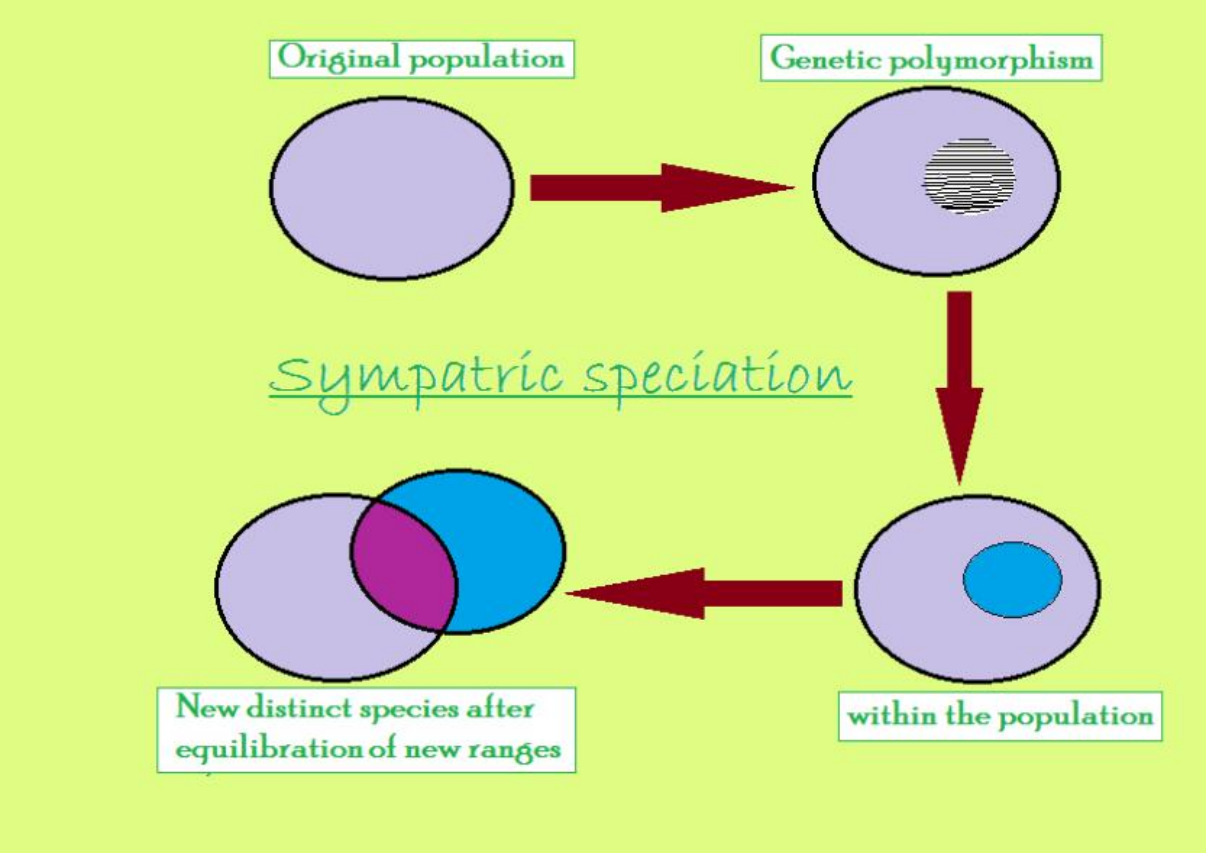


Figure: Steps of sympatric speciation

Alloparapatric Speciation:

It is specialised kind of speciation where differentiation in isolation takes place through barrier breakdown processes, as influenced by gradual environmental variation. The details of different kinds of speciation mechanisms are shown in Fig. 2.1.

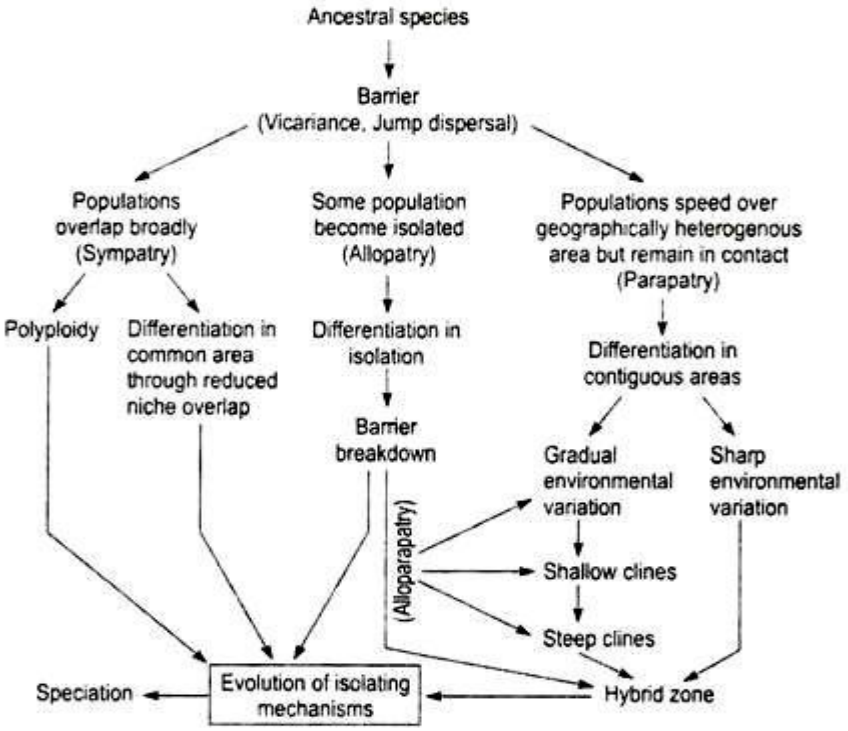


Fig. 2.1: Principal pathways of speciation: Sympatric, allopatric, parapatric and alloparapatric species (Modified after Endler, 1977)

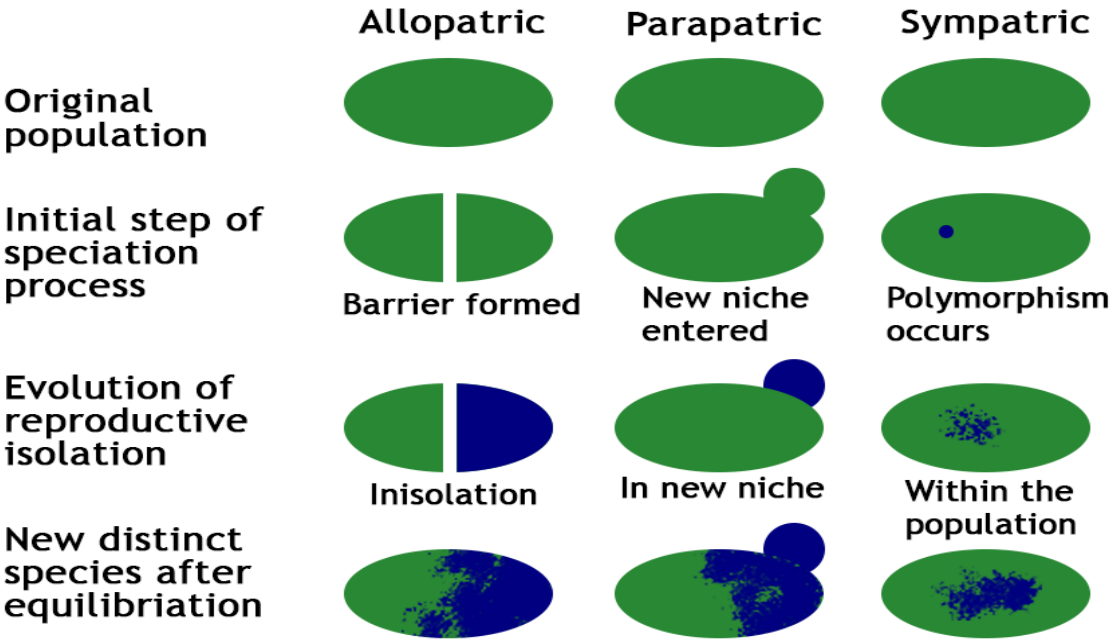


Figure: Difference between allopatric, parapatric and sympatric speciation

Most of the biologists believe that proto cells i.e. primitive life form gave rise to prokaryotes and that these in turn evolved into more complex protists, fungi, plants,animals and even man that make up the earth's stunning biodiversity.

Individuals of a species are similar and they can breed among themselves. At the same time, there are some small, but significant, differences (variations) between the individuals of a species. Heritable variations are transmitted to the offspring. These variations are important as they produce changes in the characters of that particular species. This leads to microevolution or evolution on a small scale with the emergence of new varieties or new subspecies.

To understand how such small variations lead to the formation of a new speeds, let us take the beetle's example again. Suppose there are beetles spread over a large area. If the population of beetles gets divided into two subpopulations by a barrier (say, a river or a mountain) then it will be difficult for the members of one subpopulation to go to the other side for mating. Therefore, exchange of genetic material, or gene flow, between them will decrease. They will be restricted to mate within their own subpopulations. In other words, they will be forced to inbreed, or mate with closely related individuals in their own isolated subpopulations.

In this process, the recessive mutant genes of each parent have a much greater chance of coming together. The genes will now be expressed giving benefit or harm to the offspring. These new characters, or variations, may be selected by nature and may lead to the formation of a new species. The new generations differ so much from the original population that they can no longer interbreed to produce fertile offspring. This leads to speciation, that is, the formation of one or more species from an existing species.

After a few years, if a male beetle from one isolated area and a female from another area are brought together, they may or may not mate with each other. If they mate but are unable to reproduce, then they have become two different species. If they are able to reproduce, then they are still the same species. Over many generations, different variations are accumulated in each subpopulation. Suppose, for example, in one area with a beetle subpopulation, crows are scarce due to the presence of eagles. And in another area, crows are present in large numbers.

Natural selection will not select the green variety of beetles in the first area as there are no crows to eat the beetles. But the green variety will be selected in the second area as the crows will eat the other beetles there. Thus, natural selection may operate differently on the same variations in subpopulations of different areas. Nature will select those variations that help to adapt better in a particular environment.

Over a period of time, the processes of genetic drift and natural selection will cause the two isolated subpopulations to become more and more different from each other. Microevolution is generally a consequence of gene mutation. But larger changes in the genetic make-up, like change in the number of chromosomes, may not allow the germ cells of two subpopulations to fuse together. This prevents interbreeding and causes the emergence of new species.

Speciation due to inbreeding, genetic drift and natural selection will be applicable to all sexually reproducing organisms. Geographical isolation does not play any role in the speciation of asexually reproducing organisms. It also does not play any major role in the speciation of self-pollinating plants.

Probable questions:

1. Discuss different types of pre zygotic isolating mechanism.
2. Discuss different types of post zygotic isolating mechanism.
3. Discuss characteristics and examples of allopatric speciation.
4. Discuss characteristics and examples of parapatric speciation.
5. Discuss characteristics and examples of sympatric speciation.
6. Differentiate phyletic and quantum evolution.

Suggested readings*:*

1. Snustad D P, Simmons MJ. 2009. Principles of Genetics. V Edition. John Wiley and Sons Inc
2. Strickberger M. W – Genetics; Macmillam
3. Tamarin R. H. – Principles of Genetics; McGraw Hill
4. Klug W S, Cummings MR, Spencer CA. 2012. Concepts of Genetics. Xth Ed.Benjamin Cummings

UNIT-XIII
Phylogenetic trees and molecular evolution

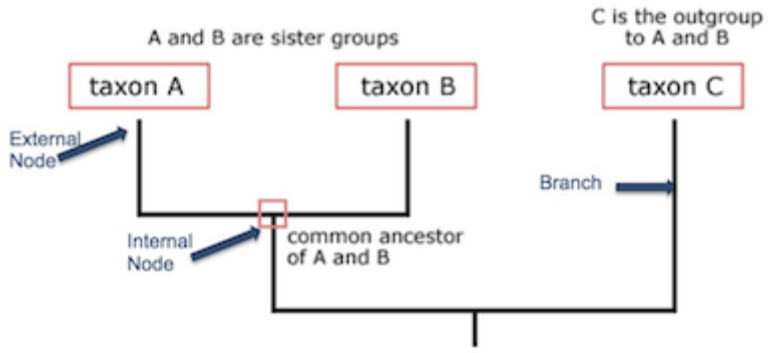
Objective: In this unit we will discuss about different aspects of phylogenetic tree and its role in explaining molecular eviolution

Phylogenetic Tree

A phylogenetic tree or **evolutionary tree** is a diagrammatic representation of the evolutionary relationship among various taxa. The phylogenetic tree, including its reconstruction and reliability assessment, is discussed in more detail in Chapter 9. The terms **evolutionary tree**, **phylogenetic tree**, and **cladogram** are often used interchangeably to mean the same thing—that is, the evolutionary relationships among taxa. The term dendrogram is also used interchangeably with cladogram, although there are subtle differences. Thus, it is important to be aware that usage of the vocabulary is not always consistent in the literature, although the context is the same, that is, representation of the evolutionary relationships of taxa.

What Is the Tree of Life?

Hennig's method of visualizing these relationships resulted in what we loosely refer to as a genealogical tree of life. The tree is constructed using a system of **nodes** and **branches**.



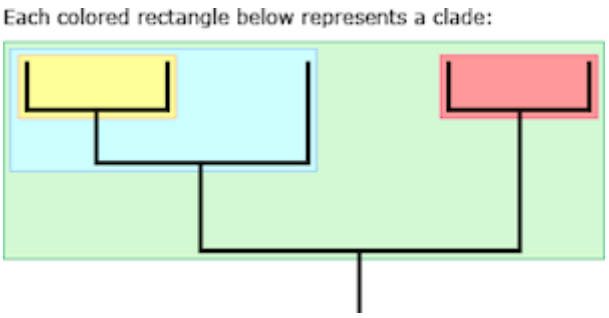
Anatomical description of the parts of a phylogenetic tree:

The term **node** refers to any terminating end of a branch (a line). **External nodes** represent the final taxon (singular form of taxa) while **internal nodes** represent a common ancestor that underwent some **speciation event** (where organisms within

that taxon stop interbreeding due to reasons like physical isolation, such as the formation of an island, or the preference of a particular physical trait that a subset of the population begins to favor through the process of sexual selection). As a result, speciation events give rise to divergent lineages of taxa and are represented by horizontal branches.

These diverging lines of taxa stem from a common ancestor, resulting in a relationship called **sister taxa** (such as taxon A and taxon B), meaning that they share the closest evolutionary relationship because they stem from the same common ancestor. In this way chimpanzees are our sister taxon, as we are more evolutionarily related to them than we are with, say, gorillas.

Taxa outside of that common ancestor are referred to as **outgroups** as they are more evolutionarily distant in relation than sister taxa are to one another, due to a more distant common ancestor. With each successive speciation event, a new clade is formed within the tree, allowing scientists to identify common ancestors between evolutionarily distant taxa.



