

# Clusters and Repeats

By  
Rasul Chaudhry

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.

**Translocation** describes the stage of nuclear import or export when a protein or RNA substrate moves through the nuclear pore.

A **gene cluster** is a group of adjacent genes that are identical or related.

**Nonreciprocal recombination (unequal crossing-over)** results from an error in pairing and crossing-over in which nonequivalent sites are involved in a recombination event. It produces one recombinant with a deletion of material and one with a duplication.

**Satellite DNA (Simple-sequence DNA)** consists of many tandem repeats (identical or related) of a short basic repeating unit.

**Minisatellite DNAs** consist of ~10 copies of a short repeating sequence. the length of the repeating unit is measured in 10s of base pairs. The number of repeats varies between individual genomes

## Unequal crossing-over changes the repeat number

ABCABCABCABCABCABCABCABC

ABCABCABCABCABCABCABCABC



ABCABCABCABCABCABCABCABC

ABCABCABCABCABCABCABCABC

©virtualtext [www.ergito.com](http://www.ergito.com)

# Gene Duplication as a source of Evolution

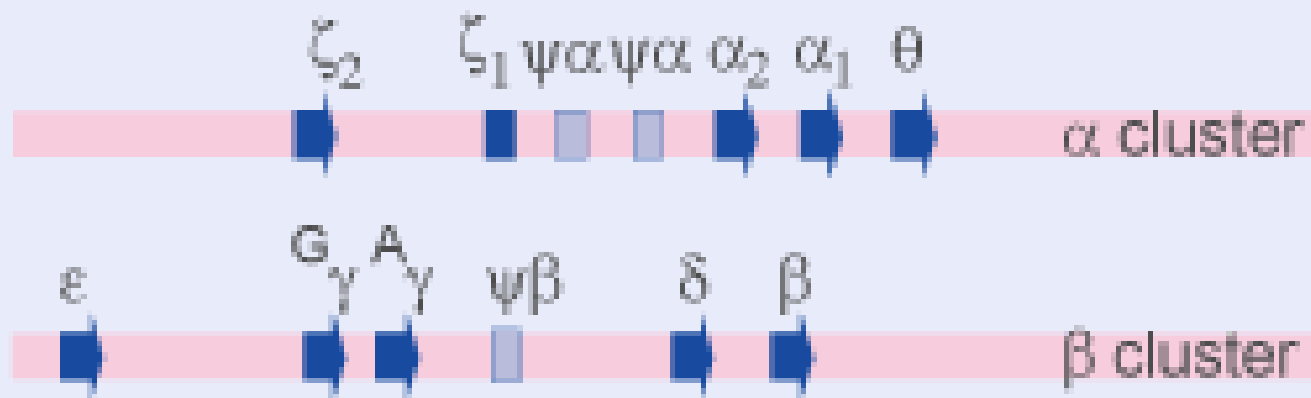
- Duplicated genes may diverge or be silenced.
- There is ~1% probability that a given gene will be included in a duplication in a period of 1 million years. After the gene has duplicated, differences develop as the result of the occurrence of different mutations in each copy. These accumulate at a rate of ~0.1% per million years



- Analysis of the human genome sequence shows that ~5% comprises duplications of identifiable segments ranging in length from 10-300 kb. These have arisen relatively recently, that is, there has not been sufficient time for divergence between them to eliminate their relationship.

# Hemoglobins change during development

Globin genes are organized in two clusters

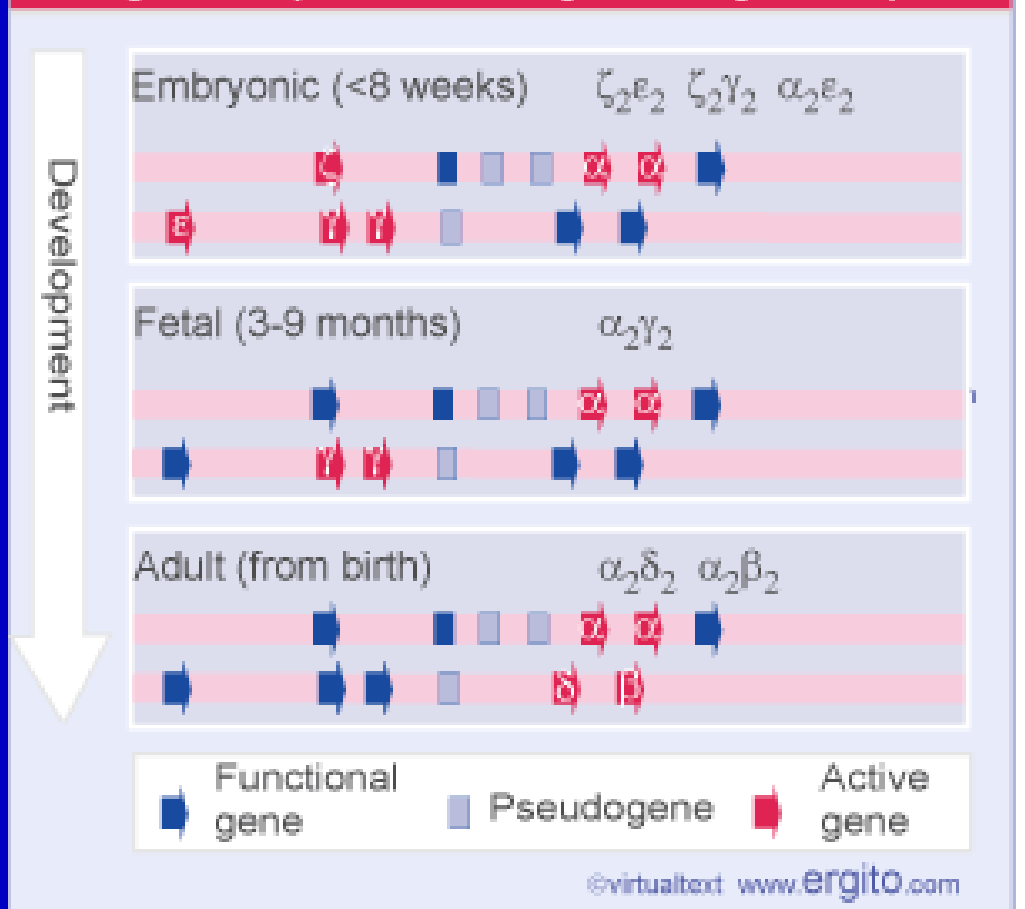


Functional gene      Pseudogene

**Nonallelic** genes are two (or more) copies of the same gene that are present at *different* locations in the genome (contrasted with alleles which are copies of the same gene derived from different parents and present at the same location on the homologous chromosomes).

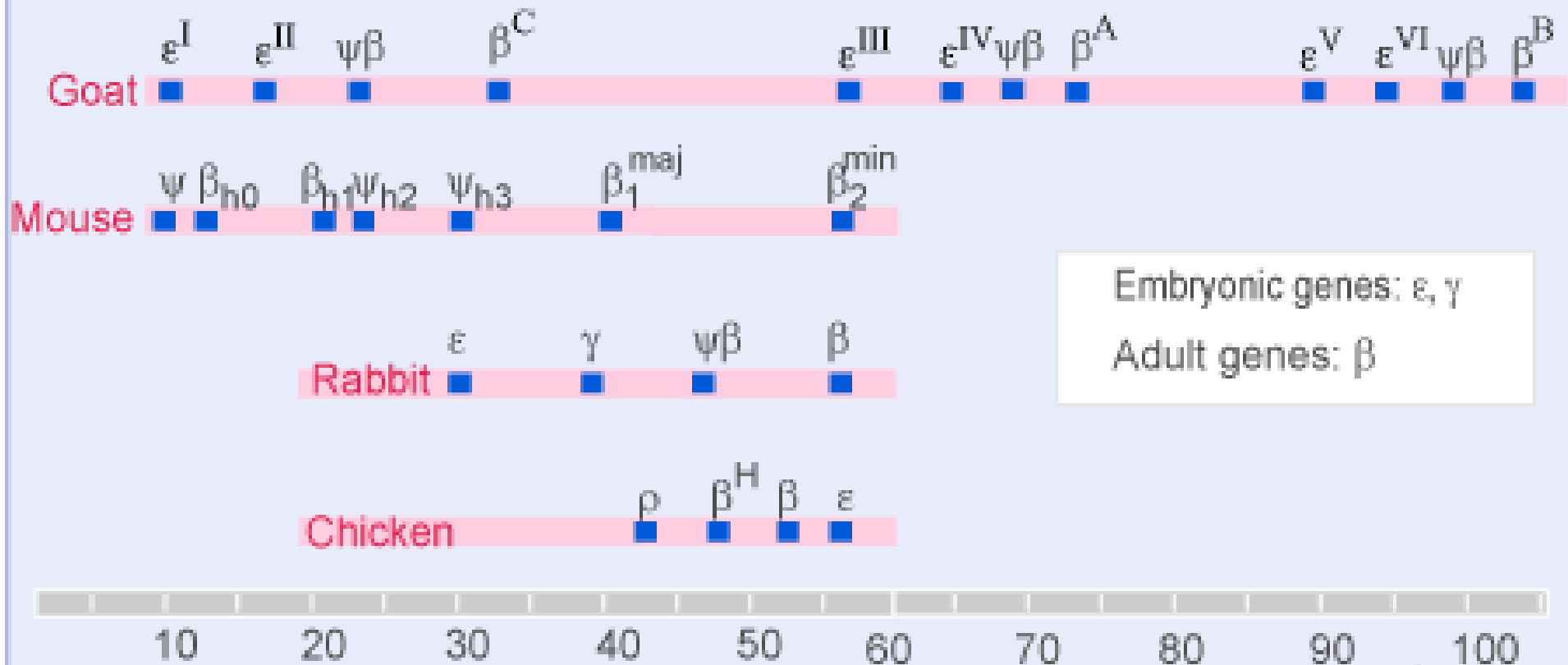
**Pseudogenes** are inactive but stable components of the genome derived by mutation of an ancestral active gene. Usually they are inactive because of mutations that block transcription or translation or both.

