

Genome Content

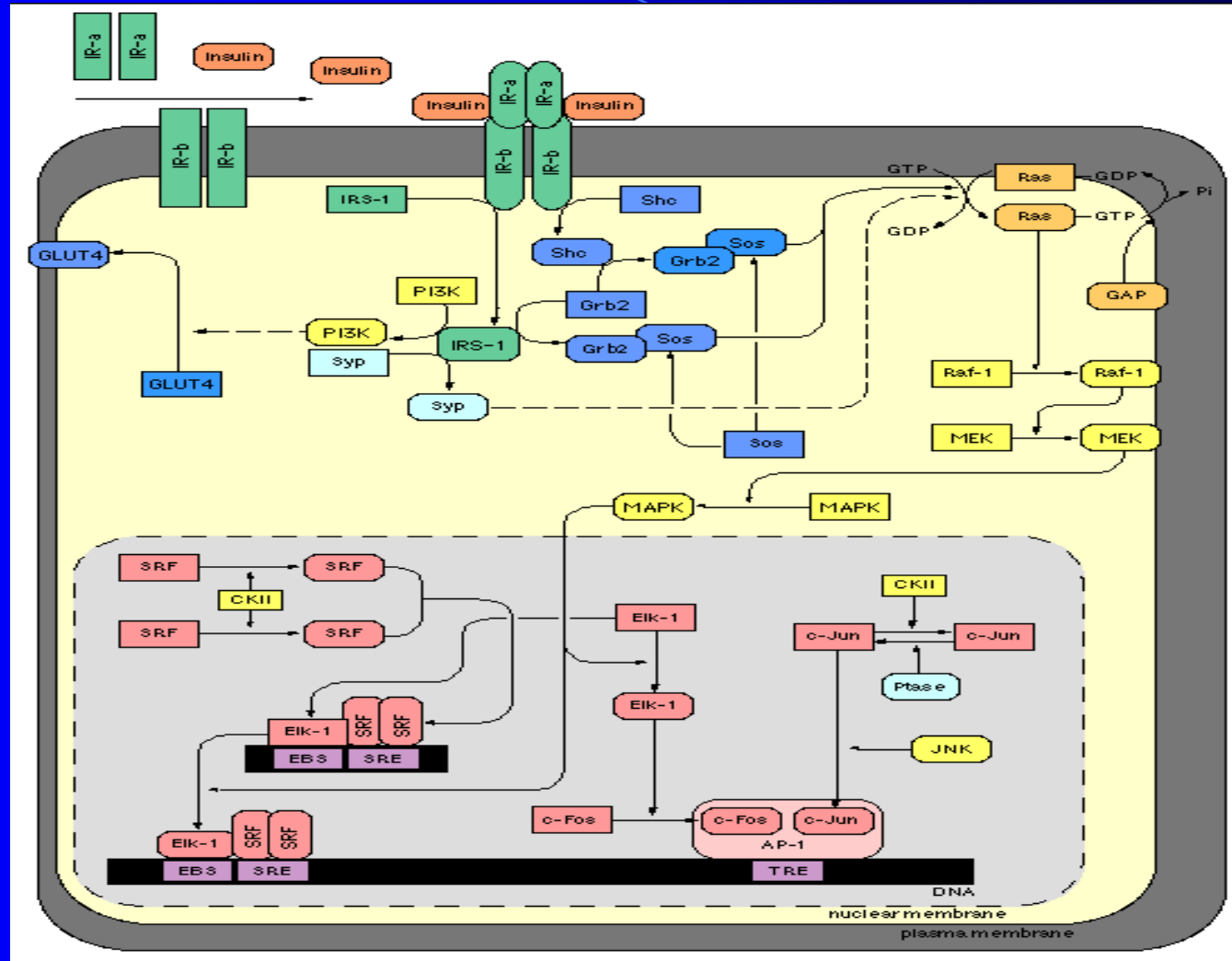
Chapter 3

By
Rasul Chaudhry

- The **genome** is the complete set of sequences in the genetic material of an organism. It includes the sequence of each chromosome plus any DNA in organelles.
- The **transcriptome** is the complete set of RNAs present in a cell, tissue, or organism. Its complexity is due mostly to mRNAs, but it also includes noncoding RNAs.
- The **proteome** is the complete set of proteins that is expressed by the entire genome. Because some genes code for multiple proteins, the size of the proteome is greater than the number of genes. Sometimes the term is used to describe complement of proteins expressed by a cell at any one time.

- **Polymorphism** (more fully genetic polymorphism) refers to the simultaneous occurrence in the population of genomes showing variations at a given position. The original definition applied to alleles producing different phenotypes. Now it is also used to describe changes in DNA affecting the restriction pattern or even the sequence. For practical purposes, to be considered as an example of a polymorphism, an allele should be found at a frequency $> 1\%$ in the population.
- **Single nucleotide polymorphism (SNP)** describes a polymorphism (variation in sequence between individuals) caused by a change in a single nucleotide. This is responsible for most of the genetic variation between individuals.
- **Restriction fragment length polymorphism (RFLP)** refers to inherited differences in sites for restriction enzymes (for example, caused by base changes in the target site) that result in differences in the lengths of the fragments produced by cleavage with the relevant restriction enzyme. RFLPs are used for genetic mapping to link the genome directly to a conventional genetic marker.

Effect of point mutations

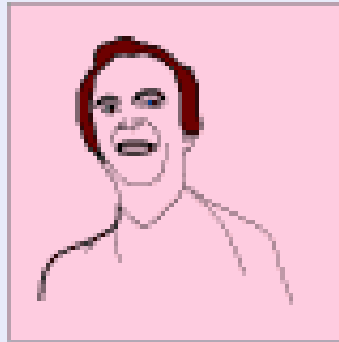


- Some point mutations change the restriction map

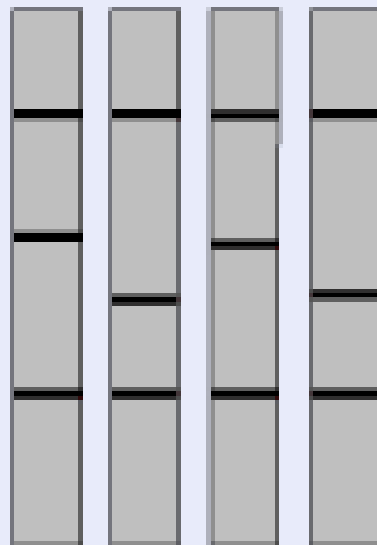
Restriction analysis can be used to identify individuals

- **DNA fingerprinting** analyzes the differences between individuals of the fragments generated by using restriction enzymes to cleave regions that contain short repeated sequences. Because these are unique to every individual, the presence of a particular subset in any two individuals can be used to define their common inheritance (e.g. a parent-child relationship).
- RFLPs and SNPs can be the basis for linkage maps and are useful for establishing parent-progeny relationships.