Example clades highlighted by color

Anatomy of a phylogenetic tree

When we draw a phylogenetic tree, we are representing our best hypothesis about how a set of species (or other groups) evolved from a common ancestor^11start superscript, 1, end superscript. As we'll explore further in the article on building trees, this hypothesis is based on information we've collected about our set of species – things like their physical features and the DNA sequences of their genes.
[Are phylogenetic trees only for species?]
In a phylogenetic tree, the species or groups of interest are found at the tips of lines referred to as the tree's **branches**. For example, the phylogenetic tree below represents

relationships between five species, A, B, C, D, and E, which are positioned at the ends of the branches:

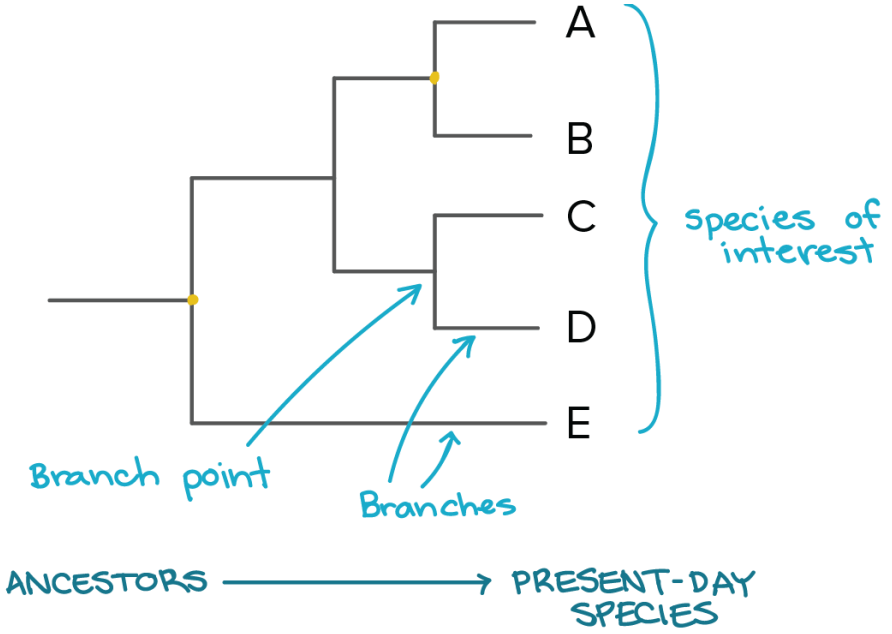
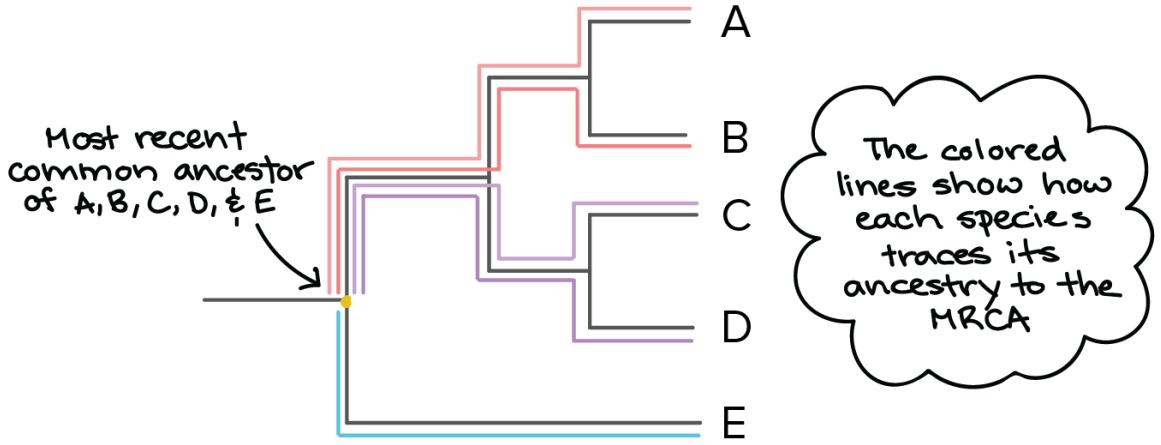
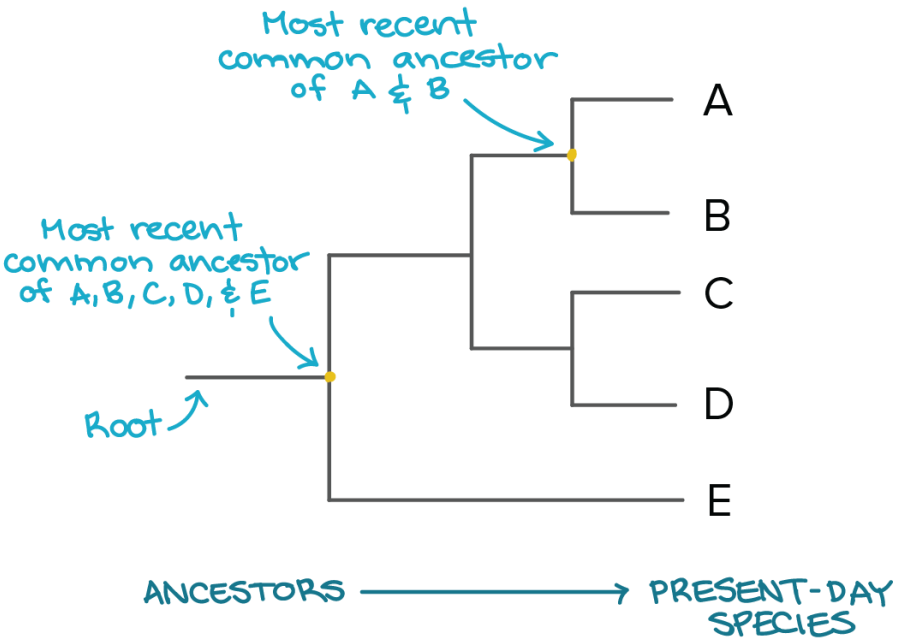


Image modified from *Taxonomy and phylogeny: Figure 2* by Robert Bear et al., CC BY 4.0

The pattern in which the branches connect represents our understanding of how the species in the tree evolved from a series of common ancestors. Each branch point (also called an **internal node**) represents a **divergence** event, or splitting apart of a single group into two descendant groups.

At each branch point lies the **most recent common ancestor** of all the groups descended from that branch point. For instance, at the branch point giving rise to species A and B, we would find the most recent common ancestor of those two species. At the branch point right above the **root** of the tree, we would find the most recent common ancestor of all the species in the tree (A, B, C, D, E).





Each horizontal line in our tree represents a series of ancestors, leading up to the species at its end. For instance, the line leading up to species E represents the species' ancestors since it diverged from the other species in the tree. Similarly, the root represents a series of ancestors leading up to the most recent common ancestor of all the species in the tree.

Which species are more related?

In a phylogenetic tree, the **relatedness** of two species has a very specific meaning. Two species are *more* related if they have a *more recent* common ancestor, and *less* related if they have a *less recent* common ancestor. We can use a pretty straightforward method to find the most recent common ancestor of any pair or group of species. In this method, we start at the branch ends carrying the two species of interest and "walk backwards" in the tree until we find the point where the species' lines converge.

For instance, suppose that we wanted to say whether A and B or B and C are more closely related. To do so, we would follow the lines of both pairs of species backward in the tree. Since A and B converge at a common ancestor first as we move backwards, and B only converges with C after its junction point with A, we can say that A and B are more related than B and C.

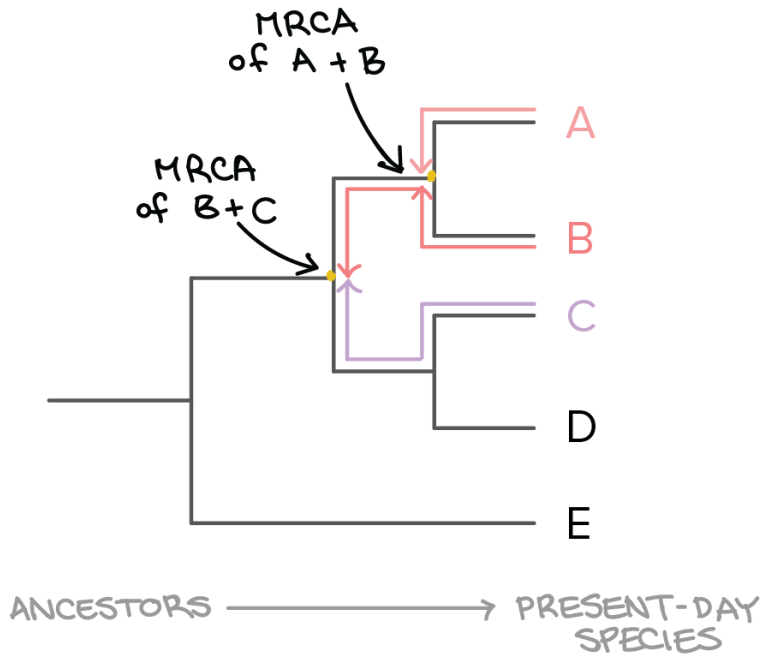


Image modified from *Taxonomy and phylogeny: Figure 2* by Robert Bear et al., CC BY 4.0

Importantly, there are some species whose relatedness we can't compare using this method. For instance, we can't say whether A and B are more closely related than C and D. That's because, by default, the horizontal axis of the tree doesn't represent time in a direct way. So, we can only compare the timing of branching events that occur on the same lineage (same direct line from the root of the tree), and not those that occur on different lineages.

Some tips for reading phylogenetic trees

You may see phylogenetic trees drawn in many different formats. Some are blocky, like the tree at left below. Others use diagonal lines, like the tree at right below. You may also see trees of either kind oriented vertically or flipped on their sides, as shown for the blocky tree.

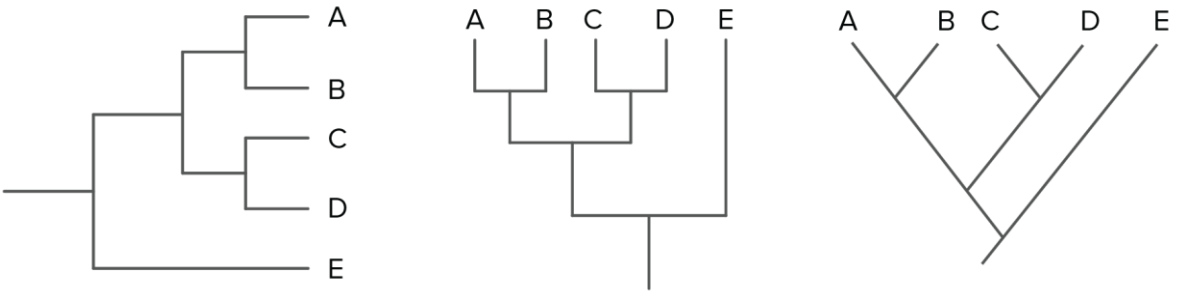
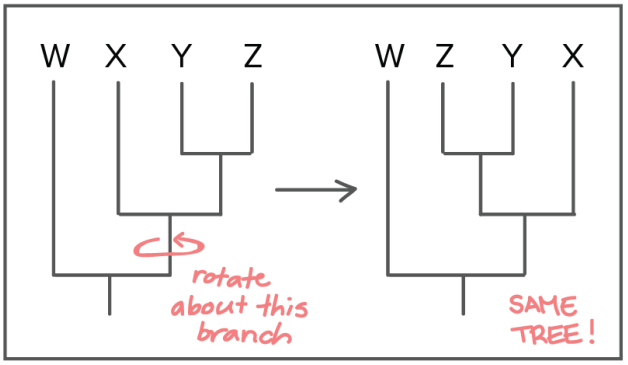


Image modified from *Taxonomy and phylogeny: Figure 2* **by Robert Bear et al.,** CC BY 4.0

The three trees above represent identical relationships among species A, B, C, D, and E. You may want to take a moment to convince yourself that this is really the case – that is, that no branching patterns or recent-ness of common ancestors are different between the two trees. The identical information in these different-looking trees reminds us that it's the branching pattern (and not the lengths of branches) that's meaningful in a typical tree.

Another critical point about these trees is that if you rotate the structures, using one of the branch points as a pivot, you don't change the relationships. So just like the two trees above, which show the same relationships even though they are formatted differently, all of the trees below show the same relationships among four species:



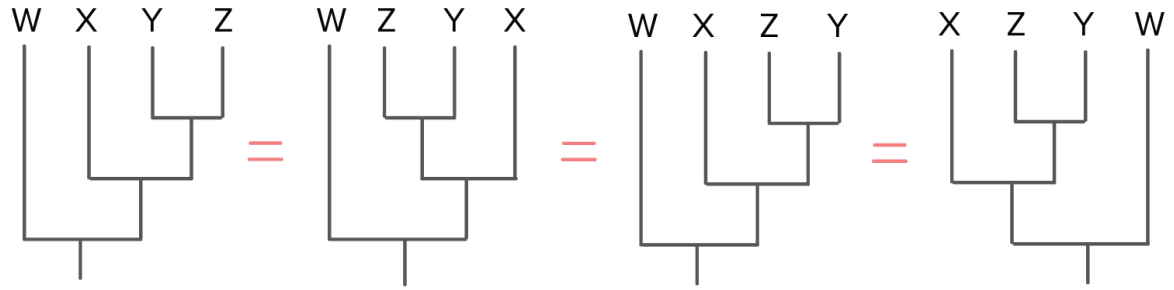


Image modified from *Taxonomy and phylogeny: Figure 3* **by Robert Bear et al.,** CC BY 4.0

If you don't see right away how that is true (and I didn't, on first read!), just concentrate on the relationships and the branch points rather than on the ordering of species (W, X, Y, and Z) across the tops of the diagrams. That ordering actually doesn't give us useful information. Instead, it's the branch structure of each diagram that tells us what we need to understand the tree.

So far, all the trees we've looked at have had nice, clean branching patterns, with just two lineages (lines of descent) emerging from each branch point. However, you may see trees with a **polytomy** (*poly*, many; *tomy*, cuts), meaning a branch point that has three or more different species coming off of it^22squared. In general, a polytomy shows where we don't have enough information to determine branching order.

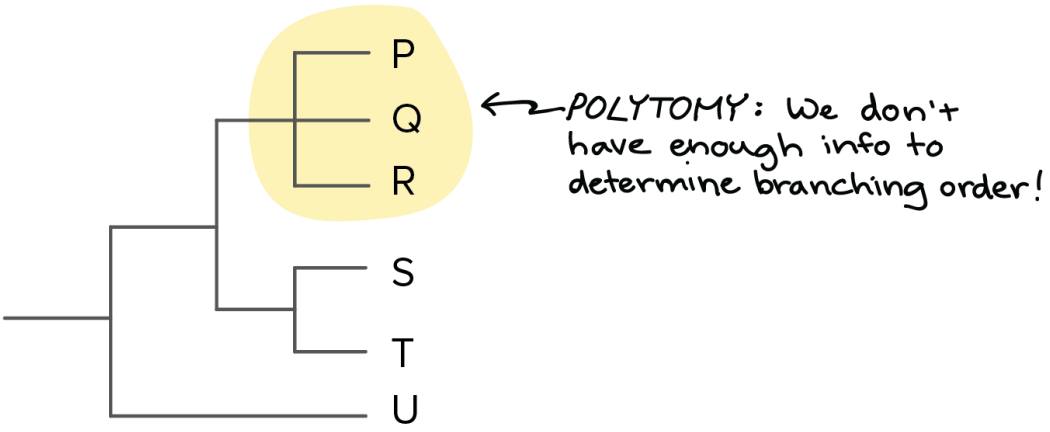


Image modified from *Taxonomy and phylogeny: Figure 2* **by Robert Bear et al.,** CC BY 4.0

If we later get more information about the species in a tree, we may be able to resolve a polytomy using the new information.

Where do these trees come from?

To generate a phylogenetic tree, scientists often compare and analyze many characteristics of the species or other groups involved. These characteristics can include external morphology (shape/appearance), internal anatomy, behaviors, biochemical pathways, DNA and protein sequences, and even the characteristics of fossils.
To build accurate, meaningful trees, biologists will often use many different characteristics (reducing the chances of any one imperfect piece of data leading to a wrong tree). Still, phylogenetic trees are hypotheses, not definitive answers, and they can only be as good as the data available when they're made. Trees are revised and updated over time as new data becomes available and can be added to the analysis. This is particularly true today, as DNA sequencing increases our ability to compare genes between species.

Construction of Phylogenetic tree:

In a phylogenetic tree, the species of interest are shown at the tips of the tree's branches. The branches themselves connect up in a way that represents the evolutionary history of the species—that is, how we think they evolved from a common ancestor through a series of divergence (splitting-in-two) events. At each branch point lies the most recent common ancestor shared by all of the species descended from that branch point. The lines of the tree represent long series of ancestors that extend from one species to the next.