### Hemoglobin expression changes during development



# $\beta$ -globin clusters vary between species

## $\beta$ -globin clusters vary between species



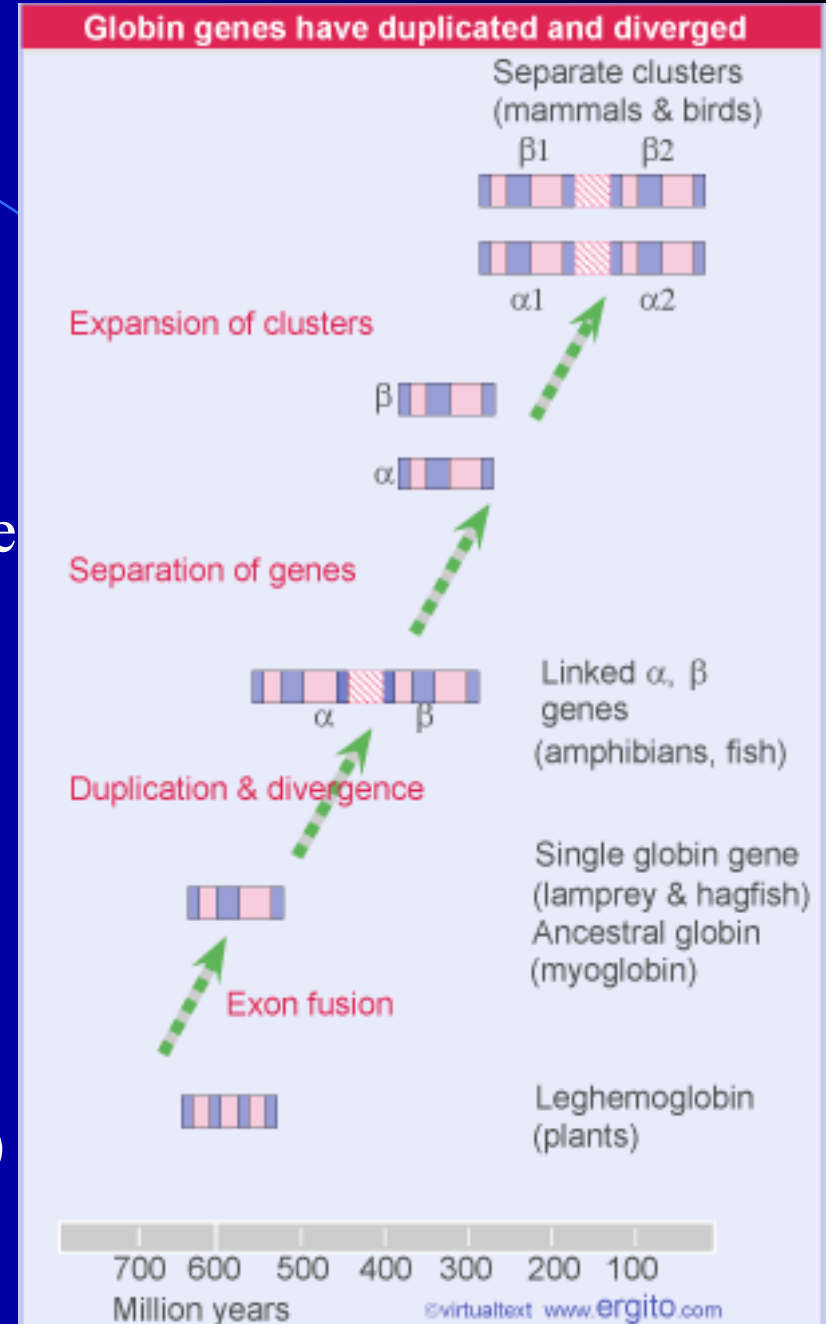
# Globin genes have duplicated and diverged

All globin genes are descended by duplication and mutation from an ancestral gene that had three exons.

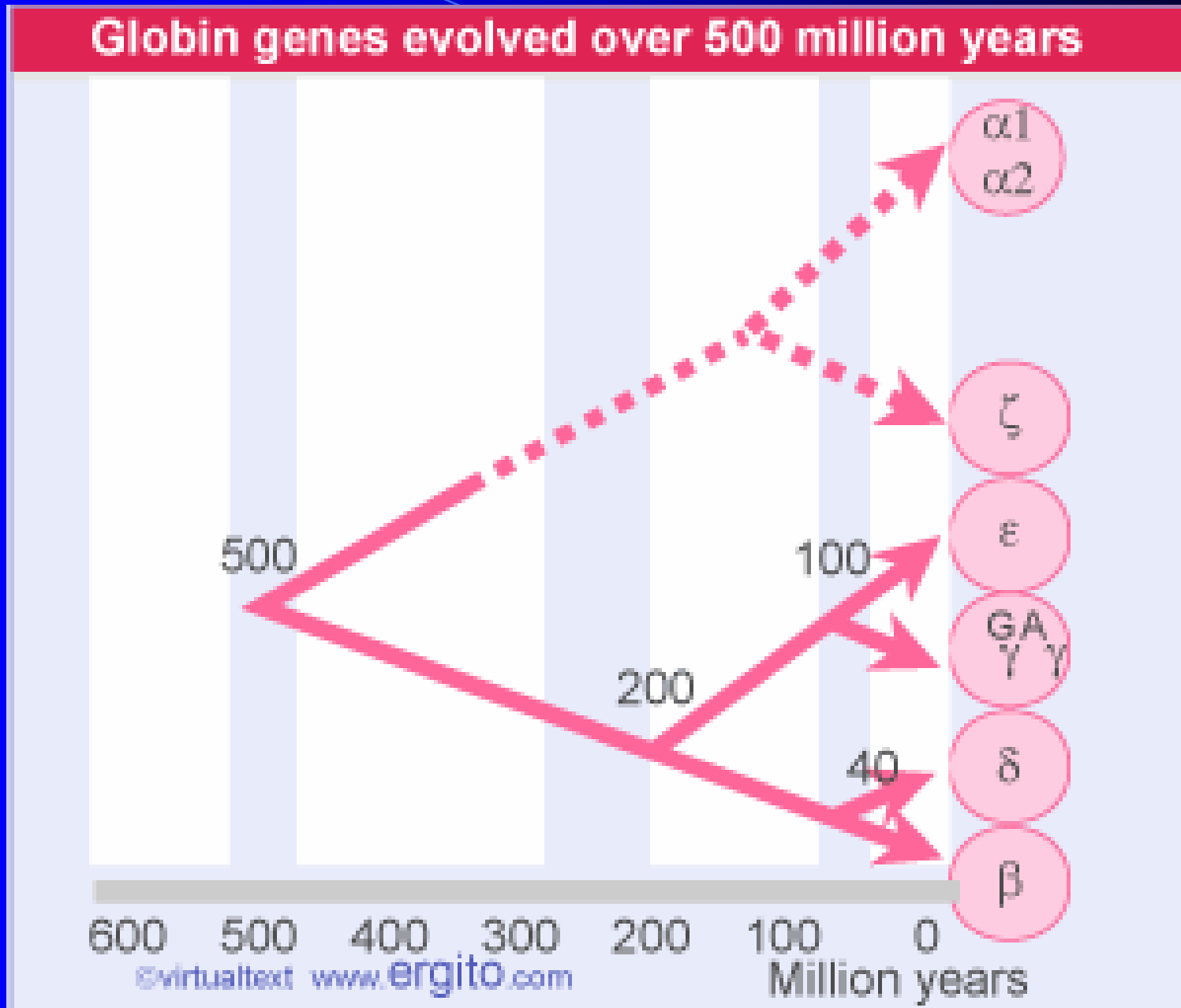
The ancestral gene gave rise to myoglobin, leghemoglobin, and  $\alpha$ - and  $\beta$ -globins.

