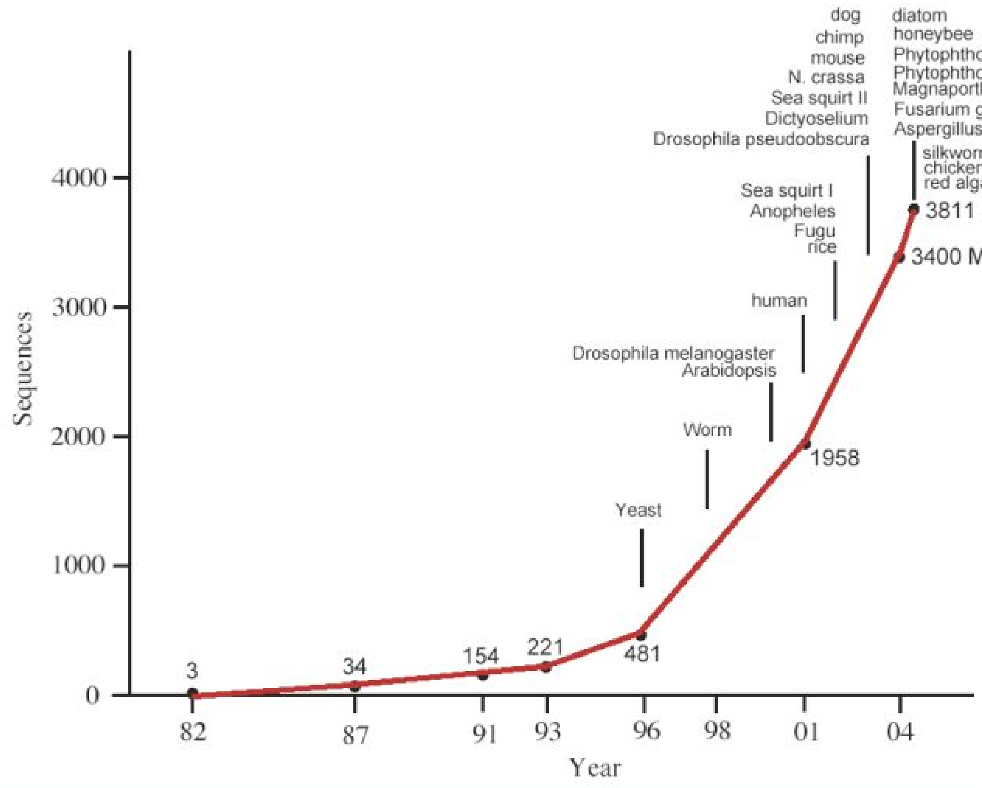
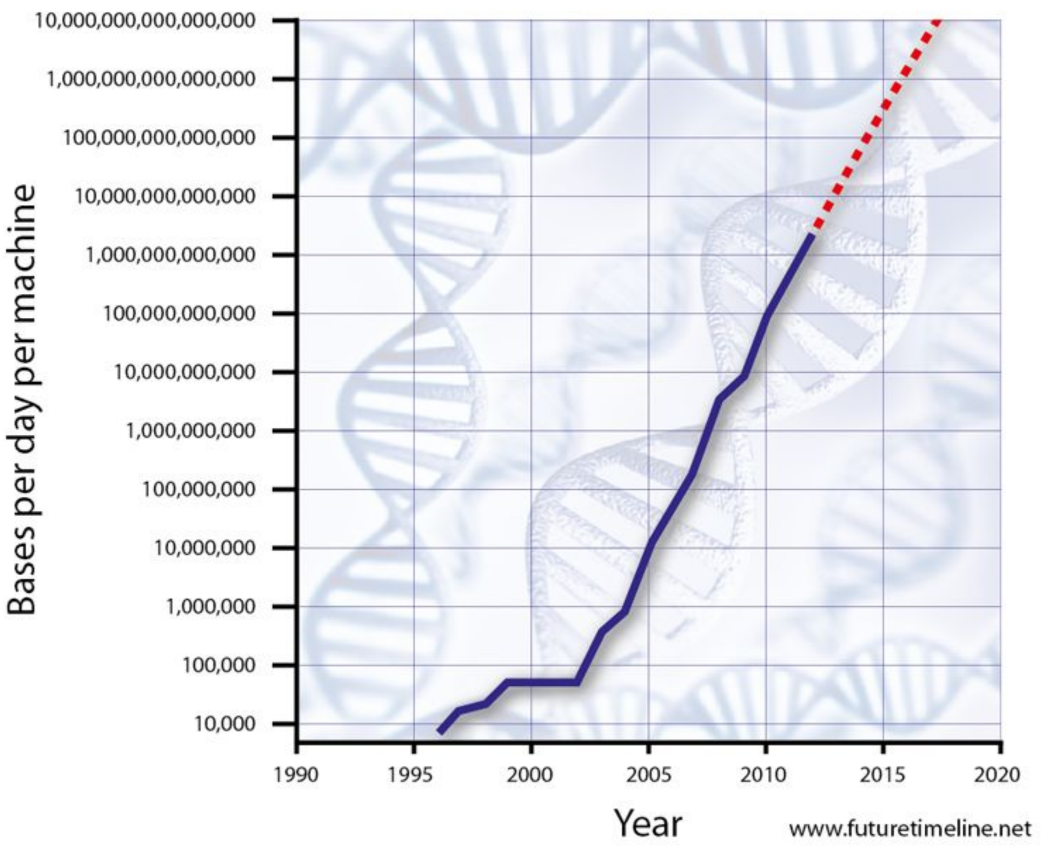