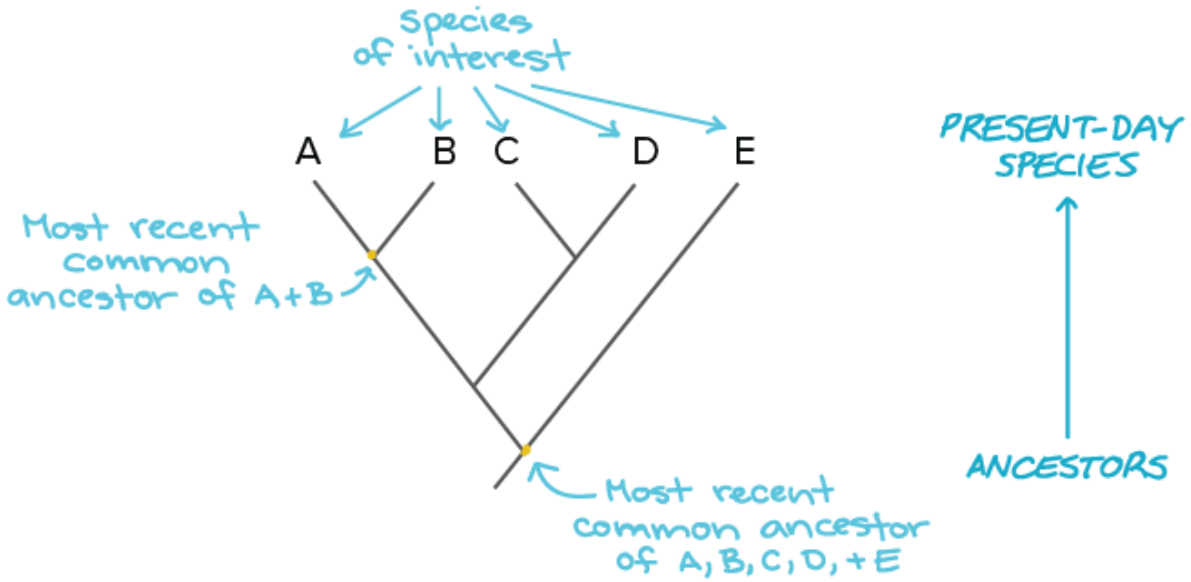
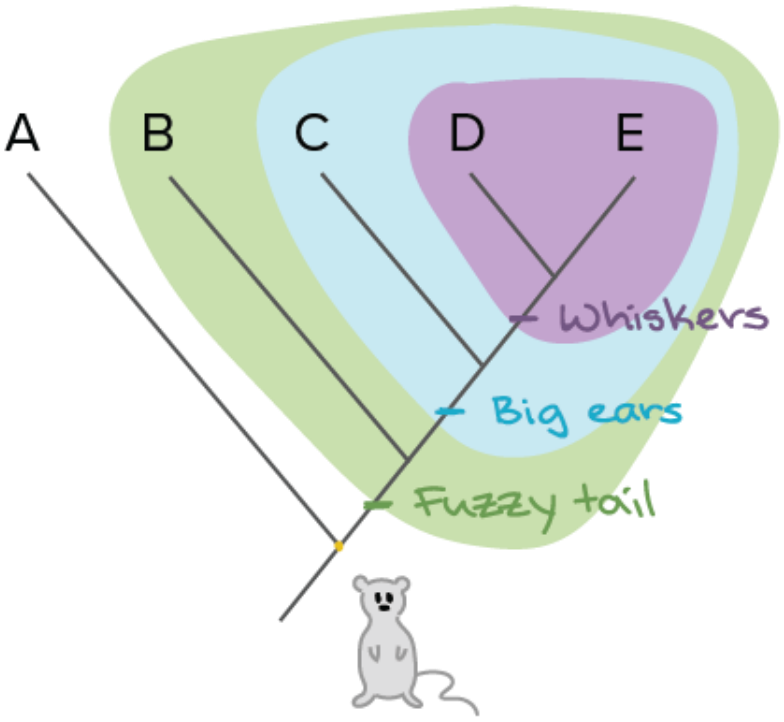


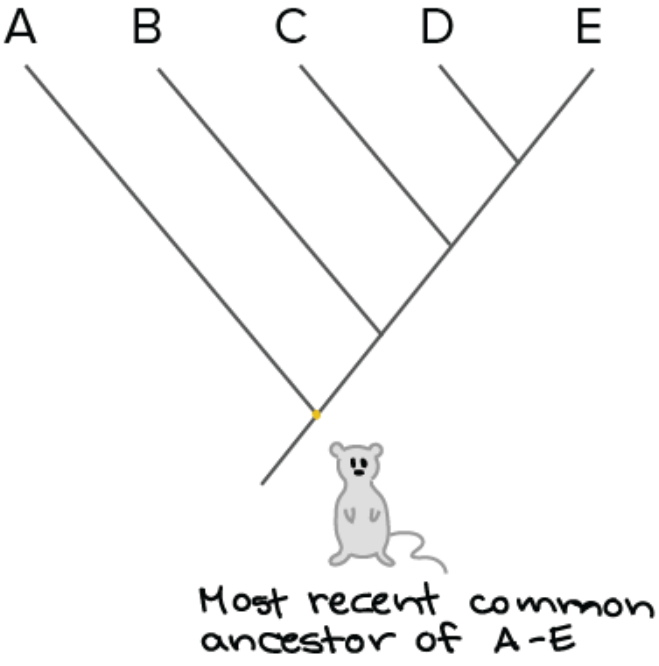
Image modified from Taxonomy and phylogeny: Figure 2**, by Robert Bear et al.,** CC BY 4.0

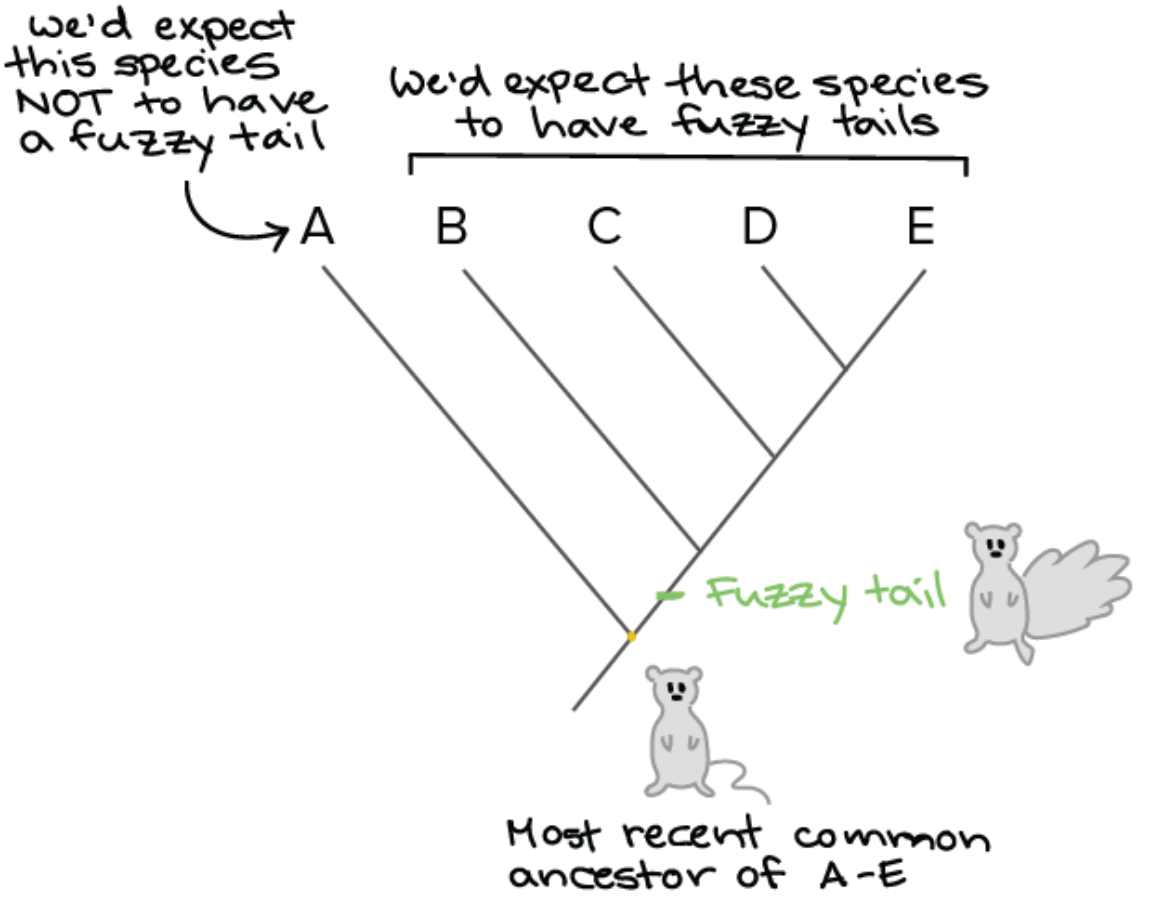
The idea behind tree construction:

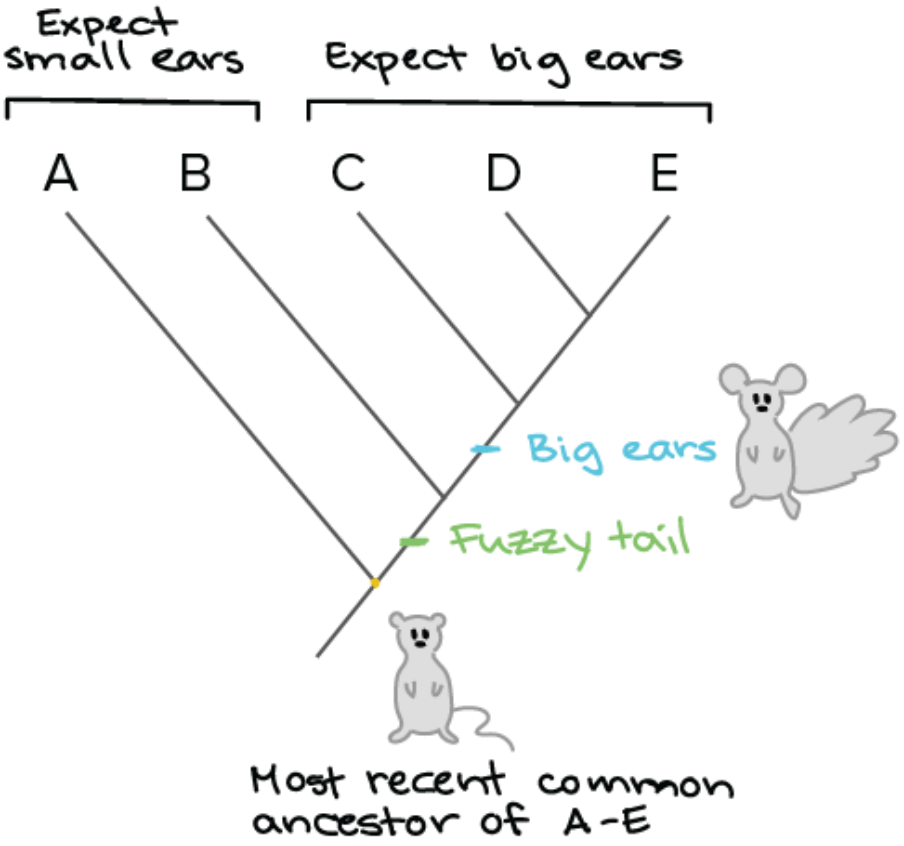
How do we build a phylogenetic tree? The underlying principle is Darwin's idea of "descent with modification." Basically, by looking at the pattern of modifications (novel traits) in present-day organisms, we can figure out—or at least, make hypotheses about—their path of descent from a common ancestor.

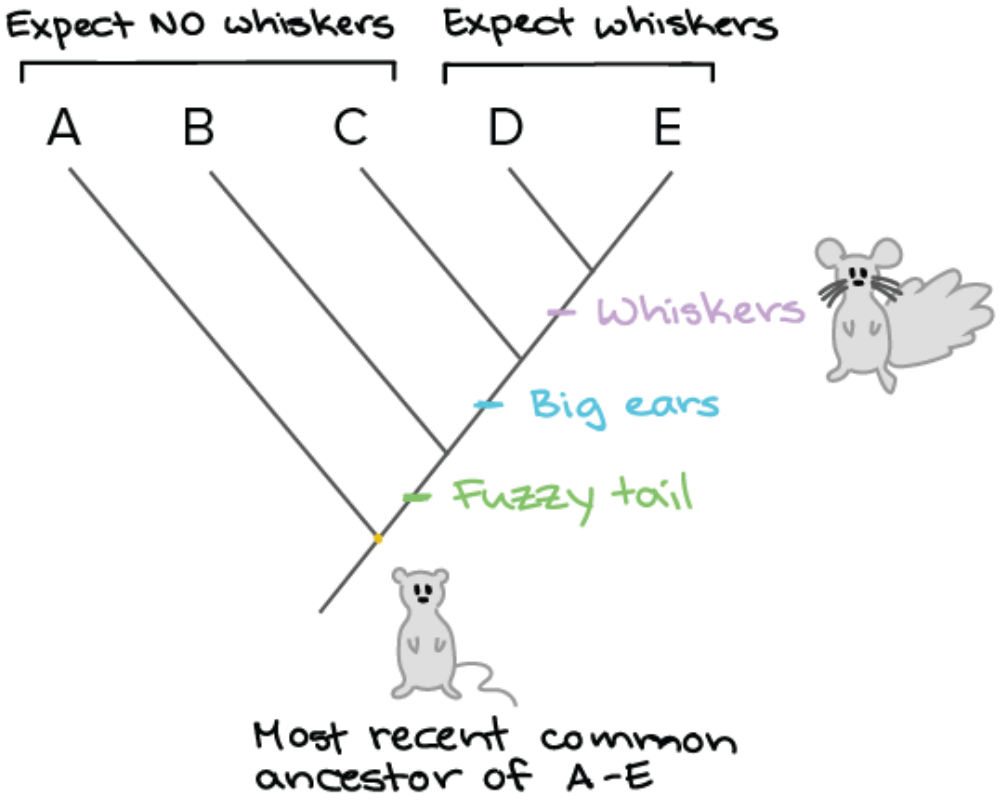
As an example, let's consider the phylogenetic tree below (which shows the evolutionary history of a made-up group of mouse-like species). We see three new traits arising at different points during the evolutionary history of the group: a fuzzy tail, big ears, and whiskers. Each new trait is shared by all of the species descended from the ancestor in which the trait arose (shown by the tick marks), but absent from the species that split off before the trait appeared.

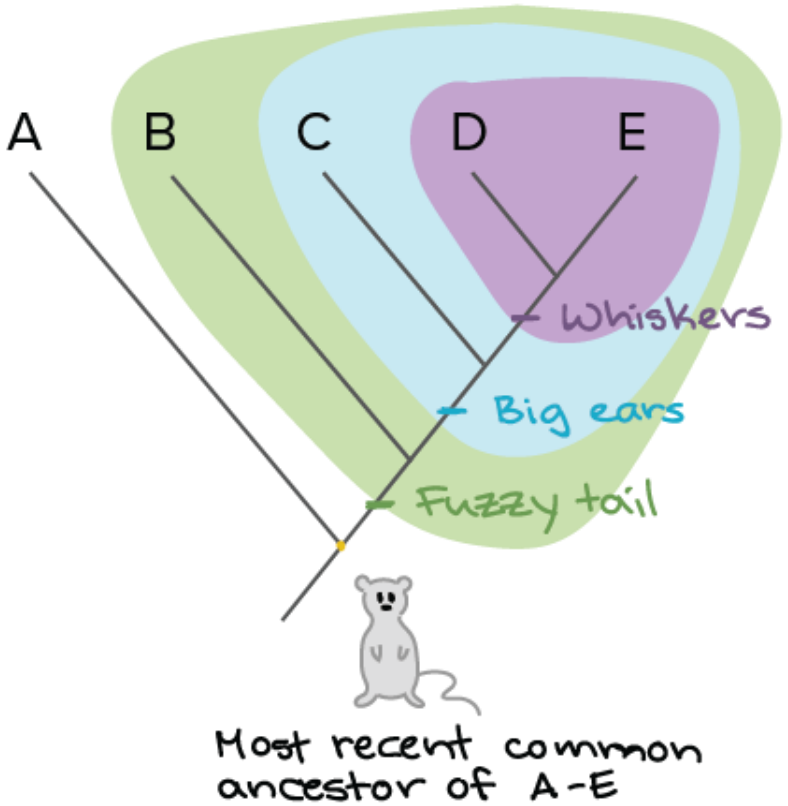

A
B
C
D
E
Whiskers
Big ears
Fuzzy tail
Most recent common ancestor of A-E


A
B
C
D
E
Most recent common ancestor of A-E


We'd expect this species NOT to have a fuzzy tail
We'd expect these species to have fuzzy tails
A
B
C
D
E
Fuzzy tail
Most recent common ancestor of A-E


Expect small ears
Expect big ears
A
B
C
D
E
Big ears
Fuzzy tail
Most recent common ancestor of A-E





When we are building phylogenetic trees, traits that arise during the evolution of a group and differ from the traits of the ancestor of the group are called **derived traits**. In our example, a fuzzy tail, big ears, and whiskers are derived traits, while a skinny tail, small ears, and lack of whiskers are **ancestral traits**. An important point is that a

derived trait may appear through either loss or gain of a feature. For instance, if there were another change on the E lineage that resulted in loss of a tail, taillessness would be considered a derived trait.
Derived traits shared among the species or other groups in a dataset are key to helping us build trees. As shown above, shared derived traits tend to form nested patterns that provide information about when branching events occurred in the evolution of the species.
When we are building a phylogenetic tree from a dataset, our goal is to use shared derived traits in present-day species to infer the branching pattern of their evolutionary history. The trick, however, is that we can't watch our species of interest evolving and see when new traits arose in each lineage.
Instead, we have to work backwards. That is, we have to look at our species of interest – such as A, B, C, D, and E – and figure out which traits are ancestral and which are derived. Then, we can use the shared derived traits to organize the species into nested groups like the ones shown above. A tree made in this way is a hypothesis about the evolutionary history of the species – typically, one with the simplest possible branching pattern that can explain their traits.