The  $\alpha$ - and  $\beta$ -globin genes separated in the period of early vertebrate evolution, after which duplications generated the individual clusters of separate  $\alpha$ -like and  $\beta$ -like genes.

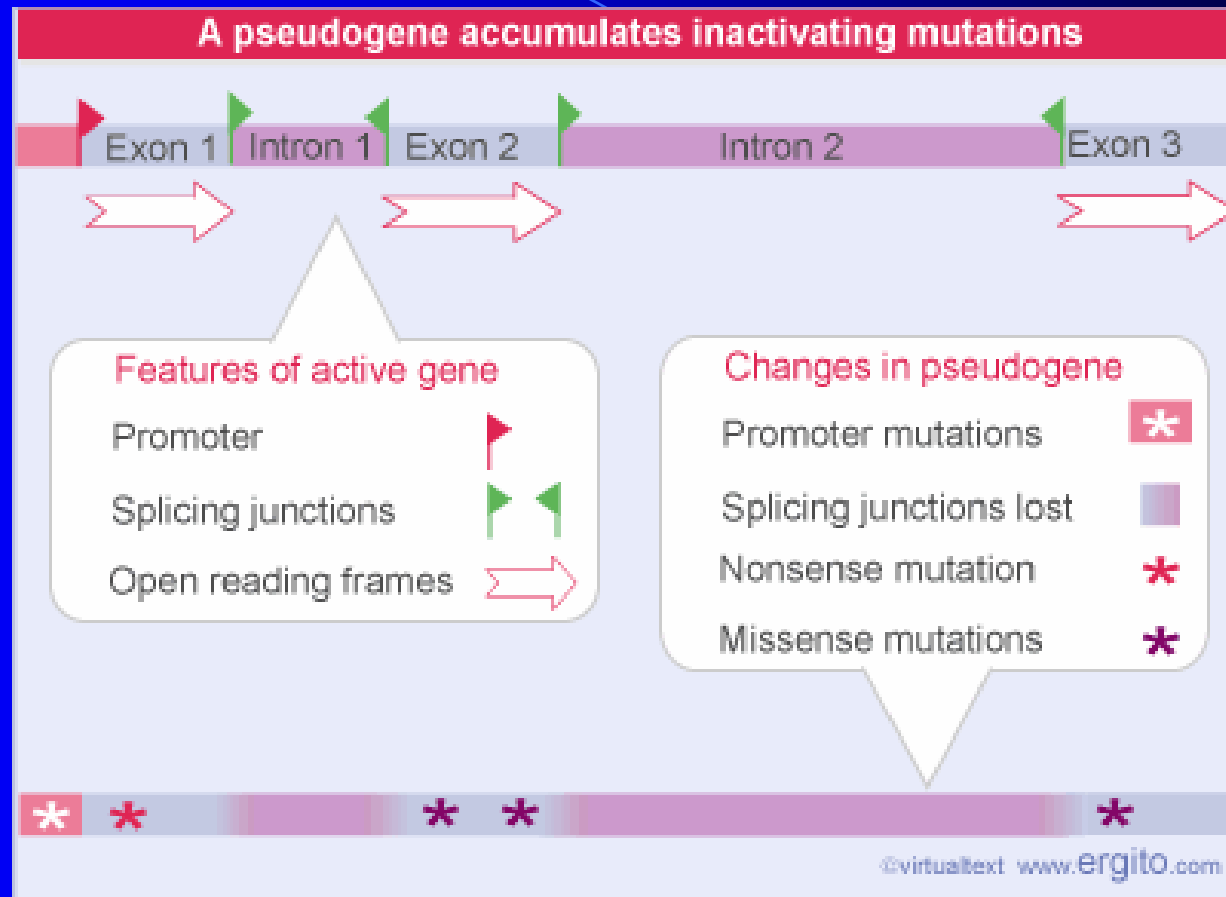
Once a gene has been inactivated by mutation, it may accumulate further mutations and become a pseudogene, which is homologous to the active gene(s) but has no functional role.



# Globin genes evolved over 500 million years



# Mechanisms of development of Pseudogenes



Pseudogenes have no coding function, but they can be recognized by sequence similarities with existing functional genes. They arise by the accumulation of mutations in (formerly) functional genes.

**Thalassemia** is disease of red blood cells resulting from lack of either  $\alpha$  or  $\beta$  globin.

**HbH** disease results from a condition in which there is a disproportionate amount of the abnormal tetramer  $\beta_4$  relative to the amount of normal hemoglobin ( $\alpha_2\beta_2$ ).

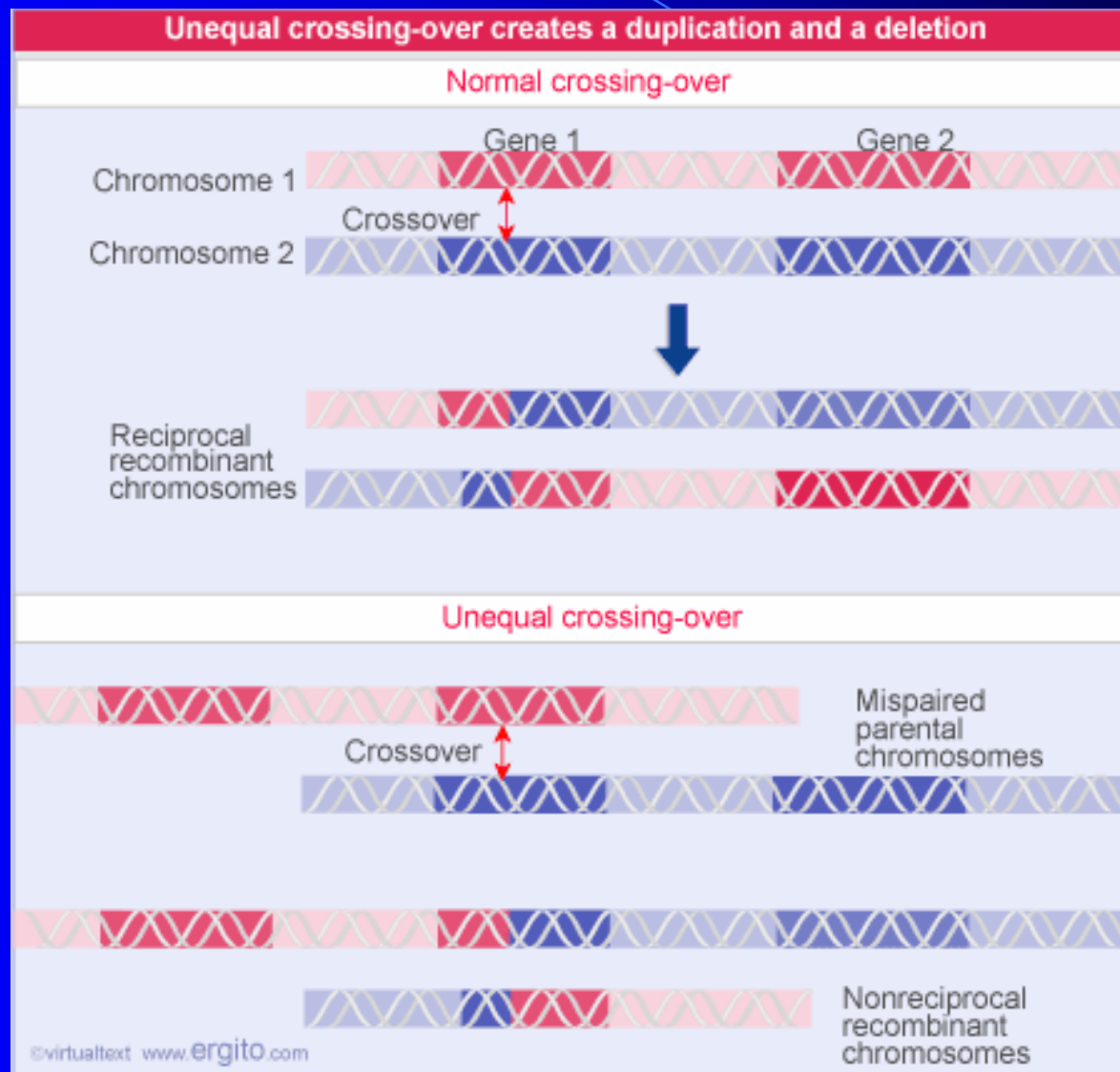
**Hydrops fetalis** is a fatal disease resulting from the absence of the hemoglobin  $\alpha$  gene.

