

# **Handouts**

**BIO301-Essentials of Genetics**

**Virtual University of Pakistan**

## Lesson 51

**Monosomy:** Loss of a single chromosome

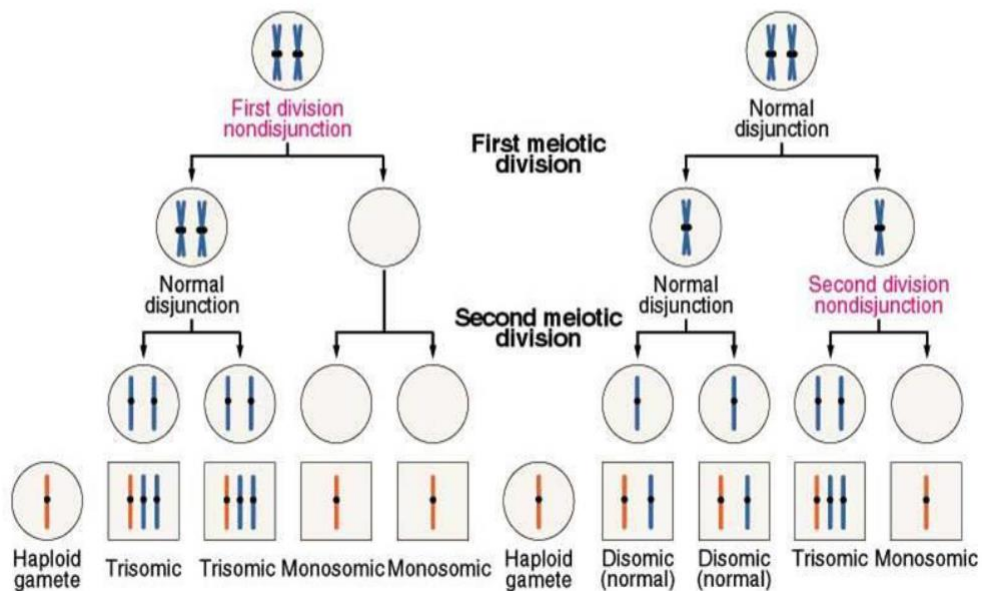
Monosomy of autosomes is lethal

Turner syndrome XO i.e. loss of sex chromosome.

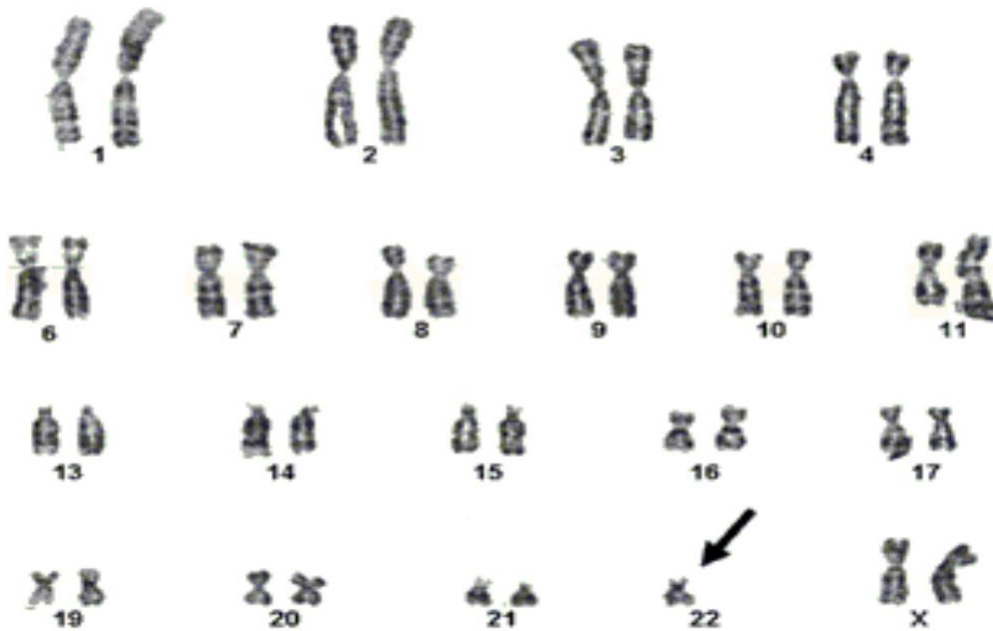
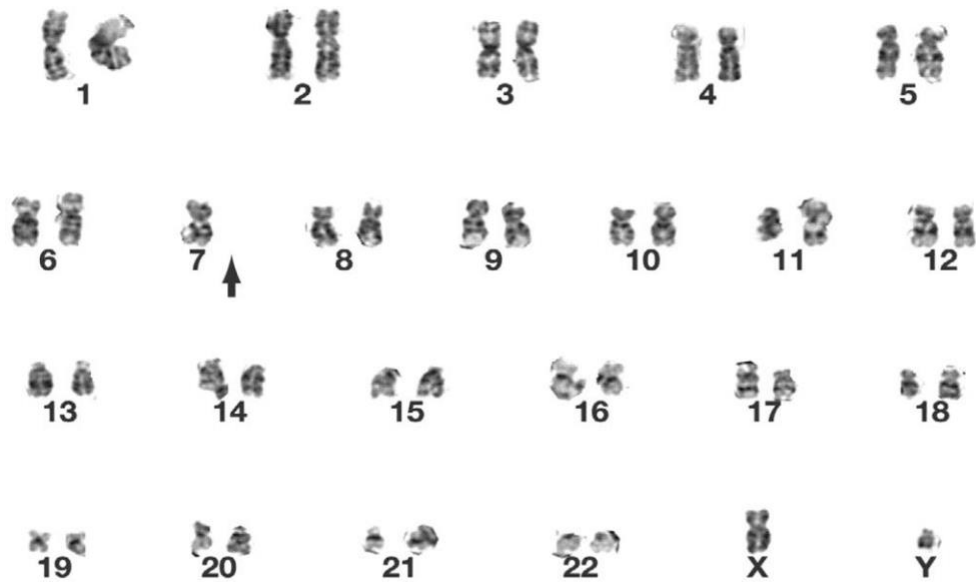
**Causes of monosomy:** followings are the causes of monosomy

- Non disjunction – one gamete receives two copies of homologous chromosomes and other will have no copy.
- Loss of chromosome. As it move towards pole of cell during anaphase.

### Fruit fly



## Monosomy – Chromosome 7



## Monosomy – chromosome 22

### Diseases due to Monosomy

- Turner syndrome
- Cri du chat syndrome

## Lesson 52

**Polyploidy:** Multiples of haploid number

Triploidy or tetraploidy etc

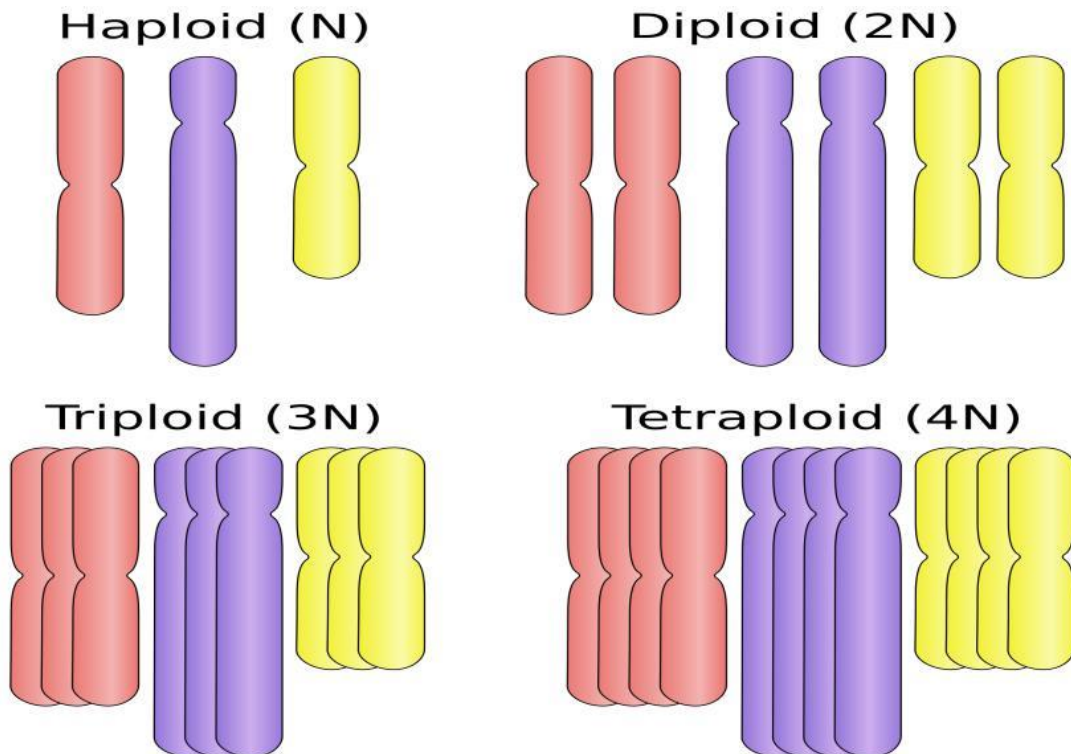
Fetus does not survive

**Two major types**

Autopolyploidy

Allopolyploidy

**Types of polyploidy**

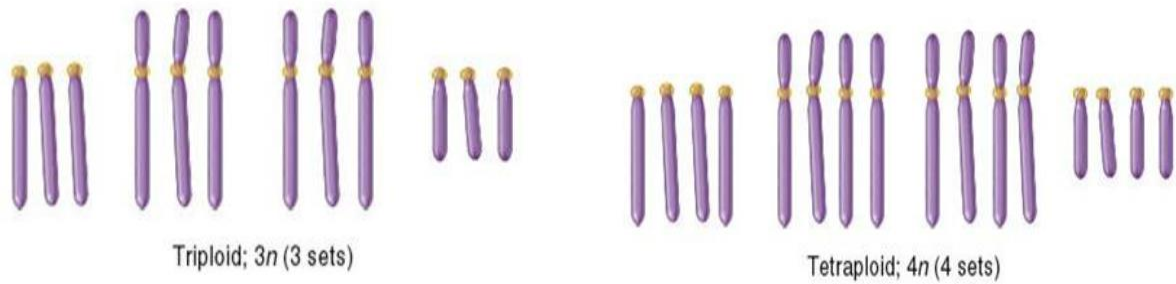


Triploid (3x) watermelon

Tetraploid (4x) - cotton

Pentaploid (5x) - Kenai Birch

Hexaploid (6x) - wheat



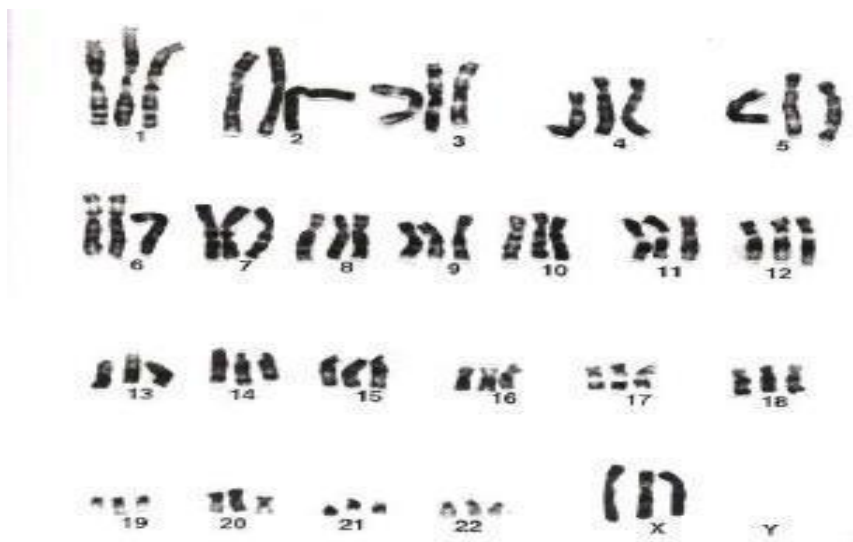
## Causes of polyploidy

Retention of polar body.

Formation of diploid sperms.

Di-spermy – fertilization by two sperms.

## Triploidy



## Polyploidy in Plants

Triploid: apple, banana, citrus, ginger, watermelon

Tetraploid: cotton, potato, tobacco, peanut

Hexaploid: oat, kiwifruit

Octaploid: strawberry

Decaploid: some sugar cane hybrids

## Lesson 53

### Allopolyploidy

#### Types of Polyploidy:

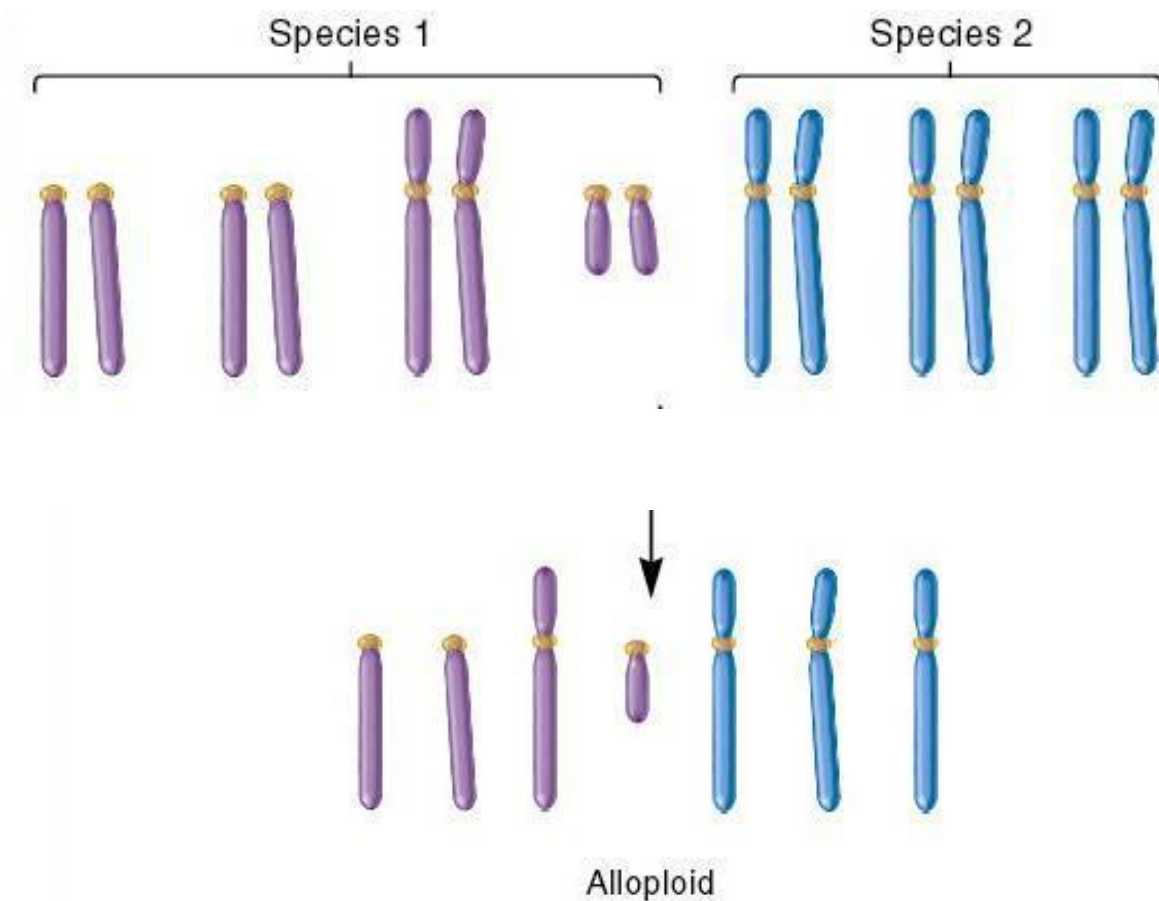
Autopolyploidy

Allopolyploidy

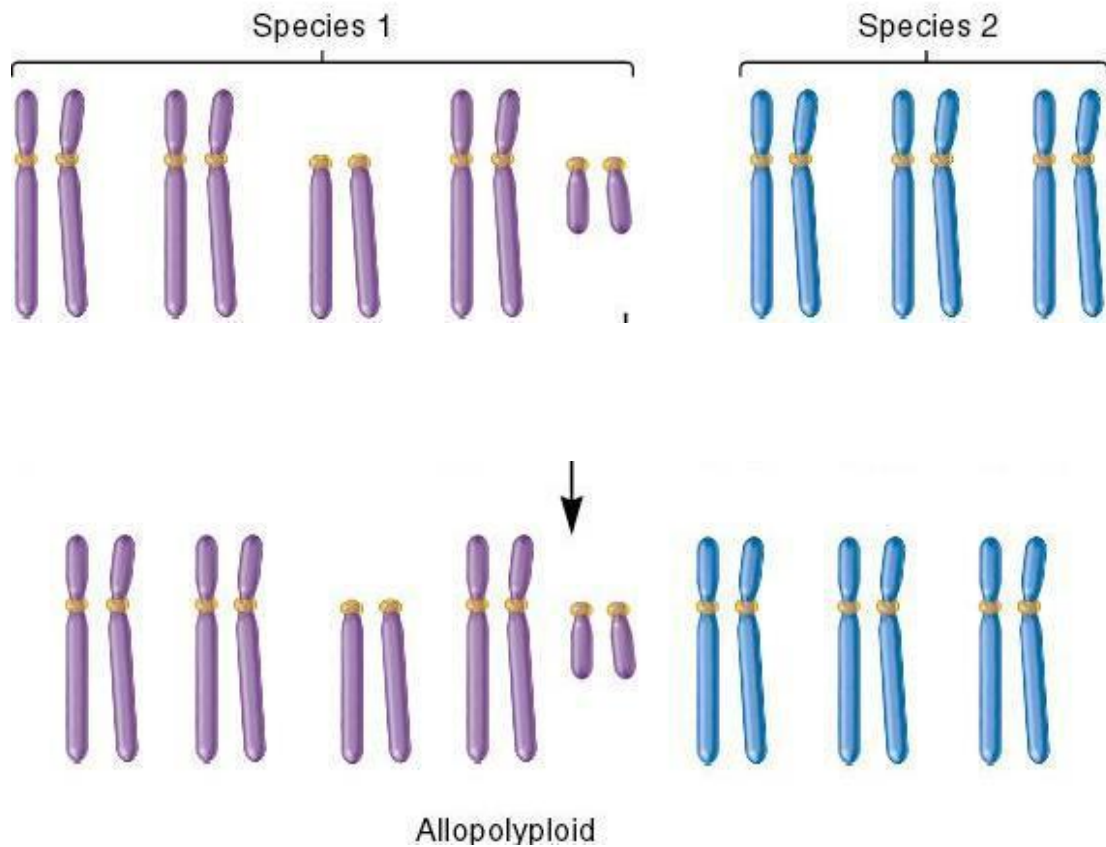
**Autopolyploidy** : Autopolyploids are polyploids with multiple chromosome sets derived from a single species.

**Allopolyploidy**: Allopolyploids are polyploids with chromosomes derived from different species. Allopolyploidy is inter-species cross Interspecies crosses can generate allopolyploids. Offspring are generally sterile.

#### Allopolyploid

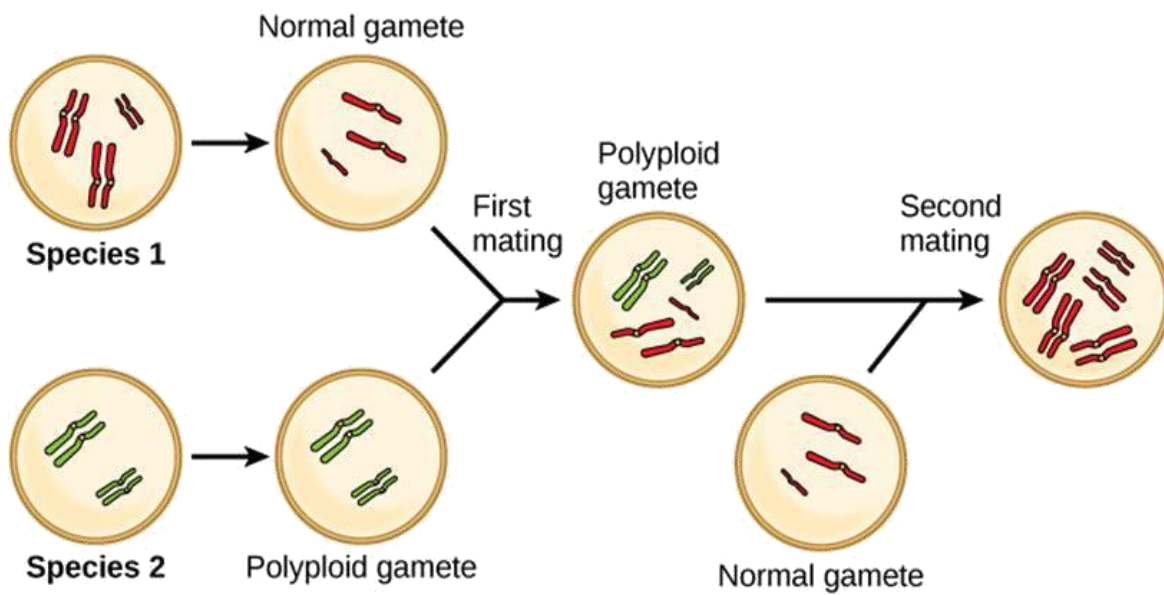


## Inter-species cross can results in Allodiploid



**Colchicine to promote polyploidy:** Polyploidy and allopolyploid plants often exhibit desirable traits. Colchicine is used to promote polyploidy Colchicine binds to tubulin, disrupting microtubule formation and blocks chromosome segregation.

## Allopolyploidy Resulting from Viable Matings between Two Species



### Allopolyploidy

An interspecies cross results in allopolyploidy.

**Mother donkey + Father horse:**

Offspring is a **hinny** (sterile)



**Domestic mare + Father zebra:**

Offspring is a **zebra hinny** (sterile)

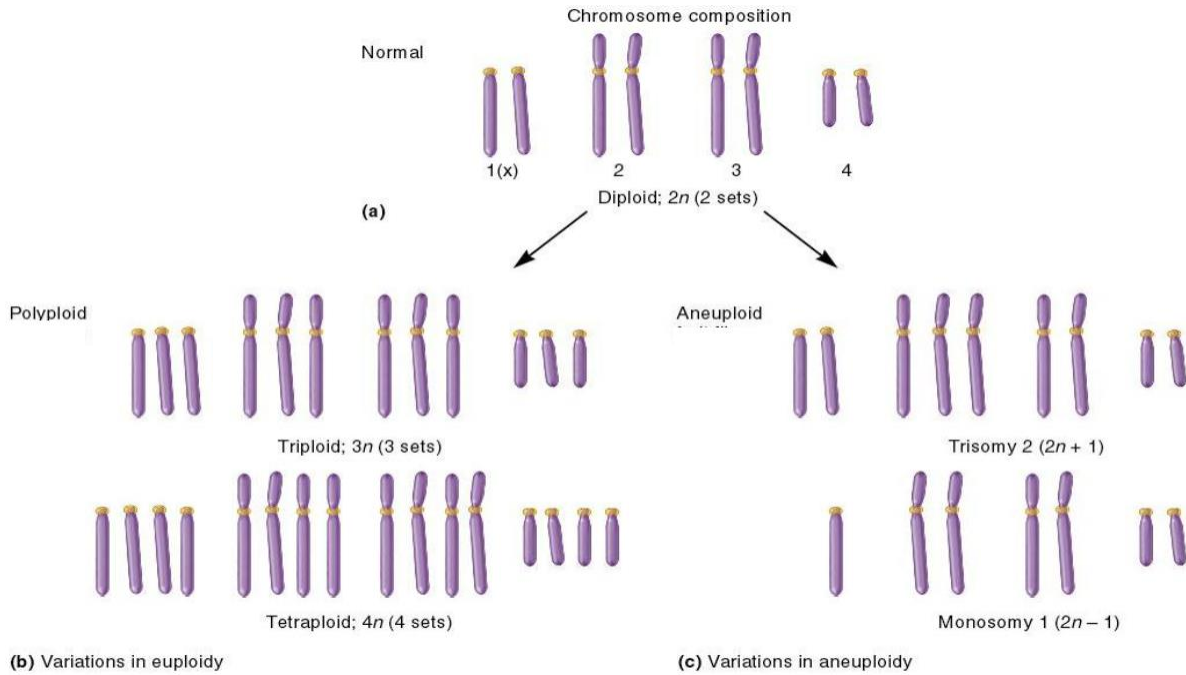




## Lesson 54

**Aneuploidy:** Change in chromosomes number can occur by the addition/deletion of chromosome or part of a chromosome.

**Euploidy:** Gain of one or more complete sets of chromosomes



### Aneuploidy

Nullisomy

Monosomy

Trisomy

Tetrasomy

### Euploidy

Triploid

Tetraploid

Pentaploid

## **Comparison**

**Aneuploidy:** change in chromosomes number.

**Euploidy:** gain of one or more complete sets of chromosomes.

## Lesson 55

### Structural Abnormalities

Chromosome breakage & subsequent reunion in a different configuration

### Types based on genetic material

Balanced – chromosome complement is complete.

Unbalanced – when there is incorrect amount of genetic material.