Example: Building a phylogenetic tree

If we were biologists building a phylogenetic tree as part of our research, we would have to pick which set of organisms to arrange into a tree. We'd also have to choose which characteristics of those organisms to base our tree on (out of their many different physical, behavioural, and biochemical features).
If we're instead building a phylogenetic trees for a class (which is probably more likely for readers of this article), odds are that we'll be given a set of characteristics, often in the form of a table, that we need to convert into a tree. For example, this table shows presence (+) or absence (0) of various features:

Feature	Lamprey	Antelope	Bald eagle	Alligator	Sea bass
Lungs	0	+	+	+	0
Jaws	0	+	+	+	+
Feathers	0	0	+	0	0
Gizzard	0	0	+	+	0
Fur	0	+	0	0	0

Table modified from *Taxonomy and phylogeny: Figure 4****, by Robert Bear et al.,*** *CC BY 4.0*

Next, we need to know which form of each characteristic is ancestral and which is derived. For example, is the presence of lungs an ancestral trait, or is it a derived trait? As a reminder, an ancestral trait is what we think was present in the common ancestor of the species of interest. A derived trait is a form that we think arose somewhere on a lineage descended from that ancestor.

Without the ability to look into the past (which would be handy but, alas, impossible), how do we know which traits are ancestral and which derived?

- In the context of homework or a test, the question you are solving may tell you which traits are derived vs. ancestral.
- If you are doing your own research, you may have knowledge that allows you identify ancestral and derived traits (e.g., based on fossils).
- You may be given information about an **outgroup**, a species that's more distantly related to the species of interest than they are to one another.

  If we are given an outgroup, the outgroup can serve as a proxy for the ancestral species. That is, we may be able to assume that its traits represent the ancestral form of each characteristic.

  For instance, in our example (data repeated below for convenience), the lamprey, a jawless fish that lacks a true skeleton, is our outgroup. As shown in the table, the lamprey lacks all of the listed features: it has no lungs, jaws, feathers, gizzard, or fur. Based on this information, we will assume that absence of these features is ancestral, and that presence of each feature is a derived trait.

Feature	Lamprey	Antelope	Bald eagle	Alligator	Sea bass
Lungs	**0**	+	+	+	0
Jaws	**0**	+	+	+	+
Feathers	**0**	0	+	0	0
Gizzard	**0**	0	+	+	0
Fur	**0**	+	0	0	0

Table modified from *Taxonomy and phylogeny: Figure 4****, by Robert Bear et al.,*** *CC BY 4.0*

Now, we can start building our tree by grouping organisms according to their shared derived features. A good place to start is by looking for the derived trait that is shared between the largest number of organisms. In this case, that's the presence of jaws: all the organisms except the outgroup species (lamprey) have jaws. So, we can start our tree by drawing the lamprey lineage branching off from the rest of the species, and we can place the appearance of jaws on the branch carrying the non-lamprey species.

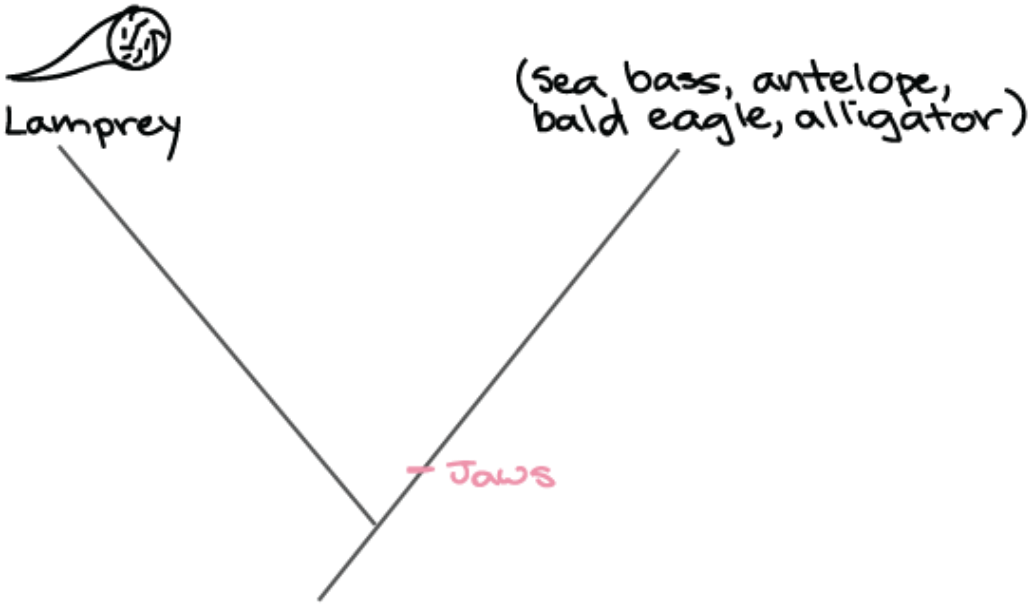


Image based on Taxonomy and phylogeny: Figure 6**, by Robert Bear et al.,** CC BY 4.0

Next, we can look for the derived trait shared by the next-largest group of organisms. This would be lungs, shared by the antelope, bald eagle, and alligator, but not by the sea bass. Based on this pattern, we can draw the lineage of the sea bass branching off, and we can place the appearance of lungs on the lineage leading to the antelope, bald eagle, and alligator.

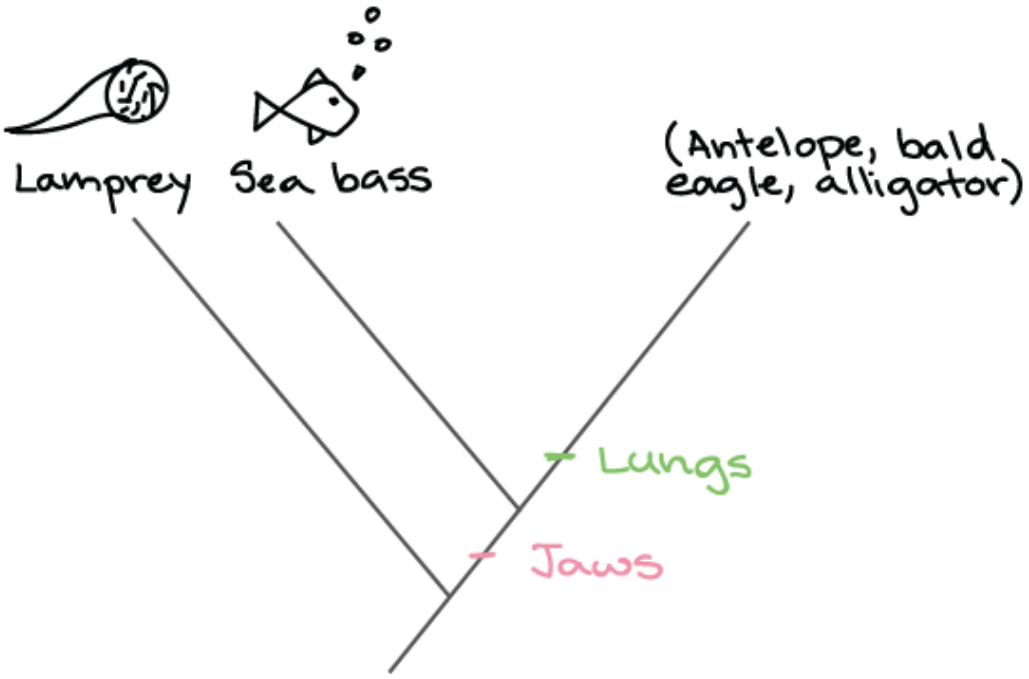


Image based on Taxonomy and phylogeny: Figure 6**, by Robert Bear et al.,** CC BY 4.0

Following the same pattern, we can now look for the derived trait shared by the next-largest number of organisms. That would be the gizzard, which is shared by the alligator and the bald eagle (and absent from the antelope). Based on this data, we can draw the

antelope lineage branching off from the alligator and bald eagle lineage, and place the appearance of the gizzard on the latter.

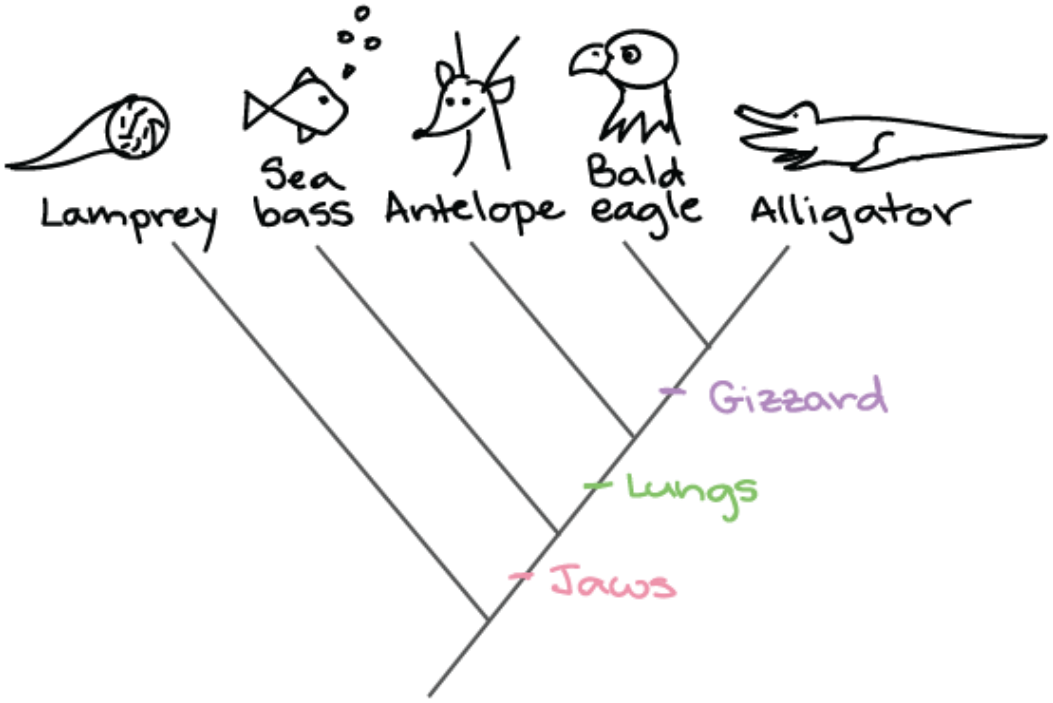


Image based on Taxonomy and phylogeny: Figure 6**, by Robert Bear et al.,** CC BY 4.0

What about our remaining traits of fur and feathers? These traits are derived, but they are not shared, since each is found only in a single species. Derived traits that aren't shared don't help us build a tree, but we can still place them on the tree in their most likely location. For feathers, this is on the lineage leading to the bald eagle (after divergence from the alligator). For fur, this is on the antelope lineage, after its divergence from the alligator and bald eagle.

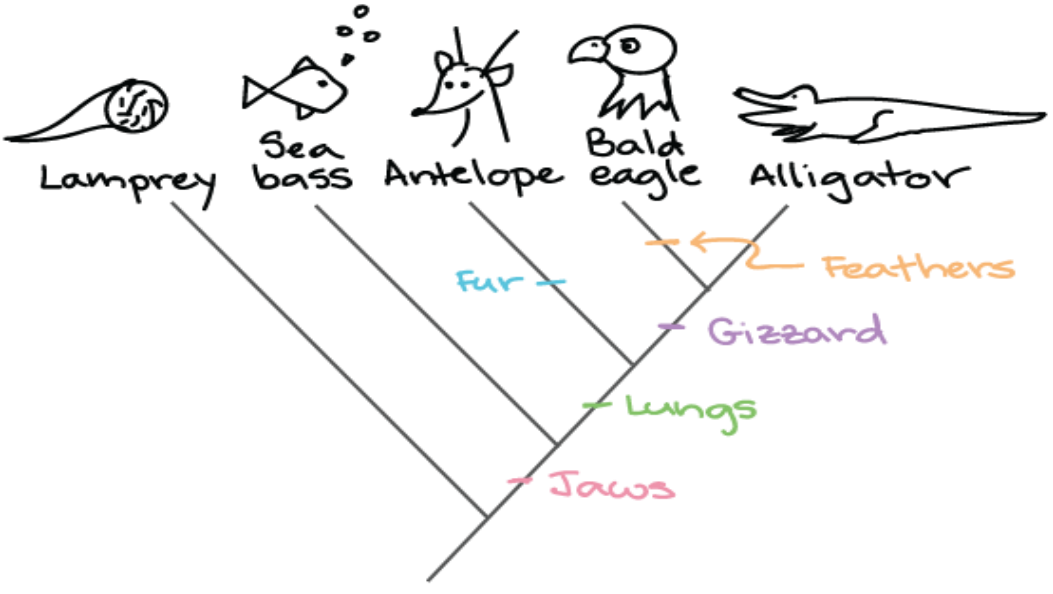


Image based on Taxonomy and phylogeny: Figure 6**, by Robert Bear et al.,** CC BY 4.0

Parsimony and pitfalls in tree construction

When we were building the tree above, we used an approach called **parsimony**. Parsimony essentially means that we are choosing the simplest explanation that can account for our observations. In the context of making a tree, it means that we choose the tree that requires the fewest independent genetic events (appearances or disappearances of traits) to take place.

For example, we *could* have also explained the pattern of traits we saw using the following tree:

Image based on Taxonomy and phylogeny: Figure 6**, by Robert Bear et al.,** CC BY 4.0

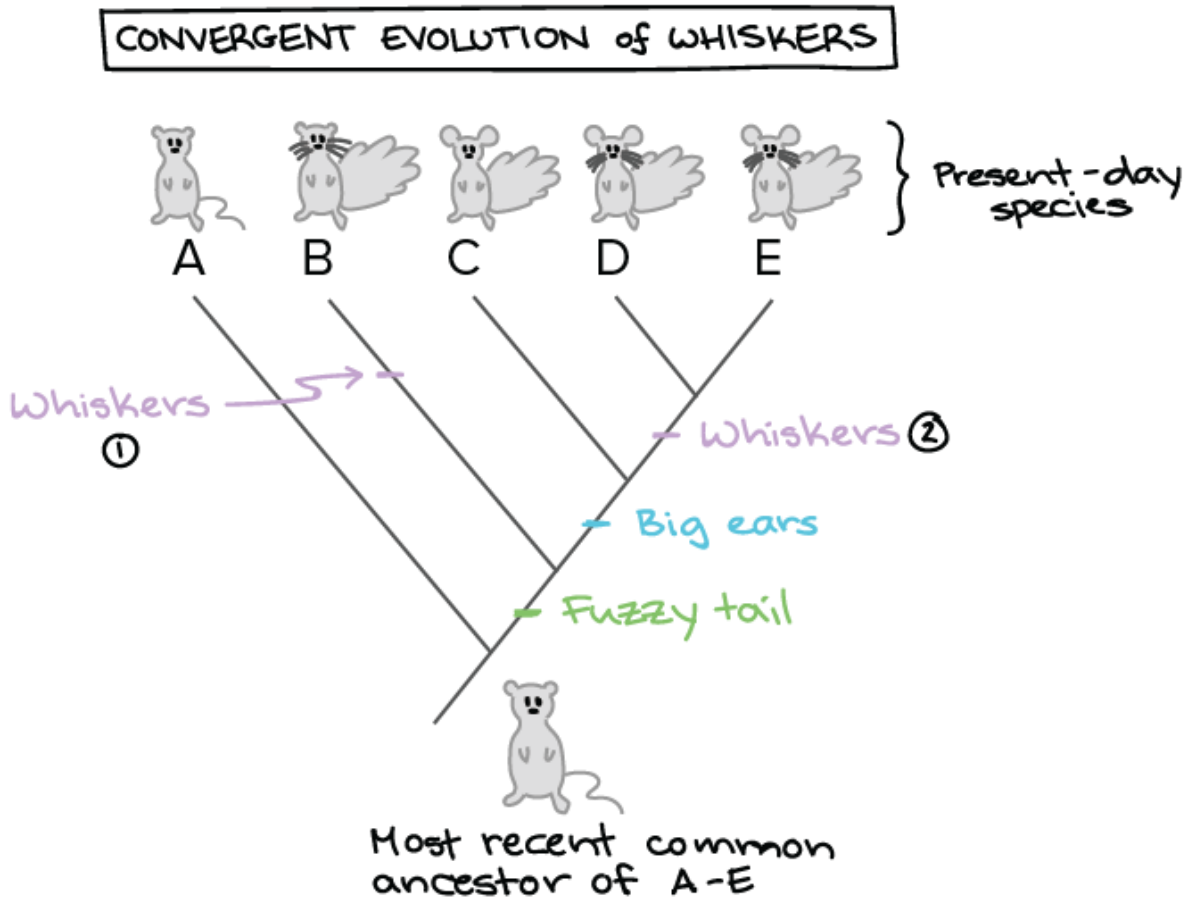
This series of events also provides an evolutionary explanation for the traits we see in the five species. However, it is *less parsimonious* because it requires more independent changes in traits to take place. Because where we've put the sea bass, we have to hypothesize that jaws independently arose two separate times (once in the sea bass lineage, and once in the lineage leading to antelopes, bald eagles, and alligators). This gives the tree a total of 666 tick marks, or trait change events, versus 555 in the more parsimonious tree above.