**Hb Lepore** is an unusual globin protein that results from unequal crossing-over between the  $\beta$  and  $\delta$  genes. The genes become fused together to produce a single  $\beta$ -like chain that consists of the N-terminal sequence of  $\delta$  joined to the C-terminal sequence of  $\beta$ .

**Hb anti-Lepore** is a fusion gene produced by unequal crossing-over that has the N-terminal part of  $\beta$  globin and the C-terminal part of  $\delta$  globin.

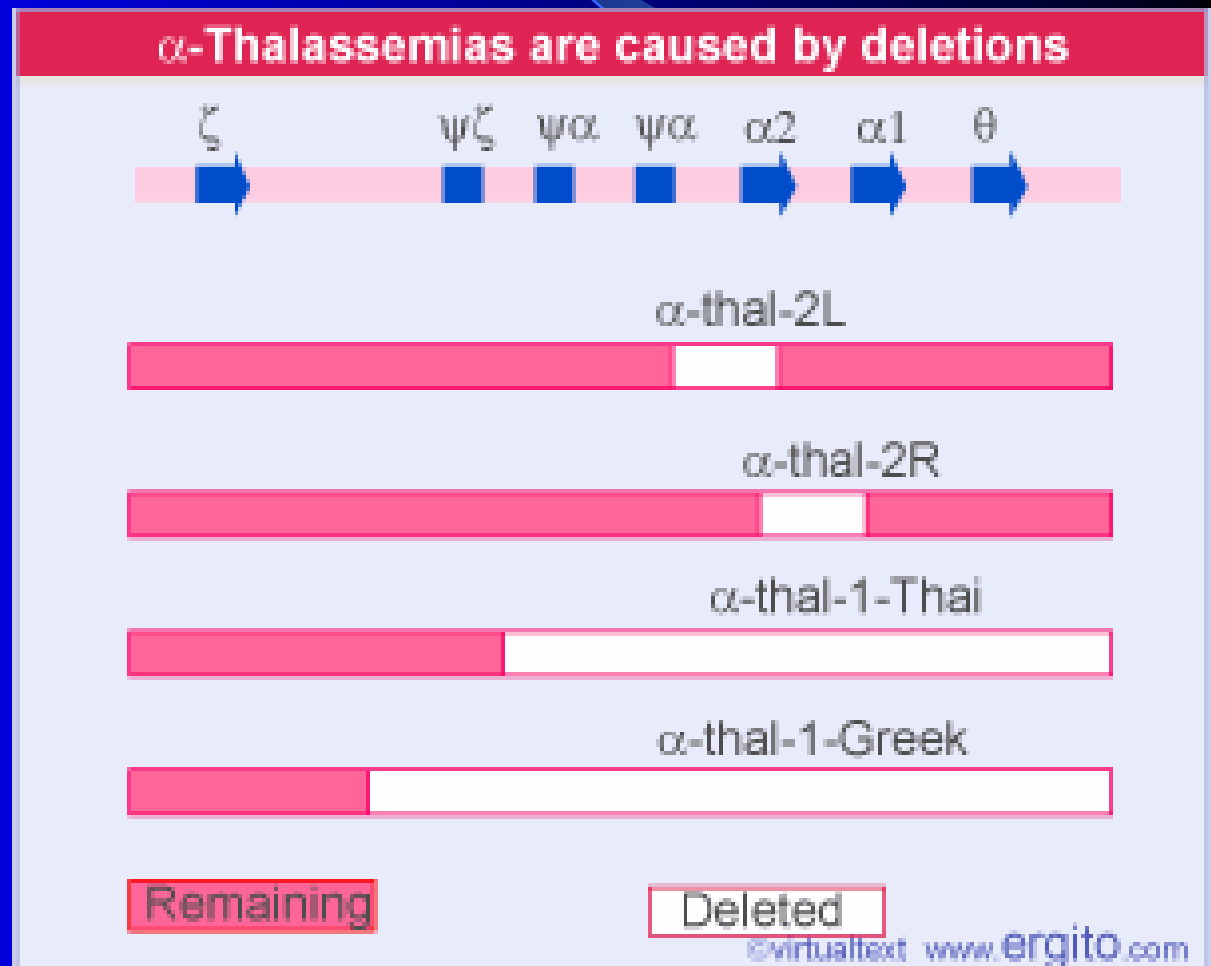
**Hb Kenya** is a fusion gene produced by unequal crossing-over between the between  $\delta$  and  $\beta$  globin genes.

# Unequal crossing-over creates a duplication and a deletion



# $\alpha$ -thalassemias are caused by deletions

The  $\alpha$ -thal-2 deletions are short and eliminate only one of the two  $\alpha$  genes. The L deletion removes 4.2 kb of DNA, including the  $\alpha 2$  gene. It probably results from unequal crossing-over, because the ends of the deletion lie in homologous regions, just to the right of the  $\psi\alpha$  and  $\alpha 2$  genes, respectively. The R deletion results from the removal of exactly 3.7 kb of DNA, the precise distance between the  $\alpha 1$  and  $\alpha 2$  genes. It appears to have been generated by unequal crossing-over between the  $\alpha 1$  and  $\alpha 2$  genes themselves.

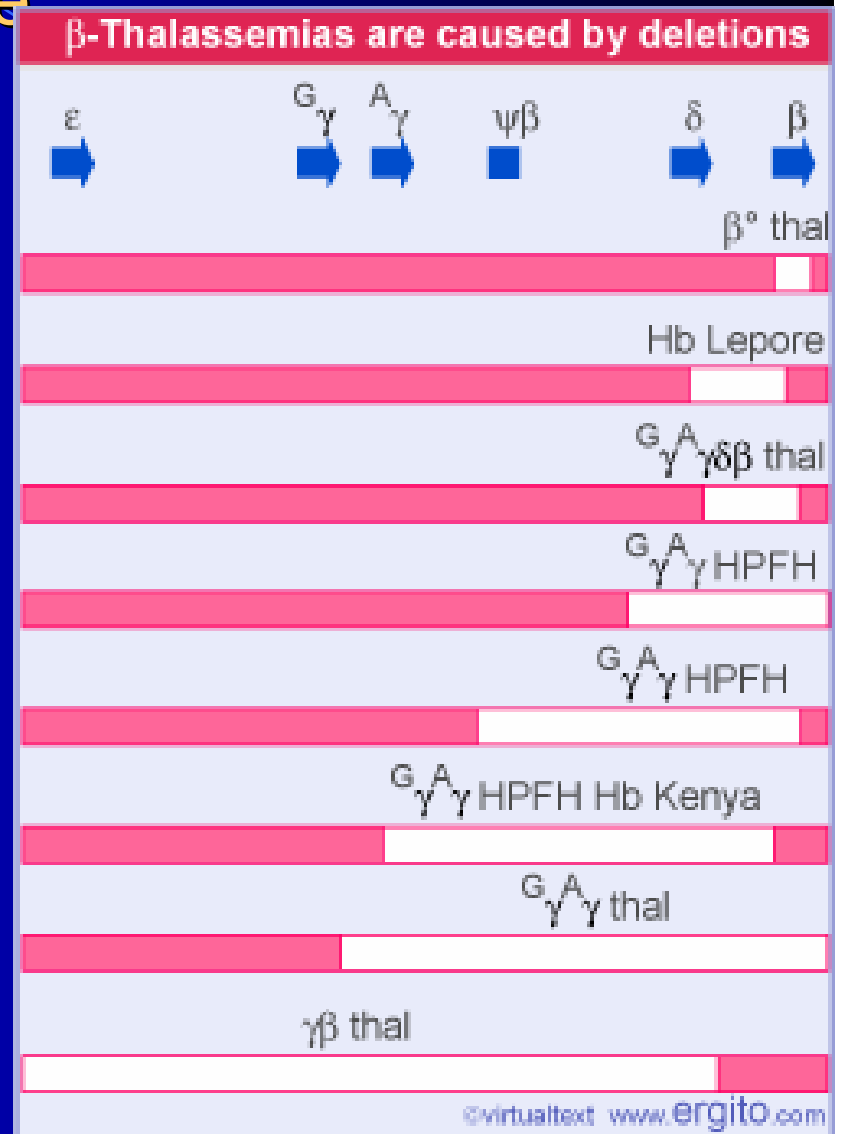


- Depending on the diploid combination of thalassemic chromosomes, an affected individual may have any number of  $\alpha$  chains from zero to three. There are few differences from the wild type (four  $\alpha$  genes) in individuals with three or two  $\alpha$  genes. But with only one  $\alpha$  gene, the excess  $\beta$  chains form the unusual tetramer  $\beta_4$ , which causes **HbH** disease. The complete absence of  $\alpha$  genes results in **hydrops fetalis**, which is fatal at or before birth.
- Variations in the number of  $\alpha$  genes are relatively frequent than  $\beta$  cluster suggesting that unequal crossing-over in the cluster must be fairly common  $\alpha$  genes. The introns in  $\alpha$  genes are much shorter, and therefore possibly present less impediment to mispairing between nonhomologous genes.