### Types

Translocations

Deletions

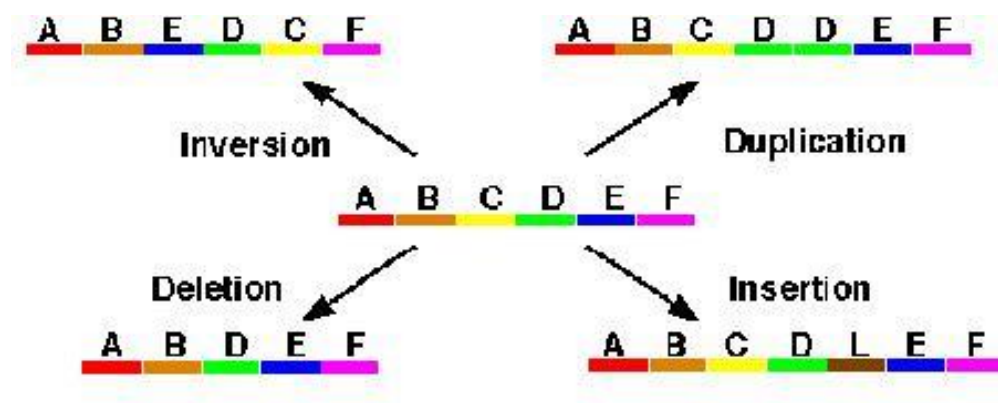
Insertions

Inversions

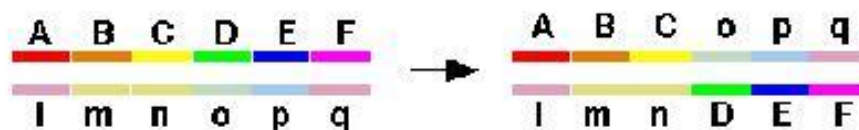
Ring chromosome

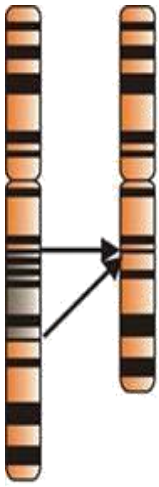
Isochromosomes

### Types of structural abnormalities

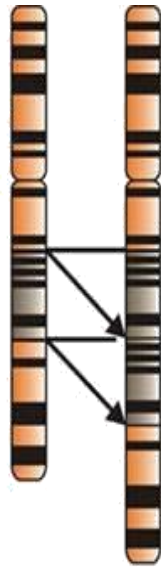


### Translocation





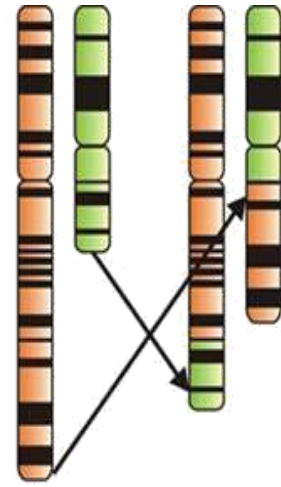
Deletion



Duplication



Inversion



Translocation

## Lesson 56

### Translocations

Definition Transfer of genetic material from one chromosome to another. There are two types of translocations

- Reciprocal translocation
- Robertsonian translocation

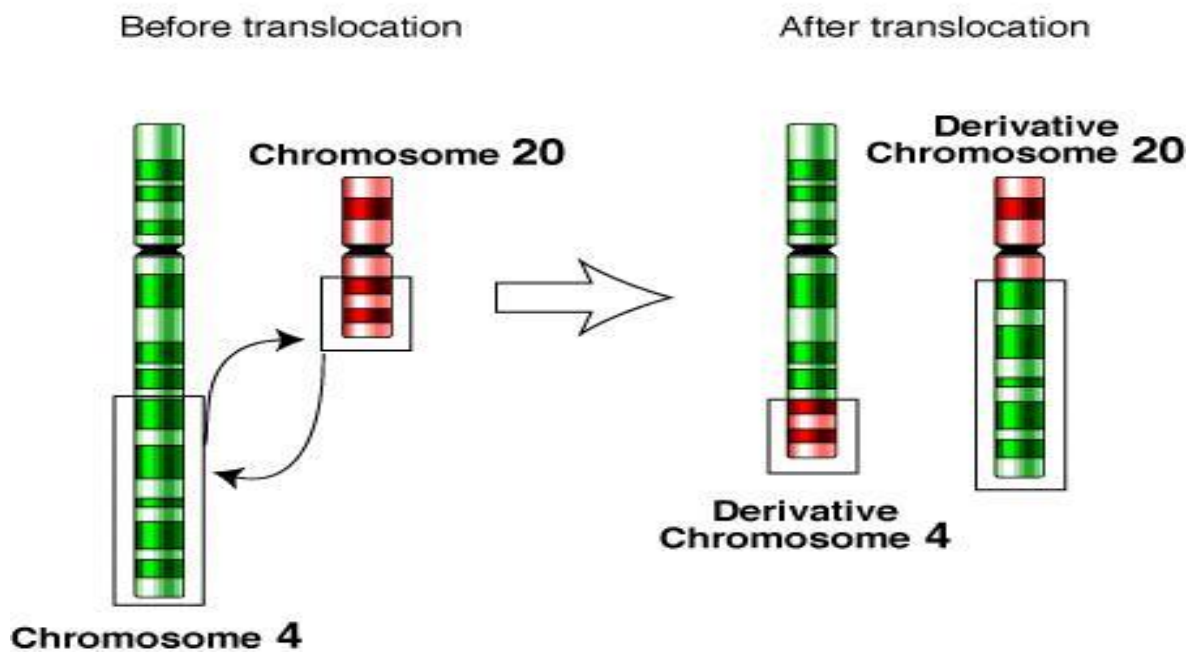
**Balanced and unbalanced** where no genetic information extra or missing genes.

Unbalanced translocations where exchange of genetic information is extra or missing some of the genes.

**Reciprocal an** exchange of material between two different chromosomes is called a reciprocal translocation. When this exchange involves no loss or gain of chromosomal material.

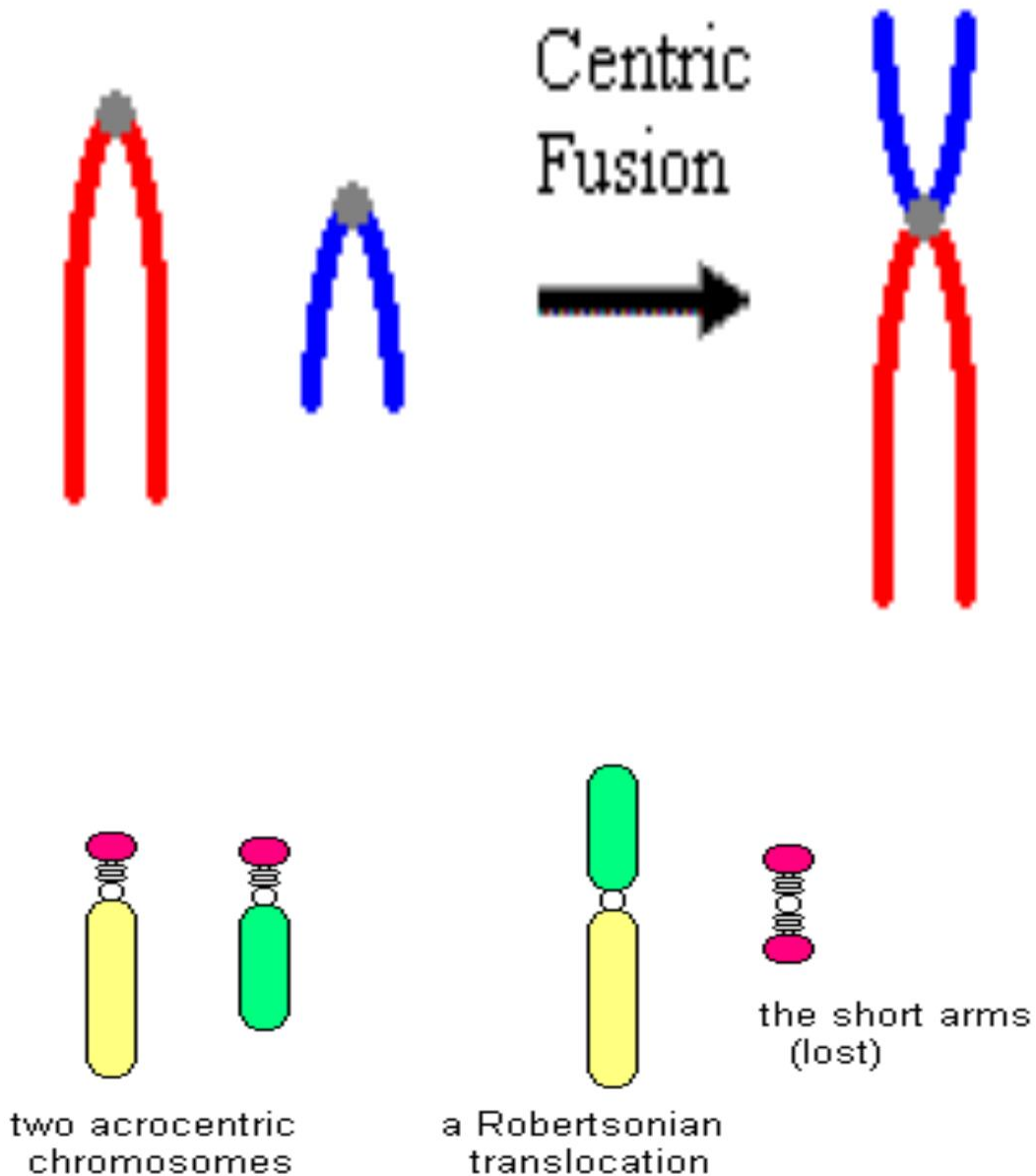
Reciprocal in non-homologous chromosomes Reciprocal translocations are usually an exchange of material between non-homologous chromosomes. One in each of 600 births in humans.

Reciprocal translocations



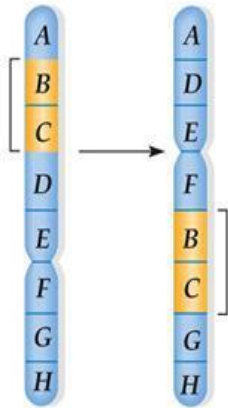
## Robertsonian Translocations

Breaks occur at the extreme ends of the short arms of two non-homologous acrocentric chromosomes. The small acentric fragments are lost. The larger fragments fuse at their centromere regions to form a single chromosome. The most common translocation in humans involves chromosomes 13 and 14 and is seen in about 0.97 / 1000 newborns.

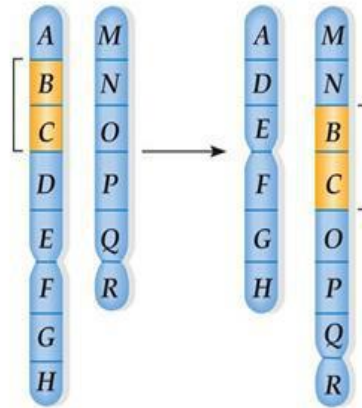


## Types of translocation

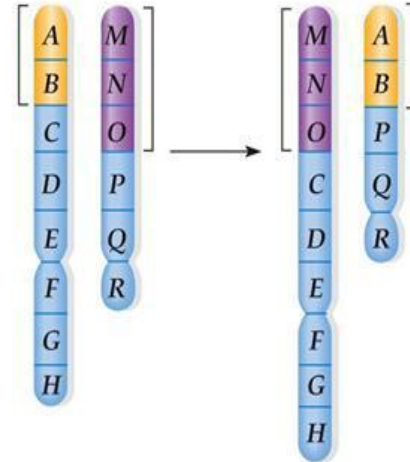
a) Nonreciprocal intrachromosomal translocation



b) Nonreciprocal interchromosomal translocation



c) Reciprocal interchromosomal translocation



## Common human diseases caused by translocation

Cancer

Infertility

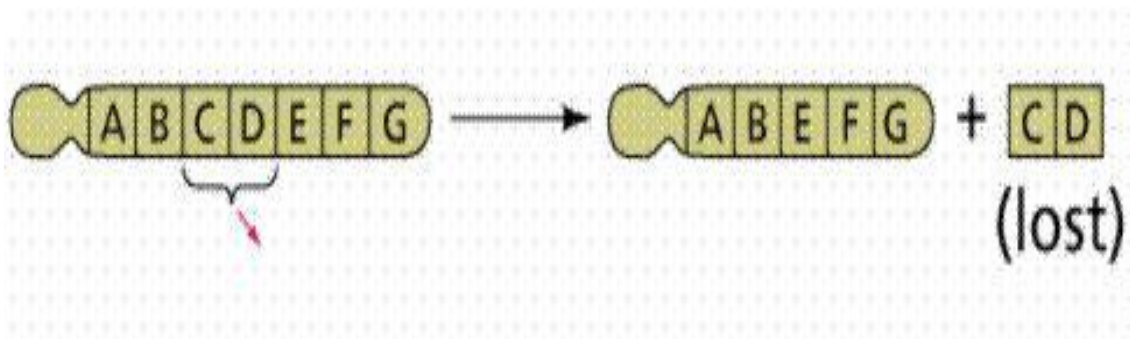
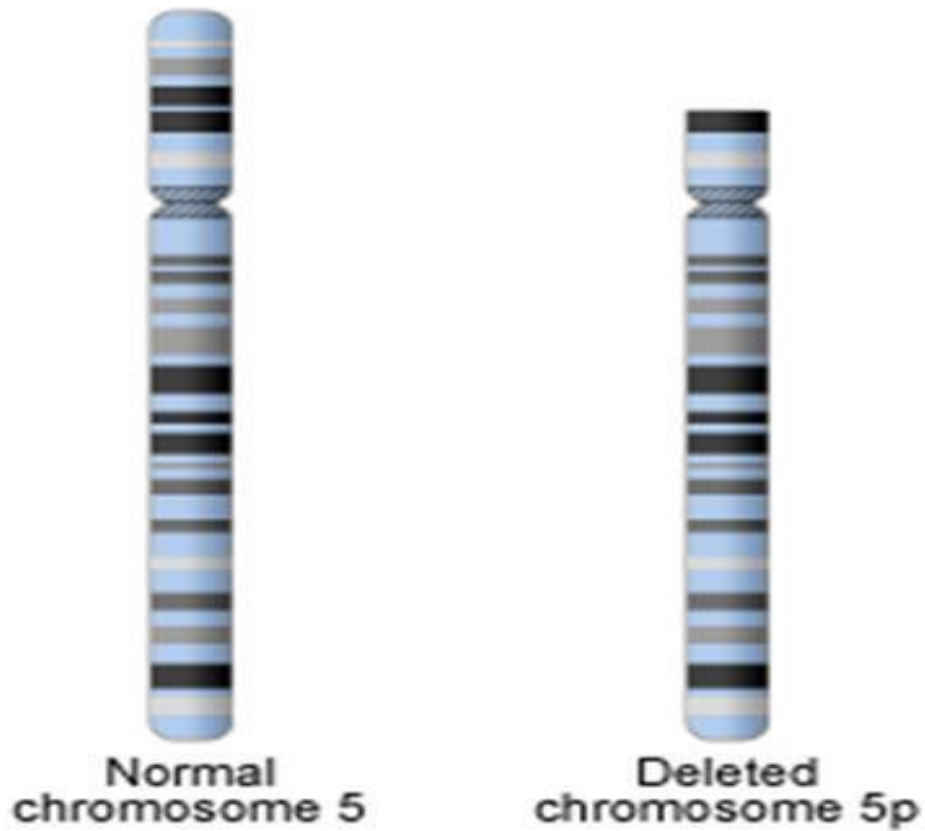
Down syndrome

Leukemia

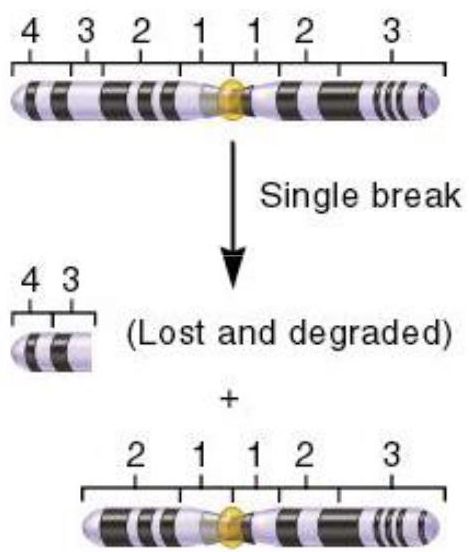
XX male syndrome

## Lesson 57

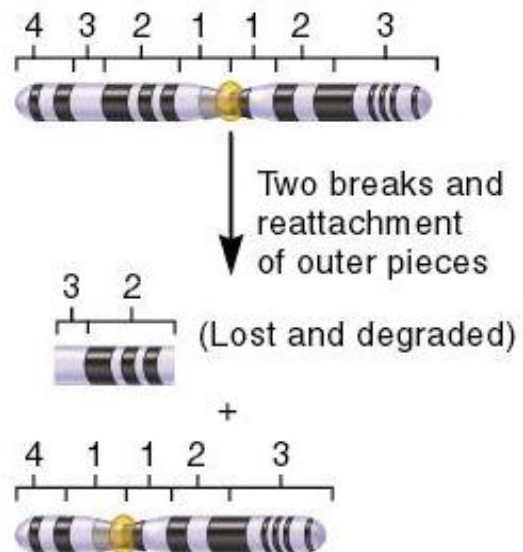
**Chromosomal deletions:** Deletions of larger portions are usually incompatible with life. 10-15% is due to balanced translocations in one parent. 85-90% are true deletions.







**(a) Terminal deficiency**



**(b) Interstitial deficiency**

Phenotypic consequences of deficiency depends on

- Size of the deletion

- Functions of genes

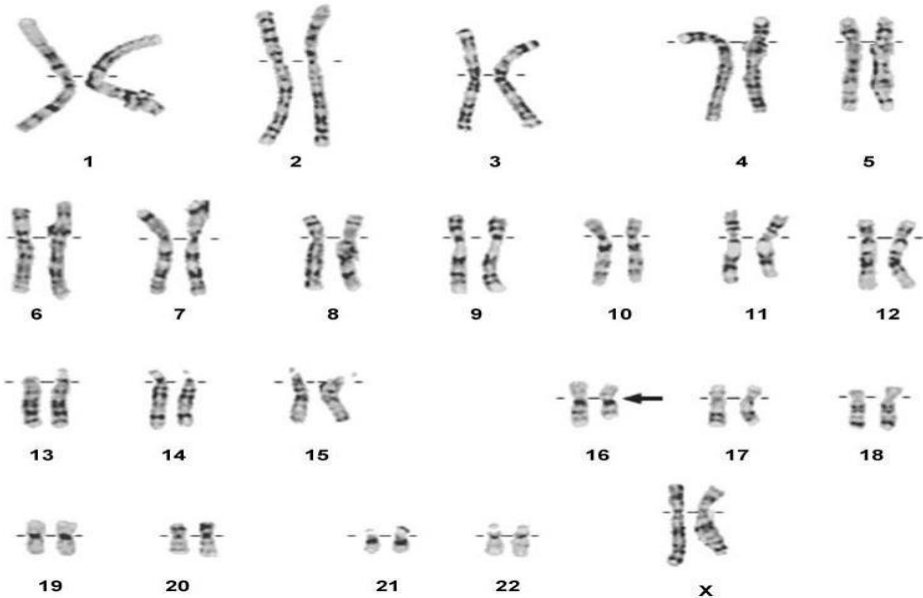
Deleted Phenotypic consequences of deficiency depends

- on - Phenotypic effect of deletions usually detrimental.

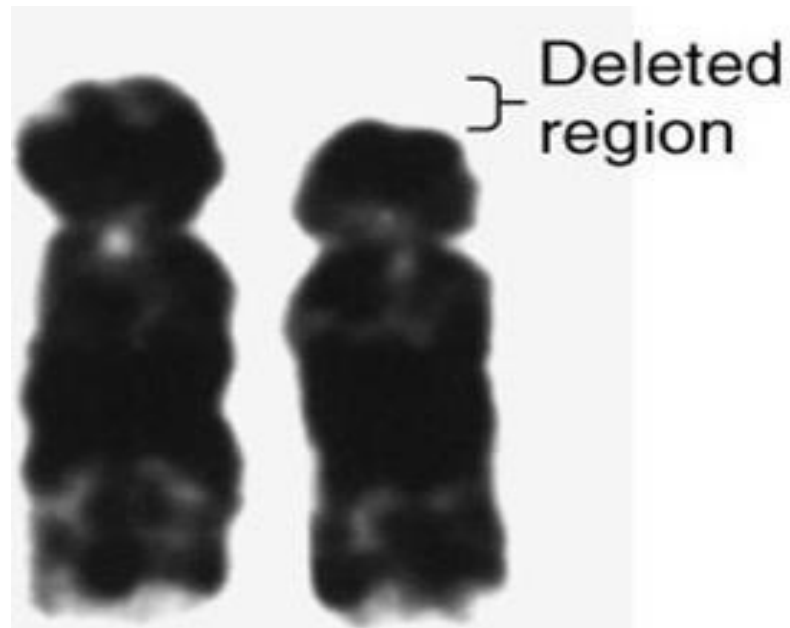
**Deletions – interstitial**



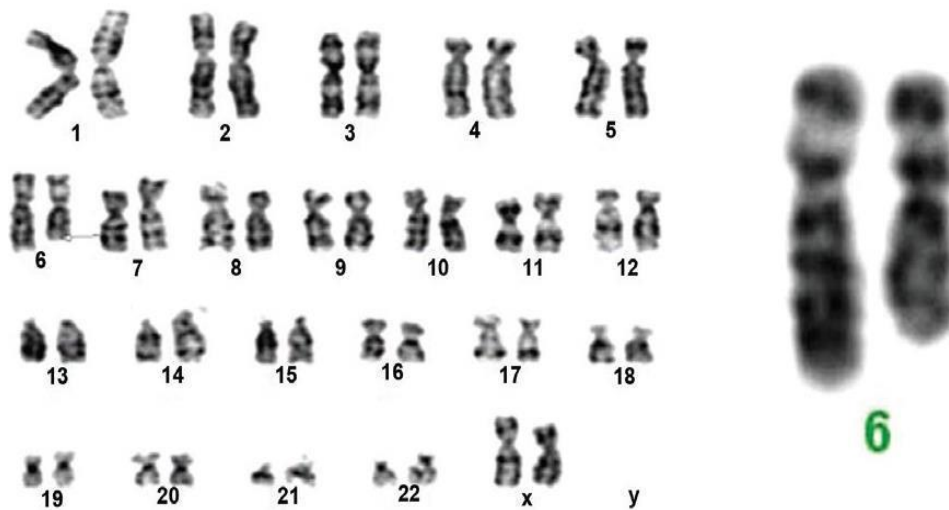
**Interstitial deletion at 16**



## Deletion – terminal in Cri-du-chat Syndrome



## Terminal deletion in chromosome 6



## Type of Deletions

Terminal – from one end

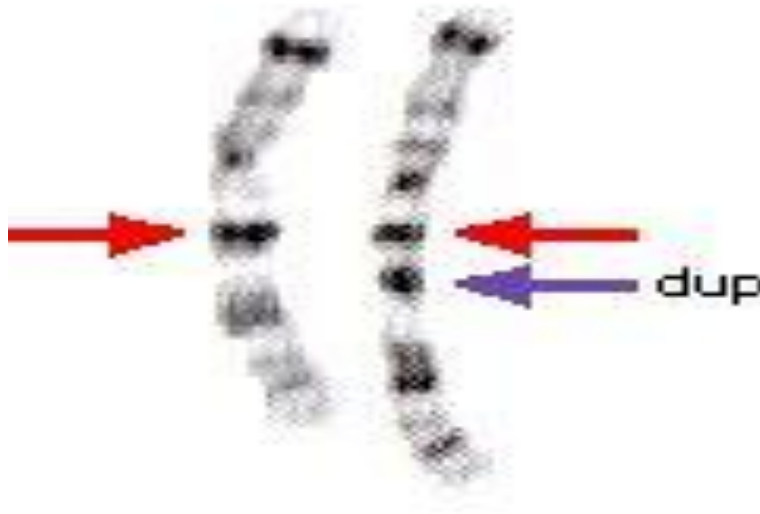
Interstitial – two breaks and middle part is lost

Micro deletions

## Lesson 58

**Chromosomal duplication:** Gene duplication (or chromosomal duplication) is a major mechanism through which new genetic material is generated during molecular evolution

### Chromosomal Duplications



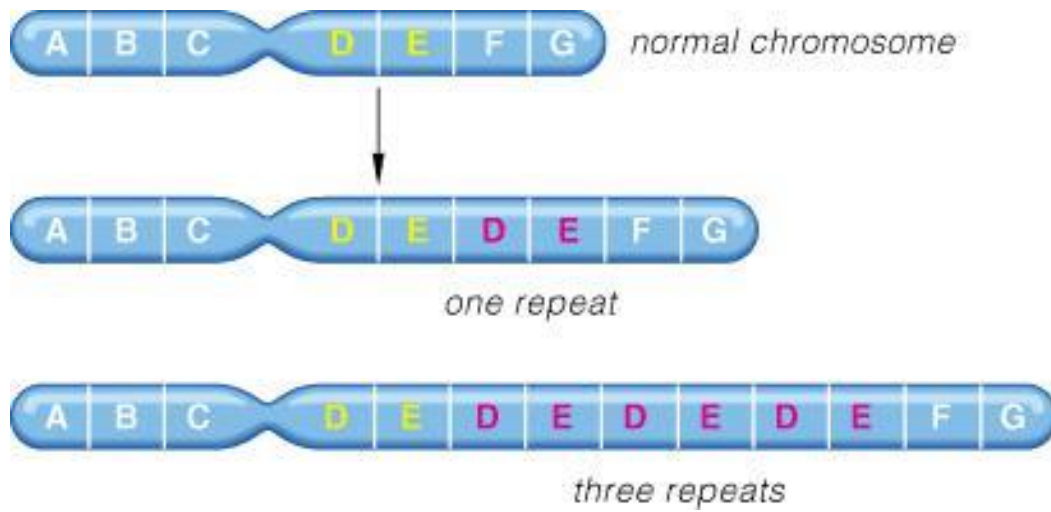
**Causes of Duplications – DNA replication:** Gene duplications can arise as products of several types of errors in DNA replication and repair machinery. Common sources of gene duplications include ectopic homologous recombination.

**Ectopic recombination - misaligned homologous** Duplications arise from an event termed unequal crossing-over that occurs during meiosis between misaligned homologous chromosomes.

**Other Causes of Duplications** Retro-transposition event, aneuploidy, polyploidy, and replication slippage are also cause of duplications.

**Replication slippage:** Replication slippage is an error in DNA replication that can produce duplications of short genetic sequences. During replication DNA polymerase begins to copy the DNA.

## Duplications



**Dispersed and tandem duplications:** Some duplication is “dispersed”, found in different locations from each other. Tandem duplications found next to each other.

**Tandem duplications become pseudo-genes:** These extra copies can then mutate to take on altered roles in the cell, or they can become pseudogenes, inactive forms of the gene, by mutation.