In this example, it may seem fairly obvious that there is one best tree, and counting up the tick marks may not seem very necessary. However, when researchers make phylogenies as part of their work, they often use a large number of characteristics, and the patterns of these characteristics rarely agree 100\%100%100, percent with one

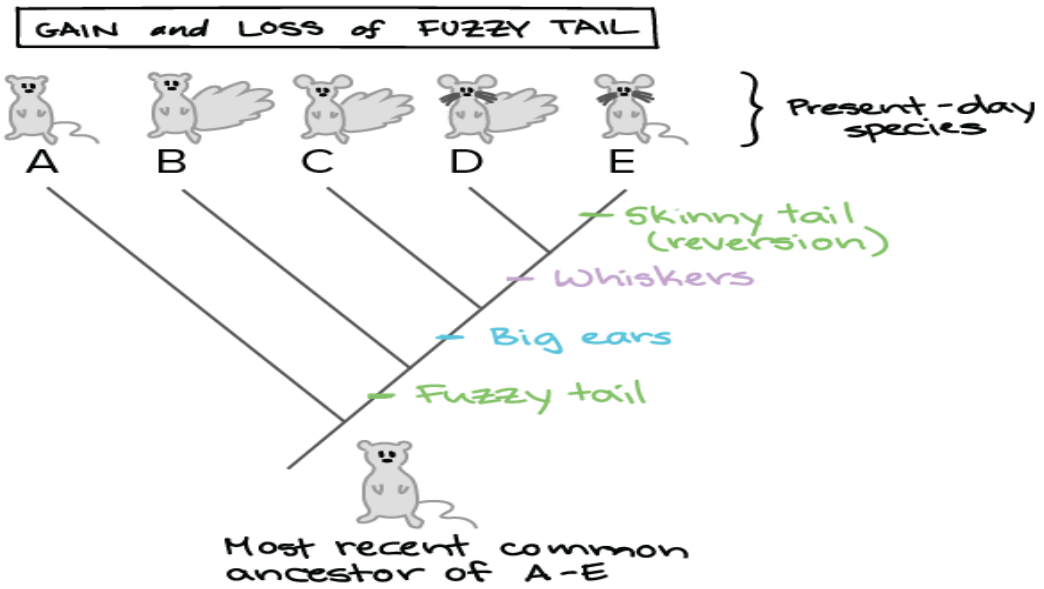
another. Instead, there are some conflicts, where one tree would fit better with the pattern of one trait, while another tree would fit better with the pattern of another trait. In these cases, the researcher can use parsimony to choose the one tree (hypothesis) that fits the data best.

You may be wondering: Why don't the trees all agree with one another, regardless of what characteristics they're built on? After all, the evolution of a group of species did happen in one particular way in the past. The issue is that, when we build a tree, we are reconstructing that evolutionary history from incomplete and sometimes imperfect data. For instance:

- We may not always be able to distinguish features that reflect shared ancestry (**homologous** features) from features that are similar but arose independently (**analogous** features arising by convergent evolution).
-



- Traits can be gained and lost multiple times over the evolutionary history of a species. A species may have a derived trait, but then lose that trait (revert back to the ancestral form) over the course of evolution.



Biologists often use many different characteristics to build phylogenetic trees because of sources of error like these. Even when all of the characteristics are carefully chosen and analysed, there is still the potential for some of them to lead to wrong conclusions (because we don't have complete information about events that happened in the past).

Using molecular data to build trees:

A tool that has revolutionized, and continues to revolutionize, phylogenetic analysis is DNA sequencing. With DNA sequencing, rather than using physical or behavioural features of organisms to build trees, we can instead compare the sequences of their orthologous (evolutionarily related) genes or proteins.

The basic principle of such a comparison is similar to what we went through above: there's an ancestral form of the DNA or protein sequence, and changes may have occurred in it over evolutionary time. However, a gene or protein doesn't just correspond to a single characteristic that exists in two states.

Instead, each nucleotide of a gene or amino acid of a protein can be viewed as a separate feature, one that can flip to multiple states (e.g., A, T, C, or G for a nucleotide) via mutation. So, a gene with 300300300 nucleotides in it could represent 300300300 different features existing in 444 states! The amount of information we get from sequence comparisons—and thus, the resolution we can expect to get in a phylogenetic tree—is much higher than when we're using physical traits. To analyse sequence data and identify the most probable phylogenetic tree, biologists typically use computer programs and statistical algorithms. In general, though, when we compare the sequences of a gene or protein between species:

- A larger number of differences corresponds to *less* related species
- A smaller number of differences corresponds to *more* related species

For example, suppose we compare the beta chain of haemoglobin (the oxygen-carrying protein in blood) between humans and a variety of other species. If we compare the

human and gorilla versions of the protein, we'll find only 111 amino acid difference. If we instead compare the human and dog proteins, we'll find 151515 differences. With human versus chicken, we're up to 454545 amino acid differences, and with human versus lamprey (a jawless fish), we see 127127127 differences. These numbers reflect that, among the species considered, humans are most related to the gorilla and least related to the lamprey.

Probable Questions:

1. Define phylogenetic tree?
2. Differentiate among cladogram, phenogram, dendogram.
3. What is rooted tree and unrooted tree?
4. How phylogenetic trees are constructed?

Suggested Readings:

1. Molecular Cell Biology by Lodish, Fourth Edition.
2. The Cell – A Molecular Approach by Cooper and Hausman, Fourth Edition
3. Principles of Genetics by Snustad and Simmons, Sixth Edition.
4. Molecular Biology of the Cell – by Bruce Alberts
5. Cell and Molecular Biology by Gerald Karp, 7th Edition.
6. Gene cloning and DNA Analysis by T. A. Brown, Sixth Edition.
7. Genetics . Verma and Agarwal.
8. Brown TA. (2010) Gene Cloning and DNA Analysis. 6th edition. Blackwell Publishing, Oxford, U.K.
9. Primrose SB and Twyman RM. (2006) Principles of Gene Manipulation and Genomics, 7th edition. Blackwell Publishing, Oxford, U.K.
10. Sambrook J and Russell D. (2001) Molecular Cloning-A Laboratory Manual. 3rd edition. Cold Spring Harbor Laboratory Pres

UNIT-XIV

Comparative genomics of bacteria, organelles and eukaryotes

Objective: In this unit we will discuss comparative genomics of bacteria, organelles and eukaryotes

Genome of Bacteria:

Bacteria are very small unicellular organisms that do not contain nuclear envelope, mitochondria, endoplasmic reticulum, mitotic apparatus and nucleolus etc., and divide by fission. Bacteria have a rigid cell wall which surrounds their cytoplasmic membrane. Their cytoplasm contains ribosomes. mesosomes and several granular inclusions. About 1/5 of the cell volume is occupied by DNA, the genetic material.

According to their external shape, bacteria are grouped into two main classes:
(1) Cocci and

(2) Bacilli.

Cocci (= berry; Latin and Greek). These are spherical in shape; they show different patterns when their cells are incompletely separated, e.g., (i) Diplococcus: cells in pairs, (ii) Streptococci: cells chains, (iii) Staphylococci: cells in clusters, and (iv) Sarcinae: cells forming tertrads (square) or cubic packets.

Bacilli (= stick; Latin): These bacteria are rod-shaped or cylindrical, and are of different types, such as, (i) Coccobacilli: short elongated cells, (ii) Fusiform bacilli: cells tapered at both ends, (iii) Filamentous bacilli: long threads, (v) Vibrios: curved small bacilli, and (v) Spirilla: long threads curved bacilli.

Bacterial Nuclear Body:

Bacterial cells do not contain a typical nucleus. But Feulgen reaction shows one, two or more discrete nuclear bodies per cell; these are called nucleoids. The bacterial genome is confined to this nucleoid, which is more or less compact structure without any membrane.

When bacterial cells are lysed in the presence of high salt concentration, nucleoids can be recovered intact. The isolated nucleoid may be membrane free or it may be

associated with membrane (mesosome). The constituents of the membrane-free nucleoid are DNA (~ 60%) RNA (-30%) and protein (- 10%) DNA forms 2-3% of the dry weight of a bacterial cell).

Bacterial Chromosomes:

Bacterial chromosome is a double-stranded circular DNA. In general, bacterial DNA ranges from 1100 pm to 1400 μm in length. An E. coli cell contains 4. 2 x 10^6kbp DNA which is about 1.3 mm (1300 μm) in length.

Such a long DNA molecule must be greatly folded to be packaged in a small space of 1.7 x 0.65 μm. The bacterial chromosome is folded into loops or domains which are about 100 in number. A chromosomal domain may be defined as a discrete structural entity within which supercoiling is independent of the other domains.

Thus different domains can maintain different degrees of supercoiling. The DNA chain is coiled on itself to produce supercoiling (Fig. 5.26). The ends of the loops or domains are bound in some way which does not allow rotational events to propagate from one domain to another.

If an endonuclease puts a nick in DNA strand of one domain, this loop becomes larger due to the uncoiling, but the other domains are not affected. Each domain contains about 40 kbp (13 μm) of DNA. The loops are bound by some mechanism that may involve proteins and/or RNA but the mechanism is not clearly understood. In E. coli, a number of proteins have been isolated which have some similarities with the eukaryotic chromosomal proteins. These proteins are HU, IHF (integration host factor). HI (H-NS) and R It is suspected that HU is involved in the nucleoid condensation.

The protein HI probably has effects on gene expression. The amino acid sequence of P has some similarity with the protamine's (DNA of certain sperms is bound with protamine's). However, the functions of the P protein are not known.

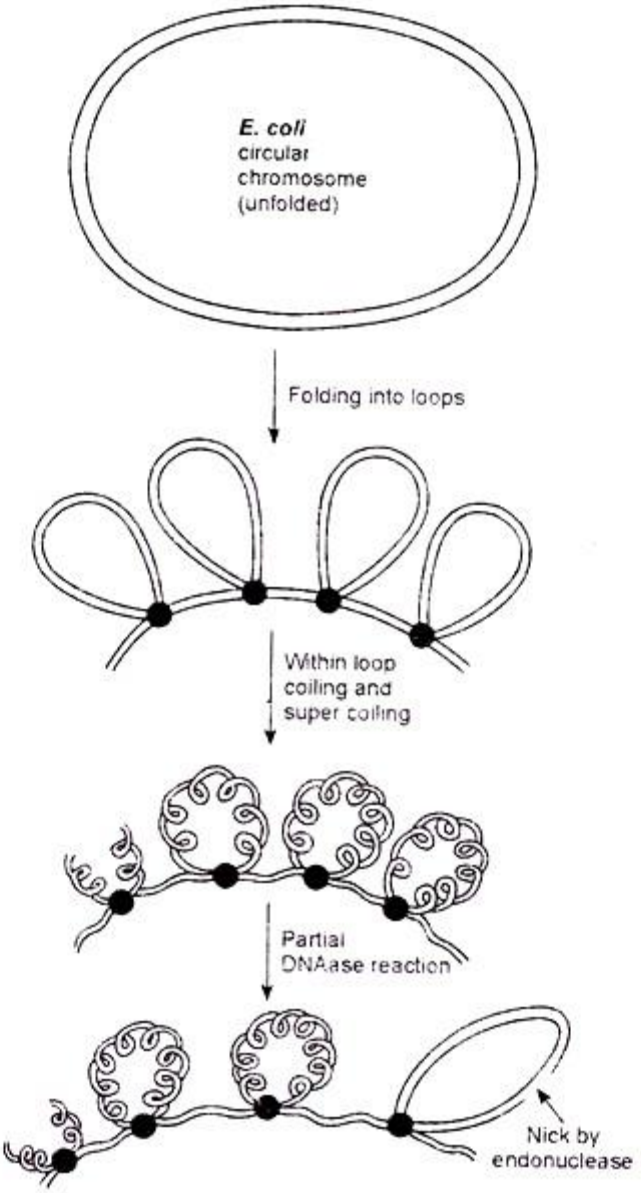


Fig. 5.26. A model of the genome of *E. coli*. The chromosome is folded into ~100 loops which undergo supercoiling. As a result, the chromosome becomes much shorter so that it is able to be packaged into the cell. When an endonuclease makes a cut in one strand of one domain only, that becomes unfolded and enlarged while the other domains remains unaffected.

Replication of Bacterial Chromosome:

Bacterial chromosome is a single replicon. Auto-radiographic studies have shown that it replicates bi-directionally in a semiconservative manner. A theta (0) shaped intermediate is formed during its replication. The rate of replication is 50,000 base pairs per minute which is 25 times faster than that of eukaryotic DNA (2000 bp/minute).

Characteristics of Prokaryotic (Bacterial) Chromosome:

1. Structural organization of bacterial chromosome is simple and is represented mainly by double-stranded DNA molecule. Although there are specific proteins associated with bacterial chromosome (not the histories) that help stabilize its supercoiled domains. Compared to eukaryotic chromosome, one can consider bacterial chromosome to be naked DNA.

2. Bacterial chromosome is covalently closed circular structure consisting of only a single molecule of DNA (with few exceptions such as Borellia burgdorfii and Streptomyces the chromosomes of which are linear).

3. Only one bacterial chromosome occurs per bacterial cell (with few exceptions such as Rhodobacter sphaeroides, a gram-negative phototroph that possesses 2 chromosomes per cell).

4. Bacterial chromosome contains only a single copy of each gene and is therefore genetically haploid.

5. Bacterial chromosomes are shorter and contain lesser number of genes.

6. Bacterial chromosome lies free in the cell cytoplasm without any membrane to separate the chromosome from the cytoplasm., Since the ribosomes also occur free in cell cytoplasm, the process of transcription and translation are not spatially separated.

7. Except few, the bacterial DNAs do not contain introns, the noncoding sequences. As a result, the protein coding genes are not interrupted by introns and synthesize a single mRNA often containing more than one coding region. Each coding region independently synthesizes one or more proteins depending upon the number of operons (Fig. 5.37).

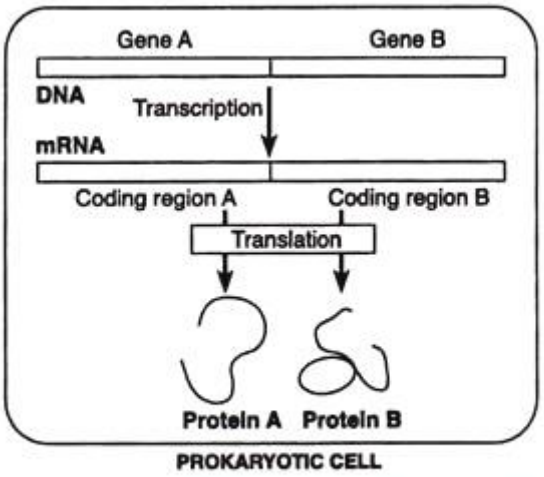


FIG. 5.37. Summarised view of synthesis of protein in absence of introns in bacterial chromosome

Organeller Genome:

The phenomenon of extra-nuclear inheritance based on transmission of visible phenotypes through mitochondria and chloroplasts. Studies in the 70s revealed presence of DNA in these organelles. Both mitochondria and chloroplasts are present only in cells of lower and higher eukaryotic organisms. Detailed studies established that DNA in these organelles is similar to the DNA in prokaryotic bacteria.

The genomes of both mitochondria and chloroplasts code for all of their RNA species and some proteins that are involved in the function of the organelles. The DNA is in the form of a circular duplex molecule, except in some lower eukaryotes in which mitochondrial DNA is linear.

Each organelle contains several copies of the genome, and because there are-multiple organelles per cell, organelle DNA constitutes a repetitive sequence. Mitochondrial DNA (mtDNA) varies enormously in size, whereas chloroplast DNA (ctDNA) ranges in size between 120 and 200 kb.

Mitochondrial genome:

Mitochondrial DNA is a double stranded circular molecule, which is inherited from the mother in all multi-cellular organisms, though some recent evidence suggests that in rare instances mitochondria may also be inherited via a paternal route. Typically, a sperm carries mitochondria in its tail as an energy source for its long journey to the egg. When the sperm attaches to the egg during fertilization, the tail falls off. Consequently, the only mitochondria the new organism usually gets are from the egg its mother provided. There are about 2 to 10 transcripts of the mt-DNA in each mitochondrion. Compared to chromosomes, it is relatively smaller, and contains the genes in a limited number.

The size of mitochondrial genomes varies greatly among different organisms, with the largest found among plants, including that of the plant Arabidopsis, with a genome of 200 kbp in size and 57 protein-encoding genes. The smallest mtDNA genomes include that of the protist Plasmodium falciparum, which has a genome of only 6 kbp and just 2 protein- encoding genomes. Humans and other animals have a mitochondrial genome size of 17 kbp and 13 protein genes.