RFLPs can be associated with disease genes



Screen DNA patterns of patients with disease



Screen DNA patterns of unaffected people as control



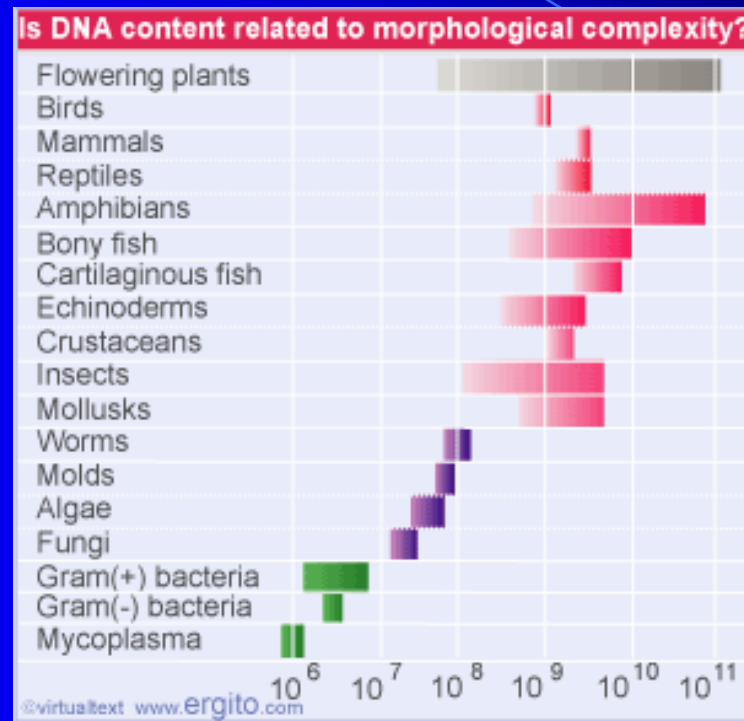
Band is same in patient and unaffected

Unlinked polymorphism varies in all samples

Band is common to patients

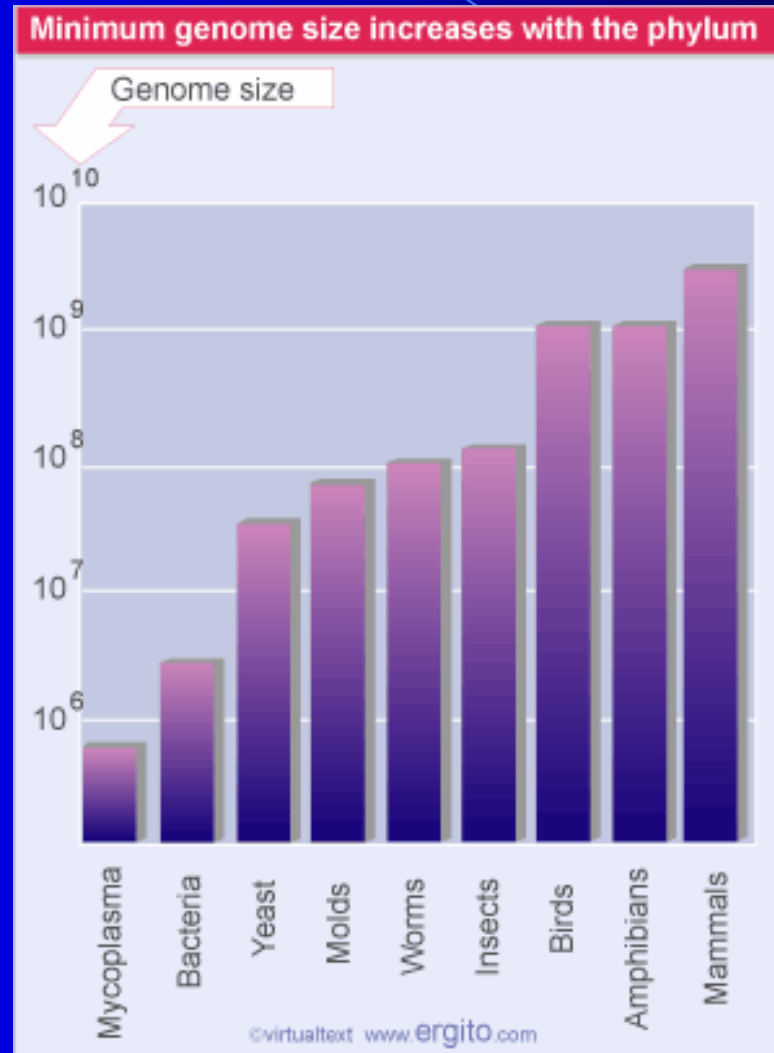
Band is common to unaffected people

Is DNA content related to morphological complexity?



- The **C-value** is the total amount of DNA in the genome (per haploid set of chromosomes).
- The **C-value paradox** describes the lack of relationship between the DNA content (C-value) of an organism and its coding potential.

Minimum genome size increases with the phylum



Useful genome sizes

Useful genome sizes		
Phylum	Species	Genome (bp)
Algae	<i>Pyrenomas salina</i>	6.6×10^5
Mycoplasma	<i>M. pneumoniae</i>	1.0×10^6
Bacterium	<i>E. coli</i>	4.2×10^6
Yeast	<i>S. cerevisiae</i>	1.3×10^7
Slime mold	<i>D. discoideum</i>	5.4×10^7
Nematode	<i>C. elegans</i>	8.0×10^7
Insect	<i>D. melanogaster</i>	1.4×10^8
Bird	<i>G. domesticus</i>	1.2×10^9
Amphibian	<i>X. laevis</i>	3.1×10^9
Mammal	<i>H. sapiens</i>	3.3×10^9