## Lesson 59

### Chromosomal Inversions

Two breaks re-arrangement

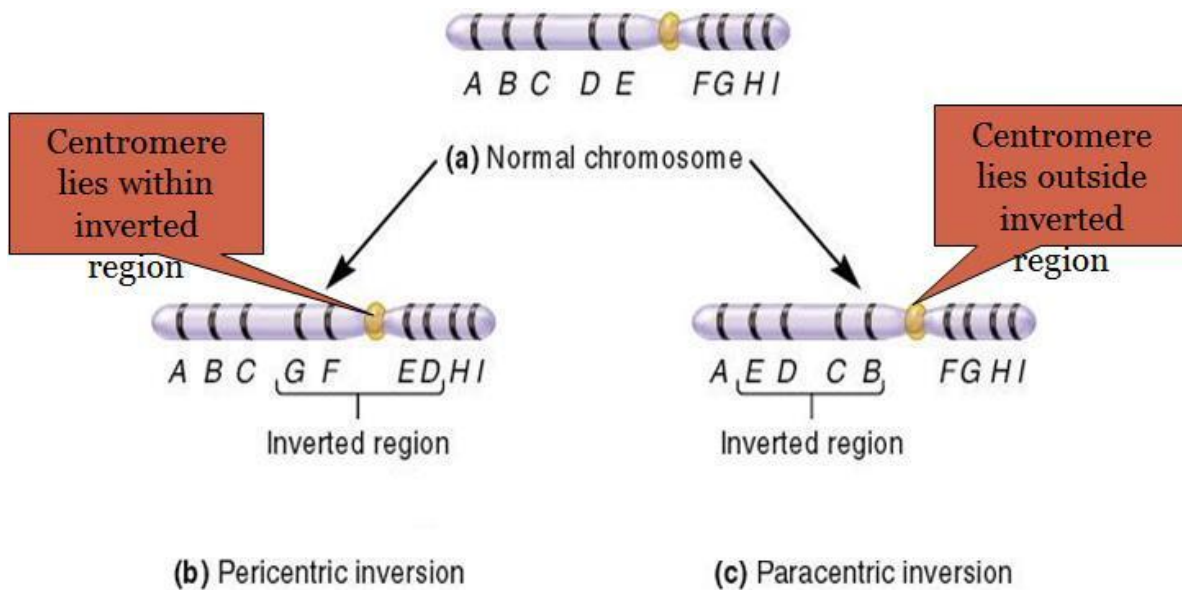
Segment is reversed

Pericentric – when centromere is involved

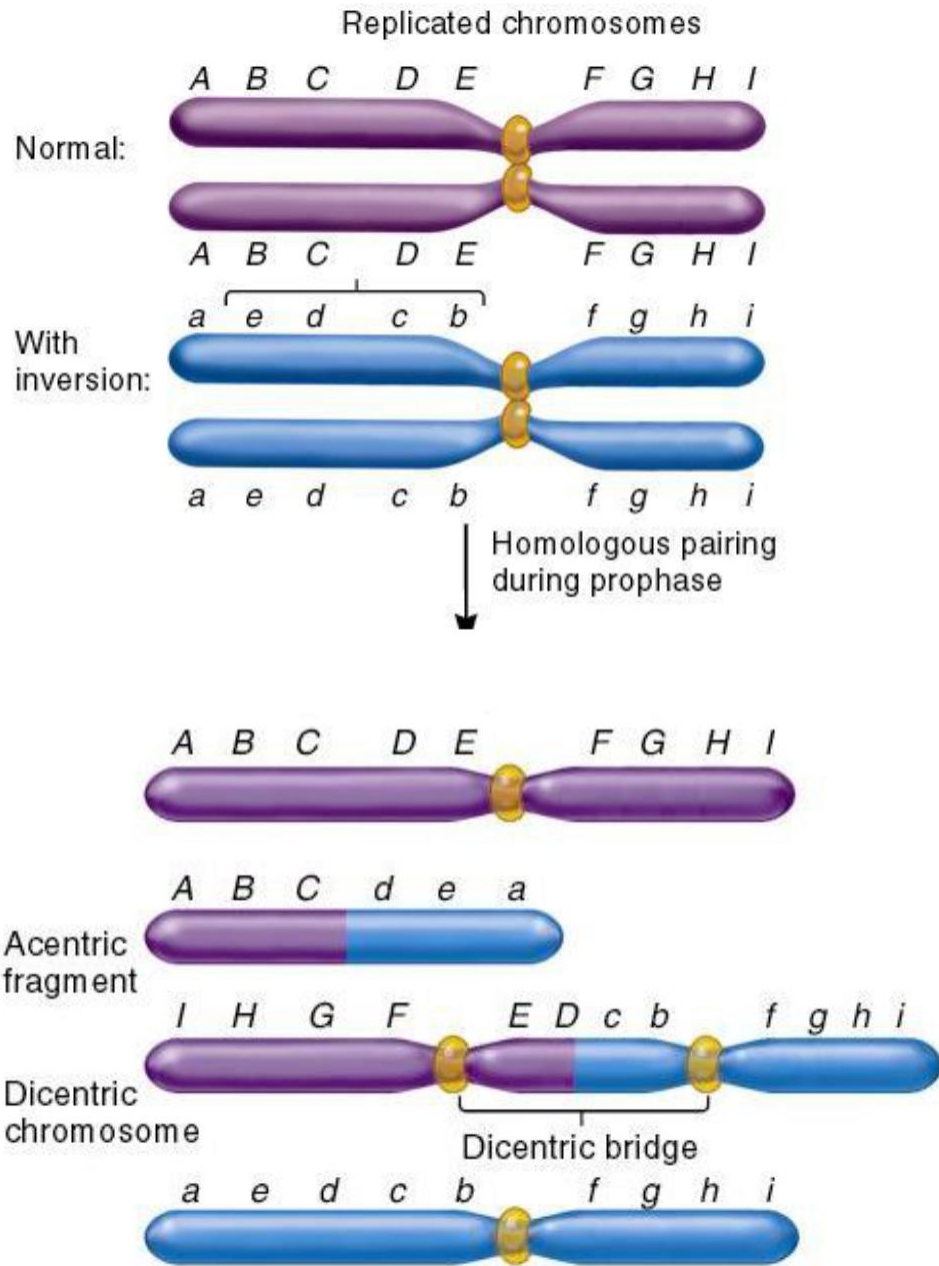
Paracentric – only one arm is involved

There is no loss of genetic information. Many inversions have no phenotypic consequences. Sometimes break point effects are within regulator or structural portion of a gene.

**Chromosomal inversions affects gene expression** Gene is re-positioned in a way that alters its gene expression. ~ 2% of the human population carries karyotypically detectable inversions.



**Inversion Heterozygotes:** Individuals with one copy of a normal chromosome and one copy of an inverted chromosome. Usually phenotypically normal



**Chromosomal Inversion:** The most common inversion seen in humans is on chromosome 9, at inv (9)(p12q13). This inversion is generally considered to have no harmful effects, but increased risk for miscarriage or infertility.



## Lesson 60

### Inheritance patterns of genetic disorders

**Autosomal Dominant:** Affected individuals are heterozygote. One allele is mutated, while other one is normal.

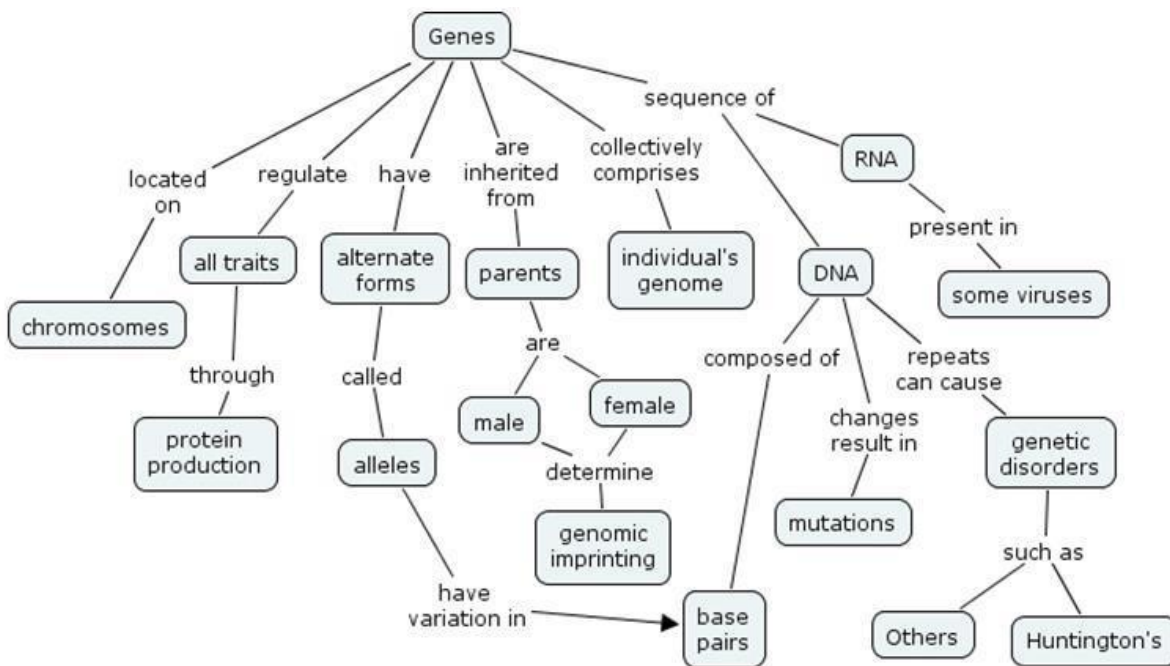
**Autosomal recessive:** Homozygous, when both alleles of the gene are mutated.

**X-linked recessive:** Mutated alleles are present on either of the sex chromosomes.

**Multifactorial:** Genes and environment both contribute.

**Multifactorial frequent, single gene disorders less frequent:** Multifactorial are common. Single gene less common like dominant/recessive pedigree patterns

### Genes and genetic disorders



### Chromosomal disorders

Thousands of genes may be involved.

Multiple organ systems affected at multiple stages in gestation.

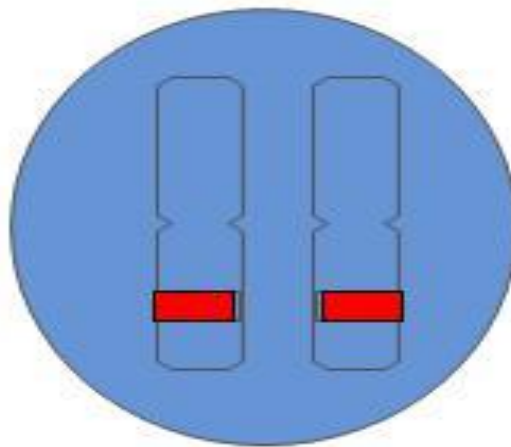
## Lesson 61

### Autosomal recessive inheritance / diseases

**Autosomal Recessive diseases:** Diseases occurs in individuals with two mutant alleles.

In general, an individual inherit one mutant allele from each of the parent.

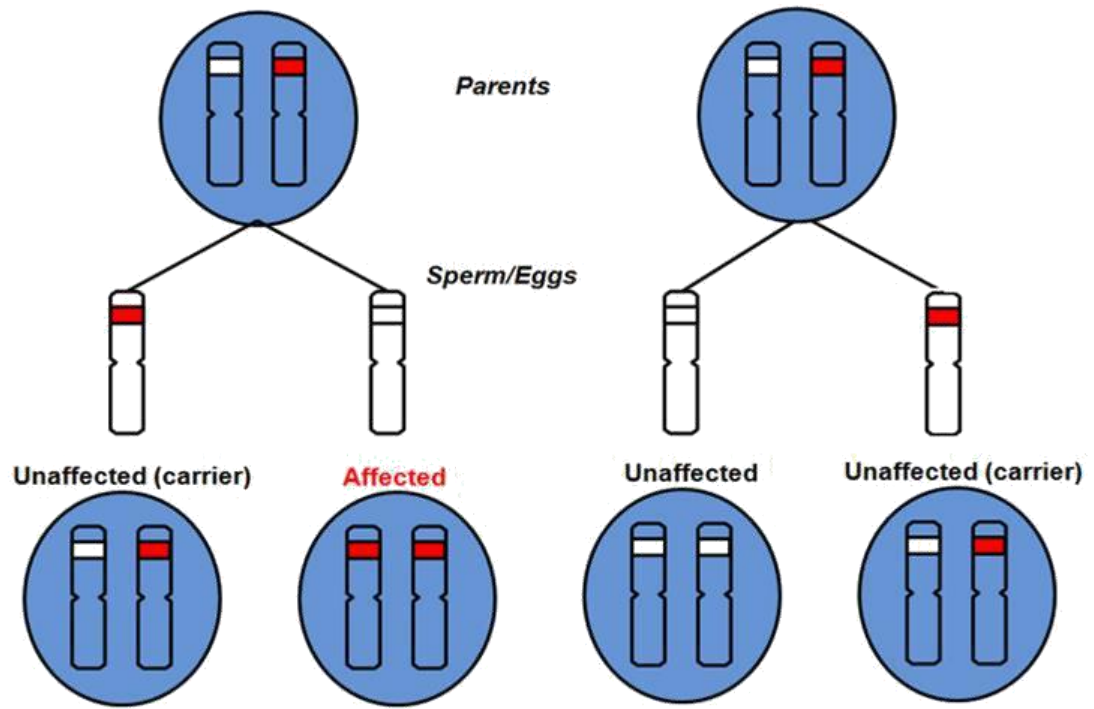
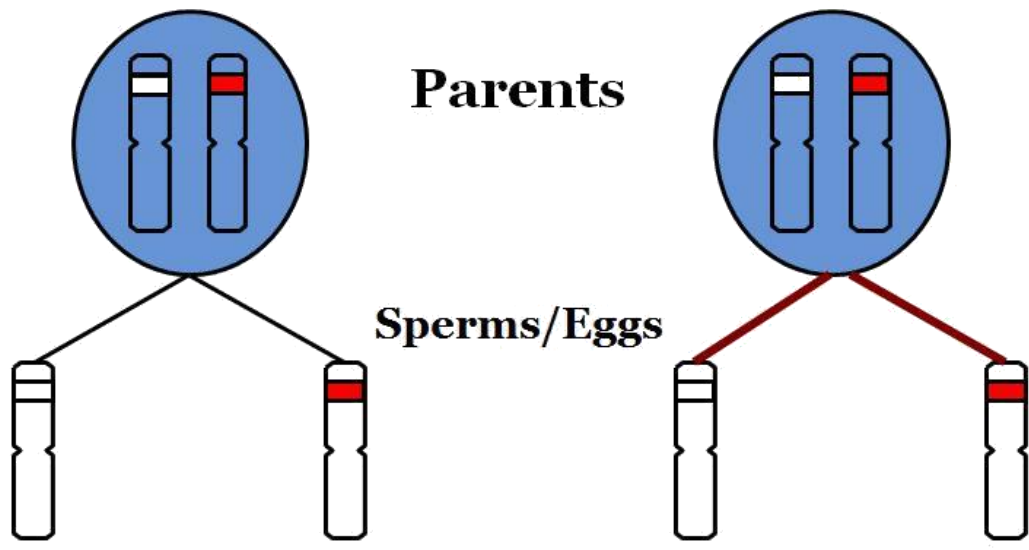
**Recessive diseases:** Homozygotes with two copies of the altered gene



**Recessive diseases:** Since each parent has two alleles. Chance of inheriting a mutant allele from one parent is 50% and for the other parent is also 50%. Net chance of inheritance of two mutant alleles is 25%.



Parent who are carriers for the same autosomal recessive condition have one copy of the normal form of the gene and one copy of an altered form of gene



Generally, the disease appears in the progeny of unaffected parents.

Affected progeny include both males and females equally.

**Phenotypic proportions are equal – autosomal recessive:** When we know that both male and female phenotypic proportions are equal, we can assume that we are dealing with autosomal inheritance.

**Common recessive disorders**

- Bloom syndrome
- Carpenter syndrome
- Cystic fibrosis
- Thalassemia
- Many forms of mental retardation
- Gaucher's disease
- Glycogen storage diseases
- Rotor syndrome
- Many of eye diseases

## **Properties of Autosomal recessive inheritance**

**Consanguinity:** These diseases appear where parents have common ancestors.

**Same allele increases chances of inheriting disorders:** The chance of inheriting identical alleles by an individual from both parents increases the chance of inheriting a recessive disorder.

**Common genetic background:** Small populations of individuals with a common genetic background may have increased risk of recessive diseases.

Ashkenazi: Gaucher disease

Phenotype found more likely in siblings of proband than in other relatives. Males and females are equally affected. In most of the cases, parents of affected individuals are asymptomatic.

### **Risk for each of the sibling being affected**

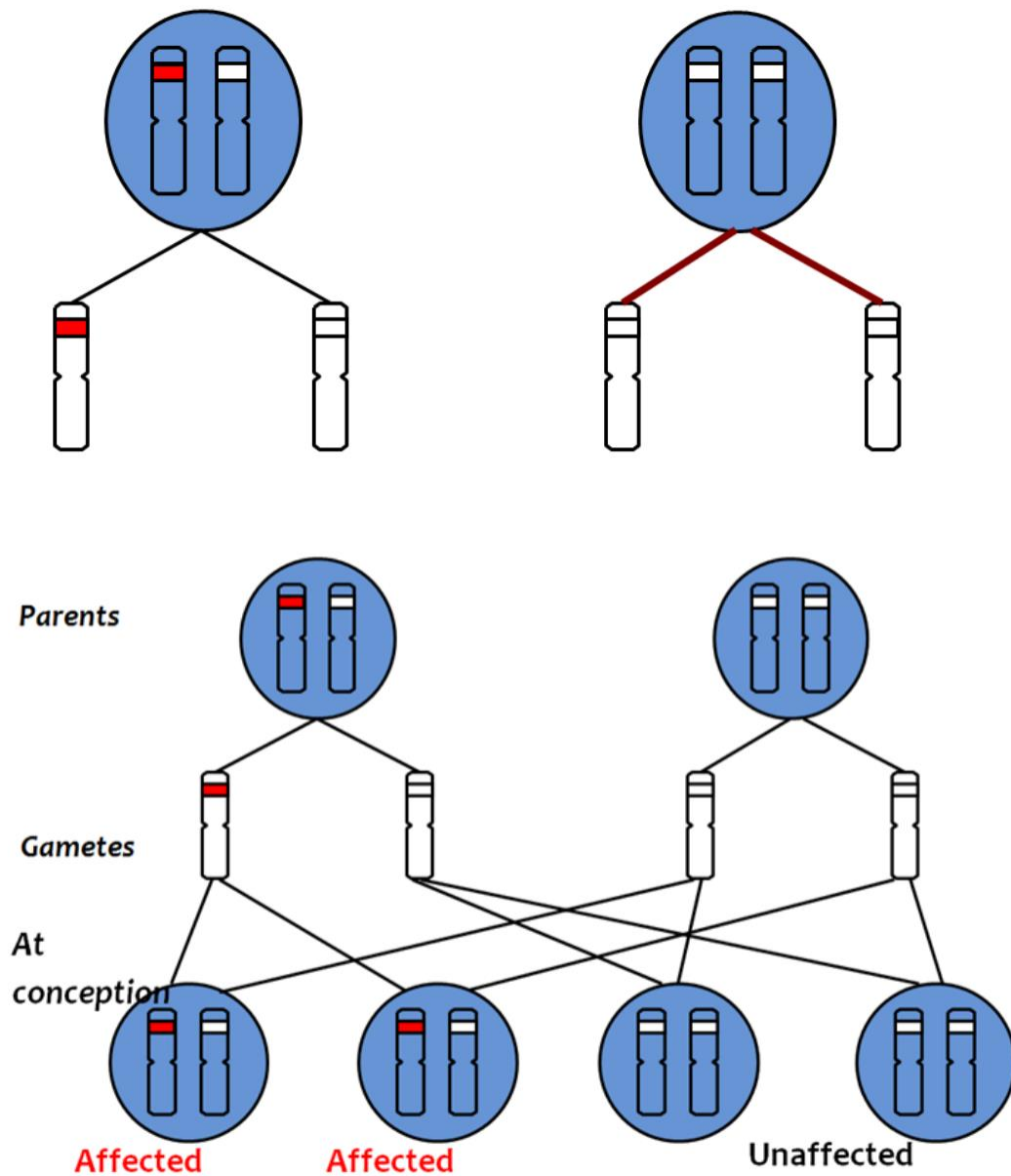
Risk of disease for each sib is 25 %.

## Lesson 62

### Autosomal dominant inheritance/ disorders

**Autosomal dominant disorders:** In autosomal dominant disorders, the normal allele is recessive and the abnormal allele is dominant. In a typical pedigree, every affected person has one affected parent. Equal numbers of affected females/males are expected. Male to male transmission is possible. One-half of the children of an affected individual are expected to have inherited the dominant allele.

### Autosomal Dominant disorders



Pedigree analysis, main clues for identifying a dominant disorder is that the phenotype tends to appear in every generation of the pedigree. Affected fathers/mothers transmit the phenotype to both sons and daughters. Trait is common in the pedigree. Trait is found in every generation.

### **Autosomal Dominant diseases**

Marfan syndrome

Huntington's disease

Retinoblastoma

Waardenburg syndrome

Myotonic dystrophy

Polycystic kidney disease

Achondroplasia

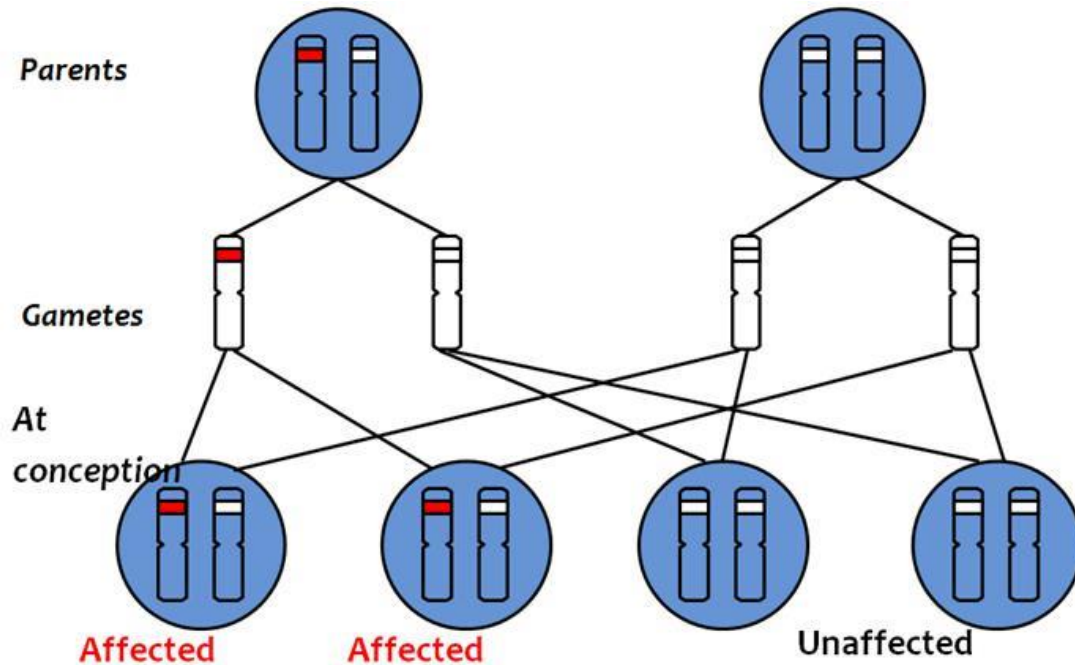
Polydactyly

Heredity hearing loss

## Variations in Autosomal dominant inheritance

**Autosomal dominant disorders:** In autosomal dominant disorders, the normal allele is recessive and the abnormal allele has dominant effect. In a typical pedigree, every affected person has one affected parent.

### Autosomal dominant disorders



### Penetrance - variation to A.D inheritance

Some people with an appropriate genotype fail to express the phenotype called as incomplete or reduced penetrance.

### Expressivity - variation to autosomal dominant inheritance

Severity of the phenotype: When phenotypic severity varies among those with same genotypes

### Pleiotropy - variation to autosomal dominant inheritance

Pleiotropy: Multiple phenotypic effects of a single gene.

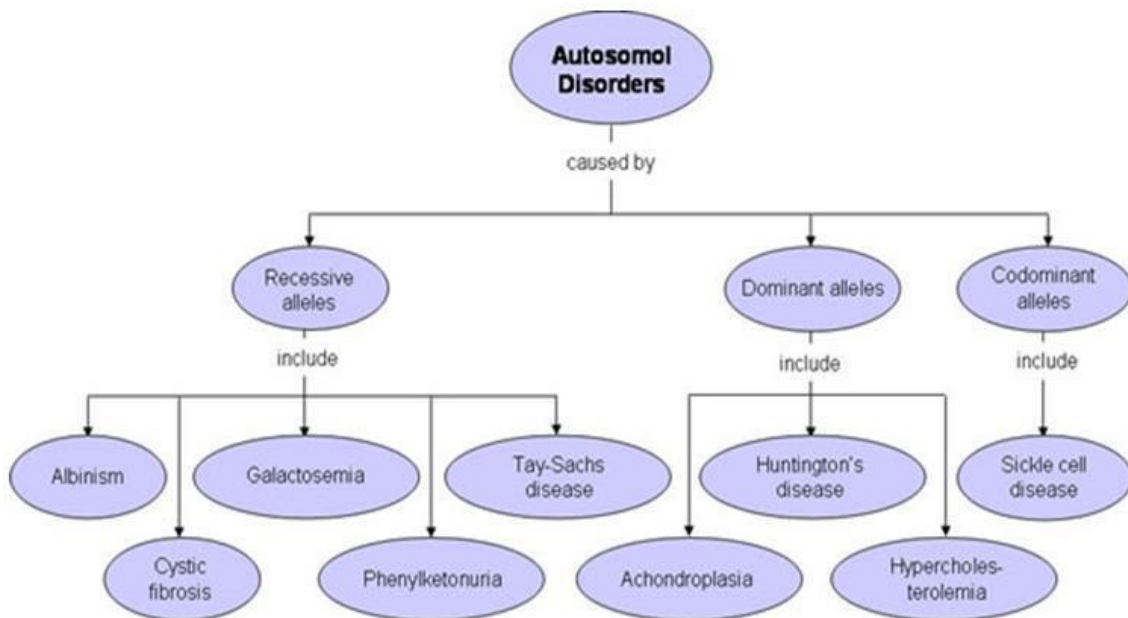


## Dominance VS. Recessiveness

**Dominance and Recessiveness** Dominance and recessiveness are the properties of characters/traits, not genes. A character is dominant if it is expressed in the heterozygote and recessive if not.

**Recessiveness:** Cystic fibrosis is recessive because only homozygotes manifest it, whereas heterozygotes show the normal phenotype.

**Hemizygous:** Males are hemizygous for loci on the X and Y chromosomes, where they have only a single



<b>Traits</b>	<b>Dominant</b>	<b>Recessive</b>
<b>Eye color</b>	<b>brown eyes</b>	<b>grey, green, hazel, blue eyes</b>
<b>Hair</b>	<b>dark hair</b>	<b>blonde, light, red hair</b>
	<b>non-red hair</b>	<b>red hair</b>
	<b>curly hair</b>	<b>straight hair</b>
<b>Facial features</b>	<b>dimples</b>	<b>No dimples</b>
	<b>unattached earlobes</b>	<b>attached earlobes</b>
	<b>broad lips</b>	<b>thin lips</b>
<b>Appendages</b>	<b>extra digits</b>	<b>normal number</b>
	<b>clubbed thumb</b>	<b>normal thumb</b>

### **Dominance and recessiveness**

Dominance and recessiveness are the properties of characters/traits, not the genes.

## Lesson 63

### X-linked recessive inheritance and disorders

**X-linked recessive:** Traits which are due to the genes which are present on either of the sex chromosomes. Many of the genes have a disease phenotype.

**Males are hemizygous and express the disease:** Males are hemizygous, will express the disease phenotype if one mutation is present.