**Thalassemias** result from mutations that reduce or prevent synthesis of either  $\alpha$  or  $\beta$  globin. The occurrence of unequal crossing-over in the human globin gene clusters is revealed by the nature

of certain thalassemias.

In some (rare) cases, only the  $\beta$  gene is affected. These have a deletion of 600 bp, extending from the second intron through the 3' flanking regions. In the other cases, more than one gene of the cluster is affected. Many of the deletions are very long, extending from the 5' end indicated on the map for >50 kb toward the right.



## **Genes for rRNA form tandem repeats**

**Ribosomal DNA (rDNA)** is usually a tandemly repeated series of genes coding for a precursor to the two large rRNAs.

The **nucleolus (nucleoli)** is a discrete region of the nucleus where ribosomes are produced.

The **nucleolar organizer** is the region of a chromosome carrying genes coding for rRNA.

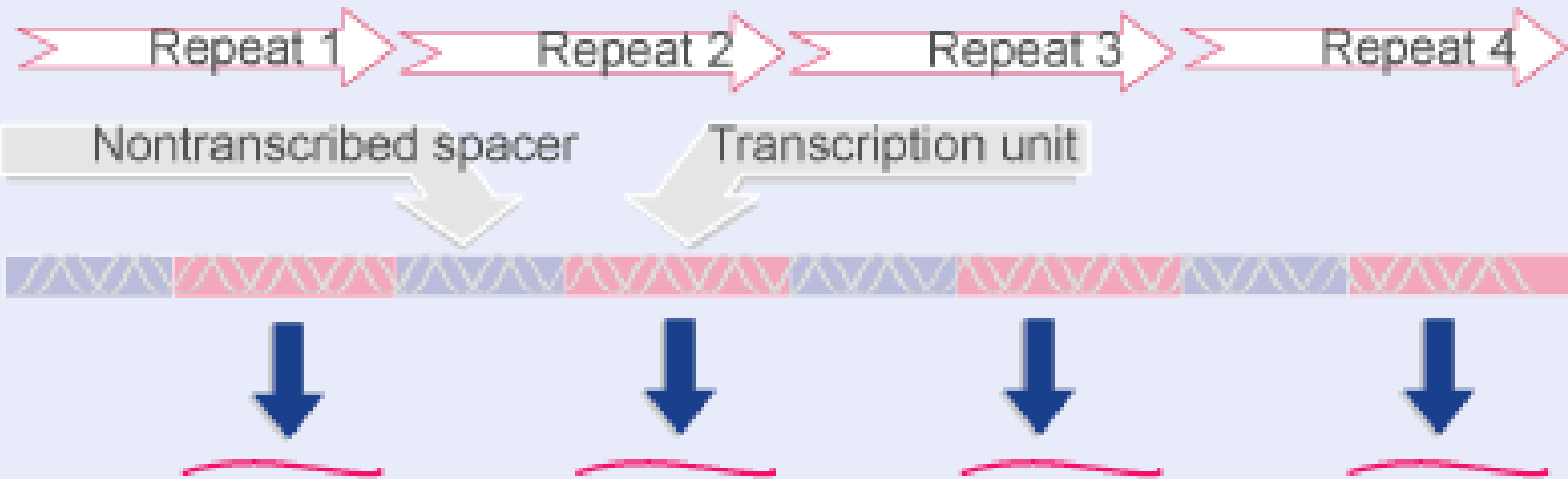
The **nontranscribed spacer** is the region between transcription units in a tandem gene cluster.

# Prokaryotic rDNA

- In bacteria, the multiple rRNA gene pairs are dispersed. In most eukaryotic nuclei, the rRNA genes are contained in a tandem cluster or clusters. Sometimes these regions are called rDNA. In some cases, the proportion of rDNA in the total DNA, together with its atypical base composition, is great enough to allow its isolation as a separate fraction directly from sheared genomic DNA.

# A repeat cluster generates a circular restriction map

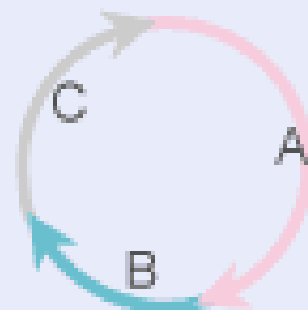
## Linear organization of cluster



## Restriction cleavage sites



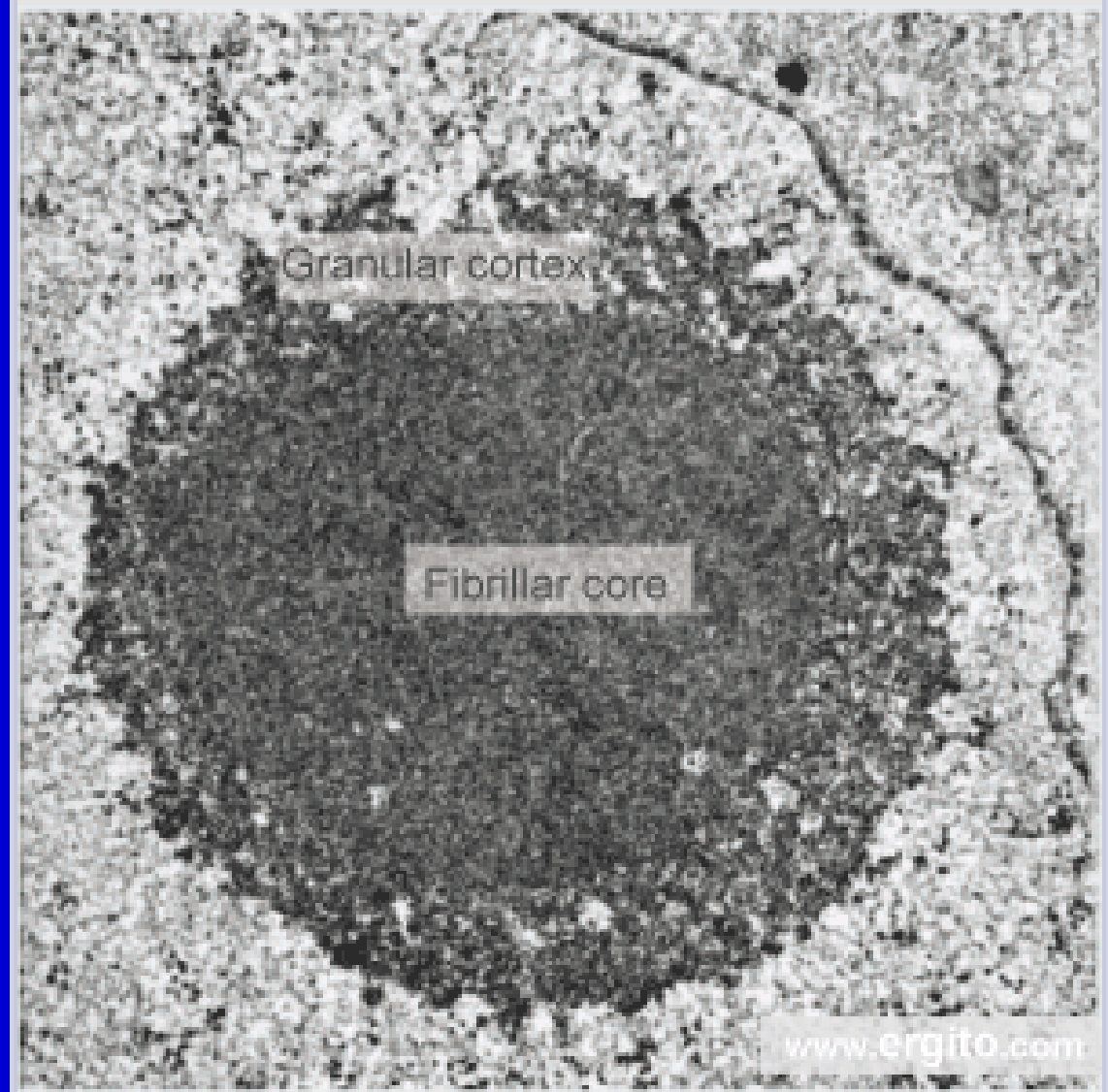
## Restriction map



Nuclear core – rDNA undergoing transcription  
Granular cortex consist of assembling ribosomal units

- The region of the nucleus where rRNA synthesis occurs has a characteristic appearance, with a core of fibrillar nature surrounded by a granular cortex. The fibrillar core is where the rRNA is transcribed from the DNA template; and the granular cortex is formed by the ribonucleoprotein particles into which the rRNA is assembled. The whole area is called the **nucleolus**.