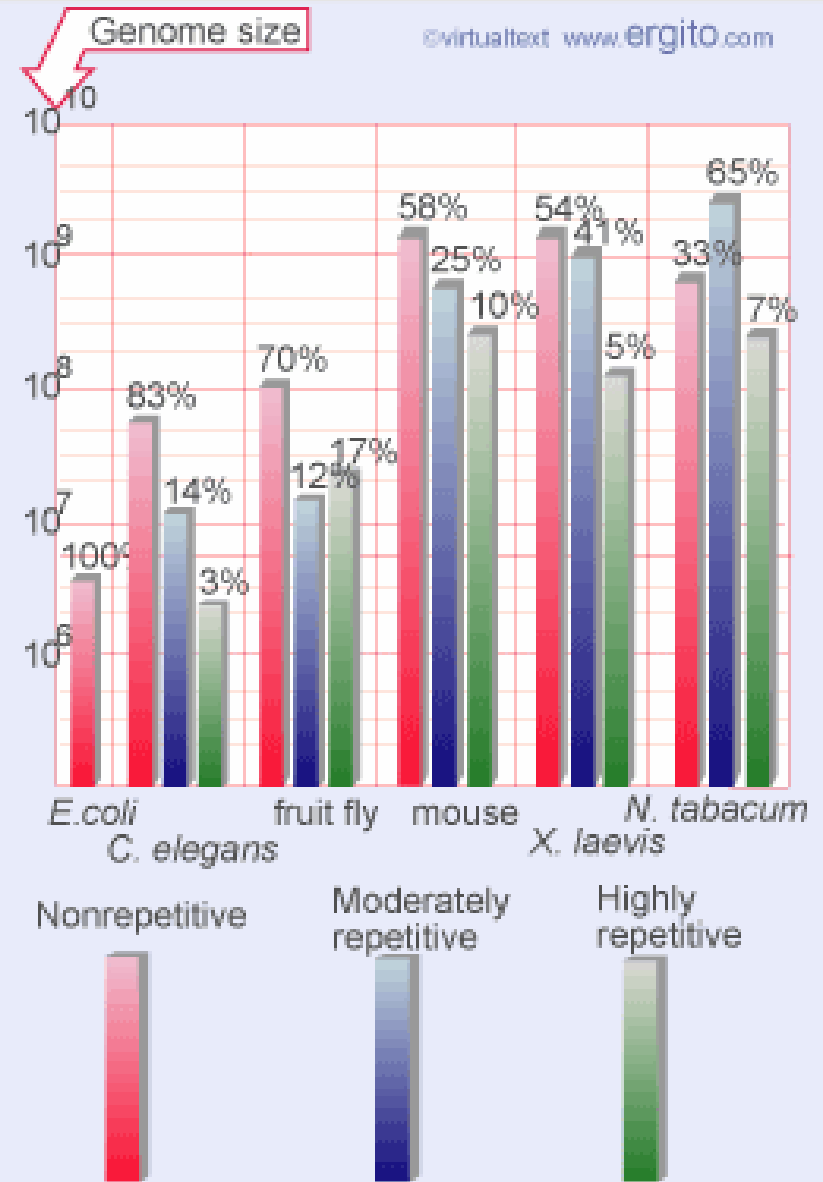
©virtualtext www.ergito.com

Eukaryotic genomes contain both nonrepetitive and repetitive DNA sequences

- **Nonrepetitive DNA** shows reassociation kinetics expected of unique sequences.
- **Repetitive DNA** behaves in a reassociation reaction as though many (related or identical) sequences are present in a component, allowing any pair of complementary sequences to reassociate.
- A **transposon (transposable element)** is a DNA sequence able to insert itself (or a copy of itself) at a new location in the genome, without having any sequence relationship with the target locus.
- **Selfish DNA** describes sequences that do not contribute to the genotype of the organism but have self-perpetuation within the genome as their sole function.

Nonrepetitive DNA is only part of the genome

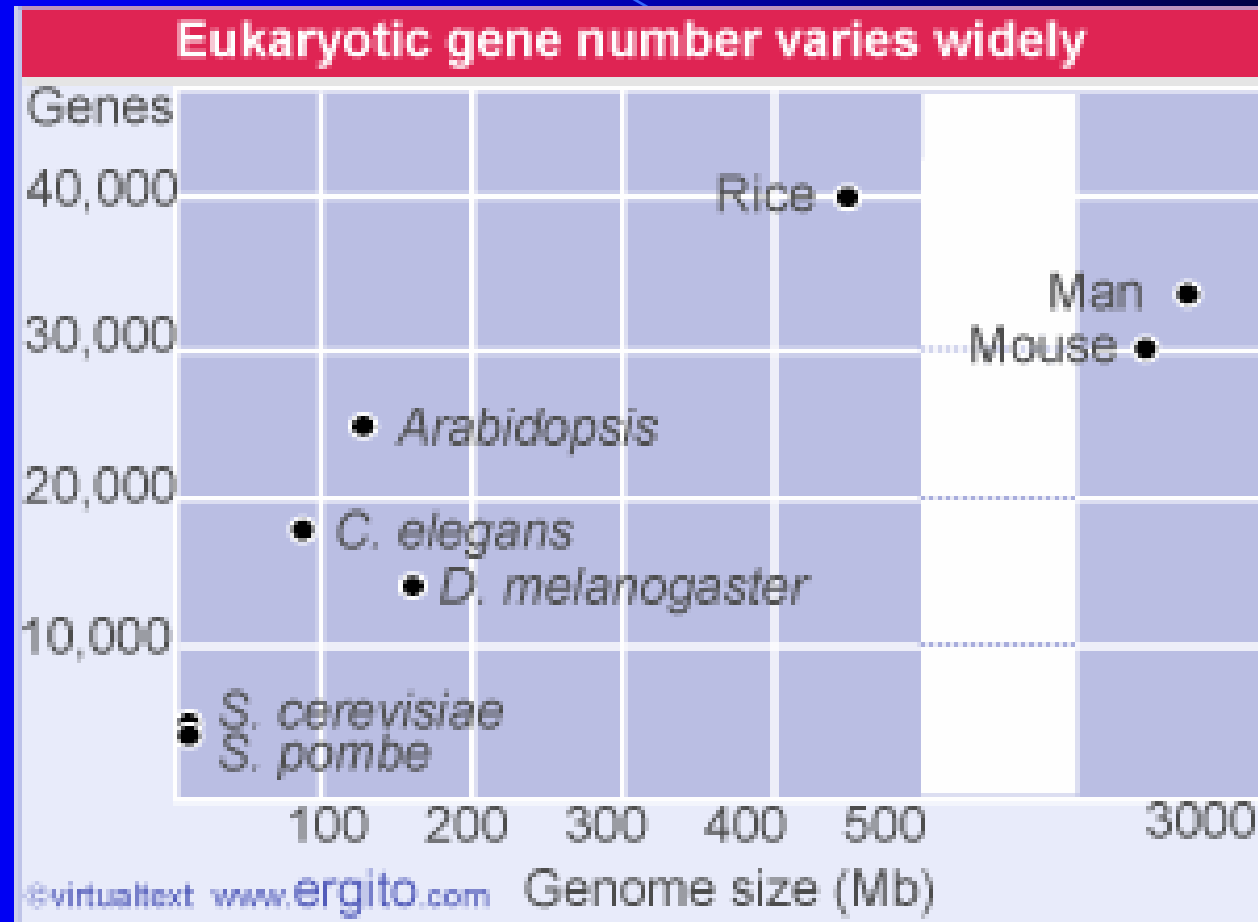
Nonrepetitive DNA is only part of the genome



Sequenced genomes vary from 470-40,000 genes

Sequenced genomes vary from 470-30,000 genes			
Species	Genome (Mb)	Genes	Lethal loci
<i>Mycoplasma genitalium</i>	0.58	470	~300
<i>Rickettsia prowazekii</i>	1.11	834	
<i>Haemophilus influenzae</i>	1.83	1,743	
<i>Methanococcus jannaschi</i>	1.66	1,738	
<i>B. subtilis</i>	4.2	4,100	
<i>E. coli</i>	4.6	4,288	1,800
<i>S. cerevisiae</i>	13.5	6,034	1,090
<i>S. pombe</i>	12.5	4,929	
<i>A. thaliana</i>	119	25,498	
<i>O. sativa</i> (rice)	466	~30,000	
<i>D. melanogaster</i>	165	13,601	3,100
<i>C. elegans</i>	97	18,424	
<i>H. sapiens</i>	3,300	~30,000	