**Homozygous Females manifest disease:** Females may be homozygous or heterozygous. Homozygous may manifest the disease.

### General characteristic

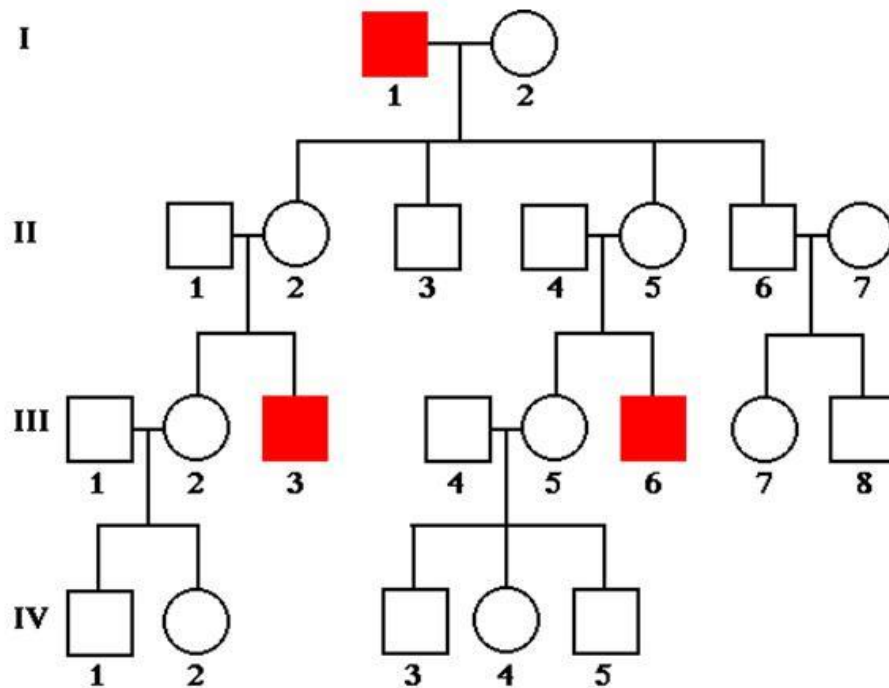
Trait is rare in pedigree.

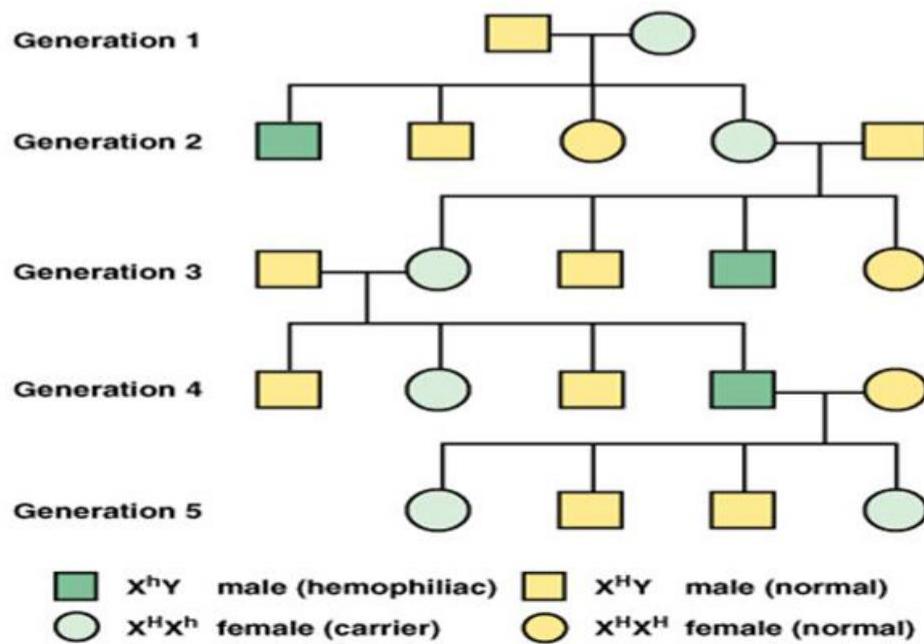
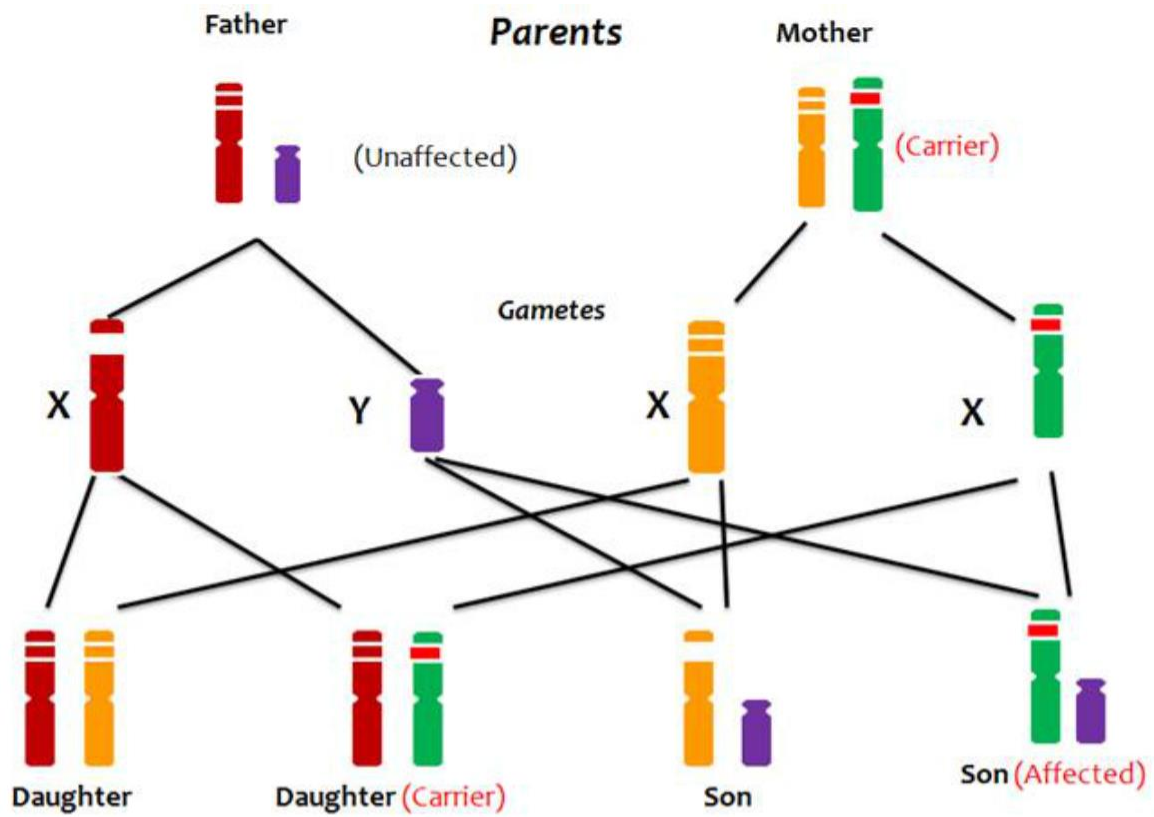
Trait skips generations.

Affected fathers do not pass to their sons.

Males are more often affected than females.

### X-linked recessive disorders





**X-linked recessive:** X-linked recessives expressed in all males but only in homozygous females.

X-linked color blindness

**Common X-linked disorders**

Duchene Muscular Dystrophy

Hunter's Disease

Menkes Disease

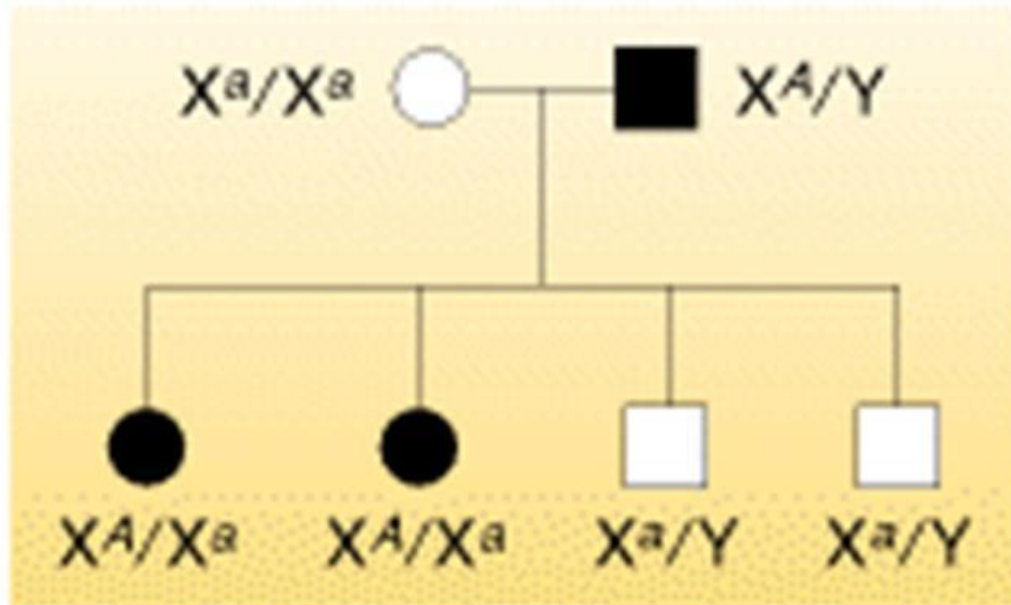
Hemophilia A and B

Color Blindness

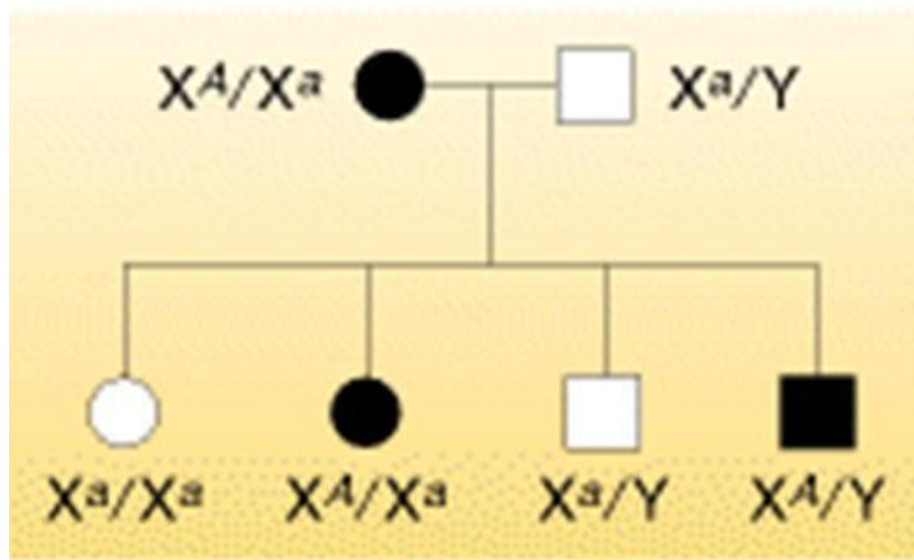
## Lesson 64

### X-linked dominant inheritance and disorders

Affected males transmit disease to all daughters but none of sons.

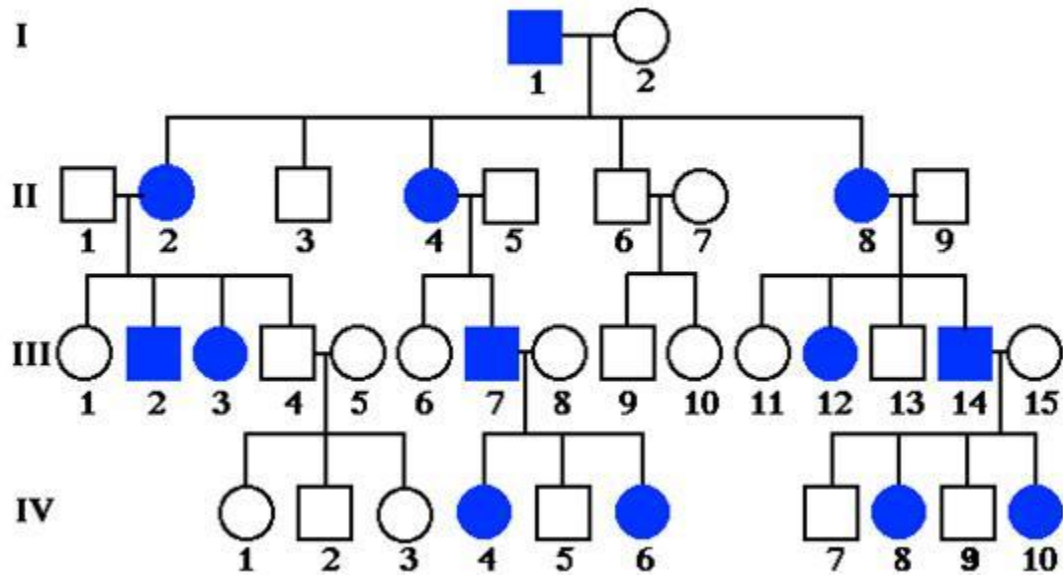


Affected heterozygous females and un-affected males pass the disease to half their sons and daughters



**X-linked dominant diseases:** Affected fathers pass to all of their daughters.

Males and females are equally likely to be affected.



### Common X-linked dominant diseases

Rett syndrome

Alport syndrome

Goltz syndrome

Fragile-X Syndrome

## Lesson 65

### Maternal inheritance of mitochondrial DNA

Mitochondrial DNA (MTDNA) is inherited through ovum not sperm

Mother could pass it to all children.

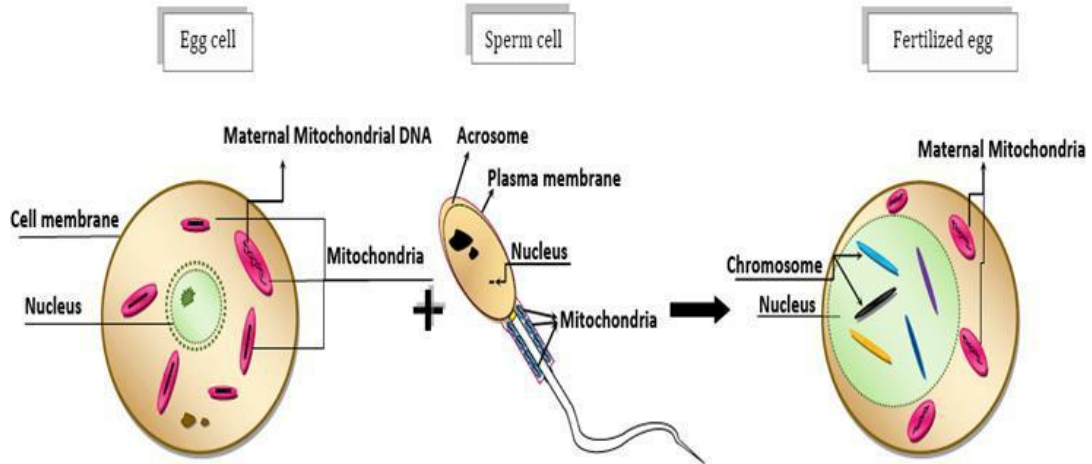
Father will pass it to none of the children.

### Mitochondrial DNA

More than one copy of mtDNA is passed.

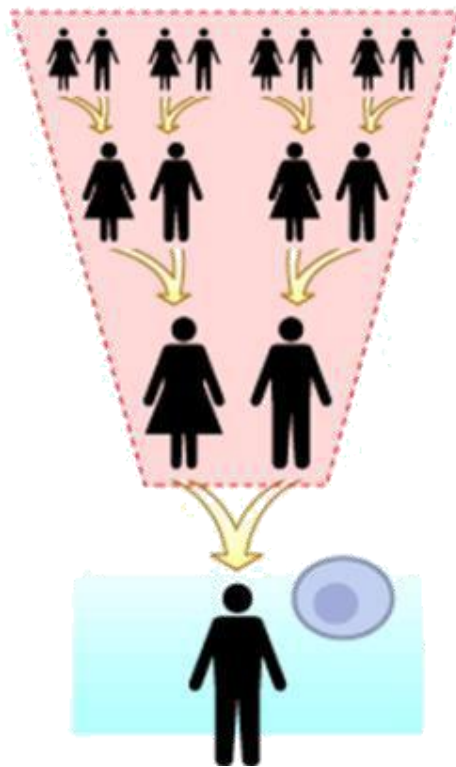
Mutations are common.

More than one type of mtDNA genome is passed.

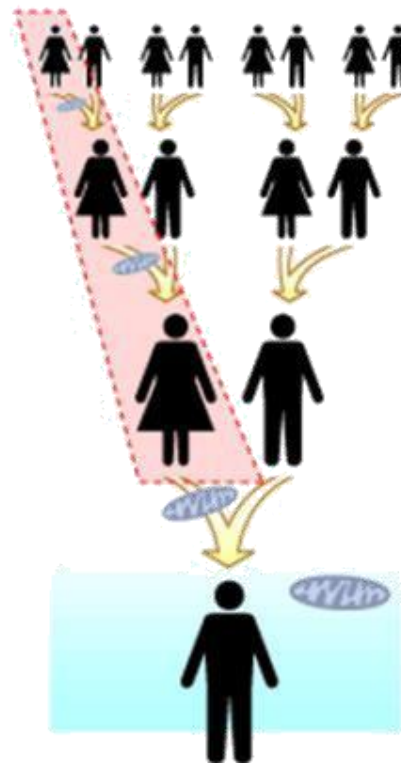




Nuclear DNA is inherited from all ancestors.



Mitochondrial DNA is inherited from a single lineage.



### Properties of Mitochondrial inheritance

Heteroplasmy

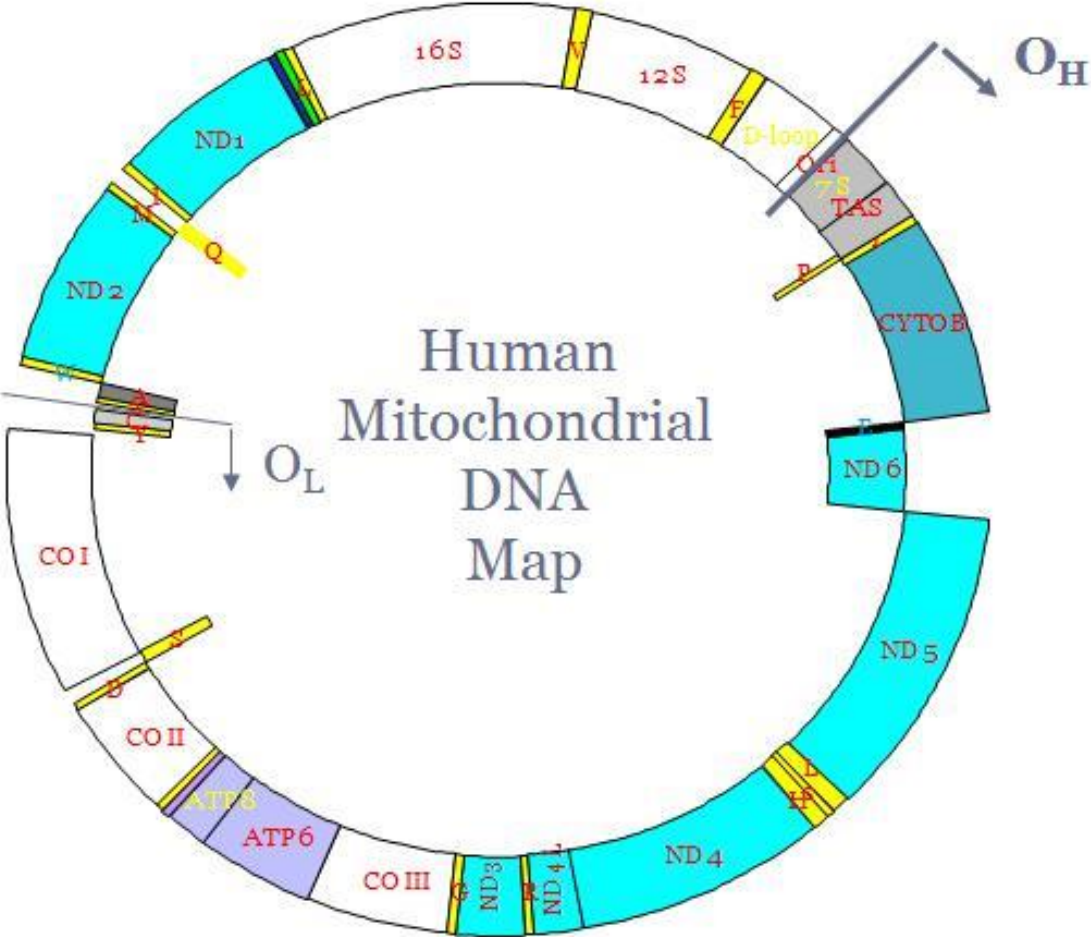
Variable expression

Pleiotropy

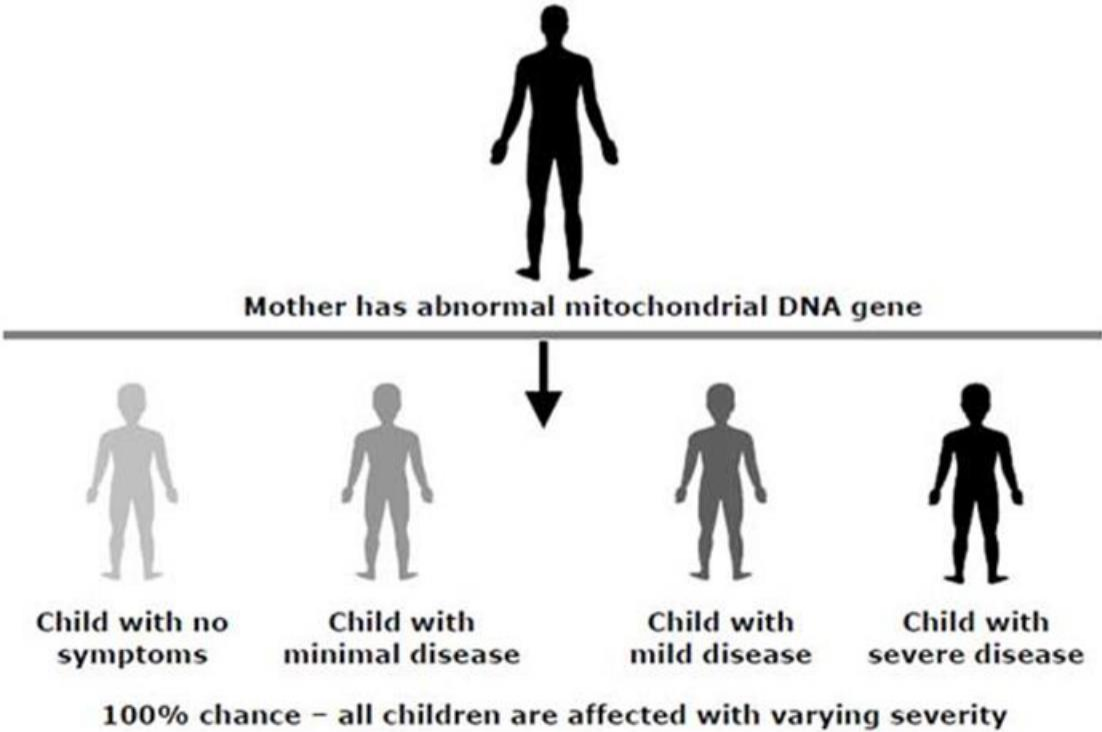
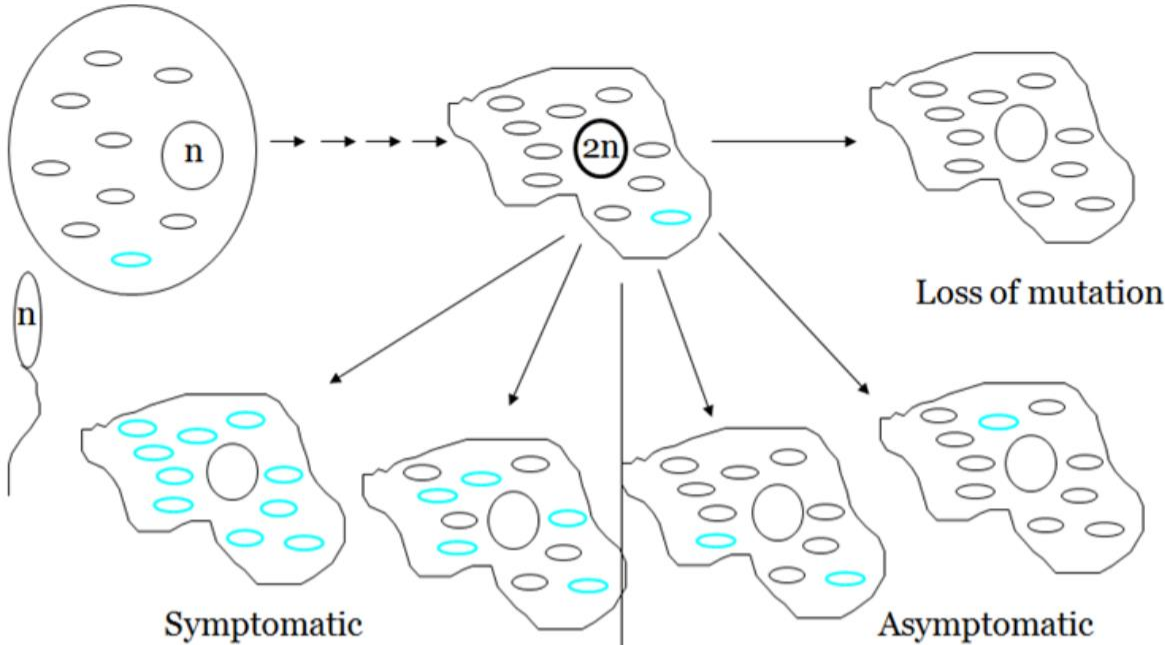
Reduced penetrance

Functional somatic mosaic

Mitochondrial DNA



**Mitochondrial DNA – Severity**



## **Common mitochondrial diseases**

Myoclonic epilepsy

Mitochondrial recessive ataxia

Leber hereditary optic neuropathy

Sensory ataxia neuropathy

## Lesson 66 AND 67

Eukaryotes have three nuclear RNA polymerases, each with distinct roles and properties.

Name	Location	Product
RNA Polymerase I (Pol I, Pol A)	nucleolus	larger ribosomal RNA (rRNA) (28S, 18S, 5.8S)
RNA Polymerase II (Pol II, Pol B)	nucleus	messenger RNA (mRNA), most small nuclear RNAs (snRNAs), small interfering RNA (siRNAs) and micro RNA (miRNA).
RNA Polymerase III (Pol III, Pol C)	nucleus (and possibly the nucleolus-nucleoplasm interface)	transfer RNA (tRNA), other small RNAs (including the small 5S ribosomal RNA (5s rRNA), snRNA U6, signal recognition particle RNA (SRP RNA) and other stable short RNAs

RNA polymerase I (Pol I) catalyses the transcription of all rRNA genes except 5S. These rRNA genes are organised into a single transcriptional unit and are transcribed into a continuous transcript. This precursor is then processed into three rRNAs: 18S, 5.8S, and 28S. The transcription of rRNA genes takes place in a specialised structure of the nucleus called the nucleolus, where the transcribed rRNAs are combined with proteins to form ribosomes.

