

13.1 Evolution by gene duplication

13.1.1 Introduction: how can a new gene emerge?

13.1.2. Internal gene duplication

13.1.3. Complete gene duplication

13.1.4. Divergence time of duplicated copies

13.1.5. Conclusion and overview of other mechanisms of new gene formation

13.1 Introduction: How new genes emerge?

Species	Haploid genome size in bp	Number of genes
Escherichia coli	4,639,221	4,377
Saccharomyces cerevisiae	12,495,682	5,770
Drosophila melanogaster	653,977	13,379
Homo sapiens	3.3×10^9	20,000

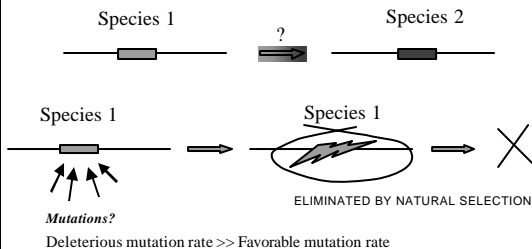
Vertebrate genomes contain many gene families that are absent in invertebrates

Question: How "new" genes appear?

13.1.1 Introduction: How new genes emerge?

Levels of evolution: - increase of number of genes
- new functions

How to get a new function without destroying the existing gene?



13.1.1 Introduction: Ohno's hypothesis

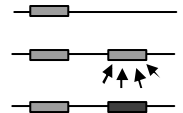


Ohno's hypothesis on the role of gene duplication in evolution

Duplications might allow mutations in the extra copy of the gene: one functional copy is kept intact.

Over time, mutations accumulate on the copy

- either not functional
- or functional -> a new gene



13.1.1 Introduction: Types of gene duplication

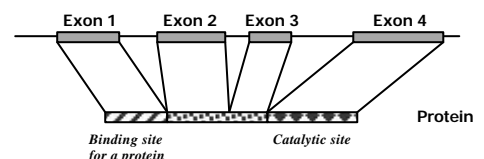
Gene duplication is one of the most important mechanisms to create new genes and new biological functions

A gene duplication can be:

- a partial or internal gene duplication
- a complete gene duplication

13.1.2 Internal duplications

In eukaryotic genomes, internal duplications of segments of genes have occurred frequently = exon duplication (or loss)



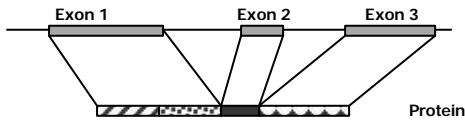
Internal duplications generally work by functional domain duplication

Example: exon 1
exon 2+3
exon 4

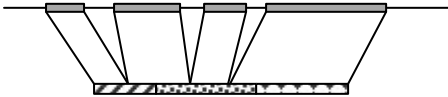
because exon 3 alone is likely to be useless

13.1.2 Internal duplications

But we can also find several domains encoded in one exon alone (ex: haemoglobin genes)

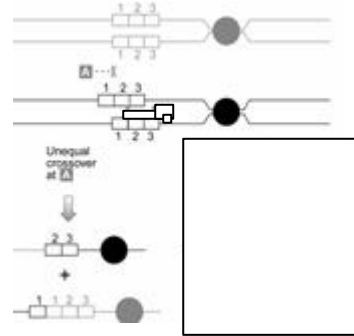


In general, there is a relationship between exon and domains
We very rarely (never) find:



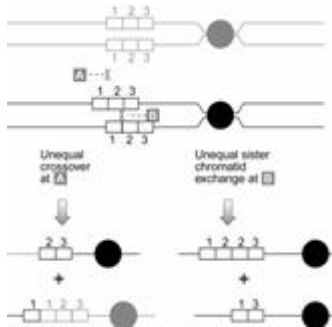
13.1.2 Internal duplications: Mechanism

Main mechanism: unequal crossing over



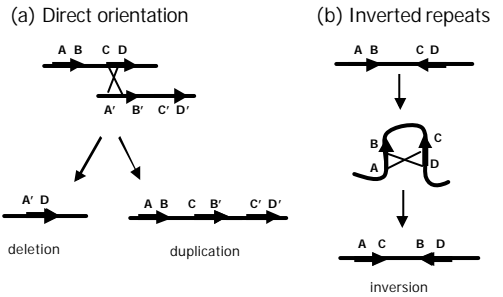
13.1.2 Internal duplications: Mechanism

Main mechanism: unequal crossing over



13.1.2 Internal duplications: Mechanism

Recombination between repeats can occur:



13.1.2 Internal duplications

Examples:

-the ovomucoid gene:

Protein present in the white of eggs, that inhibit the activity of trypsin

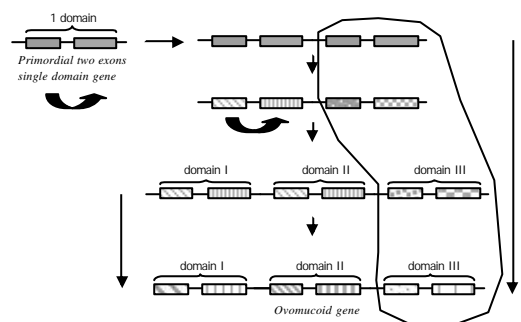
Ovomucoid polypeptid has three functional domains

The three domains share a common evolutionary origin. Each domain is composed by two exons separated by an intron. These two exons present no similarity between them.