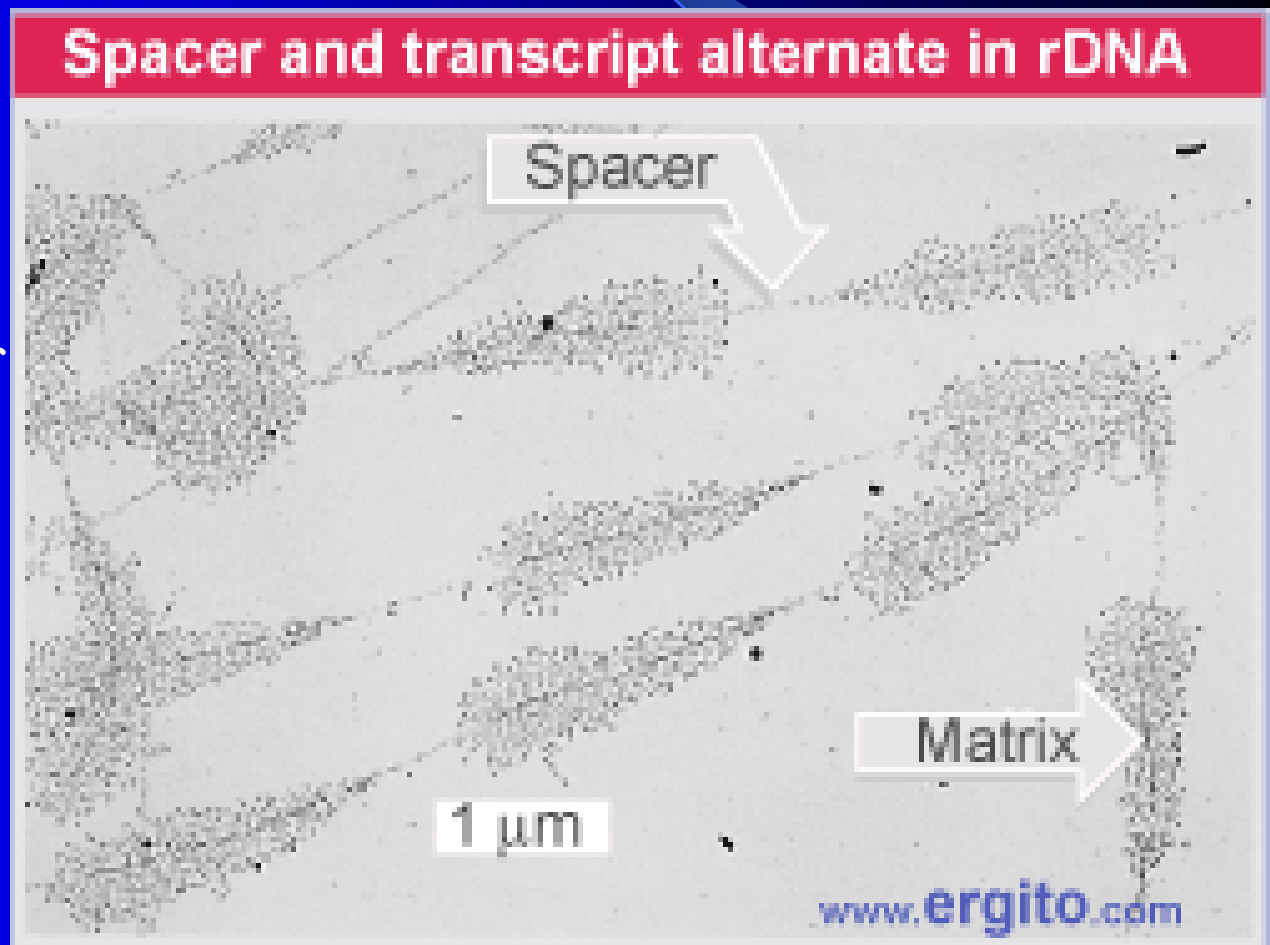
The nucleolus is a dense granular structure



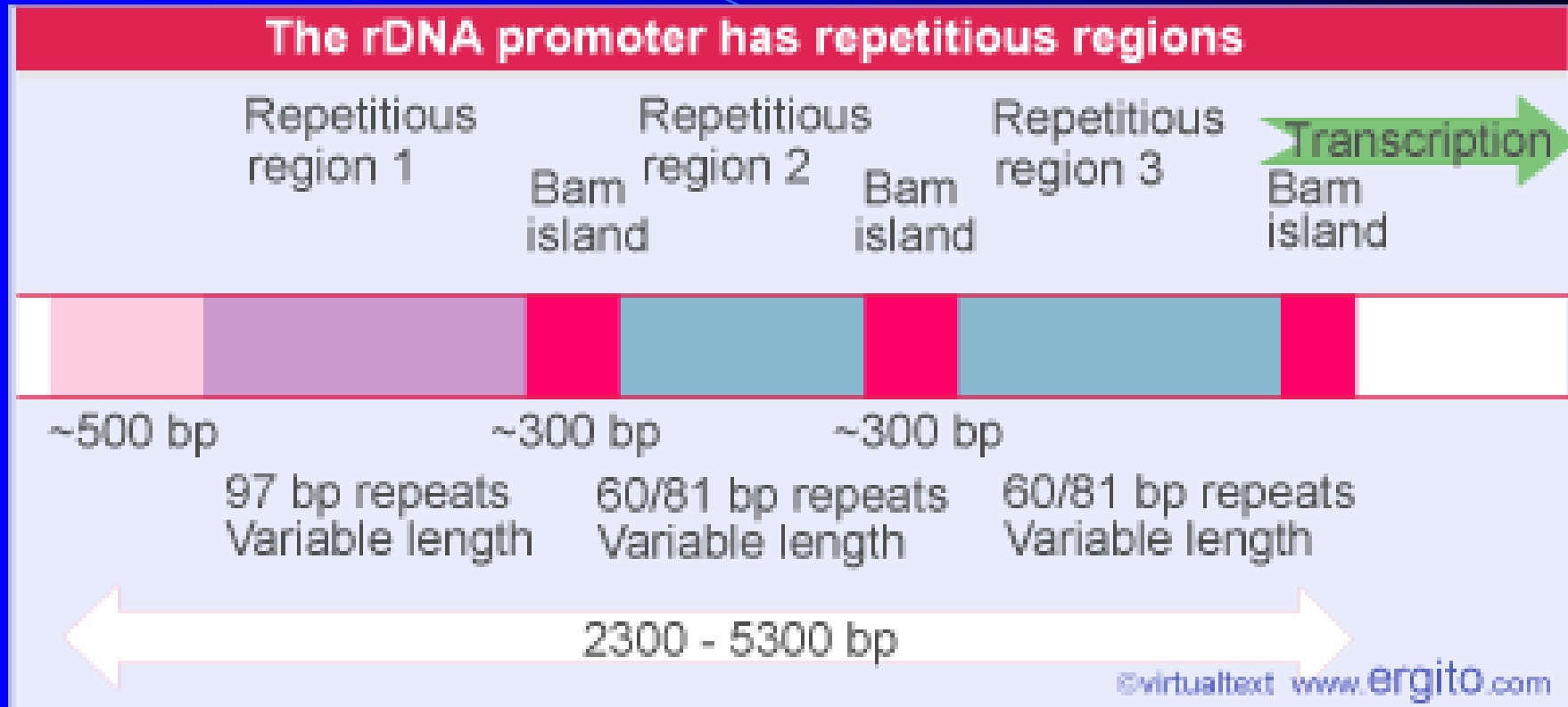
# Matrix correspond to transcription unit

- **Bam islands** are a series of short, repeated sequences found in the nontranscribed spacer of *Xenopus* rDNA genes. The name reflects their isolation by use of the BamI restriction enzyme.

In mammals the repeating unit is very much larger, comprising the transcription unit of ~13 kb and a nontranscribed spacer of ~30 kb. Usually, the genes lie in several dispersed clusters—in the case of man and mouse residing on five and six chromosomes, respectively.