Mitochondrial DNA consists of 5-10 rings of DNA and appears to carry 16,569 base pairs with 37 genes (13 proteins, 22 t-RNAs and two r-RNA) which are concerned with the production of proteins involved in respiration. Out of the 37 genes, 13 are responsible for making enzymes, involved in oxidative phosphorylation, a process that uses oxygen and sugar to produce adenosine tri-phosphate (Fig. 4.56). The other 14 genes are responsible for making molecules, called transfer RNA (t-RNA) and ribosomal RNA (r-

RNA). In some metazoans, there are about 100 – 10,000 separate copies of mt-DNA present in each cell.
Unlike nuclear DNA, mitochondrial DNA doesn't get shuffled every generation, so it is presumed to change at a slower rate, which is useful for the study of human evolution. Mitochondrial DNA is also used in forensic science as a tool for identifying corpses or body parts and has been implicated in a number of genetic diseases, such as Alzheimer's disease and diabetes. Changes in mt-DNA can cause maternally inherited diseases, which leads to faster aging process and genetic disorders.

Mitochondria convert the potential energy of food molecules into ATP by the Krebs cycle, electron transport and oxidative phosphorylation in presence of oxygen. The energy from food molecules (e.g., glucose) is used to produce NADH and $FADH_2$ molecules, via glycolysis and the Krebs cycle. The protein complexes in the inner membrane (NADH dehydrogenase, cytochrome c reductase, cytochrome c oxidase) use the released energy to pump protons (FT) against a gradient.

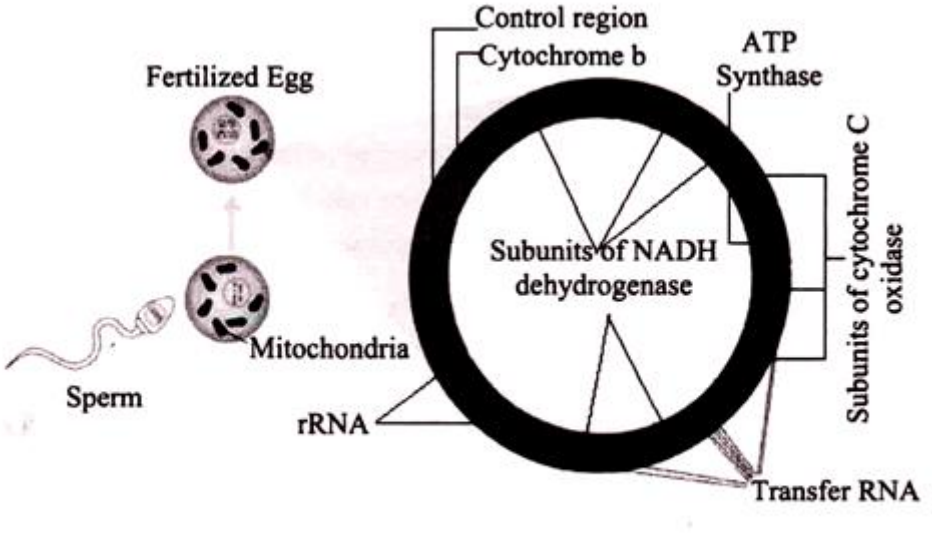


Figure 4.56: Mitochondrial DNA

The mitochondrial genome of higher plants is the largest and complex among the eukaryotes. However, the plant mitochondrial genome does not contain many genes. The complex nature of plant mitochondrial genome might be due to the presence of many chloroplast sequences in the mtDNA of higher plants.

As mitochondria have their own DNA, transcription and translation takes place for the synthesis of relatively small set of polypeptides mainly focused in ATP production. Mitochondria use slightly a different genetic code. Most of the genes for mitochondrial proteins are present in nuclear DNA, translated in cytosol and consequently to mitochondria.

The circular mitochondrial DNA in plants has a buoyant density of about 1.705 gmL^{-1} in caesium chloride. This actual figure corresponds to approximately 47% G + C. There is a

considerable size difference between the mitochondrial DNA among various organisms and ranges from 16 kb in humans to over 2010 kb in muskmelon.
The mitochondrial size in Brassica campestris is 218 kb and contains direct repeat of 2000 kb, whereas in Zea mays, there are five direct repeat and inverted repeat sequence present in 570 kb DNA (Table 5.3). The size of ribosomal RNA (26 S and 18 S) in plant mitochondria is larger than other mitochondria and there is a unique 5 S rRNA in higher plants.

The closely linked 18 S and 5 S genes in maize are separated from the gene for 25 S rRNA. The rRNA sequence in plant mitochondria has several homologies with bacterial and chloroplast rRNA genes. Several evidences have shown that DNA sequences can move from one organelle to another organelle. In maize, 12 kb chloroplast sequences have been shown to be inserted into mitochondrial DNA.

The plant mitochondrial genes that encode proteins for cytochrome C oxidase and the apoprotein of cytochrome b have been sequenced. Understanding of DNA sequence of other corresponding genes led to the conclusion that mitochondria do not use the universal genetic code and uses various other alternatives, for example, in maize mitochondria CGG codes for tryptophan. However, this codon represents for arginine in universal code. Plant mitochondrial genes seem to lack UGA termination codon.

Table 5.3 Mitochondrial DNA from Different plant species

Plant species	Base pairs	Number of molecules per organelle
Brassica sp.	218,000	3 circular
Maize	570,000	7 circular
Muskmelon	2400,000	—
Rice	490,000	—

In maize mitochondria, a gene for the apoprotein of cytochrome b is 1164 base pairs long and codes for protein of 42.9 kD and its amino acid sequence exhibits almost fifty per cent homology with other corresponding proteins in yeast. The presence of introns in maize genes was not evidenced.

The exact nature on the processing of mitochondrial DNA is not known. It is however, believed that 5′ end is not capped and no extensive poly-adenylation. Many plants fail to produce fertile pollen; as a consequence, plants exhibit male sterility, which is controlled by nuclear and mitochondrial genes.

Analysis of maize genome suggests that genes present in the mitochondria determine cytoplasmic male sterility, and its sterility problem can be reversed back by nuclear

restorer (Rf) genes. It is often difficult to identify the sequences responsible for cytoplasmic male sterility (cms) due to the larger size of mitochondrial DNA.

In Brassica napus, there are several cms-associated Open Reading Frame (ORF) associated with male sterility, for example, orf224/atp6 locus linked to male sterility for cms associated mitochondrial genes provide strong evidence that Turf 13 gene is responsible for cms in T-cytoplasm maize. Cytoplasmic male sterility has been studied in maize and has been distinguished into four general types like N, T, C and S.

The normal (N) type gives rise to functional pollen whereas T, C and S are male sterile. The mitochondrial genome of the male sterile S-cytoplasm of maize contains the repeated DNA region R, which contains two chimeric ORF. Protoplast fusion experiment has been conducted to identify cms-associated gene of petunia. Functions of genes associated with cytoplasmic male sterility.

Characteristics of Mitochondrial Genome:

Each human cell contains hundreds of mitochondria each containing multiple copies of mitochondrial DNAs (mtDNA). Mitochondria generate cellular energy through the process of oxidative phosphorylation. As a by-product they produce most of the endogenous toxic reactive oxygen species Mitochondria are also the central regulators of apoptosis or programmed cell death.

These interrelated functional systems involve activities of about 1000 genes distributed in the nuclear genome and the mitochondrial genome. Due to their dependence on the nuclear genome, mitochondria are considered as semi-autonomous.

This has been shown by experiments in which mitochondria and mtDNA could be transferred from one cell to another. The donor cell was enucleated and its mitochondria-containing cytoplast fused with a recipient cell (technique of cybrid transfer).

The genomes of mitochondria show wide variation particularly among plants and protists. Most mitochondrial DNAs (mtDNA) consist of a closed circular double stranded supercoiled DNA molecules located in multiple nucleoid regions (similar to those in bacterial cells); some protists however, have varying lengths or multiple circular molecules of DNA as in the trypanosomes. mtDNA in the protist Amoebidium parasiticum consists of several distinct types of linear molecules with terminal and subterminal repeats. Although most mtDNAs are in the size range of 15 to 60 kb, mtDNA in malarial parasite (Plasmodium spp) is only 6 kb long, while that of rice (Oryza sativa) is 490 kb, and cucurbits 2 Mb. There are about 40 to 50 coding genes in mitochondrial DNA, Plasmodium being an exception with 5 coding genes.

The large size of mitochondrial genomes in plants are due to noncoding inter-genic regions and their content of tandem repeats. Introns are present in many mtDNAs, and in some unusual cases, the genes are split into as many as 8 regions that are dispersed in the genome, and located on both strands of the DNA. Transcription takes place separately in portions of the split genes producing discrete pieces of RNA that are held together by base pairing of complementary sequences.

The mtDNA contains information for a number of mitochondrial compounds such as tRNAs, rRNA, and some of the polypeptide subunits of the proteins cytochrome oxidase, NADH- dehydrogenase and ATPase. Most of the other proteins found in mitochondria are encoded by the nuclear genome and transported into mitochondria. These include DNA polymerase and other proteins for mtDNA replication, RNA polymerase and other proteins for transcription, ribosomal proteins for ribosome assembly, protein factors for translation, and the aminoacyl-tRNA synthetases.

The mitochondrial oxidative phosphorylation complexes are composed of multiple polypeptides, mostly encoded by the nuclear DNA (nDNA). However, 13 polypeptides are encoded by mtDNA. The mtDNA also codes for 12S and 16S rRNAs and 22 tRNAs required for mitochondrial protein synthesis. The mtDNA also contains a control region consisting of approximately 1000 base pairs constituting the promoter region and the origin of replication.

The mRNAs synthesised within the mitochondria remain in the organelle and are translated by mitochondrial ribosomes that are assembled within mitochondria. Mitochondrial ribosomes have two subunits. Mitochondria in human cells have 60S ribosomes consisting of a 45S and a 35S subunit. There are only two rRNAs in mitochondrial ribosomes of most organisms, that is, 16S rRNA in large subunit and 12S rRNA in small subunit of most animal ribosomes. There is usually one gene for each rRNA in a mitochondrial genome. The proteins in mitochondrial ribosomes are encoded by the nuclear genome and transported into mitochondria from the cytoplasm.

Mitochondrial ribosomes are sensitive to most of the inhibitors of bacterial ribosome function such as streptomycin, neomycin and chloramphenicol. For protein synthesis, mitochondria of most organisms use a genetic code that shows differences from the universal genetic code. Only plant mitochondria use the universal nuclear genetic code.

Transcription of mammalian mtDNA is unusual in that each strand is transcribed into a single RNA molecule that is then cut into smaller pieces. In the large RNA transcripts that are produced, most of the genes encoding the rRNAs and the mRNAs are separated by tRNA gene. The tRNAs in the transcript are recognised by specific enzymes and are cut out, leaving only the mRNAs and the rRNAs. A poly (A) tail is then added to the 3'end of each mRNA and CCA is added to the 3'end of each tRNA. There are no 5′ caps in mitochondrial mRNAs.

Mitochondrial DNA replication is semi-conservative and uses DNA polymerases that are specific to the mitochondria. The mtDNA replicates throughout the cell cycle, independently of nuclear DNA synthesis which takes place in S phase of cell cycle. Observations on mtDNA replication in animal mitochondria in vivo have resulted in a model referred to as the displacement loop (D loop) model as follows (Fig. 17.6).

The two strands of mtDNA in most animals have different densities because the bases are not equally distributed on both strands, called H (heavy) and L (light) strands. The synthesis of a new H strand starts at the replication origin for the H strand and forms a D-loop structure (Fig. 17.6). As the new H strand extends to about halfway around the molecule, initiation of synthesis of a new L strand takes place at a second replication origin. Synthesis continues until both strands are completed. Finally, each circular DNA assumes a supercoiled form.

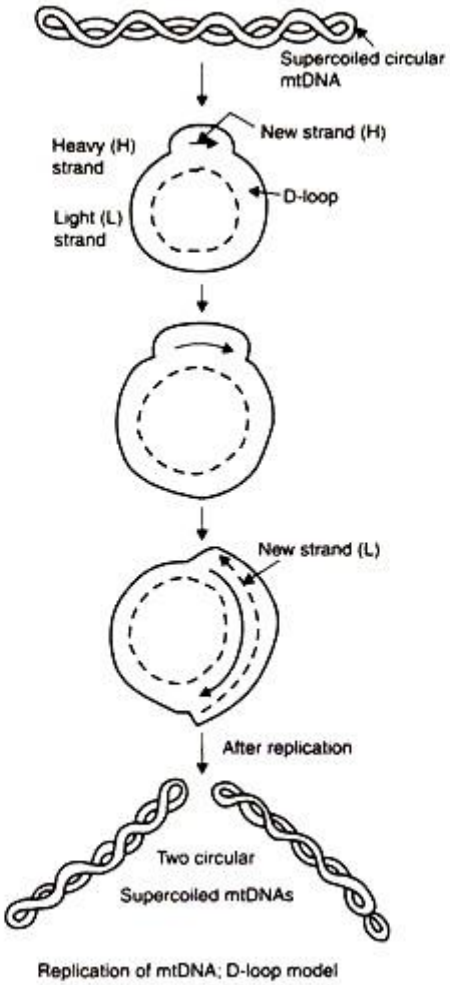


Fig. 17.6 Model for mitochondrial DNA replication by formation of a D-loop structure.

The mtDNA is maternally inherited and has a very high mutation rate. When a new mtDNA mutation occurs in a cell, a mixed intracellular population of mtDNAs is generated, known as heteroplasmy. During replication in a heteroplasmic cell, the mutant and normal molecules are randomly distributed into daughter cells.

When the percentage of mutant mtDNAs increases, the mitochondrial energy producing capacity declines, production of toxic reactive oxygen species increases, and cells become more prone for apoptosis. The result is mitochondrial dysfunction. Tissues most sensitive to mitochondrial dysfunction are brain, heart, kidney and skeletal muscle.

The mtDNA mutations are associated with a variety of neuromuscular disease symptoms, including various ophthalmological symptoms, muscle degeneration, cardiovascular diseases, diabetes mellitus, renal function and dementias.

The mtDNA diseases can be caused either by base substitutions or rearrangement mutation. Base substitution mutations can either alter protein (missense mutation) or rRNAs and tRNAs (protein synthesis mutations). Rearrangement mutations generally delete at least one tRNA and thus cause protein synthesis defects. Missense mutations are associated with myopathy, optic atrophy, dystonia and Leigh's syndrome. Base substitution mutations in protein synthesizing genes have been associated with a wide spectrum of neuromuscular diseases, and the more severe typically include mitochondrial myopathy.

Mitochondrial diseases are also associated with a number of different nuclear DNA mutations. Mutations in the RNA component of the mitochondrial RNAse have been implicated in metaphyseal chondrodysplasia or cartilage hair hypoplasia which is an autosomal recessive disorder resulting from mutation in nuclear chromosome 9 short arm position (9p13).

Chloroplast DNA:

Chloroplasts are present in green plants and photosynthetic protists. ctDNA sequence studied in a number of plants indicates uniformity in size and organisation. The differences in size are due mainly to the differences in lengths of introns and inter-genic regions as well as the number of genes. All cp DNAs contain a significant proportion of noncoding DNA sequences.

The ctDNA is double stranded circular, and devoid of histones and other proteins. In many cases, the GC content of cpDNA differs from that of nuclear DNA and mitochondrial DNA. Complete cpDNA sequences have been determined in tobacco (155, 844 bp) and rice (135, 42 bp).

Multiple copies of cpDNA are present in the nucleoid region of each chloroplast. In the green alga Chlamydomonas, one chloroplast contains 500 to 1500 cpDNA molecules. Chloroplasts divide by growing and then dividing into two daughter chloroplasts. The proportion of introns in chloroplast DNA could be high, 38% in Euglena. Among the expressed genes in chloroplast genome, 70 to 90% of the genes encode proteins including those involved in photosynthesis, four genes code for rRNAs (one each for 16S, 23S, 4.5S and 5S), and about 30 genes encode tRNAs.

Chloroplast genome also contains genes for some of the proteins required for transcription and translation of the encoded genes, and most importantly, genes for photosynthesis. Most of the proteins in chloroplasts are encoded by the nuclear genes. The mRNA transcripts of the chloroplast genes are translated according to the standard genetic code.

However, the primary structures of several RNA transcripts are found to go through editing consisting of C to U transitions, that cause mRNA sequence to deviate from the sequence in the corresponding gene. Editing makes it difficult to convert chloroplast nucleotide sequences into amino acid sequences of the corresponding protein. Most of the cpDNAs studied share a common feature, that is, a 10 to 24 kb segment present in two identical copies as an inverted repeat. The cpDNA also contains two copies of each of the rRNA genes which are located in these two identical repeat sequences in an inverted orientation.