Domain I and II are more similar to each other than either of them to domain III

So the evolutionary history of the gene seems to be the following:

13.1.2 Internal duplications



Time of divergence between domain I and II
Time of divergence between domain III and the others

13.1.2 Internal duplications

In the case of ovomucoid gene, each domain is capable of binding one molecule of trypsin, or another serine proteinase.
Here, the three domains have retained more or less the same function

It can happen that the duplicated domain sequences diverged a lot.
In this latter case, it's difficult to infer whether the gene has evolved by internal duplications.

In some cases, we can infer the common ancestry by comparing the secondary structure of the protein which is better preserved than the nt or aa sequence.

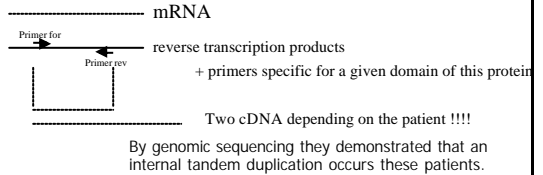
13.1.2 Internal duplications: examples

Mutations can occur in the duplicated domain, to create a new function...

... Or to damage the original gene function

Example: Myeloid leukemia

The *flt3* is composed of 24 exons and contains 4 domains
The expression of the gene was studied by Reverse Transcription-PCR
In some patients, they found a longer transcript that expected



13.1.2 Internal duplications: often deleterious

	Distance (kb)	Repeat length (bp)
Deletion		
Colour blindness	0	39,000
α-Thalassemia	3.7 or 4.2	4,000
Growth hormone deficiency	6.7	2,200
Debrisoquine sensitivity	9.3	2,800
X-linked ichthyosis	1900	20,000
Williams syndrome	~2,000	>30,000
Incontinentia Pigmenti	~4	878
Duplication		
Glucocorticoid remediable aldosteronism	45	10,000
Dup/Del		
CMT1A / HNPP, dup(17)(p11.2)	1,500	24,011
Smith-Magenis syndrome	~5,000	>200,000
Inversion		
Hunter mucopolysaccharidosis	20	3,000

13.1.1 Introduction: Types of gene duplication

Gene duplication is one of the most important mechanisms to create new genes and new biological functions

A gene duplication can be:

- a partial or internal gene duplication
- a complete gene duplication

13.1.3 Complete gene duplication

Consequence of complete gene duplication:

- Retain the original function -> larger quantity of mRNA produced (ex: tRNA, histones, rRNA)
- Functionless pseudogene
- New gene
Depending on the time of evolution, the two genes can have completely different functions (ex: fibrinogen and trypsin/ lysozyme and lactalbumin)

13.1.3 Complete gene duplication

Although the relation between similarity, function and time since the duplication does not follow a general rule

Gene pair (organism)	Amino acid similarity (%)	Time of duplication (million years)
Trypsin and chymotrypsin (human)	36	1,500
Hemoglobin and myoglobin (human)	23	800
Lactate dehydrogenase M and H chains (human)	74	600
Hemoglobin α and β chains (human)	41	500
Immunoglobulin H and L chains (human)	25	400
Lactalbumin and lysozyme (human)	37	350
Growth hormone and prolactin (human)	25	330
Chymotrypsins A and B (human)	79	270
Carbonic anhydrases B and C (human)	60	180
Insulins I and II (rat)	96	30
Growth hormone and lactogen (human)	85	23
Alcohol dehydrogenase A and S chains (horse)	98	10

Modified from Li (1983).

13.1.3 Complete gene duplication

Consequence of complete gene duplication:

- Retain the original function -> larger quantity of mRNA produced (ex: tRNA, histones, rRNA)
- Functionless pseudogene
- New gene
 - Depending on the time of evolution, the two genes can have completely different functions (ex: fibrinogen and trypsin/ lysozyme and lactalbumin)
 - or constitute a gene family (ex: β -globin gene family)

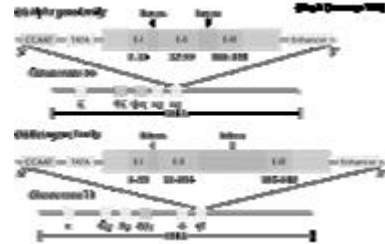
Are generally close to each other in the genome

13.1.3 Complete gene duplication

Superfamily : if proteins exhibit < 50% aa homology

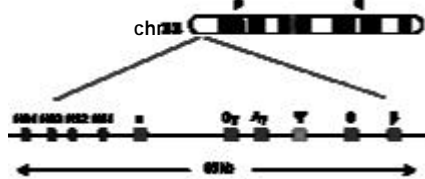
α -globin + β -globin gene + myoglobin = globin superfamily