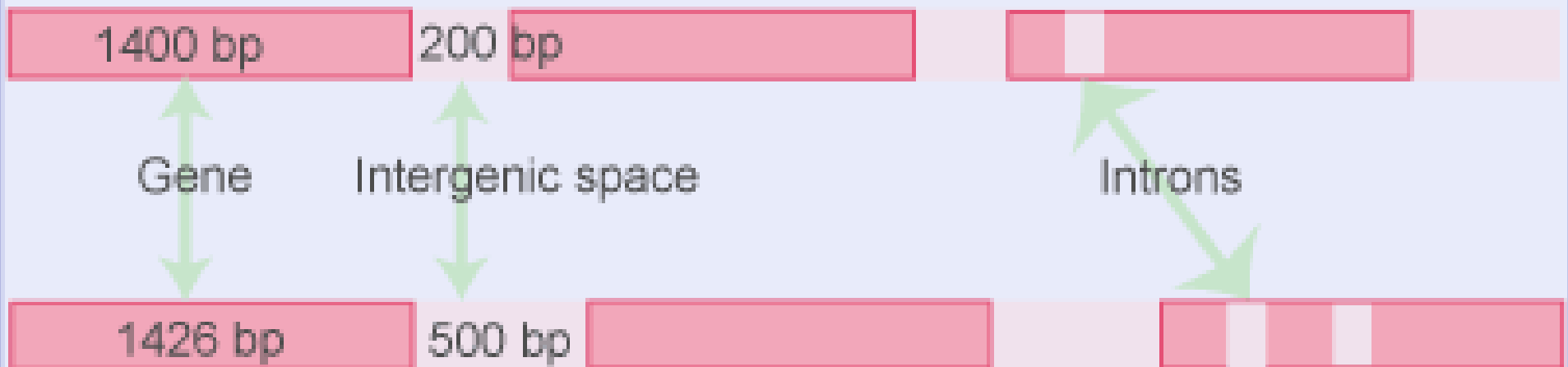
Total genes eukaryotes



Eukaryotic gene number varies widely

Yeast genomes are compact

5% of *S. cerevisiae* genes have 1 intron on average

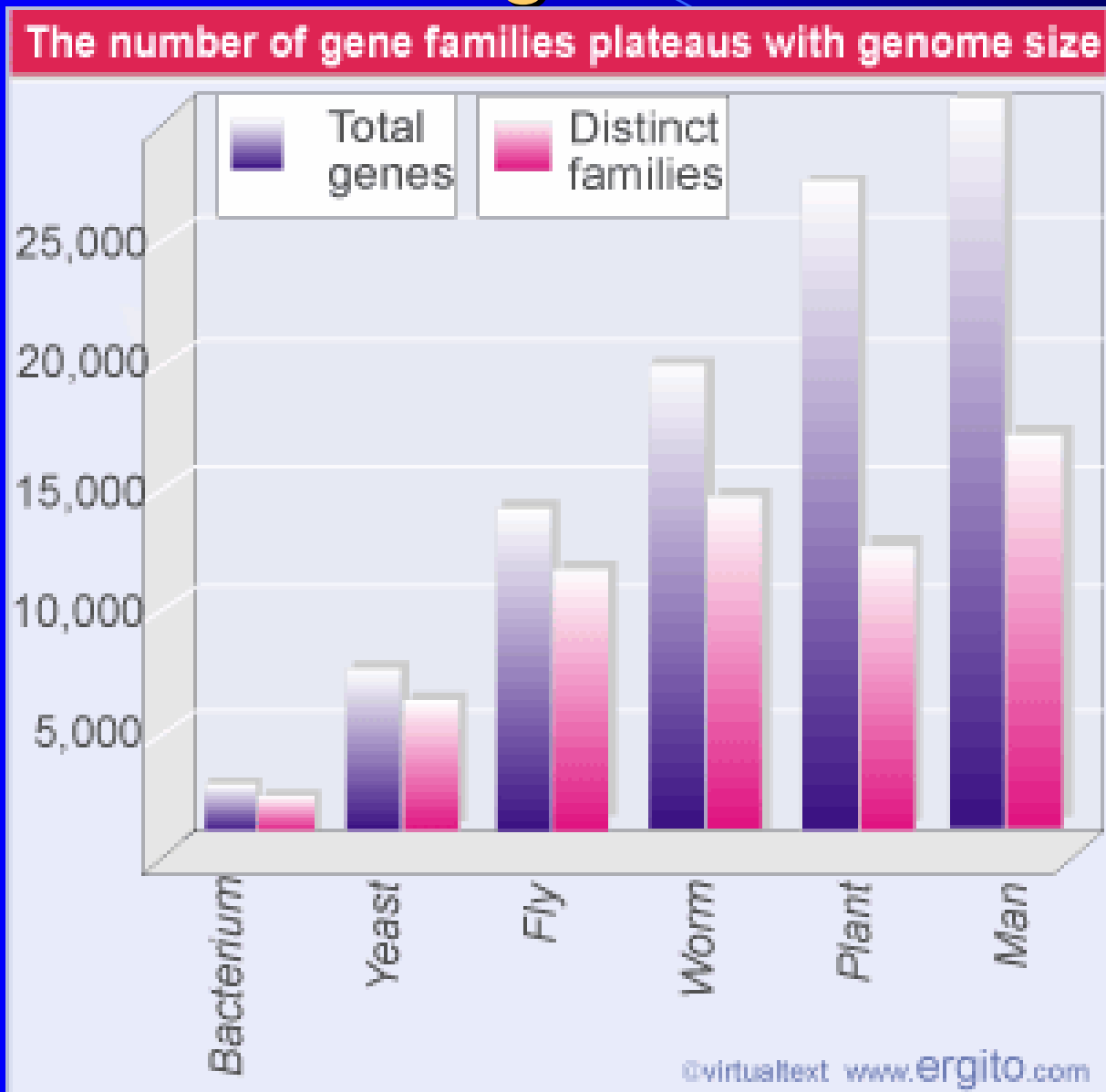


43% of *S. pombe* genes have introns
Average interrupted gene has 2 introns

Gene Family

- A **gene family** consists of a set of genes within a genome that code for related or identical proteins. The members were derived by duplication of an ancestral gene followed by accumulation of changes in sequence between the copies. Most often the members of a family are related but not identical.
- The **proteome** is the complete set of proteins that is expressed by the entire genome. Because some genes code for multiple proteins, the size of the proteome is greater than the number of genes. Sometimes the term is used to describe complement of proteins expressed by a cell at any one time.
- **Orthologs** are corresponding proteins in two species as defined by sequence homologies.