RNA polymerase II (Pol II) is responsible for the transcription of all mRNAs, some snRNAs, siRNAs, and all miRNAs. Many Pol II transcripts exist transiently as single strand precursor RNAs (pre-RNAs) that are further processed to generate mature RNAs. For example, precursor mRNAs (pre-mRNAs) are extensively processed before exiting into the cytoplasm through the nuclear pore for protein translation.

RNA polymerase III (Pol III) transcribes small non-coding RNAs, including tRNAs, 5S rRNA, U6 snRNA, SRP RNA, and other stable short RNAs such as ribonuclease P RNA.

RNA Polymerases I, II, and III contain 14, 12, and 17 subunits, respectively. All three eukaryotic polymerases have five core subunits that exhibit homology with the  $\beta$ ,  $\beta'$ ,  $\alpha$ I,  $\alpha$ II, and  $\omega$  subunits of E. coli RNA polymerase. An identical  $\omega$ -like subunit (RBP6) is used by all three eukaryotic polymerases, while the same  $\alpha$ -like subunits are used by Pol I and III. The three eukaryotic polymerases share four other common subunits among themselves. The remaining subunits are unique to each RNA polymerase. The additional subunits found in Pol I and Pol III relative to Pol II, are homologous to Pol II transcription factors.

Crystal structures of RNA polymerases I and II provide an opportunity to understand the interactions among the subunits and the molecular mechanism of eukaryotic transcription in atomic detail.

The carboxyl terminal domain (CTD) of RPB1, the largest subunit of RNA polymerase II, plays an important role in bringing together the machinery necessary for the synthesis and processing of Pol II transcripts. Long and structurally disordered, the CTD contains multiple repeats of heptapeptide sequence YSPTSPS that are subject to phosphorylation and other posttranslational modifications during the transcription cycle. These modifications and their regulation constitute the operational code for the CTD to control transcription initiation, elongation and termination and to couple transcription and RNA processing.

## Lesson 68 to 75

### Pre-mRNA Splicing

Because eukaryotic pre-mRNAs are transcribed from intron containing genes, the sequences encoded by the intronic DNA must be removed from the primary transcript prior to the RNA's becoming biologically active. The process of intron removal is called RNA splicing, or pre-mRNA splicing. The intron-exon junctions (splice-sites) in the precursor mRNA (pre-mRNA) of eukaryotes are recognized by trans-acting factors (prokaryotes RNAs are mostly polycistronic). In pre-mRNA splicing the intronic sequences are excised and the exons are ligated to generate the spliced mRNA.

Group I introns occur in nuclear, mitochondrial and chloroplast rRNA genes, group II introns in mitochondrial and chloroplast mRNA genes.

Many of the group I and group II introns are self-splicing in that no additional protein factors are necessary for the intron to be efficiently and accurately excised and the strands reattached. "The nucleotide sequence of group II self-splicing introns is highly conserved, and hence these introns fold into an evolutionarily conserved three-dimensional structure, which can undergo a self-splicing reaction in the absence of any trans-acting factors.

In contrast, the nucleotide sequences and length of nuclear pre-mRNA introns is highly variable, except for the short conserved sequences at the 5' and 3' splice sites and the branch points. Therefore nuclear pre-mRNA splicing requires trans-acting factors, which interact with these short conserved sequences, and from which the catalytically active spliceosome is assembled.

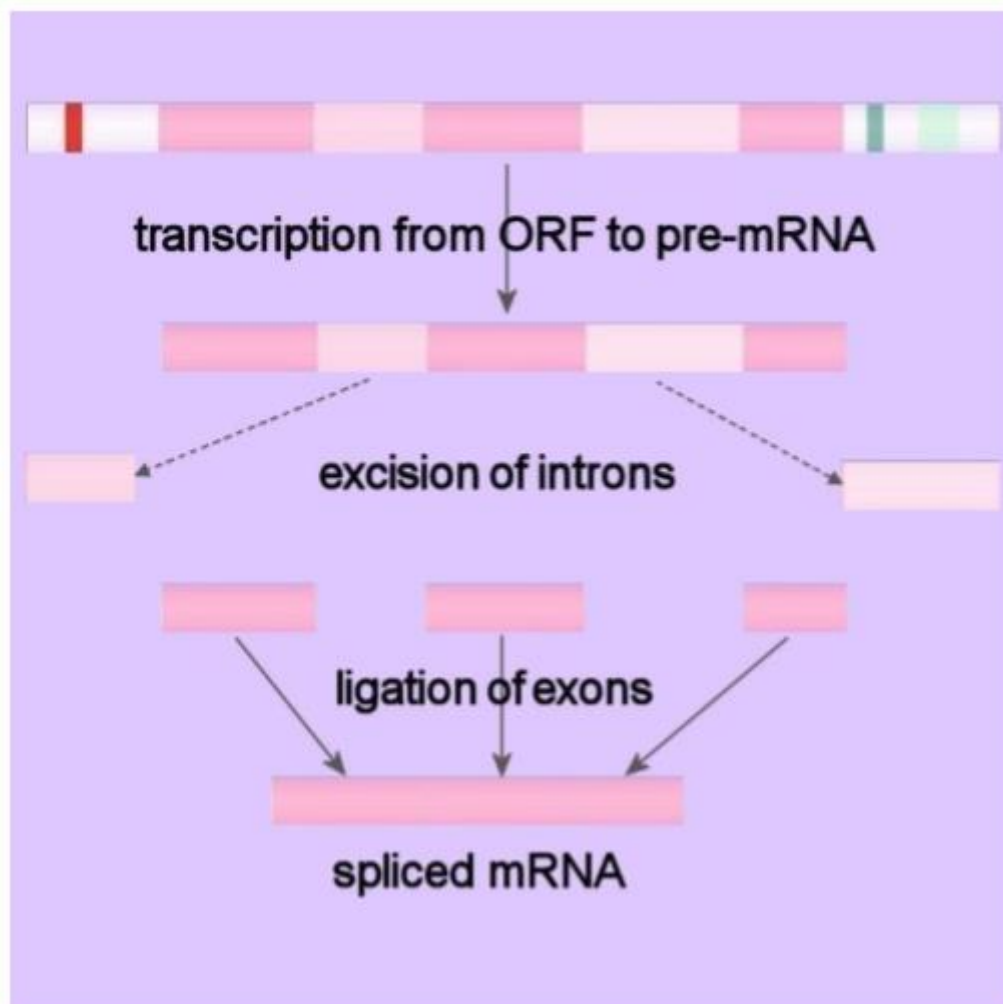
The conserved sequences are: 5' splice site = AGGuragu; 3' splice site = yyyyyyy nagG (y= pyrimidine); branch site = ynyuray (r = purine, n = nucleotide)

Expressed differently, the highly conserved, consensus sequence for the 5' donor splice site is (for RNA): (A or C)AG/GUAAGU. That is, most exons end with AG and introns begin with GU (GT for DNA). The highly conserved, consensus sequence for the 3' acceptor splice site is (for RNA): (C/U)less than 10N(C/T)AG/G, where most introns end in AG after a long stretch of pyrimidines. The branch site within introns (area of lariat formation close to the acceptor site during splicing) has the consensus sequence UAUAAC. In most cases, U can be replaced by C and A can be replaced by G. However, the penultimate (**A**) residue is fully conserved (invariant).

Group I introns require an external guanosine nucleotide as a cofactor. The 3'-OH of the guanosine nucleotide acts as a nucleophile to attack the 5'-phosphate of the intron's 5' nucleotide. The 3' end of the 5' exon is termed the splice donor site. The 3'-OH at the 3' splice donor end of the 5' exon next attacks the splice acceptor site at the 5' nucleotide of the 3' exon, releasing the intron and covalently attaching the two exons together.

Pre-mRNA processing takes place in the nucleus of eukaryotes, whereas lack of a nuclear membrane in prokaryotes permits initiation of translation while transcription is not yet complete.

Pre-mRNA processing events include capping of the 5' end on the pre-mRNA, pre-mRNA splicing to remove intronic sequences, and polyadenylation of the 3' end of the pre-mRNA.



### **SPLICEOSOME MACHINERY**

The spliceosome has been described as one of "the most complex macromolecular machines known," "composed of as many as 300 distinct proteins and five RNAs" (Nilsen, 2003). The

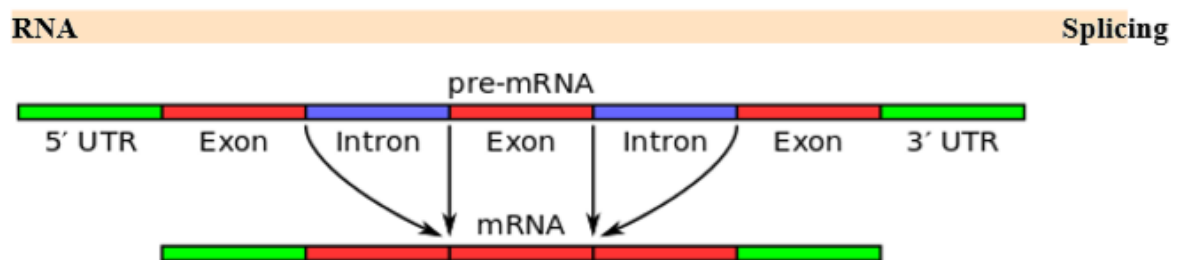


animation above reveals this astonishing machine at work on the precursor mRNA, cutting out the non-coding introns and splicing together the protein-coding exons.

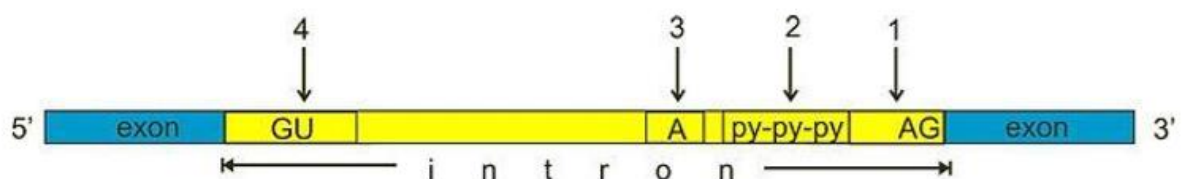
A spliceosome is a large and complex molecular machine found primarily within the splicing speckles of the cell nucleus of eukaryotic cells. The spliceosome is assembled from snRNAs and protein complexes. The spliceosome removes introns from a transcribed pre-mRNA, a kind of primary transcript. This process is generally referred to as splicing. Only eukaryotes have spliceosomes and metazoans have a second spliceosome, the minor spliceosome.

Introns (which, unlike exons, do not code for proteins) can be of considerable length in higher eukaryotes, even spanning many thousands of bases and sometimes comprising some 90% of the precursor mRNA. In contrast, lower eukaryotes such as yeast possess fewer and shorter introns, which are typically fewer than 300 bases in length. Since introns are the non-coding segments of genes, they are removed from the mRNA before it is translated into a protein. This is not to say, of course, that introns are without important function in the cell (as I discuss here).

Comprising the spliceosome, shown at right (excerpted from Frankenstein et al., 2012) are several small nuclear ribonucleoproteins (snRNPs) -- called U1, U2, U4, U5 and U6 -- each of which contains an RNA known as an snRNA (typically 100-300 nucleotides in length) -- and many other proteins that each contribute to the process of splicing by recognizing sequences in the mRNA or promoting rearrangements in spliceosome conformation. The spliceosome catalyzes a reaction that results in intron removal and the "gluing" together of the protein-coding exons.



The first stage in RNA splicing is recognition by the spliceosome of splice sites between introns and exons. Key to this process are short sequence motifs. These include the 5' and 3' splice sites (typically a GU and AG sequence respectively); the branch point sequence (which contains a conserved adenosine important to intron removal); and the polypyrimidine tract (which is thought to recruit factors to the branch point sequence and 3' splice site). These sequence motifs are represented in the illustration below:



The U1 snRNP recognizes and binds to the 5' splice site. The branch point sequence is identified and bound by the branch-point-binding protein (BBP). The 3' splice site and polypyrimidine tract are recognized and bound by two specific components of a protein complex called U2 auxiliary factor (U2AF): U2AF35 and U2AF65 respectively.

Once these initial components have bound to their respective targets, the rest of the spliceosome assembles around them. Some of the previously bound components are displaced at this stage: For instance, the BBP is displaced by the U2 snRNP, and the U2AF complex is displaced by a complex of U4-U5-U6 snRNPs. The U1 and U4 snRNPs are also released. The first transesterification reaction then takes place, and a cut is made at the 5' splice site and the 5' end of the intron is subsequently connected to the conserved adenine found in the branch point sequence, forming the so-called "lariat" structure. This is followed by the second transesterification reaction which results in the splicing together of the two flanking exons. See this page for a helpful animation of the splicing process.

### **Other important protein factors:**

Many other proteins play crucial roles in the RNA splicing process. One essential component is PRP8, a large protein that is located near the catalytic core of the spliceosome and that is involved in a number of critical molecular rearrangements that take place at the active site (for a review, see Grainger and Beggs, 2005). What is interesting is that this protein, though absolutely crucial to the RNA splicing machinery, bears no obvious homology to other known proteins.

The SR proteins, characterized by their serine/arginine dipeptide repeats and which are also essential, bind to the pre-mRNA and recruit other spliceosome components to the splice sites (Lin and Fu, 2007). SR proteins can be modified depending on the level of phosphorylation at their serine residues, and modulation of this phosphorylation helps to regulate their activity, and thus coordinate the splicing process (Saitoh et al., 2012; Plocinik et al., 2011; Zhong et al., 2009; Misteli et al., 1998). The illustration above (from here) shows the binding of SR proteins to splicing enhancer sites, which promotes the binding of U1 snRNP to the 5' splice site, and U2AF protein to the polypyrimidine tract and 3' splice site.

There are also ATPases that promote the structural rearrangements of snRNAs and release by the spliceosome of mRNA and the intron lariat. It is even thought that ATP-dependent RNA helicases play a significant role in "proofreading" of the chosen splice site, thus preventing the potentially catastrophic consequences of incorrect splicing (Yang et al., 2013; Semlow and Staley, 2012; Egecioglu and Chanfreau, 2011).

### **Composition**

Each spliceosome is composed of five small nuclear RNAs (snRNA), and a range of associated protein factors. When these small RNA are combined with the protein factors, they make an RNA-protein complex called snRNP.

The snRNAs that make up the major spliceosome are named U1, U2, U4, U5, and U6, and participate in several RNA-RNA and RNA-protein interactions. The RNA component of the

small nuclear ribonucleic protein or snRNP (pronounced "snurp") is rich in uridine (the nucleoside analog of the uracil nucleotide).

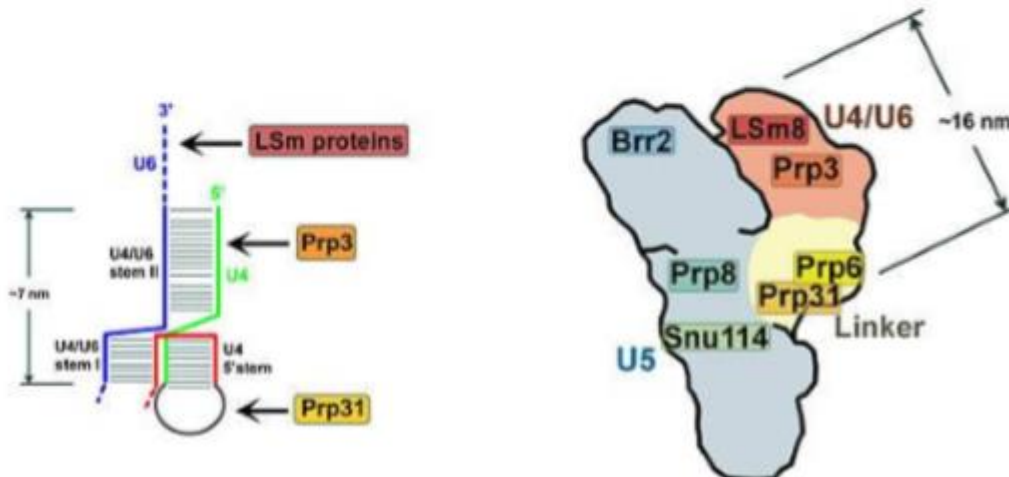
The canonical assembly of the spliceosome occurs anew on each hnRNA (pre-mRNA). The hnRNA contains specific sequence elements that are recognized and utilized during spliceosome assembly. These include the 5' end splice, the branch point sequence, the polypyrimidine tract, and the 3' end splice site. The spliceosome catalyzes the removal of introns, and the ligation of the flanking exons.

Introns typically have a GU nucleotide sequence at the 5' end splice site, and an AG at the 3' end splice site. The 3' splice site can be further defined by a variable length of polypyrimidines, called the polypyrimidine tract (PPT), which serves the dual function of recruiting factors to the 3' splice site and possibly recruiting factors to the branch point sequence (BPS). The BPS contains the conserved Adenosine required for the first step of splicing.

A group of less abundant snRNAs, U11, U12, U4atac, and U6atac, together with U5, are subunits of the so-called minor spliceosome that splices a rare class of pre-mRNA introns, denoted U12-type. The minor spliceosome is located in the nucleus like its major counterpart, though there are exceptions in some specialised cells including anucleate platelets and the dendroplasm of neuronal cells.

New evidence derived from the first crystal structure of a group II intron suggests that the spliceosome is actually a ribozyme, and that it uses a two-metal ion mechanism for catalysis.

In addition, many proteins exhibit a zinc-binding motif, which underscores the importance of zinc metal in the splicing mechanism.



Above are electron microscopy fields of negatively stained yeast (*Saccharomyces cerevisiae*) tri-snRNPs. Below left is a schematic illustration of the interaction of tri-snRNP proteins with the U4/U6 snRNA duplex. Below right is a cartoon model of the yeast tri-snRNP with shaded areas corresponding to U5 (gray), U4/U6 (orange) and the linker region (yellow).

### **Alternative splicing**

Alternative splicing (the re-combination of different exons) is a major source of genetic diversity in eukaryotes. Splice variants have been used to account for the relatively small number of genes in the human genome. For years the estimate widely varied, with top estimates reaching 100,000 genes, but now, due to the Human Genome Project, the figure is believed to be closer to 20,000 genes. One particular *Drosophila* gene (*Dscam*, the *Drosophila* homolog of the human Down syndrome cell adhesion molecule *DSCAM*) can be alternatively spliced into 38,000 different mRNA.

### **The Exon Junction Complex**

The exon junction complex (EJC) is a protein complex comprised of several protein components (RNPS1, Y14, SRm160, Aly/REF and Magoh) left behind near splice junctions by the splicing process (Hir and Andersen, 2008). Their function is to mark the transcript as processed, and thus ready for export from the nucleus to the cytoplasm, and translation at the ribosome. The EJC is typically found 20 to 24 nucleotides upstream of the splice junction.

The EJC also plays an important role in nonsense mediated decay, a surveillance system used in eukaryotes to destroy transcripts containing premature stop codons (Trinkle-Mulcahy et al., 2009; Chang et al., 2007; Gehring et al., 2005). Upon encountering an EJC during translation, the ribosome displaces the complex from the mRNA. The ribosome then continues until it reaches a stop codon. If, however, the mRNA contains a stop codon before the EJC, the nonsense mediated decay pathway is triggered. The EJC and its position thus contribute to transcript quality control.

### **The Evolution of the Spliceosome**

A popular hypothesis regarding the origins of the spliceosome is that its predecessor was self-splicing RNA introns (e.g. Valadkhan, 2007). Such a hypothesis makes sense of several observations. For example, a simpler way to achieve splicing presumably would be to bring the splice sites together in one step to directly cleave and rejoin them. The proposed scenario, however, would explain the use of a lariat intermediate, since a lariat is generated by group II RNA intron sequences (Lambowitz1 and Zimmerly, 2011; Vogel and Borner, 2002).

The hypothesis also helps to clarify why RNA molecules play such an important part in the splicing process. Examples of self-splicing RNA introns still exist today (e.g., in the nuclear rRNA genes of the ciliate *Tetrahymena*) (Hagen and Cech, 1999; Price et al., 1995; Price and Cech, 1988; Kruger et al., 1982).

These observations may be taken as evidence as to the spliceosome's evolutionary predecessor, but they are hardly helpful in elucidating a plausible scenario for transitioning from one to the

other. The spliceosome machinery is far more complex and sophisticated than autocatalytic ribozymes, involving not just five RNAs but hundreds of proteins.

## LESSON 76

### What is Genomics?

**Genome:** Total amount of DNA of a single cell of an organism (haploid cell in the case of a diploid) is called as genome. The whole hereditary information of an organism encoded by DNA is called as genome. Determination of entire genome sequence is a prerequisite to understand the complete biology of an organism. Genomics include generating physical, genetic, and sequencing maps of different genomes. It also includes sequencing the genomes of model organisms and developing new technologies for mapping/sequencing. Genomics helps in different ways such as in Functions of genes, Organizations of genomes, Structural make-up of genomes, Functions of coding and non-coding DNA. Study of Genomics helps to understand the ethical, social, and legal issues and challenges posted by genomic information.

### Genomics Sub-disciplines

- Comparative genomics
- Structural genomics
- Functional genomics
- Population genomics
- Metagenomics
- Microbial genomics

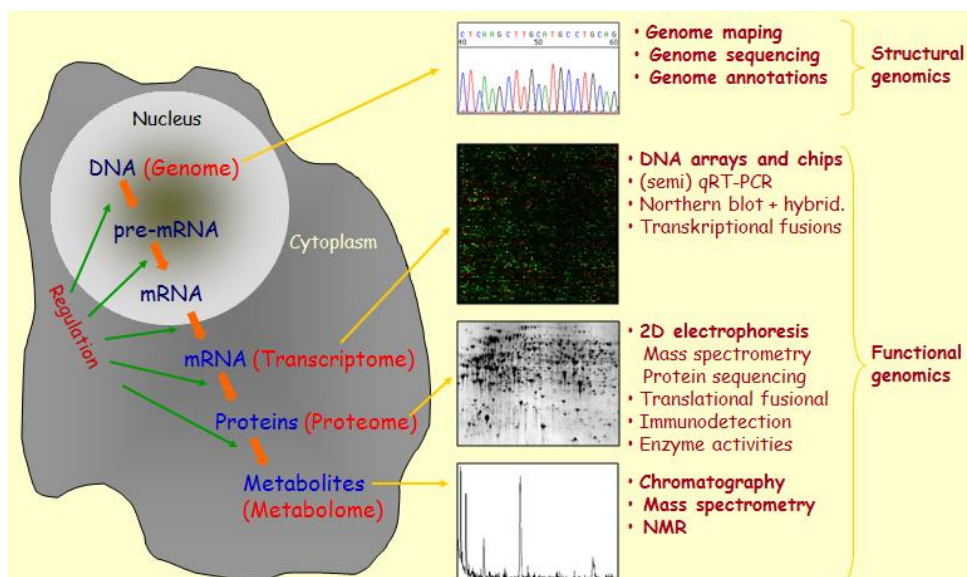
## LESSON 77

**Functional Genomics** includes Functions of genes, their regulation and end products. Functional genomics analyzes all genes in genomes to determine their functions and their gene control and expression. Classically, genetics analysis begins with a phenotype and moves for identification of genes. New approaches are needed to work in the opposite direction, from genes to phenotype.

**Functional Genomics relies on Molecular Biology, Biochemistry, Genetics and Bioinformatics tools:** Functional genomics relies on molecular biology lab research and sophisticated computer analysis by bioinformatics tools. Fusion of biology with maths and computer science is used for many things. Examples: Finding genes within a genomic sequence. Aligning DNA/proteins sequences.

Functional Genomics includes following

- Subtracted cDNA libraries
- Differential display
- Representational difference analysis
- Suppression subtractive hybridization
- cDNA Microarrays
- Serial analysis of gene expression
- 2-D Gel electrophoresis



## LESSON 78

**Structural Genomics:** The ultimate goal of genomic studies is to determine the nucleotide sequences of entire genomes of organisms. It also includes the genetic and physical mapping and sequencing of chromosomes.

**Genetic Mapping:** This includes approximate locations of genes, relative to the locations of other genes, based on the rates of recombination.

**Physical Mapping** is based on the direct analysis of DNA. Physical mapping places genes on the genomes in relation to distances measured in bp, kbp, and mbp.

**Structural Genomics:** these include

Distinct components of genomes

Abundance and complexity of mRNA

Genome sequences

Gene numbers

Coding and non-coding DNA

**Structural Genomics – Complex Genomes** have roughly 10x to 30x more DNA than is required to encode all the RNAs or proteins in the organism

**Structure of Complex Genomes:** This consists of following

Introns in genes

Regulatory elements of genes

Multiple copies of genes, including pseudogenes

Intergenic sequences

Interspersed repeats

### **Structural Genomics – Transposable Elements**

Vast majority of TEs can be classified into four families:

LINEs (Long Interspersed Nuclear Elements, autonomous)


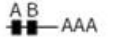

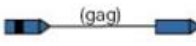


SINEs (Short Interspersed Nuclear Elements, use LINE proteins for life cycle)

LTR elements (Long Terminal Repeats; derived from retroviruses)

DNA transposons (replicate without RNA intermediary)



## Structural Genomics – Repetitive DNA in Humans

			Length	Copy number	Fraction of genome
LINEs	Autonomous		6–8 kb	850,000	21%
	Non-autonomous		100–300 bp		
Retrovirus-like elements	Autonomous		6–11 kb	450,000	8%
	Non-autonomous		1.5–3 kb		
DNA transposon fossils	Autonomous		2–3 kb	300,000	3%
	Non-autonomous		80–3,000 bp		

## LESSON 79

**Comparative Genomics:** This can be defined as comparison of gene numbers, gene locations and biological functions of genes, in the genomes of different organisms. This also includes to identify and compare groups of genes that plays a unique biological role in different organisms

**Homology** is the relationship of any two characters (genes or proteins) that have descended, usually through divergence, from a common ancestor/ ancestral character.

**Homologues** are thus components or characters (such as genes/proteins with similar sequences) that can be attributed to a common ancestor of the two organisms during evolution.

**Homologues can be** Orthologues, Paralogues, Xenologues, Analogues

**Orthologues** are homologues that have evolved from a common ancestral gene by speciation. They usually have similar functions

**Paralogues** are homologues that are related or produced by duplication within a genome followed by subsequent divergence. They often have different functions.

**Xenologues** are homologous that are related by an interspecies (horizontal transfer) of the genetic material for one of the homologues. The functions of the xenologues are quite often similar.

**Analogues** are non-homologues genes/proteins that have descended convergently from an unrelated ancestor. Analogues have similar function, different sequence or structure

## LESSON 80

**Population Genomics:** Study of genomes of a specific population, strains, varieties or organisms is called as population genetics. This is the study about the genetic diversity. It includes understanding new insights into disease and drug response with deeper insights about genomic size and further complexities in the genomes of the organisms. So this is the variations between individuals or strains. Human genome is 200 times larger than yeast but 200 times smaller than Amoeba. Less than 2% of human genome is coding sequence.