Other genes that are found in the repeated sequence are therefore, also duplicated in the chloroplast genome. The location of these repeats defines a short single copy (SSC) region and a long single copy (LSC) region in chloroplast genome. Chloroplast protein synthesis uses organelle-specific 70S ribosomes consisting of 50S and 30S subunits. The 50S subunit contains one copy each of 23S, 5S and 4.5S rRNAs, while the 30S subunit contains one copy of a 16S rRNA. Among the ribosomal proteins, some are encoded by the nuclear DNA, some by the chloroplast genome. About 100 open reading frames (ORFs), putative protein coding genes, have been identified by computer analysis. Protein synthesis is similar to that in prokaryotes.

Eukaryotic Genome:

In eukaryotic organisms — plants, animals and fungi — the major portion of DNA is present in the chromosomes which are well-organised structures and quite different from the prokaryotic counterparts. Besides the chromosomes, mitochondria of both plants and animals and the chloroplasts of green plants also contain DNA. Interesting is the fact that the organization and the nature of the DNA of these cell- organelles are similar to those of bacterial DNA. In both these organelles, the DNA is a covalently closed circular molecule. In eukaryotic cells, the chromosomes are present in a distinct, double-membrane bound structure, the nucleus which occupies on the average about 10% of the cell volume. The membrane is continuous with the endoplasmic reticulum and is provided with pores. The number of chromosome is variable, but fixed for a biological species. Chromosomes change in their physical characteristics during cell division. All these features are absent in the prokaryotic cells.

An individual eukaryotic chromosome contains a single enormously large linear ds-DNA molecule. For example, a diploid human cell containing 46 chromosomes (22 pairs of autosomes and one pair of sex chromosomes) has a total of 6×10^9 base-pairs.

The length of the DNA molecules of individual human chromosomes varies from 1.5 cm to 8.7 cm. These large molecules have to be packed into chromosomes, generally measuring a few microns in length and breadth. This is accomplished by binding DNA to proteins. The protein-DNA complex of eukaryotic cells is known as chromatin.

The DNA-binding proteins are distinguished into two main types — the histones and non-histone proteins. The histones are basic proteins, rich in basic amino acids, like lysine and arginine. Histones have large amount of positive charges and can bind tightly the negatively charged DNA molecules. These binding results in the formation of the characteristic structural units called the nucleosomes. The long ds-DNA molecule of each chromosome is folded in a very orderly way around the histones to form the nucleosomes. The nucleosomes are bead-like structures connected to each other by linker DNA. Each nucleosome consists of a histone core composed of 8 subunits (octamer) of 4 different histones — H2A, H2B, H3 and H4 with two molecules of each. The protein core is wrapped by two turns of ds-DNA molecule to form a nucleosome.
The DNA molecule runs as a continuous thread from one nucleosome to another. The intervening portion of DNA between two nucleosomes is the linker. Width of a nucleosome is 11 mm and, on the average, nucleosomes are repeated at intervals of 200 nucleotide pairs of DNA. The length of the linker between two nucleosome is variable.

These features of eukaryotic DNA are shown diagrammatically in Fig. 9.10:

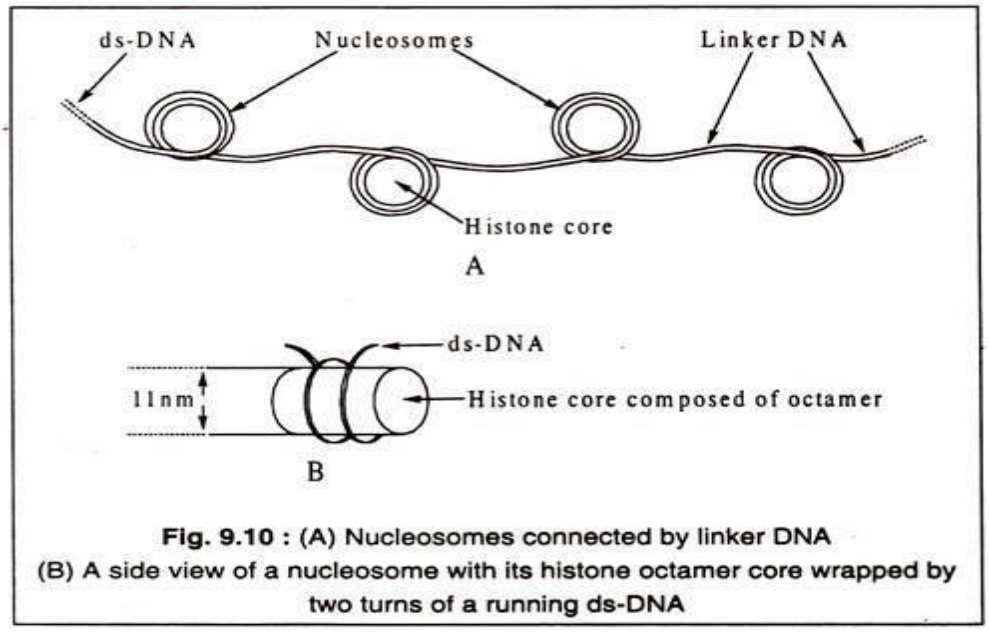


Fig. 9.10 : (A) Nucleosomes connected by linker DNA
(B) A side view of a nucleosome with its histone octamer core wrapped by two turns of a running ds-DNA

The nucleosomes are basic structures from which chromatin is made. They are further organized into closely packed 30 nm fibres of chromatin. These fibres are visible under high resolution electron microscope.

The 30 nm fibres shown in Fig. 9.11, are organized into higher orders of increasing complexity, like 300 nm fibres and 700 nm fibres to produce chromosomes:

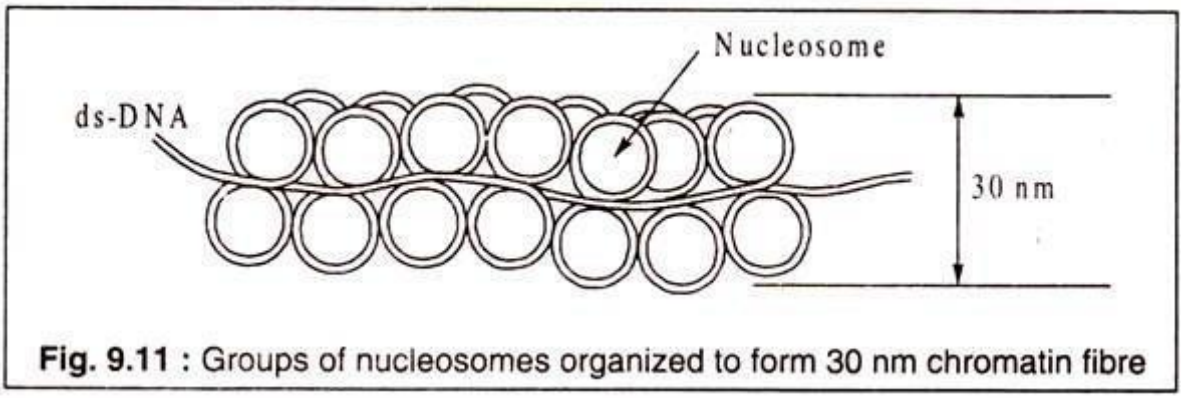


Fig. 9.11 : Groups of nucleosomes organized to form 30 nm chromatin fibre

The eukaryotic chromosomes are characterized by the presence of three types of specialized nucleotide sequences in their DNA. These sequences serve as origin of replication, as centromere which helps the daughter chromosomes to move to opposite poles, and telomere which has a number of repeating sequences functioning as template for RNA-primer in DNA synthesis.

Nucleosome Model:

The nucleosome hypothesis proposed by Roger Kornberg in 1974 was a paradigm shift for understanding eukaryotic gene expression. The assembly of DNA into chromatin involves a range of events, beginning with the formation of the basic unit, the nucleosome, and ultimately giving rise to a complex organization of specific domains within the nucleus. The first step is the assembly of the DNA with a newly synthesized tetramer (H3-H4), are specifically modified (e.g. H4 is acetylated at Lys5 and Lysl2 (H3- H4)), to form a subnucleosomal particle, which is followed by the addition of two H2A-H2B dimers. This produces a nucleosomal core particle consisting of 146 base pairs of DNA bind around the histone octamer. This core particle and the linker DNA together form the nucleosome (Figs 4.38 and 4.39).

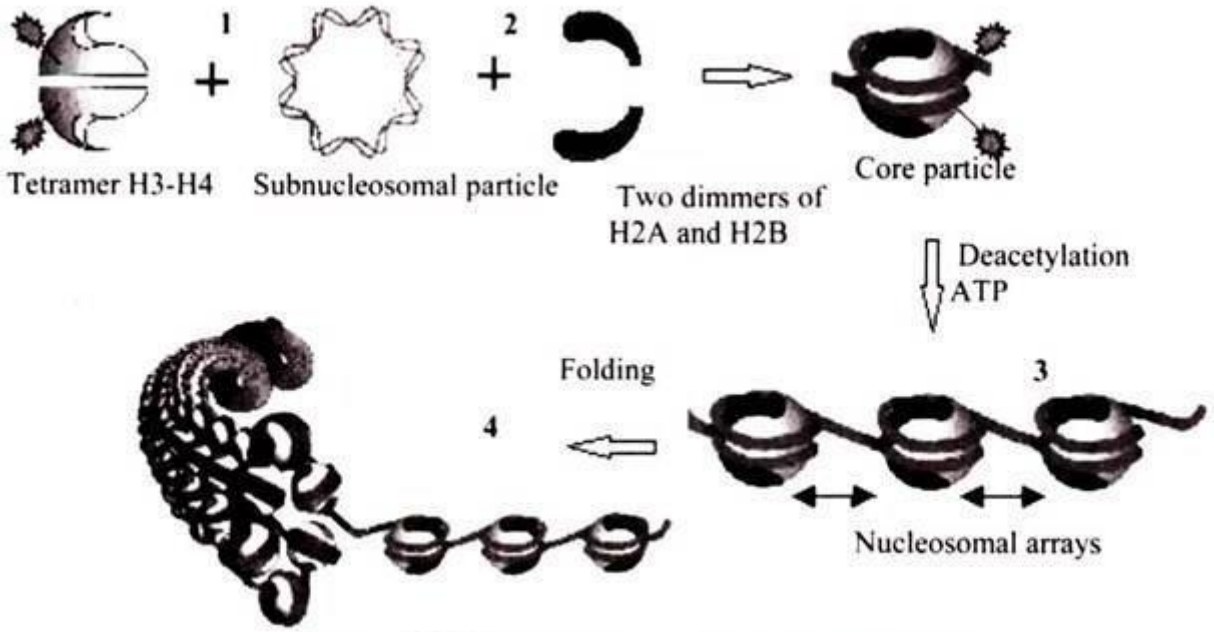


Figure 4.38: The assembly of DNA into chromatin

The next step is the maturation step that requires ATP to establish regular spacing of the nucleosome cores to form the nucleo-filament. During this step the newly incorporated histones are de-acetylated. Next the incorporation of linker histones is accompanied by folding of the nucleo-filament into the 30 nm fibre, the structure of which remains to be elucidated. Two principal models exist- the solenoid model and the zig- zag. Finally, further successive folding events lead to a high level of organization and specific domains in the nucleus.

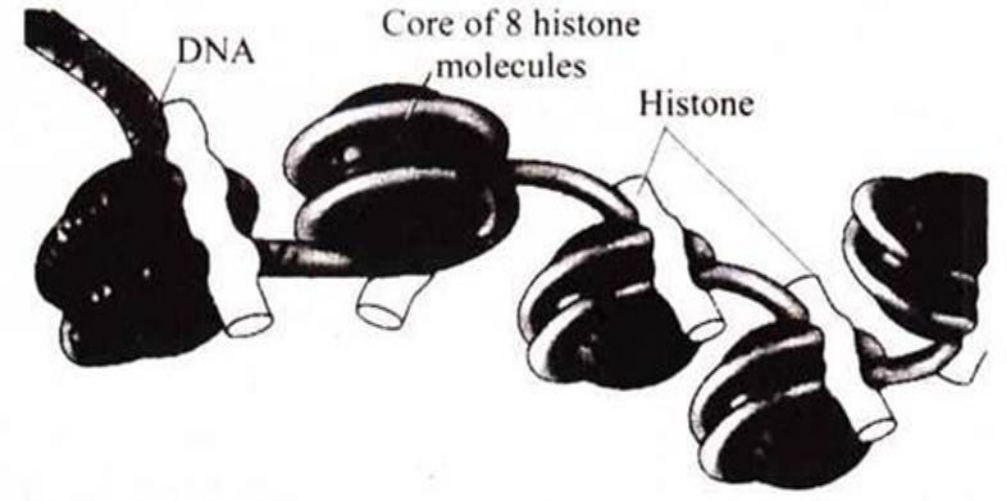


Figure 4.39: Nucleosome model of chromatin assembly

Two molecules of each of the four core histone proteins form the histone octamer via formation of one tetramer of H3 and H4 and two dimers of H2A and H2B. Each of these entities is held together by a so called hand-shake motif of protein structure, forms a "beads on a string" like structure. H1 is involved with the packing of the "beads on a string" substructures into a high order structure.

H1 is present in half the amount of the other four histones. This is because unlike the other histones, H1 does not make up the nucleosome "bead". Instead, it sits on top of the structure, keeping in place the DNA that has wrapped around the nucleosome. Specifically, the H1 protein binds to the "linker DNA" (approximately 80 nucleotides in length) region between the histone beads, helping stabilize the zig- zagged 30 nm chromatin fiber. The nucleosome together with histone HI is called a chromatosome. Chromatosomes are held together by the continuous DNA strand, thus forming linker DNA of 30-50 base pairs in length.

Genome Size and Complexity:

A genome is all the genetic information of an organism. It consists of DNA (or RNA in RNA viruses). The genome includes the genes (the coding regions) and the noncoding DNA, aswell as the genetic material of extrachromosomal origins like mitochondria and chloroplasts. Many bacterial species have only one chromosome per cell and, in nearly all cases; each chromosome contains only one copy of each gene. A very few genes, such as those for rRNAs, are repeated several times. Genes and regulatory sequences account for almost all the DNA in bacteria. Moreover, almost every gene is precisely collinear with the amino acid sequence (or RNA sequence) for which it codes. The organization of genes in eukaryotic DNA is structurally and functionally much more complex. The study of eukaryotic chromosome structure, and more recently the sequencing of entire eukaryotic genomes, has yielded many surprises. The genomes of most eukaryotes are larger and more complex than those of prokaryotes.This larger size of eukaryotic genomes is not inherently surprising, since one would expect to find more genes in organisms that are more complex. Many, if not most, eukaryotic genes have a distinctive and puzzling structural feature: their nucleotide sequences contain one or more intervening segments of DNA that do not code for the amino acid sequence of the polypeptide product. These nontranslated inserts interrupt the otherwise colinear relationship between the nucleotide sequence of the gene and the amino acid sequence of the polypeptide it encodes. Such nontranslated DNA segments in genes are called intervening sequences or introns, and the coding segments are called exons (Figure 1).

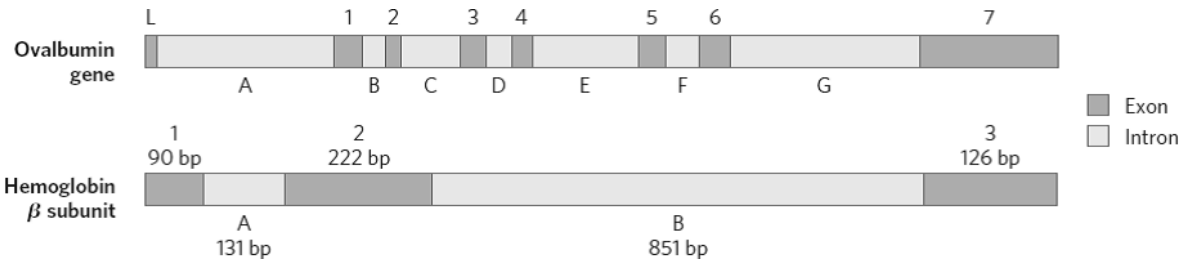


Figure 1: Structure of two eukaryotic genes.