The number of gene families plateaus with genome size



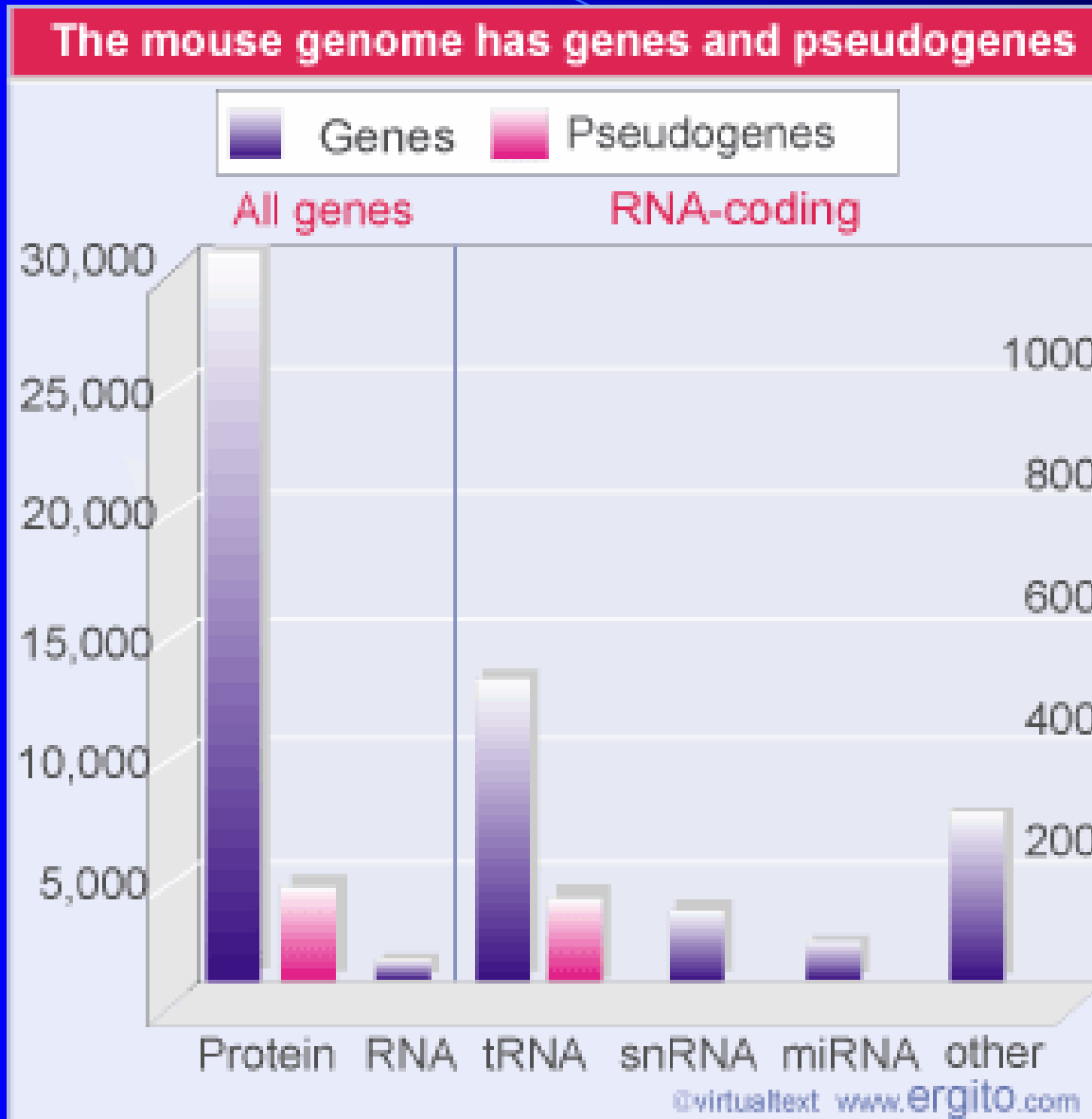
Family size increases with genome size

Family size increases with genome size			
	Unique genes	Families with 2-4 members	Families with >4 members
<i>H. influenzae</i>	89%	10%	1%
<i>S. cerevisiae</i>	72%	19%	9%
<i>D. melanogaster</i>	72%	14%	14%
<i>C. elegans</i>	55%	20%	26%
<i>A. thaliana</i>	35%	24%	41%

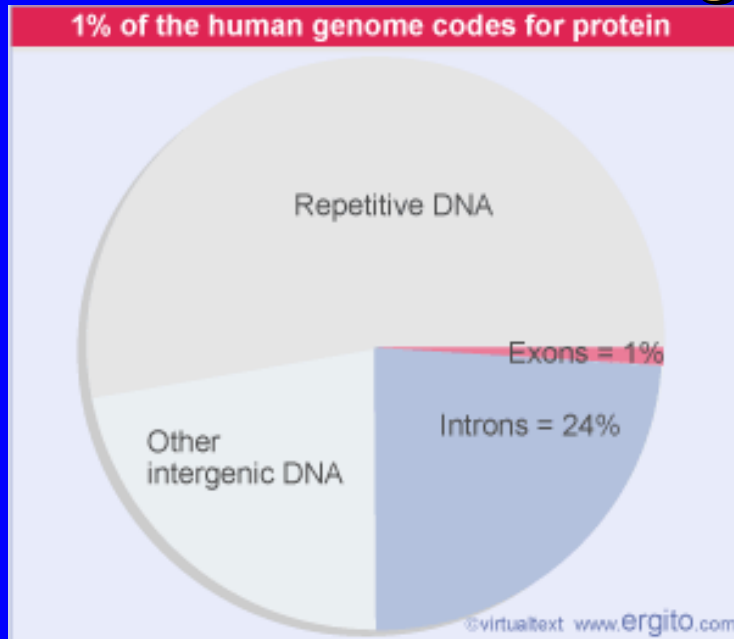
The human genome has fewer genes than expected

- 1% DNA – coding regions
- 5% of each genes are exons
- 60% genes are alternatively spliced
- 80 % of the spliced mRNA change protein sequences
- Total proteins in human proteome : 50 to 60 K
- Total genes : 30 to 40 K