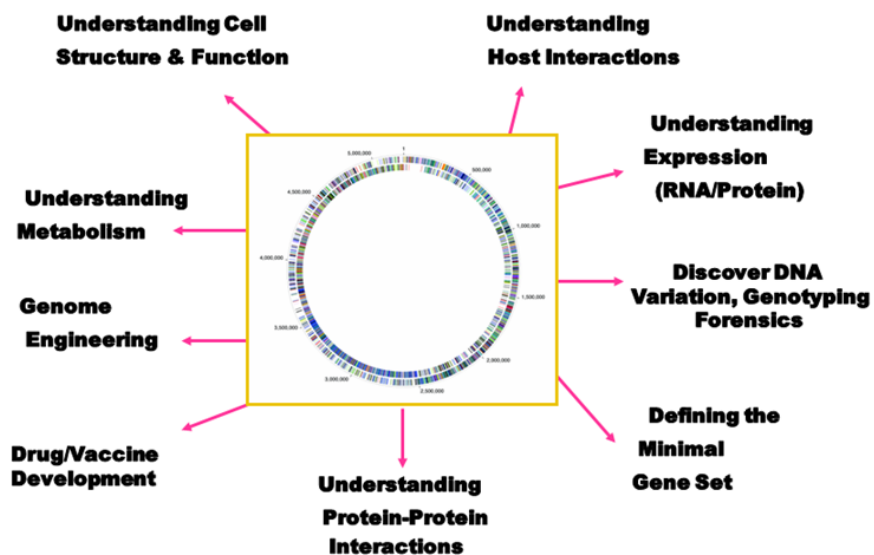
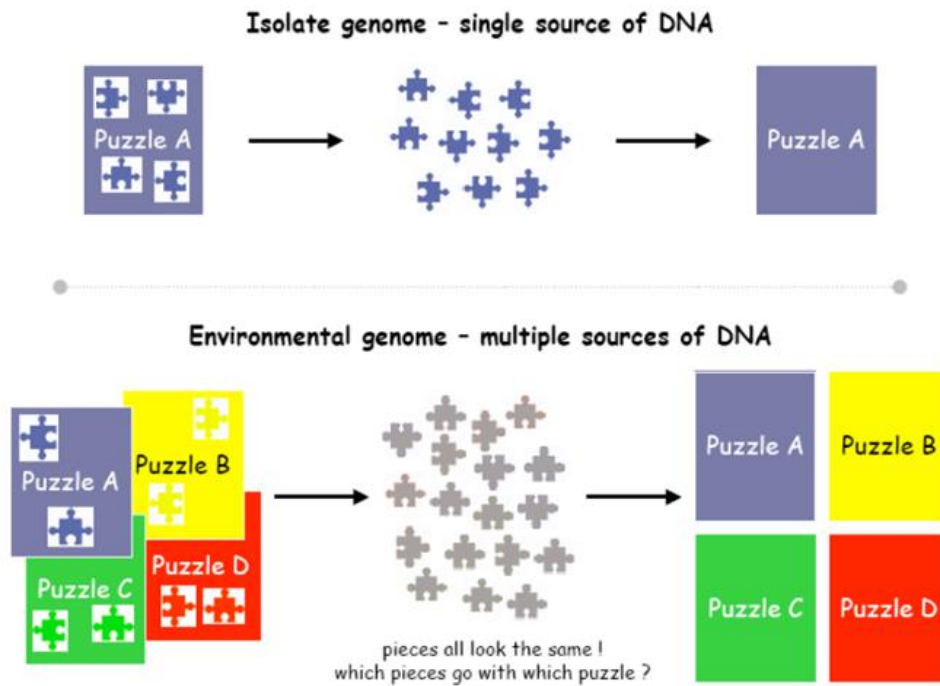
**The 1000 Genomes - Population Genomics** International research consortium sequenced the genomes of at least 1000 people from around the world. Detailed and medically useful pictures of human genome variations. Any two humans are more than 99% identical at genetic level. Genetic variations may explain individual differences in susceptibility to diseases, responses to drugs.

**HapMap Project - Population Genomics:** The HapMap project has already discovered many regions of the genome containing genetic variations associated with common human diseases.

**Goals of Population Genomics:** Followings are the goals of population genomics such as to Produce a catalog of variants present at 1% or greater frequency in the human population. Down to 0.5 percent or lower within genes. Increase sensitivity of disease discovery Provide better understanding of very rare genetic diseases. Understand contribution of common variants to common diseases like diabetes and heart diseases. Identify SNP but also large differences like rearrangements, deletions or duplications

# LESSON 81

**Metagenomics** Metagenome - Environmental genome: This can be defined as collection of genes sequenced from the environment could be analyzed in a way analogous to the study of a single genome.



**Objectives of Metagenomics:** Followings are the objectives of metagenomics

- Examining phylogenetic diversity using 16s rRNA

- Diversity patterns of microorganisms for monitoring and predicting environmental conditions/change.
- Examining genes/operons for desirable enzymes (cellulases, lipases, antibiotics, other natural products).
- Exploited for industrial or medical applications.
- Examining secretory, regulatory, and signal transduction mechanisms associated with samples or genes of interest.
- Examining bacteriophage or plasmid sequences. These potentially influence diversity and structure of microbial communities.
- Examining potential lateral gene transfer events. Knowledge of genome plasticity may give us an idea of selective pressures for gene capture and evolution within a habitat.
- Examining metabolic pathways.
- Directed approach towards designing culture media.
- Examining genes that predominate in a given environment compared to others.
- Metagenomics data can be used towards designing low and high throughput experiments focused on defining the roles of genes and microorganisms in the establishment of dynamic microbial community.

## LESSON 82

### **Genetics and Genomics - Difference**

Genetics is the study of heredity, or how the characteristics of living organisms are transmitted from one generation to the next generation through DNA. Genetics involves the study of specific and limited numbers of genes that have a known function. Genetics deals that how genes guide the body's development, cause disease or affect response to drugs.

**Genomics** in contrast, is the study of the entirety of an organism's genes – called the genome

Using high-performance computing and math techniques known as bioinformatics, genomics analyzes enormous amounts of DNA sequence data to find variations. Genomics particularly deals with genetic variants that affect health, disease or drug response. In humans that means searching through about 3 billion units of DNA across 23,000 genes Genomics is a much newer field than genetics and became possible only in the last couple of decades due to technical advances in DNA sequencing and computational biology.

Genetics: How the characteristics of living organisms are transmitted from one generation to the next generation. Genomics: study of the entirety of an organism's genes – called the genome

## LESSON 83

### **Genomics, Proteomics and Metabolomics**

**Genome and Genomics:** The complete set of DNA found in each cell is known as the genome and study is called as genomics.

**Proteome and Proteomics:** The complete set of proteins found in each cell is known as the proteome. Proteins concentration (and activity) may be different than gene expression due to post-translational modification

**Metabolomics:** The complete set of metabolites found in each cell is known as the metabolome. Use of high-throughput mass spectrometry is used to analyze the metabolic components of cell. Metabolomics is useful for determining the effects of the environment or gene transformation on the metabolism of the plants/animals.

## LESSON 84

**Why Sequence Genomes:** Sequencing genomes is necessary to identify gene numbers, their locations on genomes, and to study their functions, Genes regulation, DNA sequence , Genome organization, Chromosomal structure and organization , Noncoding DNA types, amount, distribution and functions, Coordination of gene expression, protein synthesis, and post-translational events, Interaction of proteins in complex molecular machines, Predicted vs experimentally determined gene function, Evolutionary conservation, Proteins structure and function, Proteomes (total protein content and function) in organisms, Correlation of SNPs with health and disease, Disease-susceptibility prediction based on gene sequence variation, Genes involved in complex traits and multigene diseases

### **Novel Diagnostics**

Complex systems biology, developmental genetics

To provide platform for microchips and DNA microarrays

Gene expression - RNA

Complex systems biology, developmental genetics and genomics

### **Novel Therapeutics**

Drug target discovery

Rational drug design

Molecular docking

Gene therapy

Stem cell therapy

**Understanding Metabolism** of cells and tissues within different organisms.

**Understanding mechanism of diseases:** Inherited diseases, Infectious diseases, Pathogenic bacteria and Viruses.



## LESSON 85

**Major Techniques used for Genomes Characterization:** Followings are the techniques which are used in genome characterization.

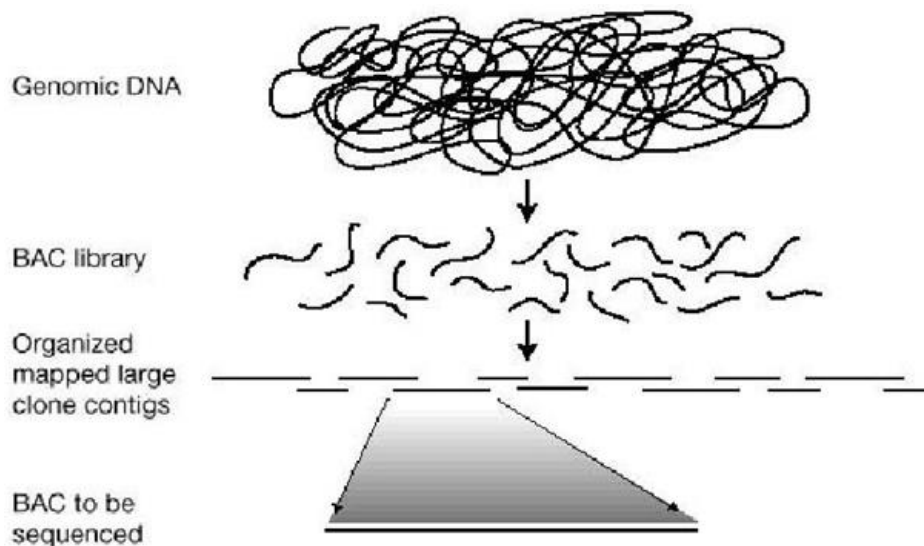
- Cloning
- Hybridization
- PCR amplification
- Sequencing
- Computational tool

### Genomes Characterization Techniques - Cloning

Genomes digested with restriction enzymes and inserted in vectors to produce genomic libraries.

BACs and YACs

**Genomes Characterization Techniques – Hybridization:** To arrange large contigs of genomes to produce genetic maps and physical maps of genomes.



**Genomes Characterization Techniques – PCR:** Technique to amplify the DNA. Different variants of the technique used

**Genomes Characterization Techniques – DNA Sequencing:** One of the important techniques used to characterize the genomes. Used to study structure and function of genomes.

**Genomes Characterization Techniques – Computational Tools:** Used to align the sequenced DNA to produce physical maps of the genomes.

## LESSON 86

**Steps of Genomes Analysis:** Followings are the steps which are used in genome analysis.

- Genome sequence assembled
- Identify repetitive sequences – mask out
- Gene prediction – train a model for each genome
- Look for EST and cDNA sequences
- Genome annotation
- Microarray analysis
- Metabolic pathways and regulation
- Protein 2D gel electrophoresis
- Functional genomics
- Gene location/gene map
- Self-comparison of proteome
- Comparative genomics
- Identify clusters of functionally related genes
- Evolutionary modeling

## LESSON 87

**Benefits of Genomes Research:** Followings are the benefits of genome research

Genomes Research – Molecular Medicine

- Improve diagnosis of disease
- Detect genetic predispositions to disease (cancer, diabetes etc)
- Create drugs based on molecular information
- Use gene therapy and control systems as drugs

**Genomes Research – Risk Assessment** Evaluate the health risks faced by individuals who may be exposed to radiations and to cancer causing chemicals and toxins.

**Genomes Research – Bioarcheology, Anthropology, Evolution and Human Migration:** Study evolution through genetic variants in lineages.

Study of migration of different populations

Study mutations on the Y chromosome to trace lineage and migration of males and Evolution of mutations with ages of populations.

**Genomes Research – DNA Forensics**

- Identify potential suspects whose DNA may match evidence left at crime scenes.
- Exonerate persons wrongly accused of crimes.
- Identify catastrophe victims.
- Establish paternity and other family relationships.
- Identify endangered and protected species as an aid to wildlife officials
- Detect bacteria and other organisms that may pollute air, water, soil and food.
- Match organ donors with recipients in transplant programs
- Determine pedigree for seed or livestock

**Genomes Research – Disease-resistant crops and disease-resistant animals**

- Grow disease/insect resistant and drought-resistant crops.
- Breed healthier, more productive, disease-resistant farm animals.

**Genomes Research – Agriculture, Livestock Breeding, and Bioprocessing**

- Develop bio pesticides.

- Incorporate edible vaccines incorporated into food products.

### **Genomes Research – Microbial Genomics**

- Rapidly detect and treat pathogens (disease-causing microbes).
- Develop new energy sources (biofuels)
- monitor environment to detect pollutants
- Protect populations from biological and chemical warfare
- Clean up toxic waste safely and efficiently

## LESSON 78

**Genes and Size of Genomes:** Genomes of most bacteria and archaea range from 1 to 6 million base pairs (Mb). Genomes of eukaryotes are usually large. Most plants and animals have genomes greater than 100 Mb. Humans have genome size of 3,000 Mb. Within each domain there is no systematic relationship between genome size and phenotype

Organism	Haploid Genome Size (Mb)	Number of Genes	Genes per Mb
<b>Bacteria</b>			
<i>Haemophilus influenzae</i>	1.8	1,700	940
<i>Escherichia coli</i>	4.6	4,400	950
<b>Archaea</b>			
<i>Archaeoglobus fulgidus</i>	2.2	2,500	1,130
<i>Methanosarcina barkeri</i>	4.8	3,600	750

Organism	Haploid Genome Size (Mb)	Number of Genes	Genes per Mb
<b>Eukaryotes</b>			
<i>Saccharomyces cerevisiae</i> (yeast, a fungus)	12	6,300	525
<i>Caenorhabditis elegans</i> (nematode)	100	20,100	200
<i>Arabidopsis thaliana</i> (mustard family plant)	120	27,000	225
<i>Drosophila melanogaster</i> (fruit fly)	165	13,700	83
<i>Oryza sativa</i> (rice)	430	42,000	98
<i>Zea mays</i> (corn)	2,300	32,000	14
<i>Mus musculus</i> (house mouse)	2,600	22,000	11
<i>Ailuropoda melanoleuca</i> (giant panda)	2,400	21,000	9

## LESSON 89

### Viral Genomes

**Genomes of Viruses:** A viral genome is the genetic material of the virus also termed as viral chromosome. Viral genomes vary in size -few thousand to more than a hundred thousand nucleotides. Viral genomes can be

ssRNA	dsRNA
ssDNA	dsDNA
Linear	Circular

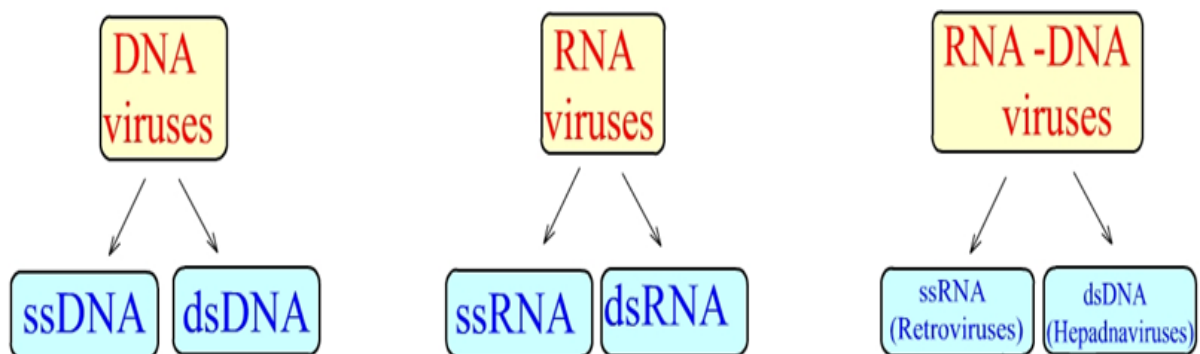
**Viruses with RNA Genomes:** Almost all plants viruses and some bacterial and animal viruses Genomes are rather small (a few thousands nucleotides)

**Viruses with DNA Genomes:** Often a circular genome lambda = 48,502 bp

**Replicative form of Viral Genomes:** All ssRNA viruses produce dsRNA molecules. Many linear DNA molecules become circular

**Viruses and Kingdoms:** Many plants viruses contain ssRNA genomes. Many fungal viruses contain dsRNA genomes. Many bacterial viruses contain dsDNA genomes.

**Genomes in Virions:** The genomes of viruses can be composed of either DNA or RNA, and some use both as their genomic material at different stages in their life cycle. However, only one type of nucleic acid is found in the virion of any particular type of virus.



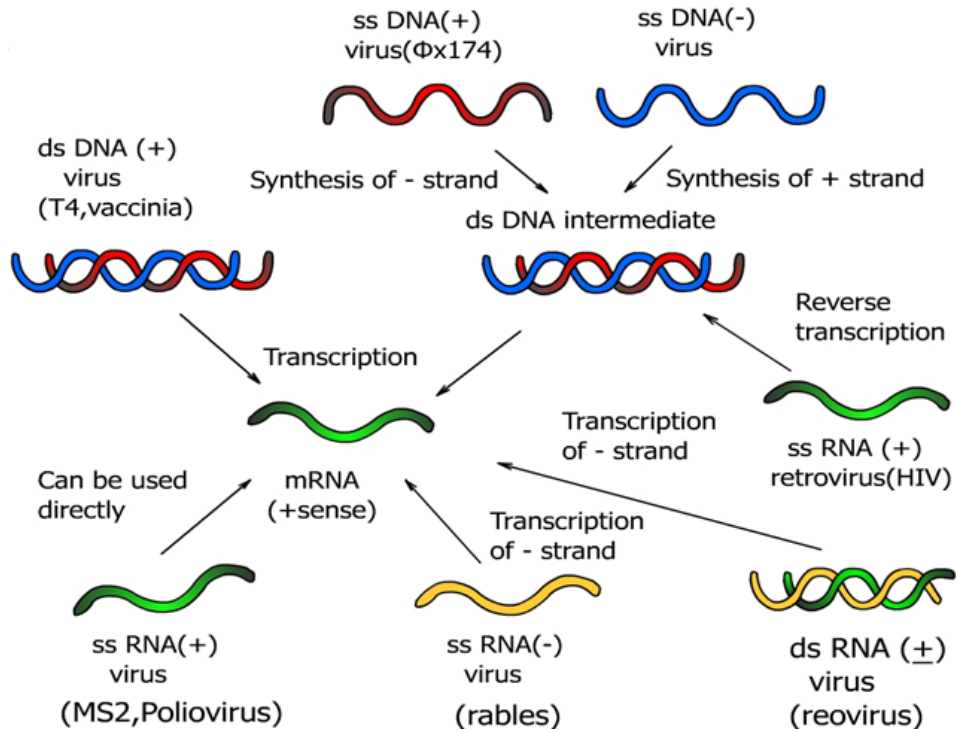
Virus	Host	Type of Nucleic Acid	Number of Genes
Parvovirus	Mammals	ssDNA	5
Phage fd	<i>E. coli</i>	ssDNA	10
Lambda	<i>E. coli</i>	dsDNA	36
T4	<i>E. coli</i>	dsDNA	>190
Q $\beta$	<i>E. coli</i>	ssRNA	4
TMV	Many plants	ssRNA	6
Influenza virus	Mammals	ssRNA	12

### Viruses and Number of Genes

Virus	Genome structure	Genome size (kb)	Number of genes
Adenovirus	Double-stranded linear DNA	36.0	30
Hepatitis B	Partly double-stranded circular DNA	3.2	4
Influenza virus	Single-stranded segmented linear RNA	22.0	12
Parvovirus	Single-stranded linear DNA	1.6	5
Poliovirus	Single-stranded linear RNA	7.6	8
Reovirus	Double-stranded segmented linear RNA	22.5	22
Retroviruses	Single-stranded linear RNA	6.0–9.0	3
SV40	Double-stranded circular DNA	5.0	5
Tobacco mosaic virus	Single-stranded linear RNA	6.4	6
Vaccinia virus	Double-stranded circular DNA	240	240

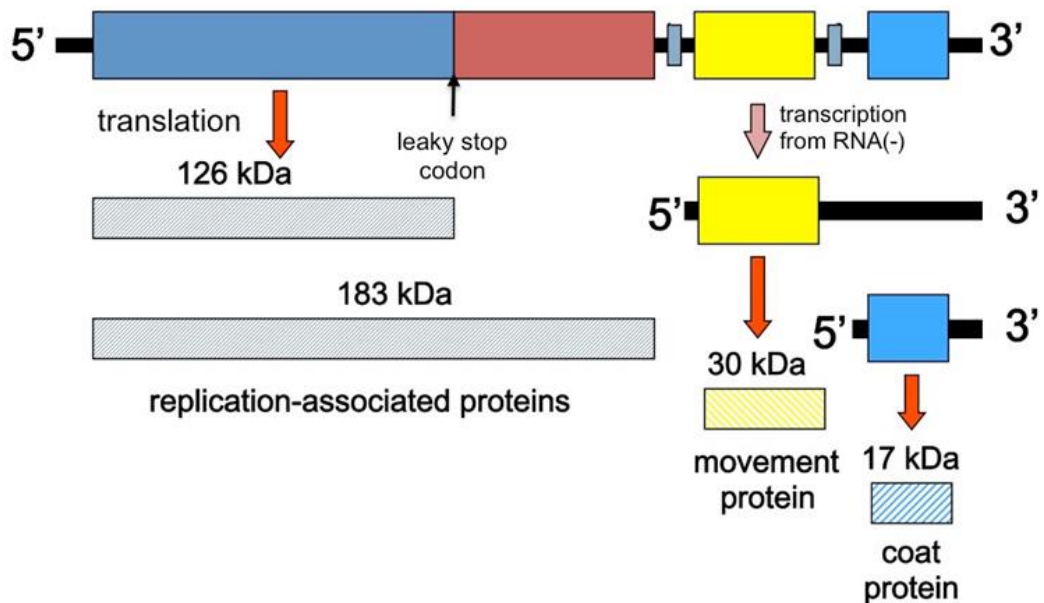


## Genomes in Virions



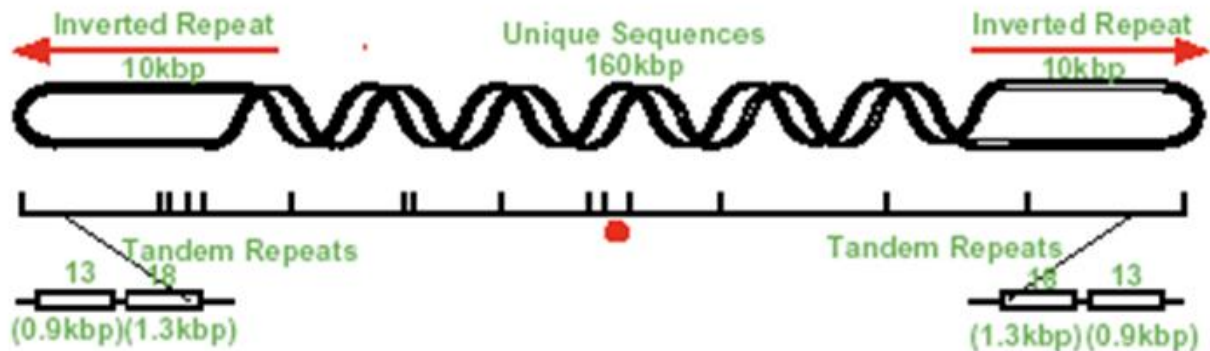
## Genome of Tobacco Mosaic Virus

- Single, 6400 nucleotides RNA, 3 Essential Genes

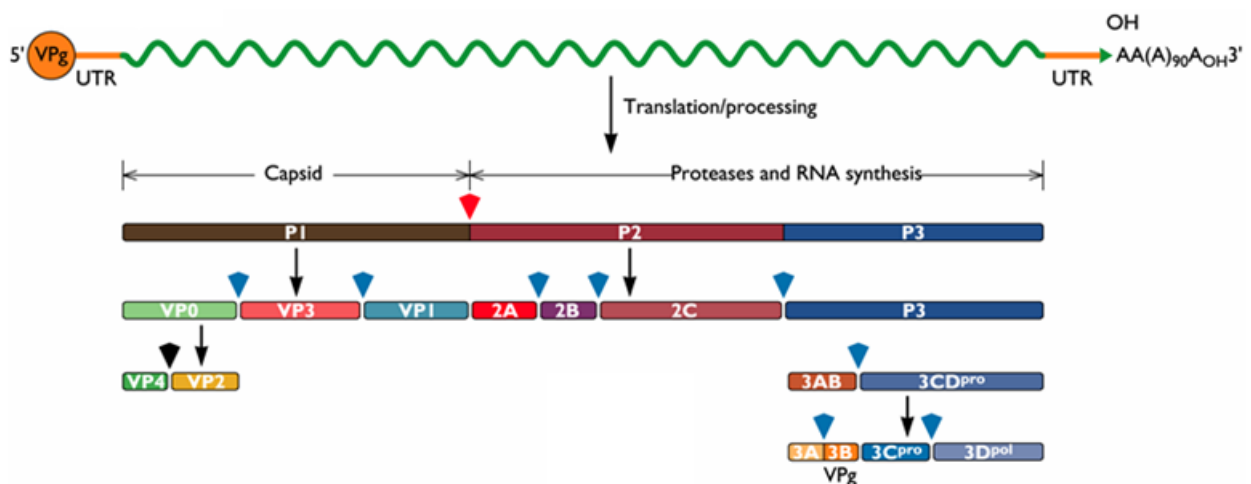


## Genome of Poxvirus –A typical large dsDNA Virus

180 kb DNA, >100 Essential Genes



**Genome of Polio Virus:** Single-stranded positive-sense RNA genome that is about 7500 nucleotides long



## Genome of Pox Virus

Linear dsDNA 130-375 kbp; covalently closed termini. Large hairpin structure at each terminus - up to 10 kb total at each end is repeat sequence. Encode 150-300 proteins. Coding regions are closely spaced, no introns. Coding regions are on both strands of genome, and are not tightly clustered with respect to time of expression or function.