# The rDNA promoter has repetitious regions



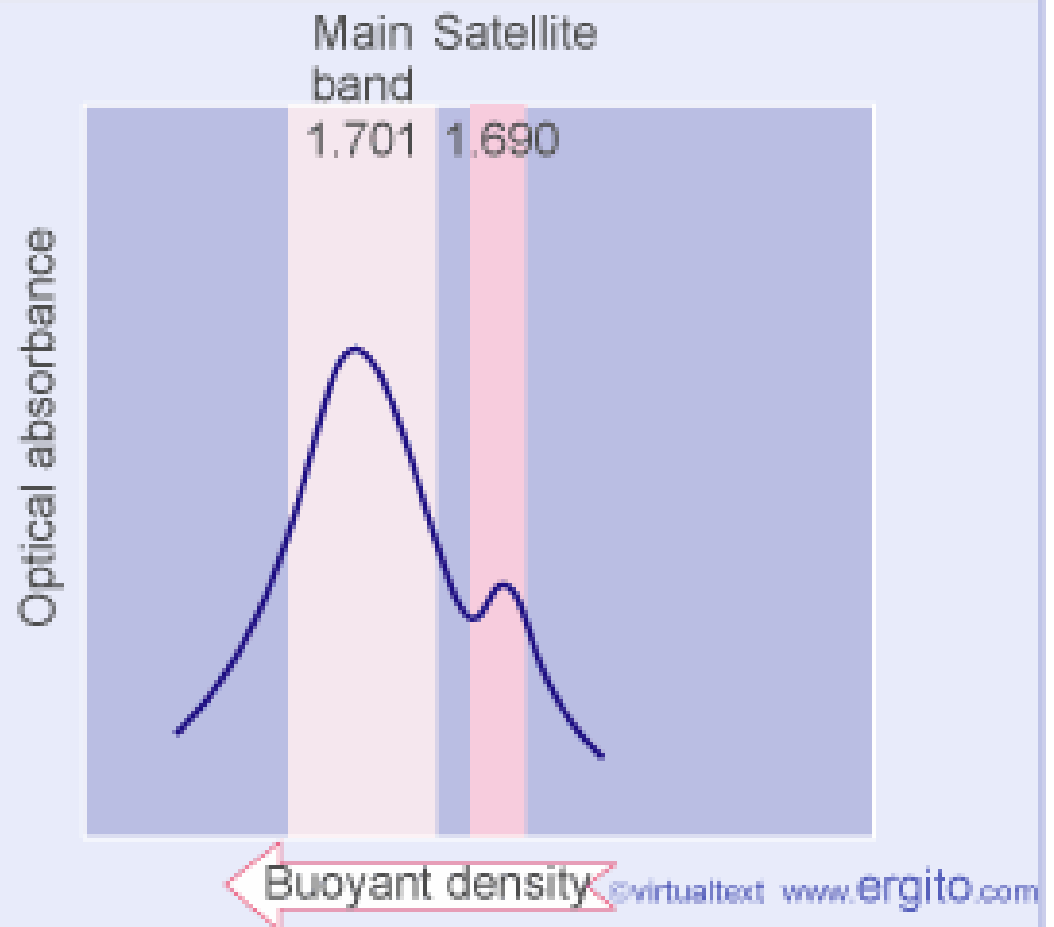
Each of the three repetitious regions (spacer) comprises a variable number of repeats of a rather short sequence. One type of repetitious region has repeats of a 97 bp sequence; the other, which occurs in two locations, has a repeating unit found in two forms, 60 bp and 81 bp long. The variation in the number of repeating units in the repetitious regions accounts for the overall variation in spacer length. The repetitious regions are separated by shorter constant sequences called **Bam islands**.

# Satellite DNAs often lie in heterochromatin

Satellites are present in many eukaryotic genomes. They may be either heavier or lighter than the main band; but it is uncommon for them to represent >5% of the total DNA.

Here the main band (a buoyant density of  $1.701 \text{ g-cm}^{-3}$ ) contains 92% of the genome (its average G•C of 42%, typical for a mammal). The smaller peak (at buoyant density of  $1.690 \text{ g-cm}^{-3}$ ) represents 8% of the genome. It contains the mouse satellite DNA, whose G•C content (30%) is much lower than any other part of the genome.

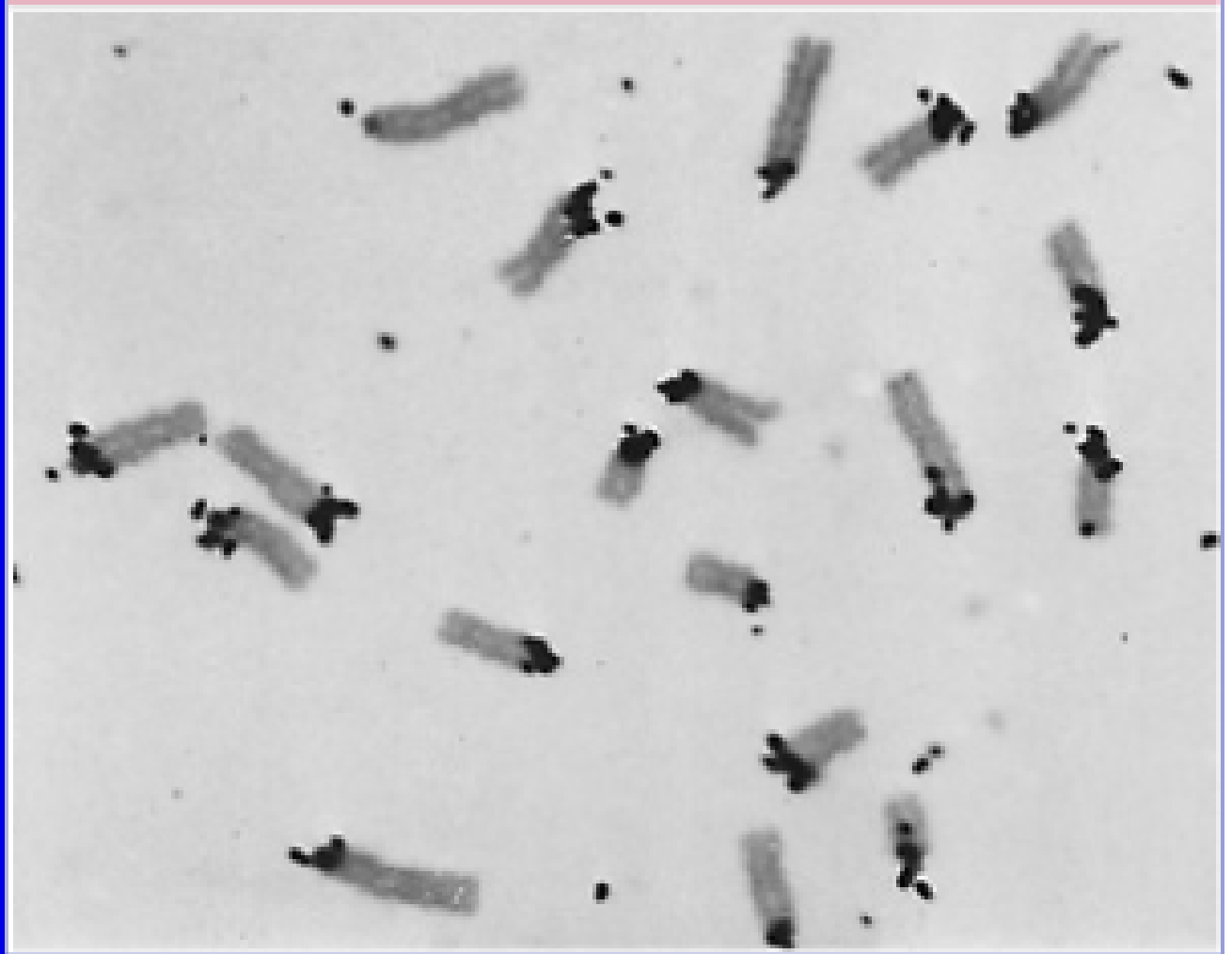
Mouse satellite DNA forms a distinct band



- **Highly repetitive DNA (Simple sequence DNA)** is the first component to reassociate and is equated with satellite DNA.
- **Satellite DNA (Simple-sequence DNA)** consists of many tandem repeats (identical or related) of a short basic repeating unit.
- A **density gradient** is used to separate macromolecules on the basis of differences in their density. It is prepared from a heavy soluble compound such as CsCl.
- A **cryptic satellite** is a satellite DNA sequence not identified as such by a separate peak on a density gradient and remains in main-band DNA.
- ***In situ* hybridization (Cytological hybridization)** is performed by denaturing the DNA of cells squashed on a microscope slide so that reaction is possible with an added single-stranded RNA or DNA; the added preparation is radioactively labeled and its hybridization is followed by autoradiography.
- **Heterochromatin** describes regions of the genome that are highly condensed, are not transcribed, and are late-replicating. Heterochromatin is divided into two types, which are called constitutive and facultative.
- **Euchromatin** comprises all of the genome in the interphase nucleus except for the heterochromatin. The euchromatin is less tightly coiled than heterochromatin, and contains the active or potentially active genes.

In the technique of *in situ* hybridization, the chromosomal DNA is denatured by treating cells that have been squashed on a cover slip. Then a solution containing a radioactively labeled DNA or RNA probe is added. The probe hybridizes with its complements in the denatured genome. The location of the sites of hybridization can be determined by autoradiography

Mouse centromeres contain satellite DNA



*Drosophila virilis* has three major satellites and also a cryptic satellite, together representing >40% of the genome.