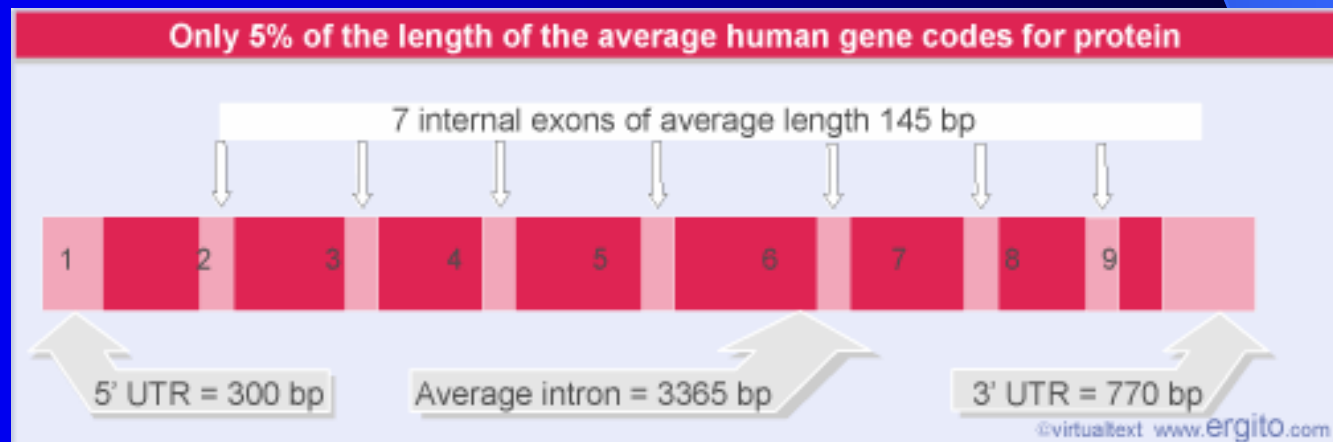
Pseudogenes



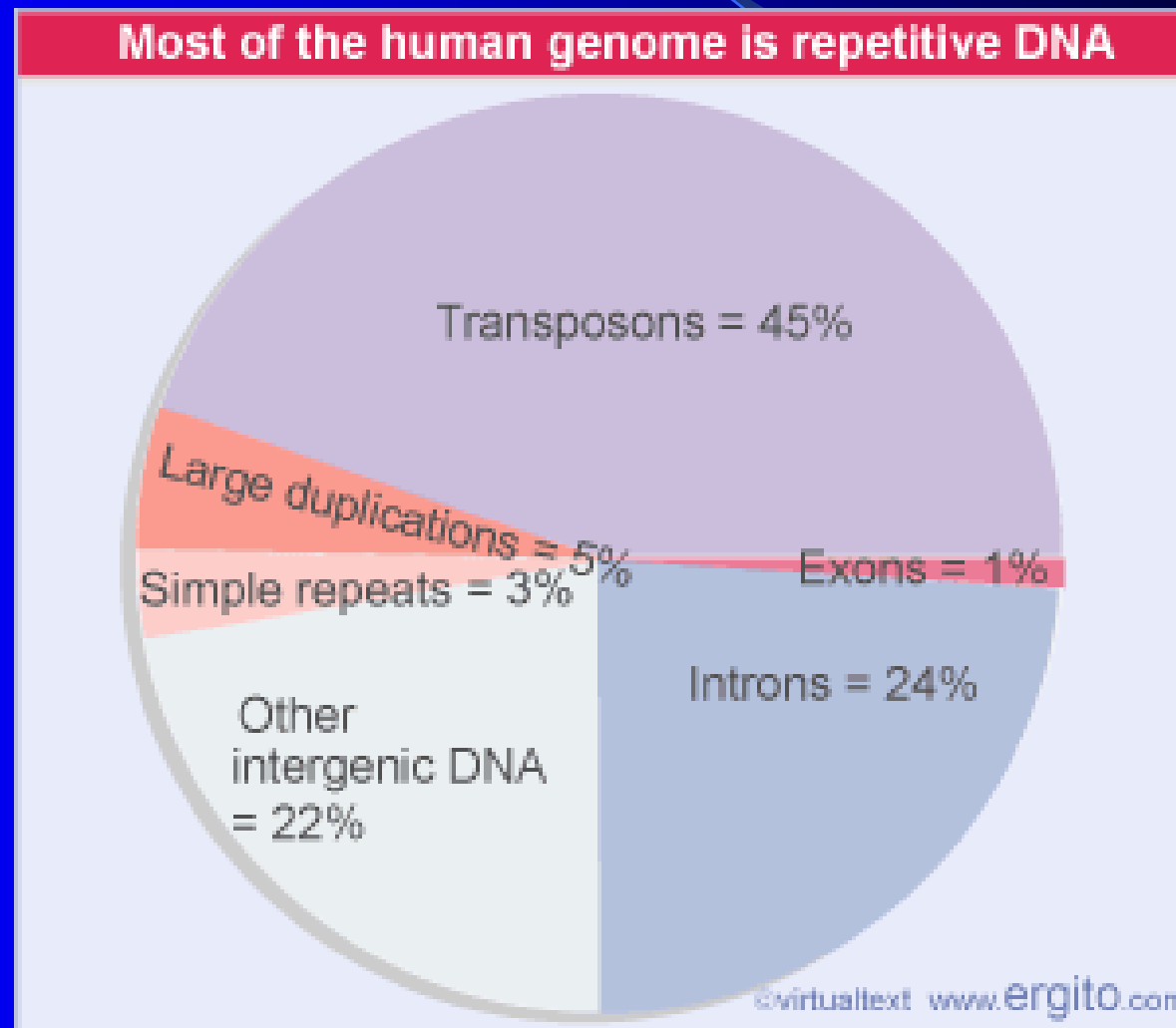
1% of the human genome codes for protein



Only 4% of the length of the average human gene codes for protein

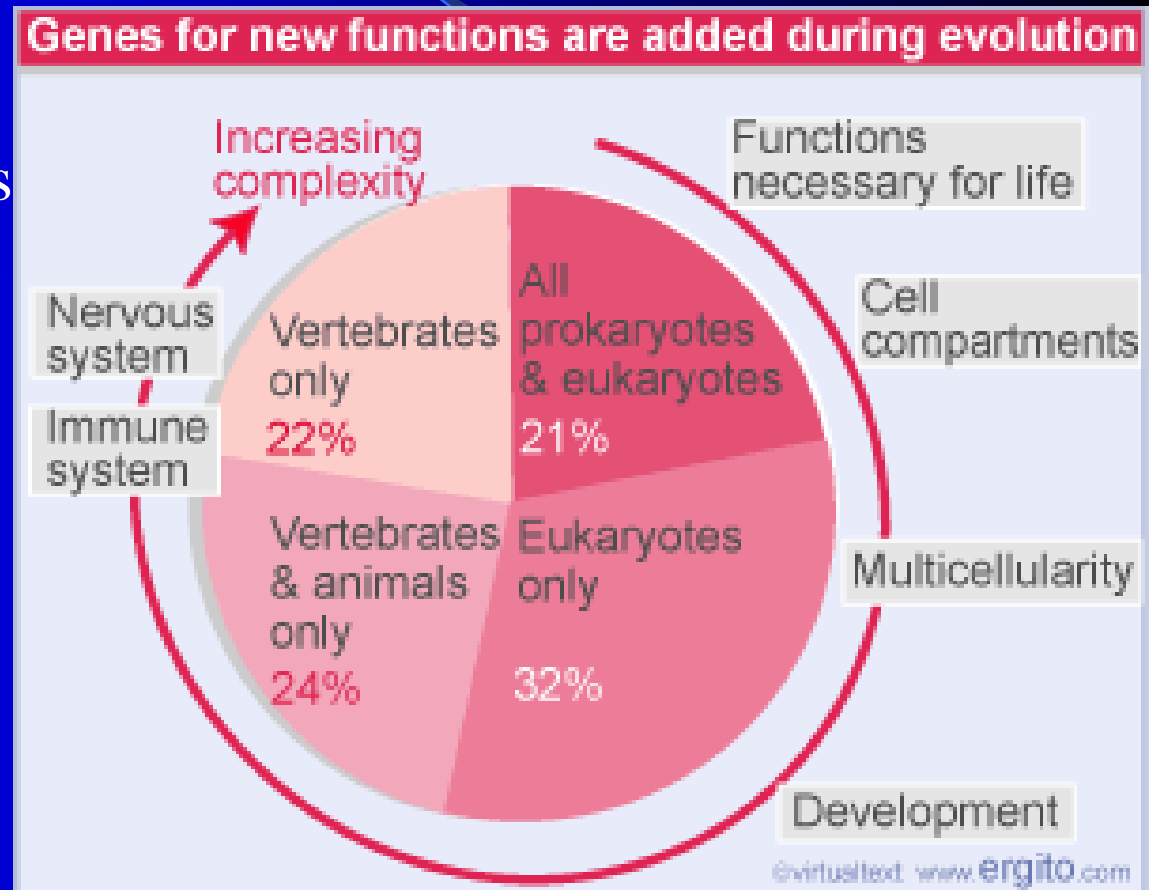


Most of the human genome is repetitive DNA



More complex species evolve by adding new gene functions

- Comparisons of different genomes show a steady increase in gene number as additional genes are added to make eukaryotes, make multicellular organisms, make animals, and make vertebrates.
- Most of the genes that are unique to vertebrates are concerned with the immune or nervous systems.



How many genes?