However, the genome size of many eukaryotes does not appear to be related to genetic complexity (Figure 2 and 3). For example, the genomes of salamanders and lilies contain more than ten times the amount of DNA that is in the human genome, yet these organisms are clearly not ten times more complex than humans. This apparent paradox was resolved by the discovery that the genomes of most eukaryotic cells contain not only functional genes butalso large amounts of DNA sequences that do not code for proteins. The difference in the sizes of the salamander and human genomes thus reflects larger amounts of noncoding DNA, rather than more genes, in the genome of the salamander. The presence of large amounts of noncoding sequences is a general property of the genomes of complex eukaryotes. Thus the thousand fold greater size of the human genome compared to that of E. coli is not due solely to a larger number of human genes. The human genome is thought to contain 20,000-25,000 genes- only about 5 times more than E.coli has. Much of the complexity of eukaryotic genomes thus results from the abundance of several different types of noncoding sequences, which constitute most of the DNA of higher eukaryotic cells. Thus several kinds of noncoding DNA contribute to the genomic complexity of higher eukaryotes. The lack of precise correlation between the complexity of an organism and the size of its genome was looked on as a bit of a puzzle, the so called **C-value paradox**. In fact the answer is quite simple: space is saved in the genomes of less-complex organisms because the genes are more closely packed together (Figure 2 and 3).

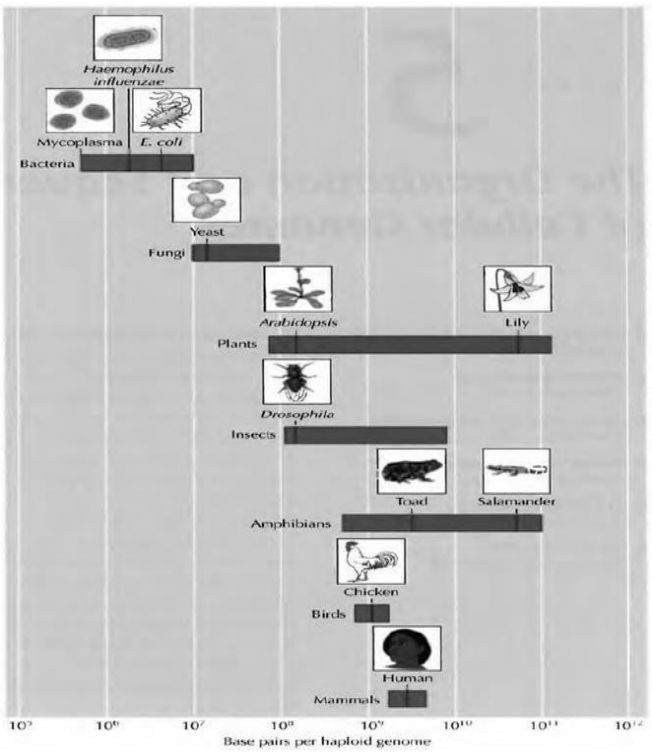


Figure 2: Genome size of representative groups of organisms.

In bacterial genomes, most of the DNA encodes proteins. For example, the genome of E. coli is approximately 4.6 x 10^6 base pairs long and contains about 4000 genes, with nearly 90% of the DNA used as protein-coding sequence. The yeast genome, which consists of 12 x 10^6 base pairs, is about 2.5 times the size of the genome of E. coli, but is still extremely compact. Only 4% of the genes of Saccharomyces cerevisiae contain introns, and these usually have only a single small intron near the start of the coding sequence. Approximately 70% of the yeast genome is used as protein-coding sequence, specifying a total of about 6000 proteins. The relatively simple animal genomes of C. elegans and Drosophila are about 10 times largerthan the yeast genome, but contain only 2-3 times more genes. Instead, these simple animal genomes contain more introns and more repetitive sequence, so that protein-coding sequences correspond to only about 25% of the C. elegans genome and about 13% of the genome of Drosophila. The genome of the model plant Arabidopsis contains a similar number of genes, with approximately 26% of the genome corresponding to protein-coding sequence. Thegenomes of higher animals (such as humans) are approximately 20-30 times larger than those of C. elegans and Drosophila. However, a major surprise from deciphering the humangenome sequence was the discovery that the human genome contains only 20,000 to 25,000 gene (Figure 2 and 3). It appears that only about 1.2% of the human genome consists of protein-coding sequence. Approximately, 20% of the genome consists of introns, and more than 60% is composed of

various types of repetitive and duplicated DNA sequences, with the remainder corresponding to pseudogenes, to nonrepetitive spacer sequences between genes, and to exon sequences that are present at the 5' and 3' ends of mRNAs but are not translated into protein. The increased size of the genomes of higher eukaryotes is thus due far more to the presence of large amounts of repetitive sequences and introns than to an increased numberof genes (Figure 1 and 2).

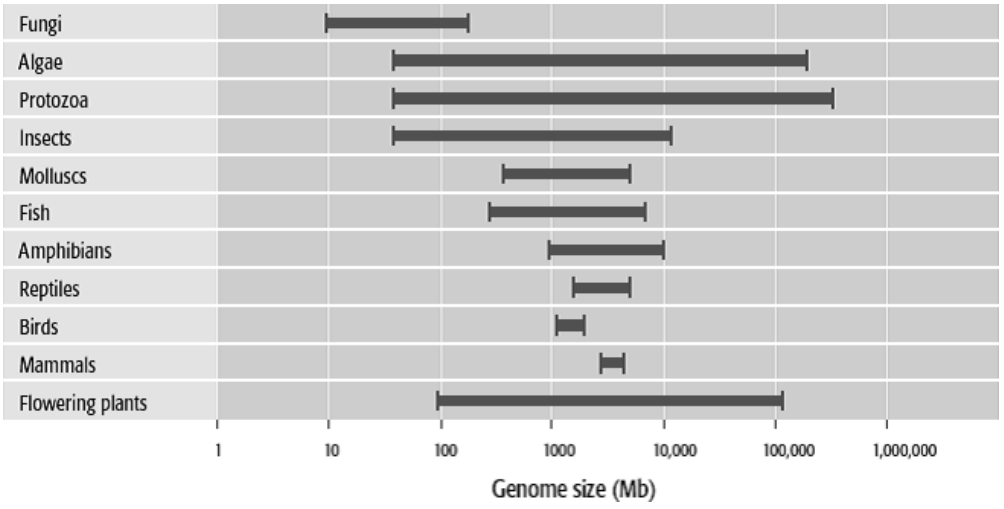


Figure 3: Genome size of eukaryotes.

Complexity of eukaryotic DNA sequences and repetitive DNA sequence:

The term **sequence complexity** refers to the number of times a particular base sequence appears throughout the genome. Unique or nonrepetitive sequences are those found once or a few times within the genome. Structural genes are typically unique sequences of DNA. The vast majority of proteins in eukaryotic cells are encoded by genes present in one or a few copies. In the case of humans, unique sequences make up roughly 41% of the entire genome. Apart from unique DNA sequences there are repetitive DNAs, that is, sequences that are similar or identical to sequences elsewhere in the genome. Most large genomes are filled with repetitive sequences; for example, nearly half of the human genome is covered by repeats, many of which have been known about for decades. Although some repeats appear to be non-functional, others have played a part in human evolution at times creating novel functions, but also acting as independent, 'selfish' sequence elements. Repeats arise from a variety of biological mechanisms that result in extra copies of a sequence being produced and inserted into the genome. Repeats come in all shapes and sizes: they can be widely interspersed

repeats, tandem repeats or nested repeats, they may comprise just two copies or millions of copies, and they can range in size from 1–2 bases (mono-and dinucleotide repeats) to millionsof bases.

One approach that has proven useful in understanding genome complexity has come from renaturation studies. These kinds of experiments were first carried out by Roy Britten and David Kohne in 1968. In a renaturation study, the DNA is broken up into pieces containing several hundred base pairs. The double-stranded DNA is then denatured (separated) into single-stranded pieces by heat treatment. When the temperature is lowered, the pieces of DNA that are complementary can reassociate, or renature, with each other to form double-stranded molecules.

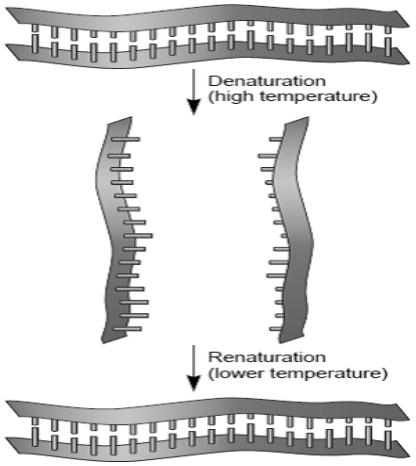


Figure 4: Denaturation and renaturation of DNA strands

The rate of renaturation of complementary DNA strands provides a way to distinguish between unique, moderately repetitive, and highly repetitive sequences. For a given category of DNA sequences, the renaturation rate depends on the concentration of its complementary partner. Highly repetitive DNA sequences renature much faster because many copies of the complementary sequences are present. In contrast, unique sequences, such as those found within most genes, take longer to renature because of the added time it takes for the unique sequences to find each other (Figure 5).

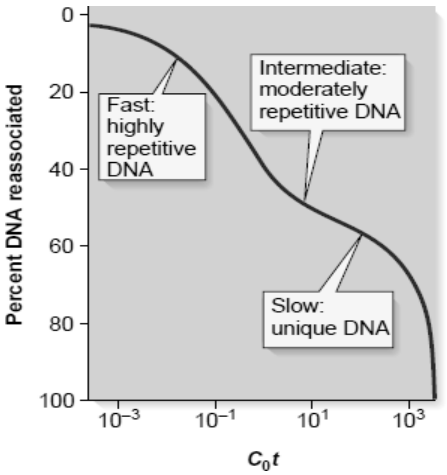


Figure 5: Renaturation of human DNA sequence comple

Highly repeated DNA sequences:

The highly repeated fraction (also called **tandem repeats**) constitutes anywhere from about 1to 10 percent of the total DNA (Figure 6). These sequences are typically short (a few hundred nucleotides at their longest) and present in clusters in which the given sequence repeats itself over and over again without interruption. A sequence arranged in this end-to-end manner is said to be present in tandem. Approximately 3% of the human genome consists of **highly repetitive** sequences, also referred to as **simplesequence DNA** or **simple sequence repeats (SSR)**. These short sequences, generally less than 10 bp long, are sometimes repeated millions of times per cell. The simple sequence DNA has also been called **satellite DNA**, so named because its unusual base composition often causes it to migrate as "satellite" bands (separated from the rest of the DNA) when fragmented cellular DNA samples are centrifuged in a cesium chloride density gradient (Figure 7).

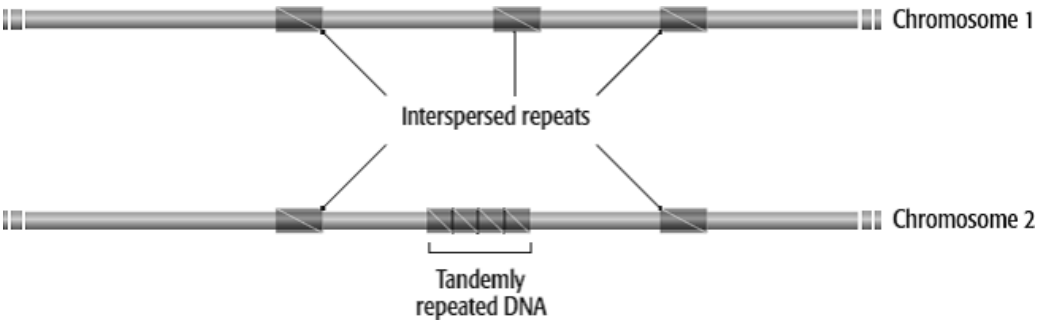


Figure 6: Two different types of repetitive sequence.

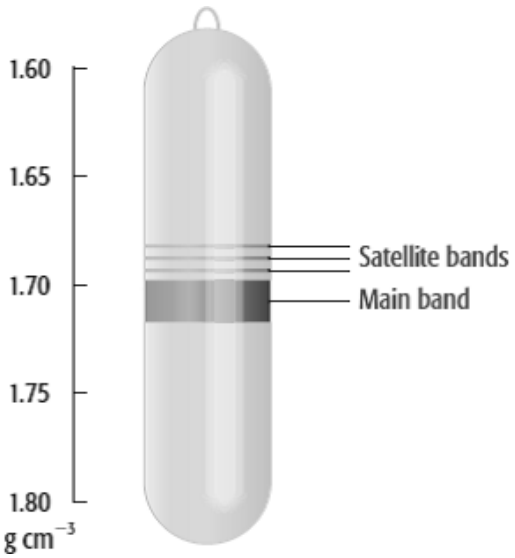


Figure 7: Satellite DNA from human genome

Highly repeated sequences fall into several overlapping categories, including satellite DNAs, minisatellite DNAs, and microsatellite DNAs. The localization of satellite DNAs are found mostly within the centromeres and telomeres of chromosomes.

Satellite DNAs: Satellite DNAs consist of short sequences (about five to a few hundred base pairs in length) that form very large linear arrays, each containing up to several million base pairs of DNA. In many species, the base composition of these DNA segments is sufficiently different from the bulk of the DNA that fragments containing the sequence can be separated into a distinct "satellite" band during density gradient centrifugation (Figure 7). Satellite DNAs tend to evolve very rapidly, causing the sequences of these genomic elements to vary even between closely related species.

Minisatellite DNAs: Minisatellite sequences range from about 10 to 100 base pairs in length and are found in sizeable clusters containing as many as 3000 repeats. Thus, minisatellite sequences occupy considerably shorter stretches of the genome than do satellite sequences. Minisatellites tend to be unstable, and the number of copies of a particular sequence often increases or decreases from one generation to the next, most likely as the result of unequal crossing over. Consequently, the length of a particular minisatellite locus is highly variable in the population, even among members of the same family. Because they are so variable (or polymorphic) in length, minisatellite sequences form the basis for the technique of DNA fingerprinting, which is used to identify individuals in criminal or paternity. **Microsatellite DNAs:** Microsatellites are the shortest sequences (1 to 9 base pairs long) and are typically present in small clusters of about 10 to 40 base pairs in length, which

are scattered quite evenly through the genome. DNA replicating enzymes have trouble in copying regions of the genome that contain these small, repetitive sequences, which causes these stretches of DNA to change in length through the generations. Because of their variable lengths within the population, microsatellite DNAs have been used to analyze the relationships between different human populations.

Moderately repeated DNA sequence

The moderately repeated fraction of the genomes of plants and animals can vary from about 20 to more than 80 percent of the total DNA, depending on the organism. This fraction includes sequences that are repeated within the genome anywhere from a few times to tens of thousands of times (Figure 5). Included in the moderately repeated DNA fraction are some sequences that code for known gene products, either RNAs (such as rRNAs) or proteins (including histones), but the bulk of this DNA fraction lacks a coding function. Rather than occurring as clusters of tandem sequences, these noncoding elements are scattered (i.e., interspersed) throughout the genome (Figure 6). Most of these repeated sequences can be grouped into two classes that are referred to as SINEs (short interspersed elements) or LINEs (long interspersed elements). SINEs and LINEs sequences are discussed later.

Nonrepeated DNA sequence

As initially predicted by Mendel, classical studies on the inheritance patterns of visible traits led geneticists to conclude that each gene was present in one copy per single (haploid) set of chromosomes. When denatured eukaryotic DNA is allowed to reanneal, a significant fraction of the fragments are very slow to find partners, so slow in fact that they are presumed to be present in a single copy per genome (Figure 5). This fraction comprises the non-repeated (or single-copy) DNA sequences, which includes the genes that exhibit Mendelian patterns of inheritance. Because they are present in a single copy in the genome, non-repeated sequences localize to a particular site on a particular chromosome. Included within the non-repeated fraction are the DNA sequences that code for virtually all proteins other than histones. Even though these sequences are not present in multiple copies, genes that code for polypeptides are usually members of a family of related genes. This is true for the globins, actins, myosins, collagens, tubulins, integrins, and most other proteins in a eukaryotic cell. Each member of a multigene family is encoded by a different but related sequence.

Probable questions:

1. What is C-Value paradox?
2. What is Satellite DNA?
3. Distinguish between microsatellite and minisatellite DNA?
4. What is moderately repeated and non repeated DNA sequence?
5. The yeast genome is 0.004 times the size of the human genome and yet it contains approximately 0.2 times fewer genes. Give explanation.
6. What are introns and exons?
7. What differences in gene distribution and repetitive DNA content are seen when yeast and human chromosomes are compared?
8. Describe nucleosome model of genome organization with suitable diagram.
9. Describe different types of Histone proteins.
10. Write three characteristics of bacterial genome.
11. Write three characteristics of mitochondrial genome.
12. Write three characteristics of chloroplast genome.

Suggested Readings/References:

1. James D. Watson, Tania A. Baker, Stephen P. Bell-Molecular Biology of Gene
2. Robert J. Brooker-Genetics Analysis and Principles
3. Benjamin Lewin-Genes IX
4. Harvey F Lodish et al-Molecular Cell Biology
5. Gerald Karp-Cell and Molecular Biology
6. Geoffrey M. Cooper, Robert E. Hausman-The Cell:A Molecular Approach
7. T.A. Brown-Genome.

DISCLAIMER: This Self Learning Material (SLM) has been compiled from various authentic books, Journals articles, e-journals and other web sources.