The three major satellites have closely related sequences. A single base substitution is sufficient to generate either satellite II or III from the sequence of satellite I.

### *D. virilis* has four related satellites

Satellite	Predominant Sequence	Total Length	Genome Proportion
I	ACAAACT TGTTTGA	$1.1 \times 10^7$	25%
II	ATAAACT TATTTGA	$3.6 \times 10^6$	8%
III	ACAAATT TGTTTAA	$3.6 \times 10^6$	8%
Cryptic	AATATAG TTATATC		

- **Heavy strands** and light strands of a DNA duplex refer to the density differences that result when there is an asymmetry between base representation in the two strands such that one strand is rich in T and G bases and the other is rich in C and A bases. This occurs in some satellite and mitochondrial DNAs.

Mouse satellite DNA has evolved by duplication and mutation of a short repeating unit to give a basic repeating unit of 234 bp in which the original half, quarter, and eighth repeats can be recognized.

Mouse satellite DNA has evolved by duplication and mutation of a short repeating unit to give a basic repeating unit of 234 bp in which the original half, quarter, and eighth repeats can be recognized.

Half repeats of mouse satellite DNA are closely related

10	20	30	40	50	60	70	80	90	100	110	
GGACGTGGAAATATGGCGAGAAAACTGAAAAATCATGGAAAAATGAGAAATACACACTTTACGGACGTGAAATATGCGGACGAAACTGAAAAACGTGGAAAAATTAGAAATGTCCACTGTAA											
GGACGTGGAAATATGGCAAGAAAACTGAAAAATCATGGAAAAATGAGAAACATCCACTTTGACGACITGAAAAATGACGAAATGCACTAAAAAACGTGAAAAATGAGAAATGCACACTGTAA											
120	130	140	150	160	170	180	190	200	210	220	230

©virtualtext www.ergito.com

By writing the 234 bp sequence so that the first 117 bp are aligned with the second 117 bp, we see that the two halves are quite well related. They differ at 22 positions, corresponding to 19% divergence.

Within the 117 bp unit, we can recognize two further subunits. Each of these is a quarter-repeat relative to the whole satellite.

Mouse satellite DNA can be organized into quarter-repeats

	10	20	30	40	50
	GGACCT	GGAATAT	GGCGAGAA	AACTGAAAAT	CACGGAAAAT
					GAGAAATACACACTTTA
60	70	80	90	100	110
	GGACGT	GAAATAT	GGCGAGAA	AACTGAAAAAGGT	GGAAAAT
					TAGAAATGTCCACTGTA
120	130	140	150	160	170
	GGACGT	GGAATAT	GGCAAGAA	AACTGAAAAT	CATGGAAAAT
					GAGAAACATCCACTTGA
180	190	200	210	220	230
	CGACTT	GAAAAT	GACGAAAT	CACTAAAAACGT	GAAAAT
					GAGAAATGCACACTGAA

The quarter-repeats are consist of two related subunits (one-eighth-repeats), shown as the  $\alpha$  and  $\beta$  sequences.

One eighth repeats identify the mouse satellite ancestral unit	
$\alpha 1$	GGACCTGGAATATGGCGAGAA AACTGAA
$\beta 1$	AATCACGGAAAATGA GAAATACACACTTTA
$\alpha 2$	GGACGTGGAATATGGCGAGAA <sup>G</sup> AACTGAA
$\beta 2$	AAAGGTGGA <sup>T</sup> AAAATTA GAAATGTCCACTGTA
$\alpha 3$	GGACGTGGAATATGGCAAGAA AACTGAA
$\beta 3$	AATCATGGAAAAATGA GAAACATCCACTTGA
$\alpha 4$	CGACTTGAAAAATGACGAAAT CACTAAA
$\beta 4$	AAACGTGAAAAATGA GAAATGCACACTGAA
Consensus	AAACGTGAAAAATGA GAAAT CACTGAA
Ancestral?	AAACGTGAAAAATGA GAAATGCACACTGAA

©virtualtext www.ergito.com

The  $\alpha$  sequences all have an insertion of a C, and the  $\beta$  sequences all have an insertion of a trinucleotide, relative to a common consensus sequence.

The current satellite sequence can be treated as derivatives of a 9 bp sequence.

G A A A A A C G T

G A A A A A T G A

G A A A A A A C T

The mouse satellite DNA consensus is 9 bp

```

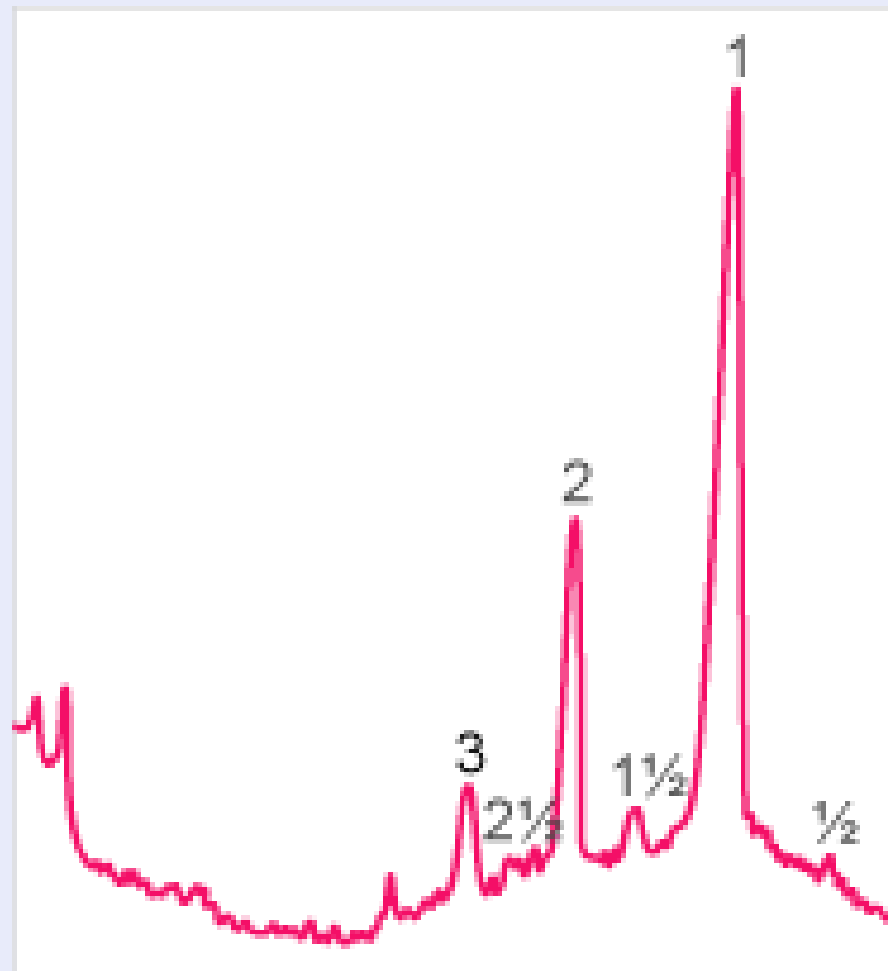
      G G A C C T
G G A A T A T G G C
G A G A A A A C T
G A A A A T C A C
G G A A A A T G A
G A A A T C A C T
T T A G G A C G T
G A A A T A T G G C
G A G AG A A A C T
G A A A A A G G T
G G A A A A TT T A
G A A A T* C A C T
G T A G G A C G T
G G A A T A T G G C
A A G A A A A C T
G A A A A T C A T
G G A A A A T G A
G A A A C* C A C T
T G A C G A C T T
G A A A A A T G A C
G A A A T C A C T
A A A A A A C G T
G A A A A A T G A
G A A A T* C A C T
G A A

```

G<sub>20</sub> A<sub>16</sub> A<sub>21</sub> A<sub>20</sub> A<sub>12</sub> A<sub>17</sub> T<sub>8</sub> G<sub>11</sub> A<sub>5</sub>  
 T<sub>7</sub> C<sub>5</sub> A<sub>8</sub> C<sub>9</sub> T<sub>15</sub>  
 C<sub>7</sub>

\* indicates inserted triplet in β sequence  
 C in position 10 is extra base in α sequence

## Mouse satellite DNA has repeats and half-repeats



← Size ©virtualtext www.ergito.com

- **Microsatellite** DNAs consist of repetitions of extremely short (typically <10 bp) units.
- **Minisatellite** DNAs consist of ~10 copies of a short repeating sequence. the length of the repeating unit is measured in 10s of base pairs. The number of repeats varies between individual genomes.
- **VNTR** (variable number tandem repeat) regions describe very short repeated sequences, including microsatellites and minisatellites.
- **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)