## LESSON 90

**Bacterial Genomes:** Small organisms carry high coding density (85-90%). 1 gene per 1000 bases in prokaryotes. Large variation in genome size between bacteria

### Genomes of Bacteria – Large Variation

*Tremblaya princeps* 140kb, 121 coding sequences

*Sorangium cellulosum*

14000kb

11599 coding sequences

### Comparison of regulatory genes in bacterial genomes

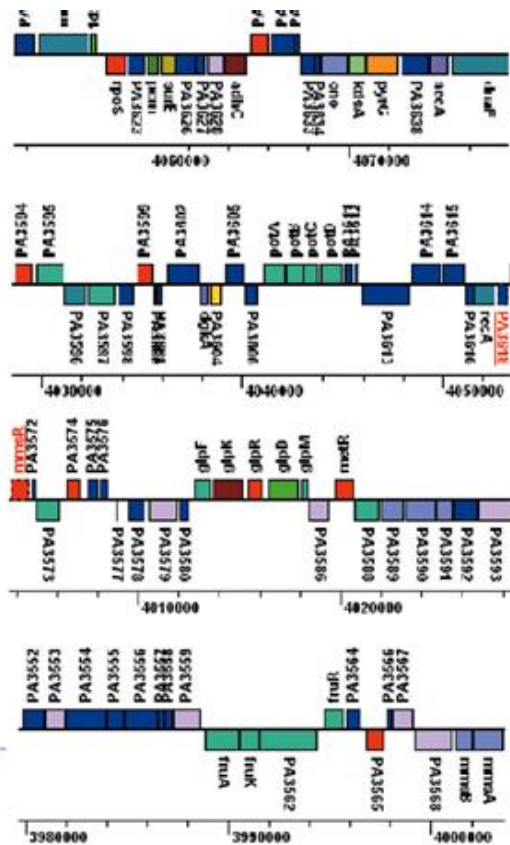
Microorganism	# Genes in the Genome	# Regulatory Proteins	% of Total
<i>Pseudomonas aeruginosa</i>	5570	468	8.4
<i>Escherichia coli</i>	4289	250	5.8
<i>Bacillus subtilis</i>	4100	217	5.3
<i>Mycobacterium tuberculosis</i>	3918	117	3.0
<i>Helicobacter pylori</i>	1566	18	1.1

### Distribution of genes among selected bacterial genomes and their sizes

Organism	Genome Size (Mbp)	No. of ORFs (% coding)	Unknown Function	Unique ORFs
<i>Aeropyrum pernix</i> K1	1.67	1,885 (89%)		
<i>A. aeolicus</i> VF5	1.50	1,749 (93%)	663 (44%)	407 (27%)
<i>A. fulgidus</i>	2.18	2,437 (92%)	1,315 (54%)	641 (26%)
<i>B. subtilis</i>	4.20	4,779 (87%)	1,722 (42%)	1,053 (26%)
<i>B. burgdorferi</i>	1.44	1,738 (88%)	1,132 (65%)	682 (39%)
<i>Chlamydia pneumoniae</i> AR39	1.23	1,134 (90%)	543 (48%)	262 (23%)
<i>Chlamydia trachomatis</i> MoP <sub>n</sub>	1.07	936 (91%)	353 (38%)	77 (8%)
<i>C. trachomatis</i> serovar D	1.04	928 (92%)	290 (32%)	255 (29%)
<i>Deinococcus radiodurans</i>	3.28	3,187 (91%)	1,715 (54%)	1,001 (31%)
<i>E. coli</i> K-12-MG1655	4.60	5,295 (88%)	1,632 (38%)	1,114 (26%)
<i>H. influenzae</i>	1.83	1,738 (88%)	595 (35%)	237 (14%)
<i>H. pylori</i> 26695	1.66	1,589 (91%)	744 (45%)	539 (33%)
<i>Methanobacterium thermotautotrophicum</i>	1.75	2,008 (90%)	1,010 (54%)	496 (27%)

Organism	Genome Size (Mbp)	No. of ORFs (% coding)		Unknown Function		Unique ORFs	
<i>Methanococcus jannaschii</i>	1.66	1,783	(87%)	1,076	(62%)	525	(30%)
<i>M. tuberculosis</i> CSU#93	4.41	4,275	(92%)	1,521	(39%)	606	(15%)
<i>M. genitalium</i>	0.58	483	(91%)	173	(37%)	7	(2%)
<i>M. pneumoniae</i>	0.81	680	(89%)	248	(37%)	67	(10%)
<i>N. meningitidis</i> MC58	2.24	2,155	(83%)	856	(40%)	517	(24%)
<i>Pyrococcus horikoshii</i> OT3	1.74	1,994	(91%)	589	(42%)	453	(22%)
<i>Rickettsia prowazekii</i> Madrid E	1.11	878	(75%)	311	(37%)	209	(25%)
<i>Synechocystis</i> sp.	3.57	4,003	(87%)	2,384	(75%)	1,426	(45%)
<i>T. maritima</i> MSB8	1.86	1,879	(95%)	863	(46%)	373	(26%)
<i>T. pallidum</i>	1.14	1,039	(93%)	461	(44%)	280	(27%)
<i>Vibrio cholerae</i> El Tor N1696	4.03	3,890	(88%)	1,806	(46%)	934	(24%)
	50.60	52,462	(89%)	22,358	(43%)	12,161	(23%)

### Genes in a portion of a bacterial genome



## LESSON 91

**Yeast Genome:** The nuclear genome consists of 16 chromosomes. In addition, there is a mitochondrial genome and a plasmid, 2 micron circle. The haploid yeast genome consists of ~ 12.1 Mb. Yeast genome was completely sequenced by 1996.

**Yeast Genome – Characteristics:** Followings are the characteristics of yeast genome.

Small and compact with small intergenic séquences

Few transposable elements with few introns

Limited RNA interference

The yeast genome is predicted to contain about 6,200 genes and 274 tRNA and 287 introns

Small percentage of yeast genes have introns. The intergenic space between genes is only between 200bp - 1,000bp

The largest known regulatory sequences are spread over about 2,800bp , MUC1/FLO11

Yeast genes have names consisting of three letters and up to three numbers GPD1, HSP12, PDC6 . Usually they are meaningful

### Yeast Genome: Genome of Yeast Cell

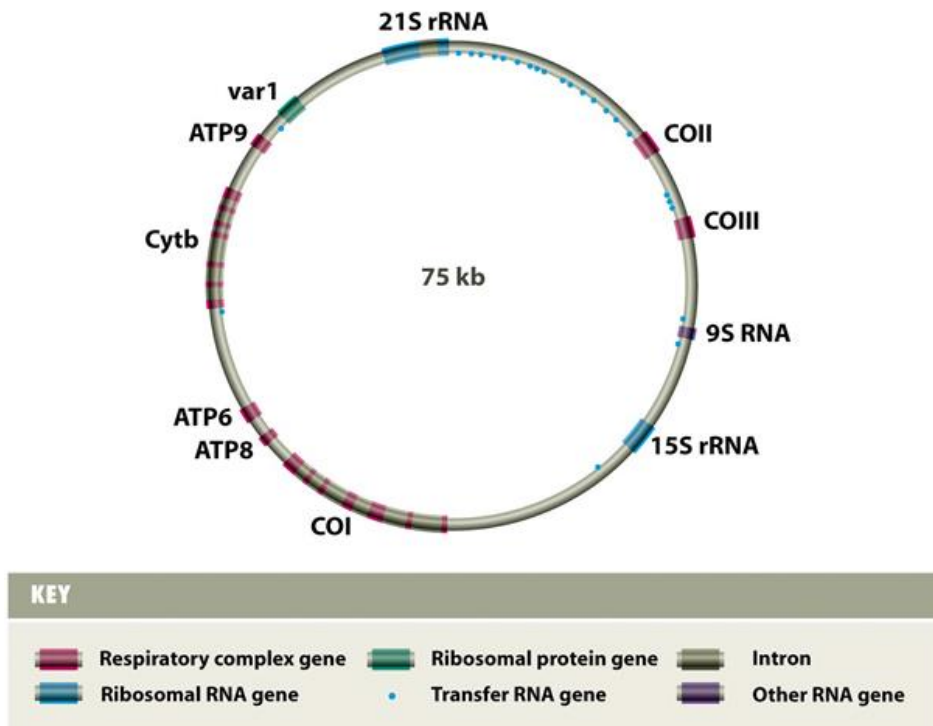
<b>Characteristic</b>	<b>Chromosomes</b>	<b>Plasmid</b>	<b>Mitochondria</b>
<b>Relative amount (%)</b>	<b>85</b>	<b>5</b>	<b>10</b>
<b>Number of copies</b>	<b>2 x 16</b>	<b>60-100</b>	<b>~50 (8-130)</b>
<b>Size (kbp)</b>	<b>~ 12,100</b>	<b>6.318</b>	<b>70-76</b>

### Yeast Genome – Genes Nomenclature

Wild type genes are written with capital letters in italics: *TPS1*, *RHO1*, *CDC28*. Recessive mutant genes are written with small letters in italics: *tps1*, *rho1*, *cdc28*. Three letters provides information about a function, mutant phenotype, or process related to that gene.

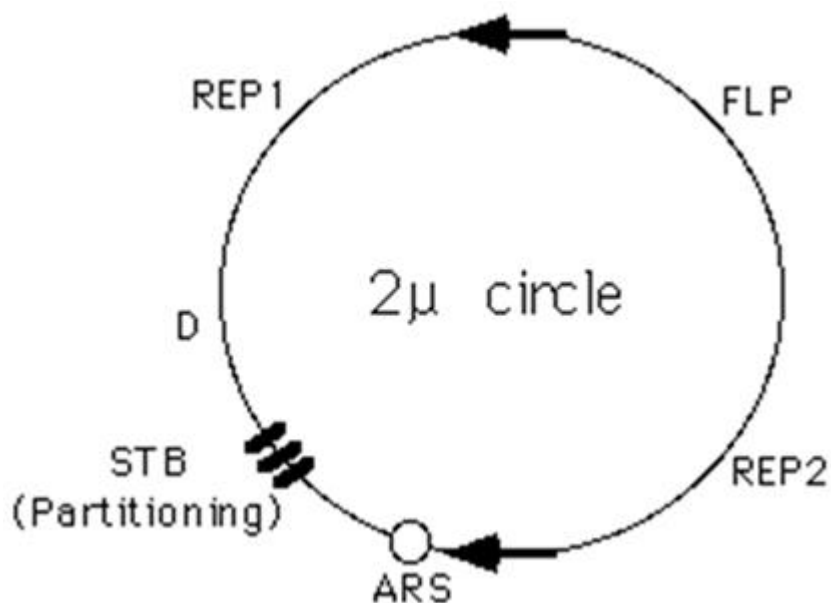
CDC - Cell Division Cycle; ADE-ADenine biosynthesis

## Yeast – Mitochondrial DNA



## Yeast – Plasmid DNA

The 2 $\mu$  circle is a 6.3 kb. 50 to 100 copies per haploid genome of the yeast cells. ARS, the FLP gene, the three genes which encode proteins required for regulation of FLP expression (REP2, REP1, and D). Set of small direct repeats (called "STB") required for partitioning into daughter cells during mitosis and meiosis.



## LESSON 92

**Mitochondrial Genome:** Multiple identical circular chromosomes

~15-16 Kb in animals

~ 200 kb to 2,500 kb in plants

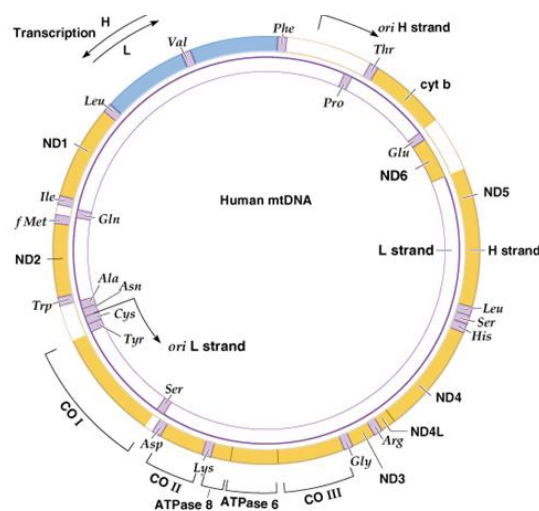
Over 95% of mitochondrial proteins are encoded in the nuclear genome.

Often A+T rich genomes

**Human Mitochondrial Genome:** Circular, double stranded 16.6 kb. The two strands are notably different in base composition, leading to one strand being heavy (H strand) and the other light (L strand). Both strands encode genes, although more are on the H strand. A short region (1121 bp), the D loop is a DNA triple helix: two overlapping copies of the H strand. The D loop is also the site where most of replication and transcription is controlled. Genes are tightly packed, with almost no non-coding DNA outside of D loop. Human mitochondrial genes contain no introns, although introns are found in the mitochondria of other groups (plants)

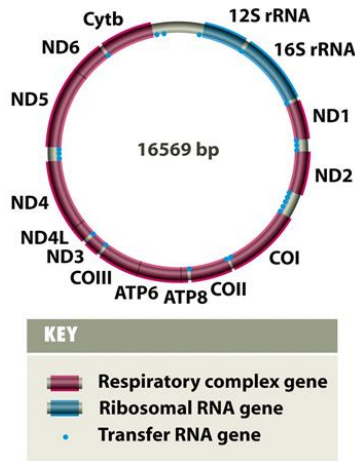
- 37 Genes
- 22 tRNAs
- 2 rRNAs
- 13 polypeptides
- tRNA: only 60 of the 64 codons code for amino acids.

### Mitochondrial Genome - Humans

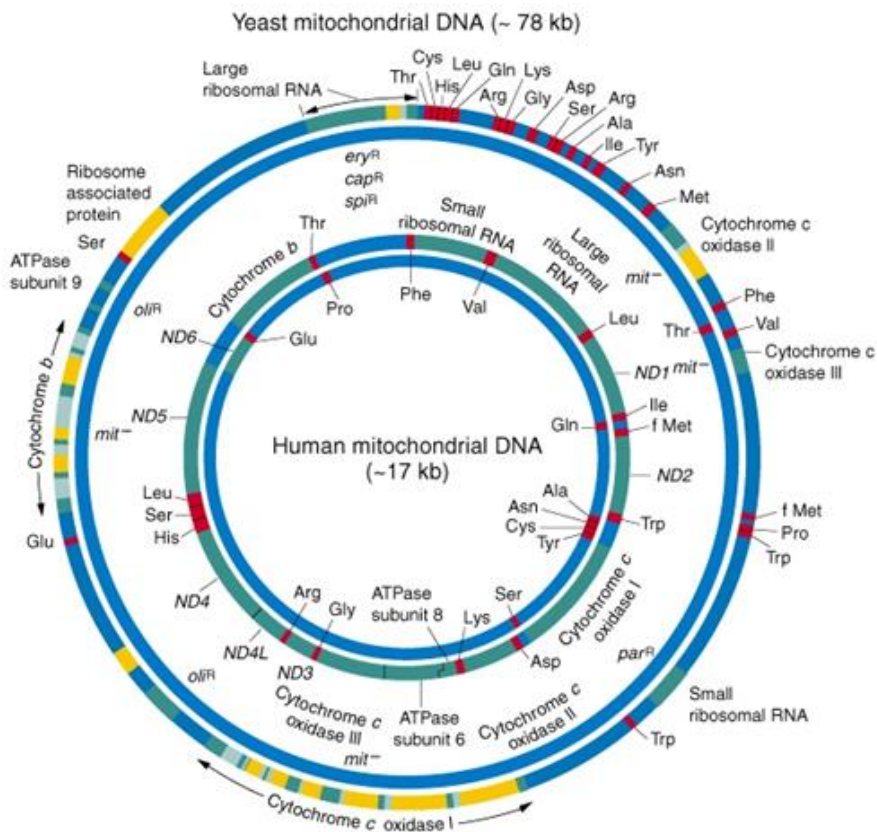


**Mitochondrial**

**Genome – Yeast**



### Mitochondrial Genome – Human and Yeast





## Mitochondrial Genomes

Feature	<i>Plasmodium falciparum</i>	<i>Chlamydomonas reinhardtii</i>	<i>Homo sapiens</i>	<i>Saccharomyces cerevisiae</i>	<i>Arabidopsis thaliana</i>	<i>Reclinomonas americana</i>
Total number of genes	5	12	37	35	52	92
Types of genes						
Protein-coding genes	3	7	13	8	27	62
Respiratory complex	3	7	13	7	17	24
Ribosomal proteins	0	0	0	1	7	27
Transport proteins	0	0	0	0	3	6
RNA polymerase	0	0	0	0	0	4
Translation factor	0	0	0	0	0	1
Functional RNA genes	2	5	24	27	25	30
Ribosomal RNA genes	2	2	2	2	3	3
Transfer RNA genes	0	3	22	24	22	26
Other RNA genes	0	0	0	1	0	1
Number of introns	0	1	0	8	23	1
Genome size (kb)	6	16	17	75	367	69

## Universal Code and Mitochondrial Code

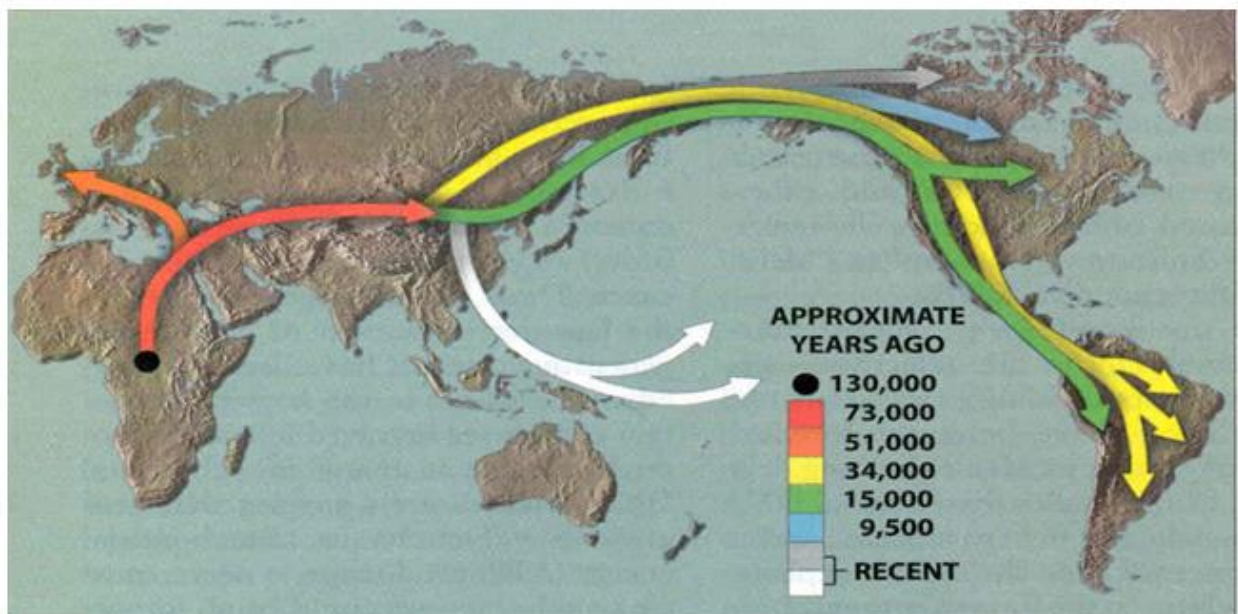
Mitochondrial code						Universal code							
		Second letter						Second letter					
		U	C	A	G			U	C	A	G		
First letter	U	Phe	Ser	Tyr	Cys	U	UUU	Phe	UCU	UAU	Tyr	UGU	Cys
		Phe	Ser	Tyr	Cys		UUC	UCC	UAC	UGC	U		
		Leu	Ser	Stop	(Stop) Trp		UUA	UCA	UAA	UGA	Stop	A	
		Leu	Ser	Stop	Trp		UUG	UCG	UAG	UGG	Trp	G	
C	Leu	Pro	His	Arg	U	CUU	Leu	CCU	CAU	His	CGU	Arg	
		Pro	His	Arg		CUC		CCC	CAC	CGC	C		
		Pro	Gin	Arg		CUA		CCA	CAA	CGA	A		
		Pro	Gin	Arg		CUG		CCG	CAG	CGG	G		
A	Ile (Met)	Thr	Asn	Ser	U	AUU	Ile	ACU	AAU	Asn	AGU	Ser	
		Thr	Asn	Ser		AUC		ACC	AAC	AGC	C		
		(Ile) Met	Lys	(Arg) Stop		AUA		ACA	AAA	AGA	A		
		Ile	Lys	(Arg) Stop		AUG		ACG	AAG	AGG	G		
G	Val	Ala	Asp	Gly	U	GUU	Val	GCU	GAU	Asp	GGU	Gly	
		Ala	Asp	Gly		GUC		GCC	GAC	GGC	C		
		Ala	Glu	Gly		GUA		GCA	GAA	GGA	A		
		Ala	Glu	Gly		GUG		GCG	GAG	GGG	G		

## Mitochondrial Genome – Tumors due to mutations



## Mitochondrial DNA polymorphisms track human migrations

All humans descend from a small group of Africans. This group originated in central Africa ~200,000 years ago. The founding group was small (102-104 people)

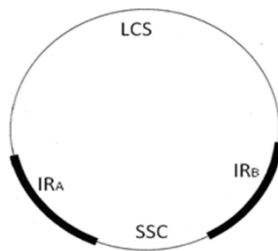


## LESSON 93

**Chloroplast Genome (cpDNA):** Multiple circular molecules. Size ranges from 70 kb to 2000 kb. Land plants typically 120 – 170 kb. Similar to mtDNA

- ~ 70 kb – Epifagus
- ~2,000 kb – Acetabularia

Many chloroplast proteins are encoded in the nucleus (separate signal sequence). Double stranded DNA molecule. Chloroplasts genomes are relatively larger. Multiple copies of genome are present. Large enough to code 50-100 proteins as well as rRNAs & tRNAs. cpDNA regions includes Large Single Copy (LSC), Small Single Copy (SSC) regions, and Inverted Repeats (IRA & IRB).



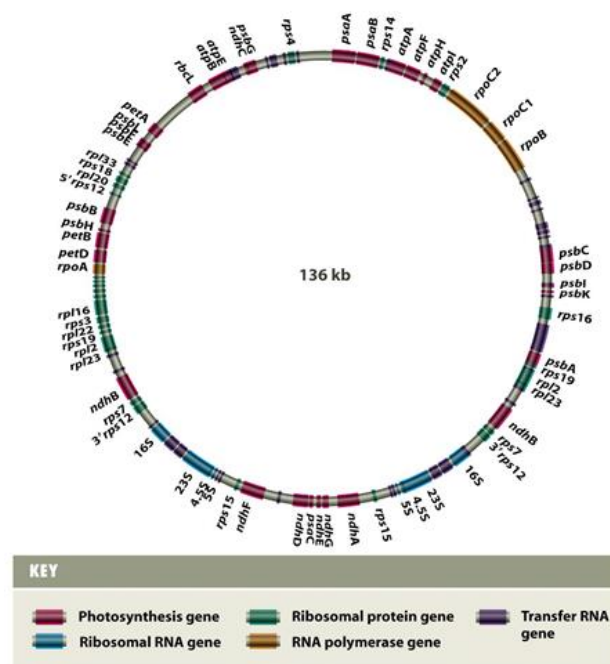
### Chloroplast Genomes: Size in different Taxa

Taxa	Genome size (In kb)	Inverse duplication (In kb)
<b>Angiospermae</b>		
<i>Nicotiana tabacum</i>	156	25
<i>Spinacia oleracea</i>	150	24
<i>Pelargonium hortorum</i>	217	76
<i>Pisum sativum</i>	120	Not present
<i>Epifragus virginiana</i>	70	22
<i>Oryza sativa</i>	134	21
<b>Gymnospermae</b>		
<i>Pinus</i>	120	Not present
<i>Ginkgo biloba</i>	158	17
<b>Pteridophyta</b>		
<i>Osmunda cinnamomea</i>	144	10
<b>Bryophyta</b>		
<i>Marchantia polymorpha</i>	121	10
<b>Chlorophyta</b>		
<i>Codium fragile</i>	85	Not present
<i>Chlamydomonas reinhardtii</i>	195	22
<i>Chlamydomonas moewusii</i>	292	41
<b>Rhodophyta</b>		
<i>Cyanophora paradoxa</i>	127	10
<b>Chromophyta</b>		
<i>Dictyota dichotoma</i>	123	5

## Chloroplast Genome Size

Species	Type of organism	Genome size (kb)
<b>Chloroplast genomes</b>		
<i>Pisum sativum</i>	Flowering plant (pea)	120
<i>Marchantia polymorpha</i>	Liverwort	121
<i>Oryza sativa</i>	Flowering plant (rice)	136
<i>Nicotiana tabacum</i>	Flowering plant (tobacco)	156
<i>Chlamydomonas reinhardtii</i>	Green alga	195

## Chloroplast Genome of Rice



**Functions of Chloroplast Genes:** Most cp genes fall into two functional groups:

Genes involved in replication, transcription, translation

Genes involved in photosynthesis

**Genes Nomenclature** Based on bacterial naming system, which uses lower case letters, and a descriptive prefix, based on the probable function. If the gene product is part of a multi-subunit complex, a letter of the alphabet is used to denote different subunits.

- psa for genes of photosystem I (psaA, psaB, etc.)
- psb for genes of photosystem II (psbA, psbB, etc.)

**Properties of Chloroplast Genome**

- Non-mendelian inheritance
- Self-replication
- Somatic segregation in plants
- Inherited independently of nuclear genes
- Conservative rate of nucleotide substitution enables to resolve plant phylogenetic relationships at deep levels of evolution.