Structure of the gene suggests the duplication origin of the two gene families



13.1.3 Complete gene duplication

Gene family: if proteins exhibit = or > 50% aa homology



β -globin gene family includes β , δ , $\Delta\gamma$, $\Gamma\gamma$ and ϵ . Ψ is a pseudogene. HS1 to HS4 are regulatory elements

13.1.3 Complete gene duplication

Summary:

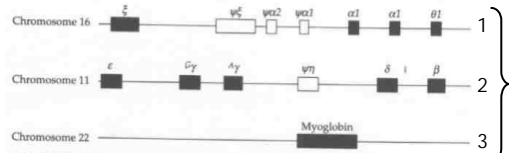


FIGURE 6.11 The chromosomal arrangement of the three gene families belonging to the globin superfamily of genes in humans: the α -globin family on chromosome 16, the β -globin family on chromosome 11, and myoglobin on chromosome 22. Solid black boxes denote functional genes; empty boxes denote pseudogenes.

13.1.3 Complete gene duplication

The globin super family has experienced all the possible evolutionary pathways:

- Retention of original function
- Acquisition of a new function
- Loss of functions in some duplicates

13.1.3 Complete gene duplication

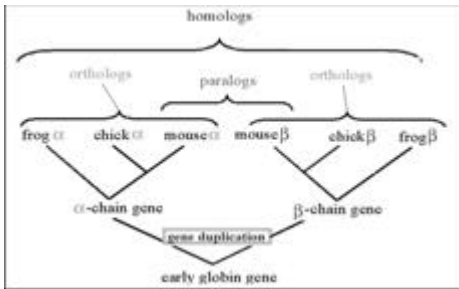
Formation of functional copies occurs but..
A duplicated gene is more likely to become an unprocessed pseudogene

They are copies that present one or several of the following characteristics:

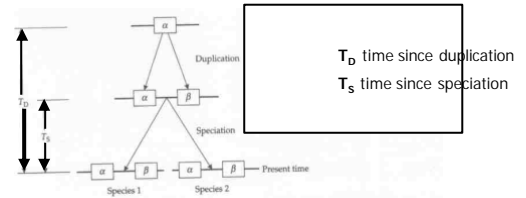
- Loss of TATA box
- Loss of the initiation codon
- Frameshift
- Premature/alterd stop codon
- etc...

13.1.3 Terminology

Orthology, paralogy and homology



13.1.4 Dating time of divergence



We can estimate T_D if we know the rate of rate of substitution K in genes α and β , that is the number of substitutions between the orthologous α and β , in conjunction with the time of divergence T_S

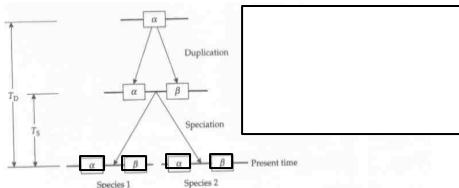
$$r_\alpha = \frac{K_\alpha}{2T_S} \quad r_\beta = \frac{K_\beta}{2T_S} \quad r = (r_\alpha + r_\beta) / 2 \quad T_D = \frac{\overline{K_{\alpha\beta}}}{2r}$$

13.1.4 Dating time of divergence

To estimate T_D , we need to know $K_{\alpha\beta}$, the number of substitution per site between genes α and β

To obtain this number, we can compute four comparisons:

- Gene a from species 1 and gene b species 2
- Gene b from species 1 and gene a species 2
- Genes a and b from species 1
- Genes a and b from species 2



13.1.4 Dating time of divergence

It is possible to estimate the nonfunctionalization time of an unprocessed pseudogene.

It corresponds to the time since mutations accumulate neutrally

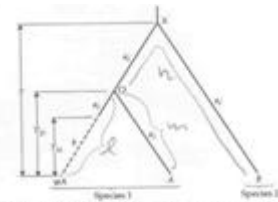


FIGURE 6.30 Schematic phylogenetic tree used to estimate the nonfunctionalization time of an unprocessed pseudogene. T denotes the divergence time between species 1 and 2, T_D denotes the time since duplication of gene A, and T_S denotes the time since nonfunctionalization of pseudogene α . K is the rate of nucleotide substitution per site per year at the 0th codon position in the functional genes, and K_p is the rate of substitution for the pseudogenes. The node connecting the orthologous genes is denoted by X, and the node connecting the paralogous genes is marked by O. Modified from Li et al. (1992).

13.1.4 Dating time of divergence

Can be difficult in case of concerted evolution

The members of a repeated-sequence family are very similar to each other within a species, and very divergent from the members of the same family even in a very related species

That is: within a species, a mechanism conserves the copies similar faster than mutations occur.

13.1.5 Conclusion: new genes formation

Partial or complete gene duplication is one of the main way to create new genes and new biological functions

But they are other important mechanisms, like:

Exon shuffling (insertion of an exon from ANOTHER protein)

Overlapping genes (several ORF in the same gene)

Alternative splicing

Gene sharing

...