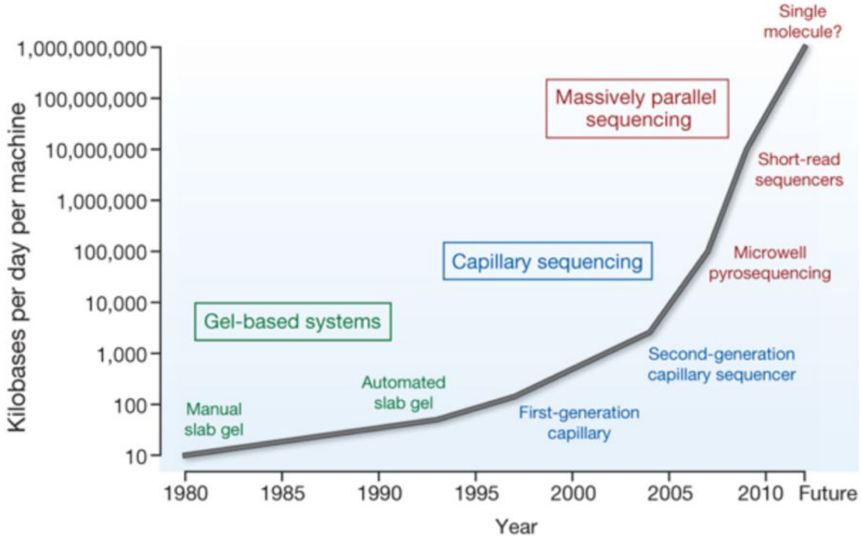
Genome Organization