## LESSON 94

**Eukaryotic Genomes:** Located on several chromosomes; relatively low gene density. Carry organelles genome in addition to nuclear genome. Eukaryotic genomes contains repetitive sequences like

- SINEs (short interspersed elements)
- LINEs (long interspersed elements)

### LINES

- 1-5 kb
- 10-10,000 copies

### SINES

- 200-300 bp
- 100,000 copies

Highly repetitive: Minisatellites, Microsatellites, Telomeres

Minisatellites: Repeats of 14-500 bp. 1-5 kb long. Scattered throughout the genome

Microsatellites: Repeats up to 13 bp

Telomeres: Short repeats (6 bp). 250-1,000 at ends of chromosomes

### Elements of Eukaryotic Genomes

**Chromosomes:** linear, centromeres, telomeres, origins of replication, replicons

**Protein-coding genes** and **spliceosomal introns**

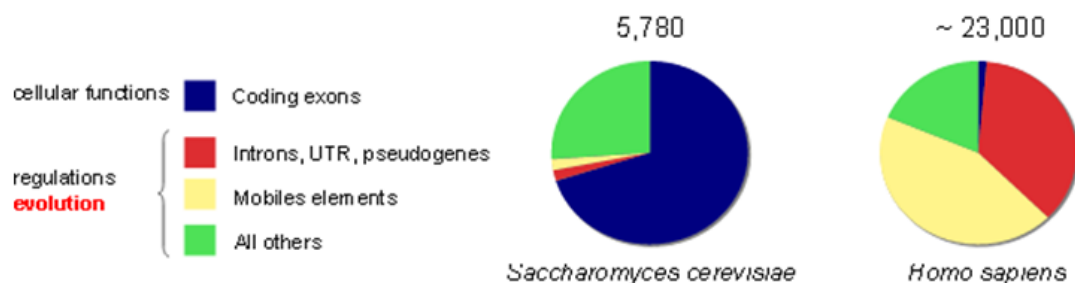
**Genes for non coding RNAs:** rRNAs, tRNAs, snoRNAs, snRNAs, microRNAs

**Mobile genetic elements:** and their remnants

**Pseudogenes:** and processed pseudogenes

**Satellite DNAs:** micro-, minisatellites, repeated sequences

**Fragments of organellar DNAs:** NUMTs and NUPTs



### Comparison - Prokaryotic & Eukaryotic Genomes

### **Prokaryotic**

- Usually circular
- Smaller
- Found in the nucleoid region
- Less elaborately structured and folded

### **Eukaryotic**

- Complex with a large amount of protein to form chromatin
- Highly extended and tangled during interphase
- Found in the nucleus

## LESSON 95

### Genomes Comparisons

**Genomes vary in size** Genomes of most bacteria and archaea range from 1 to 6 million base pairs (Mb). Most plants and animals have genomes greater than 100 Mb; humans have 3,000 Mb

**Genomes vary in genes numbers:** Free-living bacteria and archaea have 1,500 to 7,500 genes. Fungi have about 5,000 genes and multicellular eukaryotes upto 40,000 genes. Number of genes is not correlated to genome size

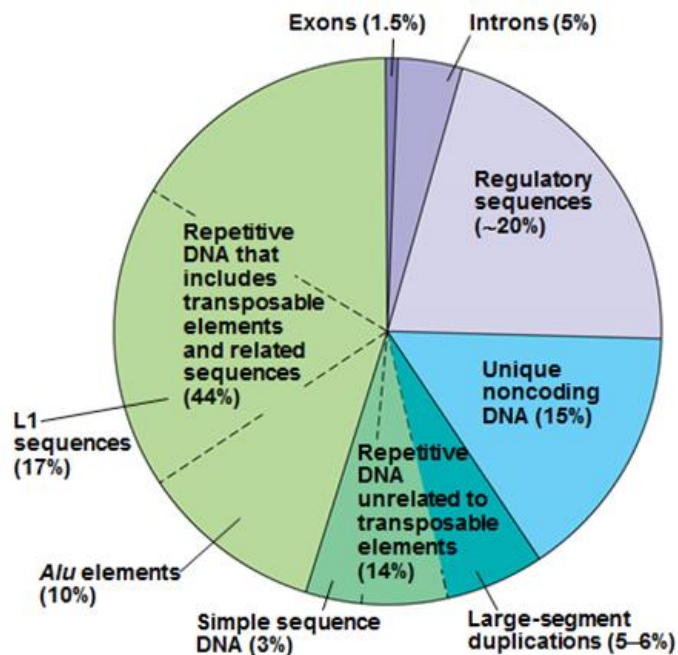
Nematode *C. elegans* has 100 Mb and 20,000 genes, while *Drosophila* has 165 Mb and 13,700 genes. Vertebrate genomes can produce more than one polypeptide per gene because of alternative splicing of RNA transcripts

**Humans and Mammals have low gene density:** Humans and other mammals have the lowest gene density, or number of genes, in a given length of DNA. Multicellular eukaryotes have many introns within genes and noncoding DNA between genes

### Multicellular eukaryotes have much noncoding DNA and multigene families

Most of eukaryotic genomes neither encode proteins nor functional RNAs. Evidence indicates that noncoding DNA plays important roles in the cell

### Human Genome: Distribution of coding and non-coding DNA



### Comparing Genomes



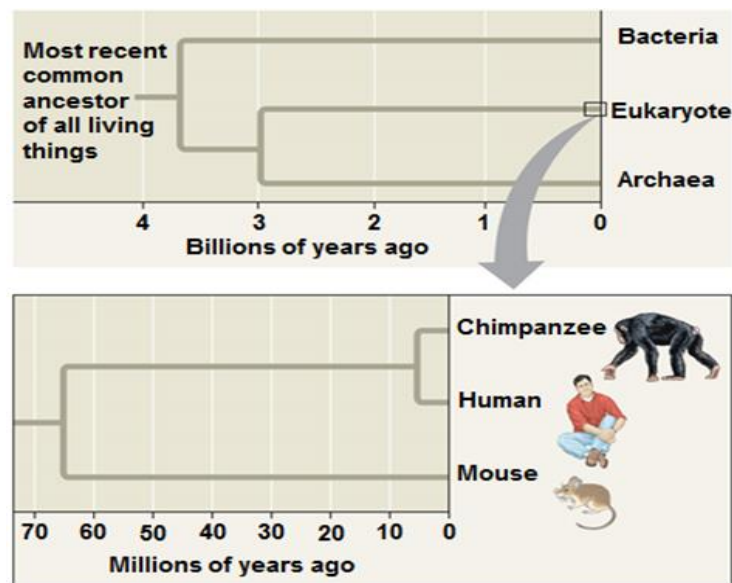
Significant similarity between genomes of “distant” species like (Man – Yeast: 23%)

Similarity increases for taxonomically close species.

Closely related species help us understand recent evolutionary events

Distantly related species help us understand ancient evolutionary events

**Comparing Genomes: Bacteria, archaea, and eukaryotes diverged from each other between 2 and 4 billion years ago**



	Bacteria	Archaea	Eukaryotes
Genome size	Most are 1–6 Mb		Most are 10–4,000 Mb, but a few are much larger
Number of genes	1,500–7,500		5,000–40,000
Gene density	Higher than in eukaryotes		Lower than in prokaryotes (Within eukaryotes, lower density is correlated with larger genomes.)
Introns	None in protein-coding genes	Present in some genes	Unicellular eukaryotes: present, but prevalent only in some species Multicellular eukaryotes: present in most genes
Other noncoding DNA	Very little		Can be large amounts; generally more repetitive noncoding DNA in multicellular eukaryotes

**Comparing Genomes:** Human and chimpanzee genomes differ by 1.2%, at single base-pairs, and by 2.7% because of insertions and deletions

## LESSON 96

### Comparing Distantly/Closely Related Species

**Comparing distantly related species:** Highly conserved genes have changed very little over time. These help to clarify relationships among species that diverged from each other long ago. Bacteria, archaea, and eukaryotes diverged from each other 2 and 4 billion years ago. Highly conserved genes can be studied in one model organism.

**Comparing closely related species:** Genetic differences between closely related species can be correlated with phenotypic differences. Genetic comparison of several mammals with non-mammals helps to identify what make mammals. Human and chimpanzee genomes differ by 1.2%, at single base-pairs, and by 2.7% because of insertions and deletions. Several genes are evolving faster in humans than chimpanzees. Genes involved in defense against malaria and tuberculosis and in regulation of brain size, genes code for transcription factors. Humans and chimpanzees differ in the expression of the *FOXP2* gene, whose product turns on genes involved in vocalization. Differences in the *FOXP2* gene may explain why humans but not chimpanzees communicate by speech.

**Conclusion:** Highly conserved genes have changed very little over the time. These help to clarify relationships among species that diverged from each other long ago

# LESSON 97

## Anatomy and Organization of Genomes

**Genome Anatomy:** Anatomy of different genomes differs from each other. Eukaryotes and prokaryotes genomes differ very significantly

Size of genomes; there can be 1000 fold difference between eukaryotes and prokaryotes genomes

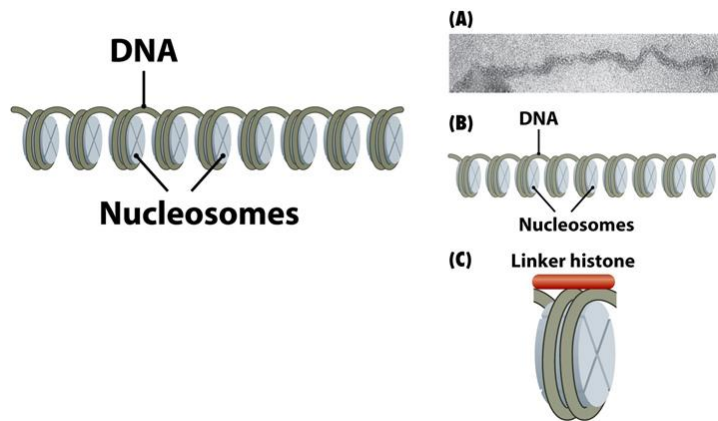
~ 30 fold between genomes of different eukaryotes

In humans ~ 23,000 while bacterial genomes ~ 1,500 – 2,000 genes

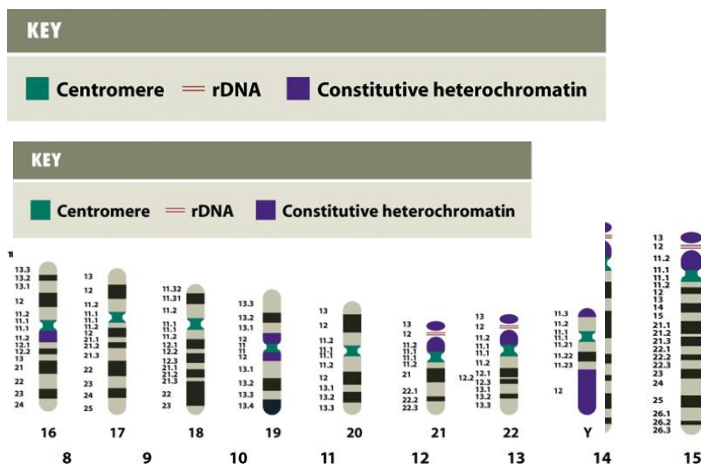
**Genome Anatomy – Eukaryotes:** Eukaryotic genomes are full of simple repeats, numerous types of transposable elements and other sequences.

**Genome Anatomy – Prokaryotes:** Prokaryotes have a few repeats and transposable elements and their genomes consist of mainly the genes.

## Genome Organization



## Genome Organization in Prokaryotes



## Genome Organization in Prokaryote

Species	Size of genome (Mb)	Approximate number of genes
<b>Bacteria</b>		
<i>Mycoplasma genitalium</i>	0.58	500
<i>Streptococcus pneumoniae</i>	2.16	2300
<i>Vibrio cholerae</i> El Tor N16961	4.03	4000
<i>Mycobacterium tuberculosis</i> H37Rv	4.41	4000
<i>Escherichia coli</i> K12	4.64	4400
<i>Yersinia pestis</i> CO92	4.65	4100
<i>Pseudomonas aeruginosa</i> PA01	6.26	5700
<b>Archaea</b>		
<i>Methanococcus jannaschii</i>	1.66	1750
<i>Archaeoglobus fulgidus</i>	2.18	2500

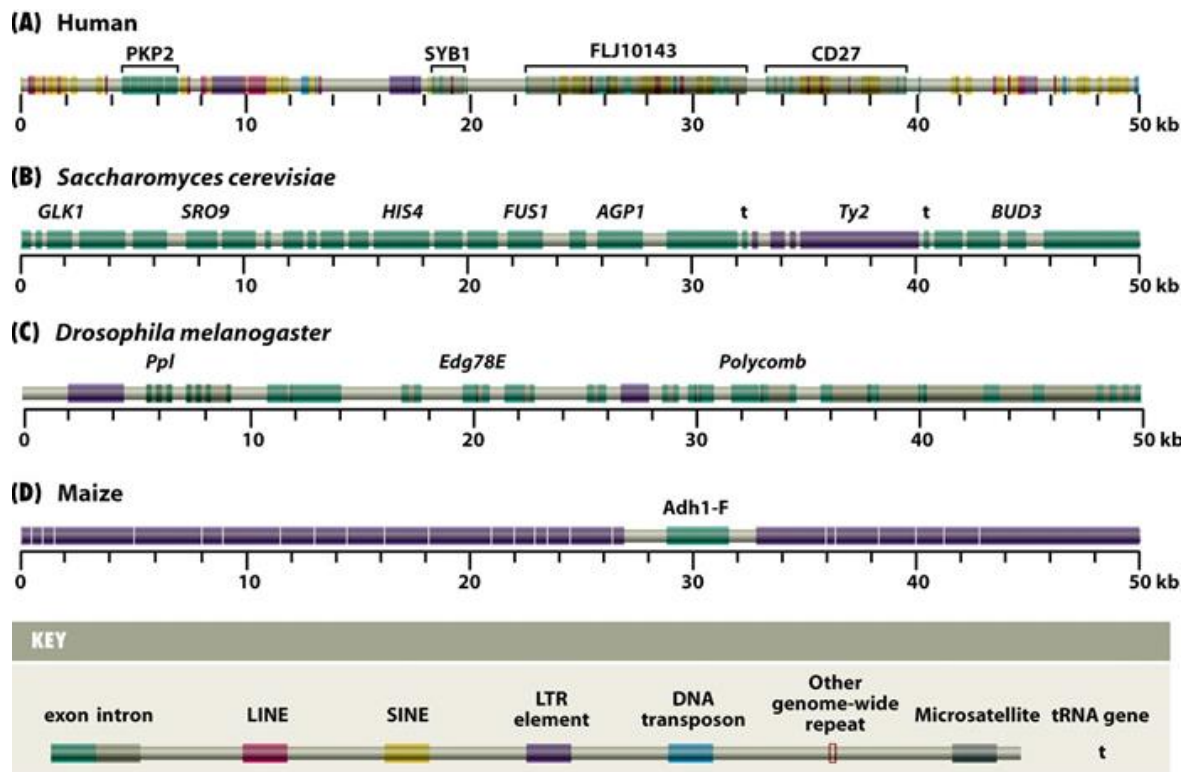
## Genome Organization: Comparisons

Species	DNA molecules	Genome organization	
		Size (Mb)	Number of genes
<i>Escherichia coli</i> K12	One circular molecule	4.639	4405
<i>Vibrio cholerae</i> El Tor N16961	Two circular molecules		
	Main chromosome	2.961	2770
	Megaplasmid	1.073	1115
<i>Deinococcus radiodurans</i> R1	Four circular molecules		
	Chromosome 1	2.649	2633
	Chromosome 2	0.412	369
	Megaplasmid	0.177	145
	Plasmid	0.046	40

Species	Genome size (Mb)
<b>Fungi</b>	
<i>Saccharomyces cerevisiae</i>	12.1
<i>Aspergillus nidulans</i>	25.4
<b>Protozoa</b>	
<i>Tetrahymena pyriformis</i>	190
<b>Invertebrates</b>	
<i>Caenorhabditis elegans</i>	97
<i>Drosophila melanogaster</i>	180
<i>Bombyx mori</i> (silkworm)	490
<i>Strongylocentrotus purpuratus</i> (sea urchin)	845
<i>Locusta migratoria</i> (locust)	5000

Species	Genome size (Mb)
<b>Vertebrates</b>	
<i>Takifugu rubripes</i> (pufferfish)	400
<i>Homo sapiens</i>	3200
<i>Mus musculus</i> (mouse)	3300
<b>Plants</b>	
<i>Arabidopsis thaliana</i> (vetch)	125
<i>Oryza sativa</i> (rice)	466
<i>Zea mays</i> (maize)	2500
<i>Pisum sativum</i> (pea)	4800
<i>Triticum aestivum</i> (wheat)	16,000
<i>Fritillaria assyriaca</i> (fritillary)	120,000

### Genome Organization: Human, Yeast, Fruit Fly , Maize



## Compactness of Genomes

Feature	Yeast	Fruit fly	Human
Gene density (average number per Mb)	496	76	11
Introns per gene (average)	0.04	3	9
Amount of the genome that is taken up by genome-wide repeats	3.4%	12%	44%

## LESSON 98

### Gene Anatomy

**What is Gene?** : A piece of DNA (or RNA) that contains the primary sequence to produce a functional biological gene product (RNA or protein). Entire nucleic acid sequence necessary for the synthesis of a functional polypeptide (protein chain) or functional RNA.

**Genetic information is stored in DNA.** Segments of DNA that encode proteins or other functional products are called genes. Gene sequences are transcribed into messenger RNA (mRNA).

**mRNA is translated into Proteins.** mRNA intermediates are translated into proteins that perform most of the life functions.

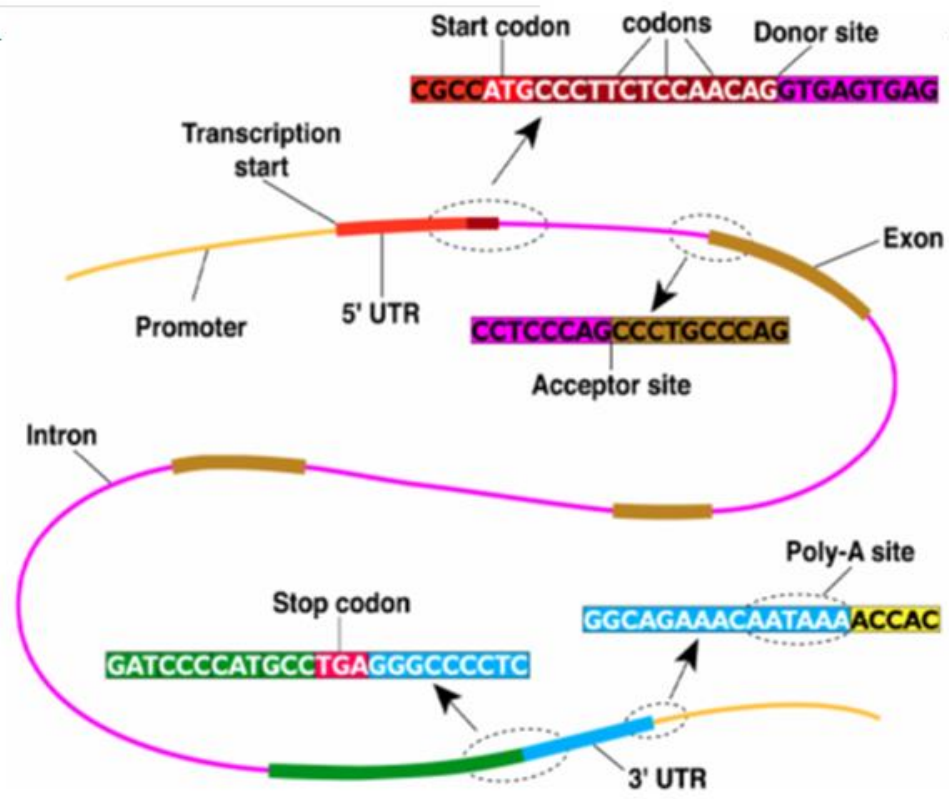
**Gene Anatomy:** Three components

- Open Reading Frame: From start codon (ATG) to stop (TGA, TAA, TAG)
- Upstream region with binding site. (e.g. TATA box, GC box, CAAT box)
- Poly-a tail

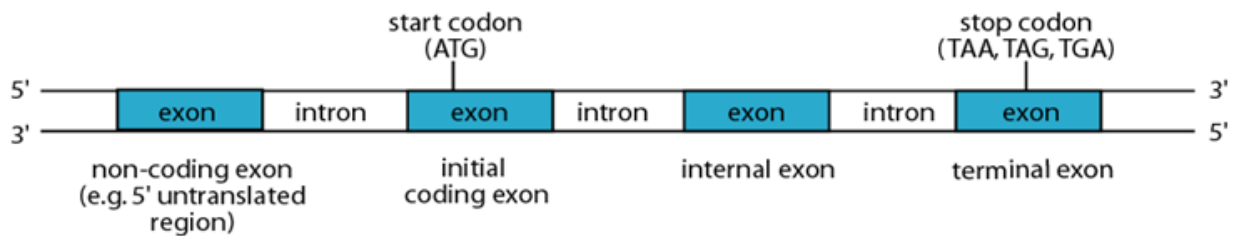
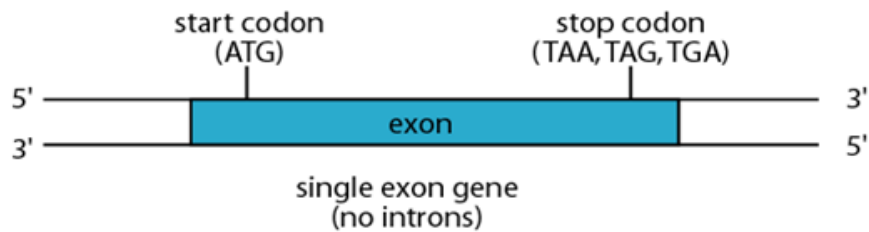
**Gene Anatomy – Typical Prokaryotic Gene**



**Gene Anatomy – Typical Eukaryotic Gene**



### Single Exon Gene and Multiple Exons Gene



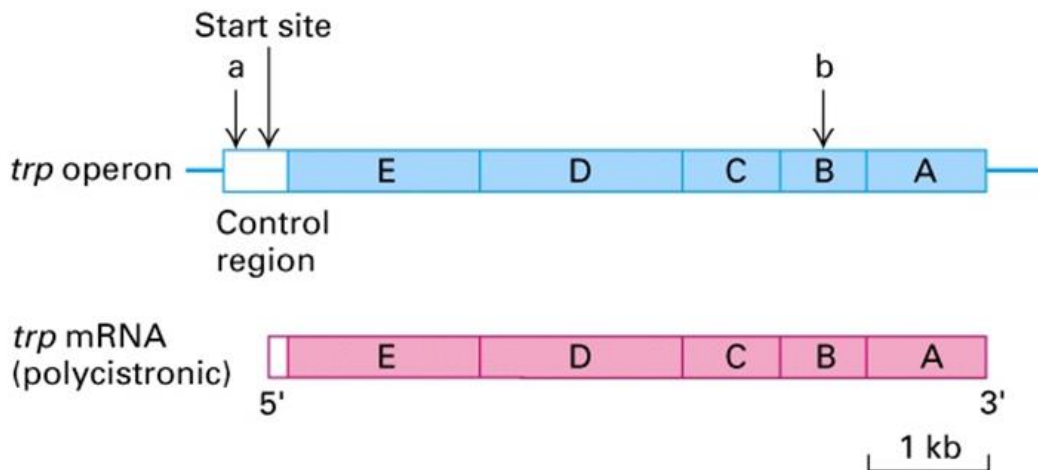


# LESSON 99

## Prokaryotic Gene and Eukaryotic Gene

**Bacterial Gene:** Most of the bacterial genes do not have introns. Many are organized in operons: contiguous genes, transcribed as a single polycistronic mRNA, which encode proteins with related functions. Polycistronic mRNA encodes several proteins

**Bacterial Gene:** Polycistronic mRNA encodes several proteins



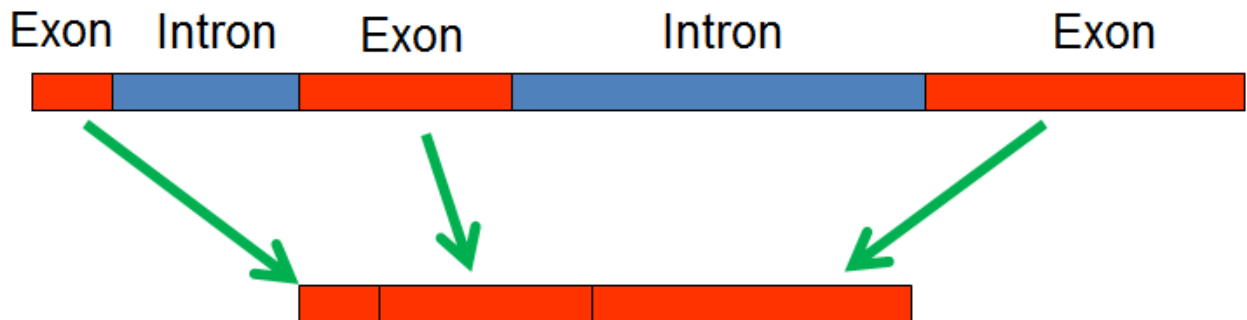
## Eukaryotic Gene: Exons and Introns

Introns: intervening sequences within a gene that are not translated into a protein sequence.

Exons: sequences within a gene that encode protein sequence.

Splicing: Removal of introns from the mRNA molecule

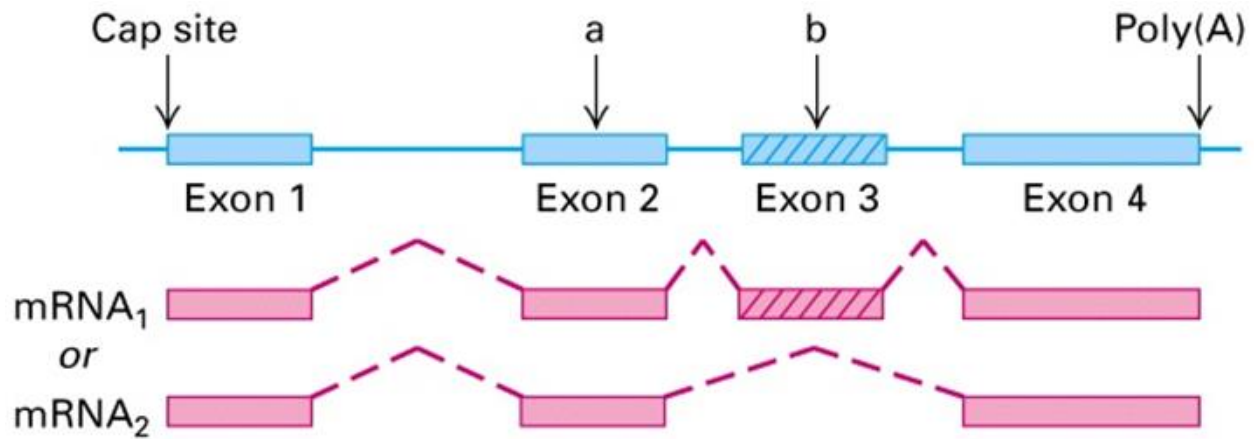
**Eukaryotic Gene :** Splicing: Removal of introns from the mRNA molecule



**Eukaryotic Gene:** Organize expression of genes' (function calls). Promoter region (binding site), usually near coding region. Binding can block (inhibit) expression

**Alternative Splicing in Eukaryotic Genes:** Most have introns. They produce monocistronic mRNA. They are large in size.

### Eukaryotic Gene: Alternative Splicing



## LESSON 100

### Types of Eukaryotic DNA

- Protein coding genes
- Tandemly repeated genes
- Repeated DNA
- Unclassified spacer DNA

Protein coding genes can also be in the form of:

- Duplicated and diverged genes
- Functional gene families and non-functional pseudo-genes
- Tandemly repeated genes encoding rRNA, 5sRNA, tRNA and histones.

### Repetitive DNA

- Simple sequence DNA
- Moderately repeated DNA (mobile DNA elements)
- Transposons
- Retrotransposons
- Long interspersed elements
- Short interspersed elements
- Unclassified spacer DNA

### Major Classes of Eukaryotic DNA in Human Genome

Class	Length	Copy Number in Human Genome	Fraction of Human Genome, %
Protein-coding genes			
Solitary genes	Variable	1	≈15* (0.8) <sup>†</sup>
Duplicated or diverged genes in gene families	Variable	2–1000	≈15* (0.8) <sup>†</sup>
Tandemly repeated genes encoding rRNAs, tRNAs, snRNAs, and histones	Variable	20–300	0.3
Repetitious DNA			
Simple-sequence DNA	1–500 bp	Variable	3
Interspersed repeats			
DNA transposons	2–3 kb	300,000	3
LTR retrotransposons	6–11 kb	440,000	8
Non-LTR retrotransposons			
LINES	6–8 kb	860,000	21
SINEs	100–300 bp	1,600,000	13
Processed pseudogenes	Variable	1–100	≈0.4
Unclassified spacer DNA	Variable	n.a. <sup>‡</sup>	≈25