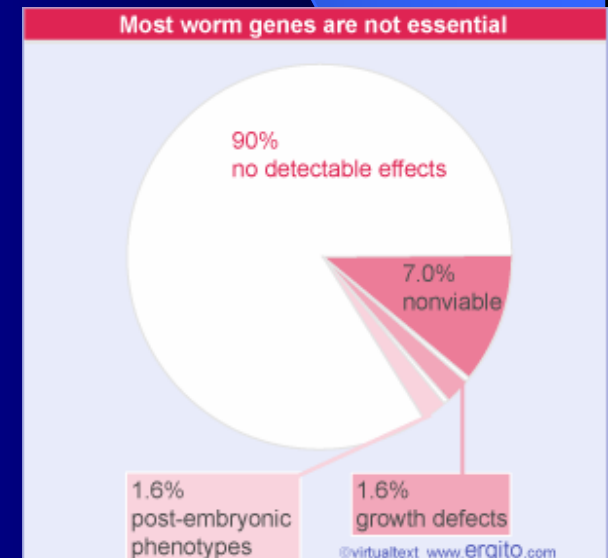
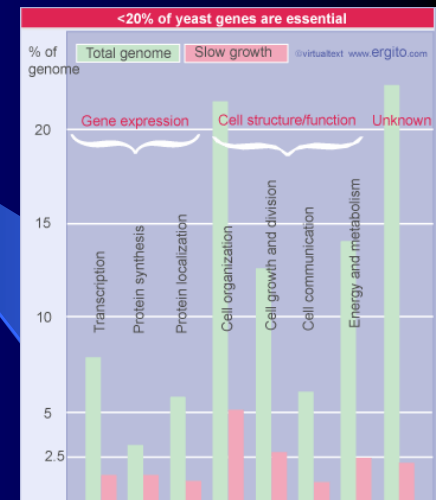
Not all genes are essential

- >20% of yeast genes are essential
- Most worm genes are not essential

The **transcriptome** is the complete set of RNAs present in a cell, tissue, or organism. Its complexity is due mostly to mRNAs, but it also includes noncoding RNAs.

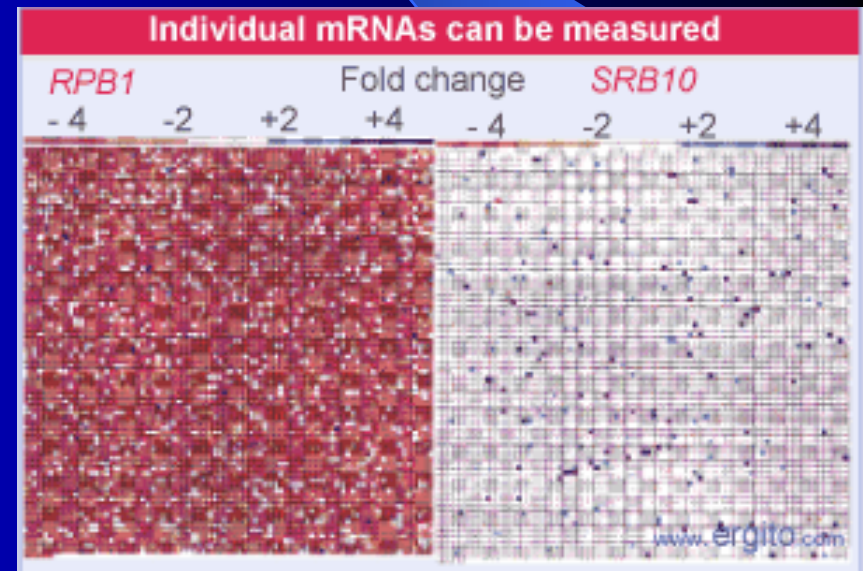
Housekeeping gene (Constitutive gene)s are those (theoretically) expressed in all cells because they provide basic functions needed for sustenance of all cell types.

Luxury genes are those coding for specialized functions synthesized (usually) in large amounts in particular cell types.



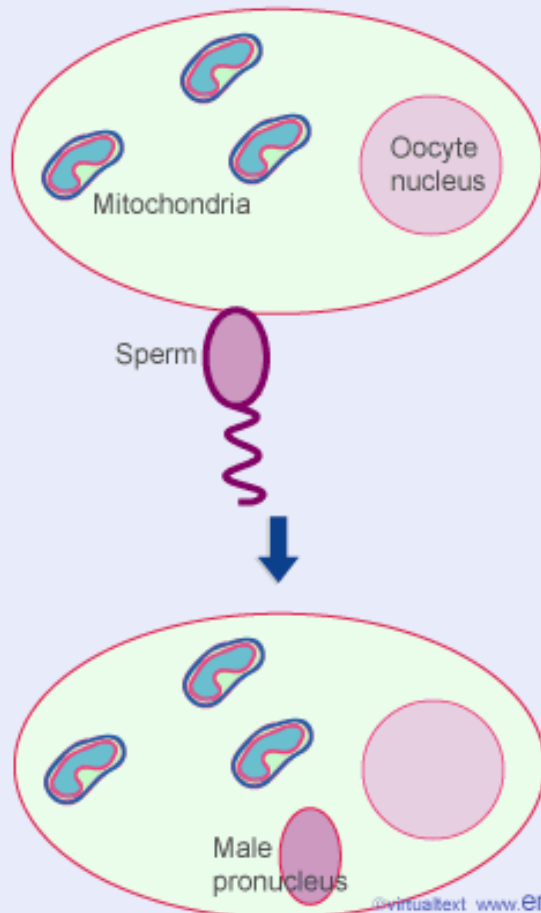
How can the gene expression be measured

- "Chip" technology allows a snapshot to be taken of the expression of the entire genome in a yeast cell.
- ~75% (~4500 genes) of the yeast genome is expressed under normal growth conditions.
- Chip technology allows detailed comparisons of related animal cells to determine (for example) the differences in expression between a normal cell and a cancer cell.



Organelle DNA

Animal mtDNA is inherited from the mother



- **Maternal inheritance** describes the preferential survival in the progeny of genetic markers provided by one parent.
- **Extranuclear genes** reside outside the nucleus in organelles such as mitochondria and chloroplasts.
- **Cytoplasmic inheritance** is a property of genes located in mitochondria or chloroplasts.