- How many genes?
- Which genes?
- How many essential genes?

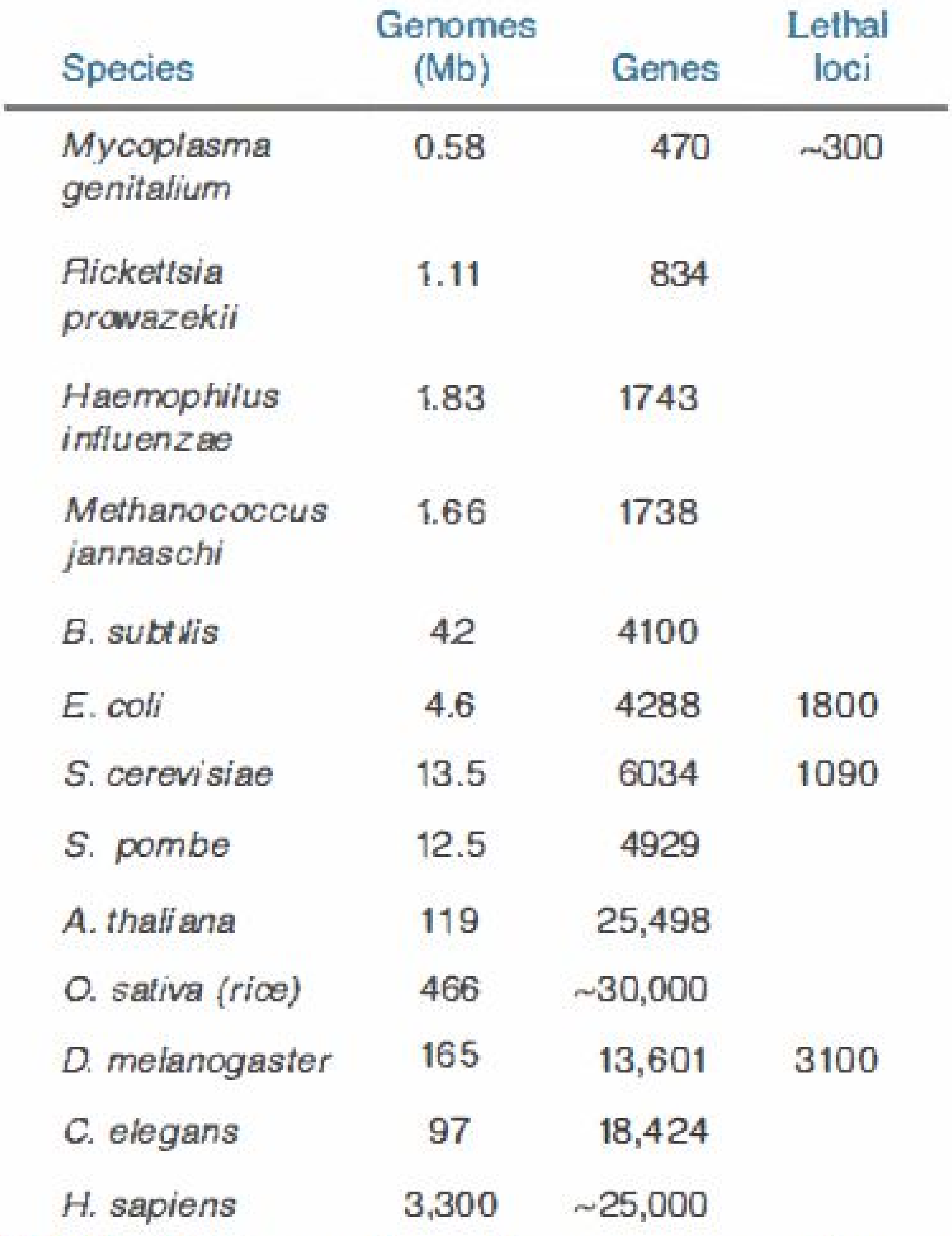




Improvements in the Rate of DNA Sequencing



Reprinted by permission from Macmillan Publishers Ltd: Nature 458, 719-724 (2009)



Species	Genomes (Mb)	Genes	Lethal loci
Mycoplasma genitalium	0.58	470	~300
Rickettsia prowazekii	1.11	834	
Haemophilus influenzae	1.83	1743	
Methanococcus jannaschi	1.66	1738	
B. subtilis	42	4100	
E. coli	4.6	4288	1800
S. cerevisiae	13.5	6034	1090
S. pombe	12.5	4929	
A. thaliana	119	25,498	
O. sativa (rice)	466	~30,000	
D. melanogaster	165	13,601	3100
C. elegans	97	18,424	
H. sapiens	3,300	~25,000	

FIGURE 6.2 Genome sizes and gene numbers are known from complete sequences for several organisms. Lethal loci are estimated from genetic data.

FIGURE 6.1 The minimum gene number required for any type of organism increases with its complexity.