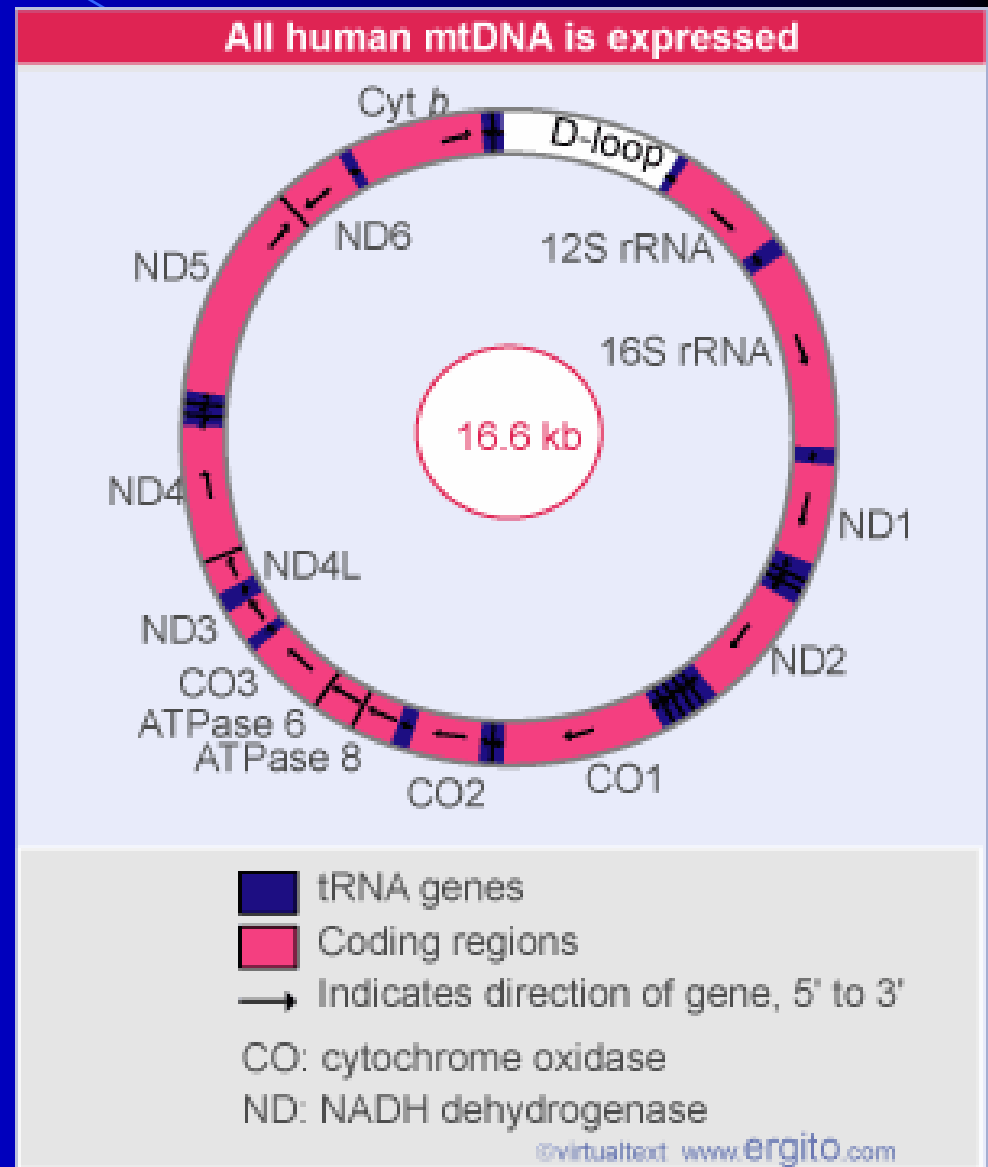
Mitochondrial DNA (mtDNA) is an independent DNA genome, usually circular, that is located in the mitochondrion. **Chloroplast DNA (ctDNA)** is an independent genome (usually circular) found in a plant chloroplast.

Mitochondria code for RNAs and proteins

Species	Size (kb)	Protein-coding genes	RNA-coding genes
Fungi	19-100	8-14	10-28
Protists	6-100	3-62	2-29
Plants	186-366	27-34	21-30
Animals	16-17	13	4-24

Mt DNA are variable

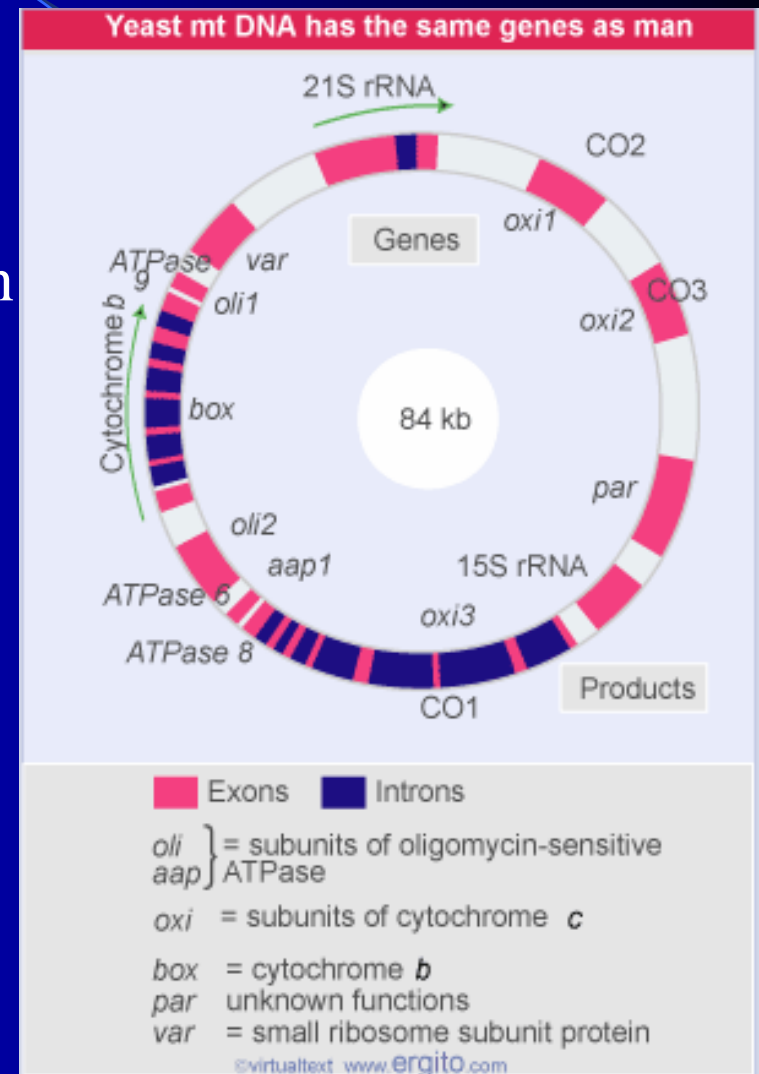
- Animal cell mitochondrial DNA is extremely compact and typically codes for 13 proteins, 2 rRNAs, and 22 tRNAs.
- Yeast mitochondrial DNA is 5× longer than animal cell mtDNA because of the presence of long introns.



Yeast mt DNA has the same genes as man

The two most prominent loci are the interrupted genes *box* (coding for cytochrome *b*) and *oxi3* (coding for subunit 1 of cytochrome oxidase). Together these two genes are almost as long as the entire mitochondrial genome in mammals! Many of the long introns in these genes have open reading frames in register with the preceding exon.

The remaining genes are uninterrupted. They correspond to the other two subunits of cytochrome oxidase coded by the mitochondrion, to the subunit(s) of the ATPase, and (in the case of *var1*) to a mitochondrial ribosomal protein. The total number of yeast mitochondrial genes is unlikely to exceed ~25.



Chloroplasts have >100 genes

- Introns in chloroplasts fall into two general classes. Those in tRNA genes are usually (although not inevitably) located in the anticodon loop, like the introns found in yeast nuclear tRNA genes.
- Those in protein-coding genes resemble the introns of mitochondrial genes
- The role of the chloroplast is to undertake photosynthesis. Many of its genes code for proteins of complexes located in the thylakoid membranes. The constitution of these complexes shows a different balance from that of mitochondrial complexes. Although some complexes are like mitochondrial complexes in having some subunits coded by the organelle genome and some by the nuclear genome, other chloroplast complexes are coded entirely by one genome.

Chloroplasts have >100 genes

Genes

RNA-coding

16S rRNA
23S rRNA
4.5S rRNA
5S rRNA
tRNA

Gene Expression

r-proteins
RNA polymerase
Others

Chloroplast functions

Rubisco & thylakoids
NADH dehydrogenase

Total