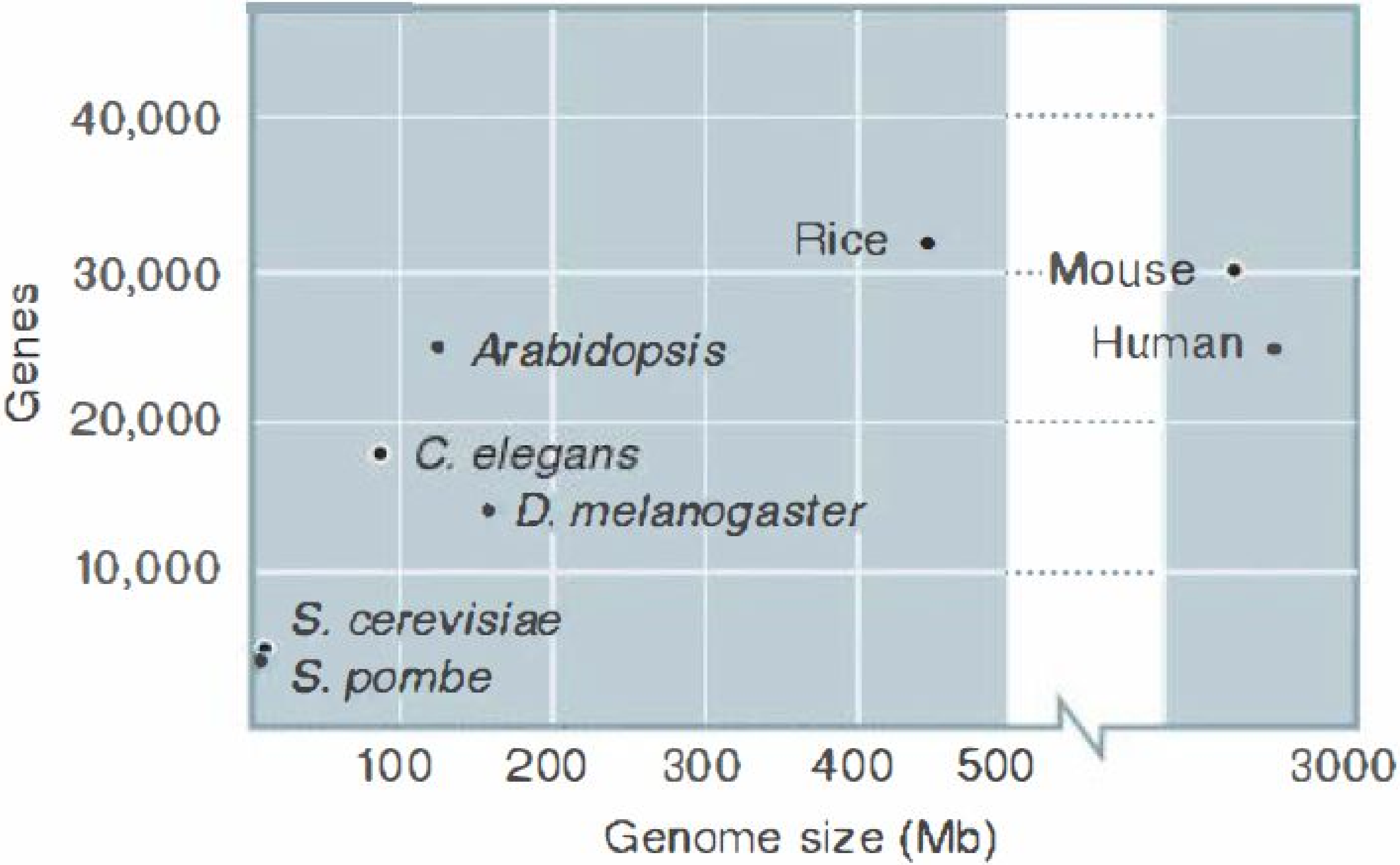


FIGURE 6.4 The number of genes in a eukaryote varies from 6000 to 32,000 but does not correlate with the genome size or the complexity of the organism.

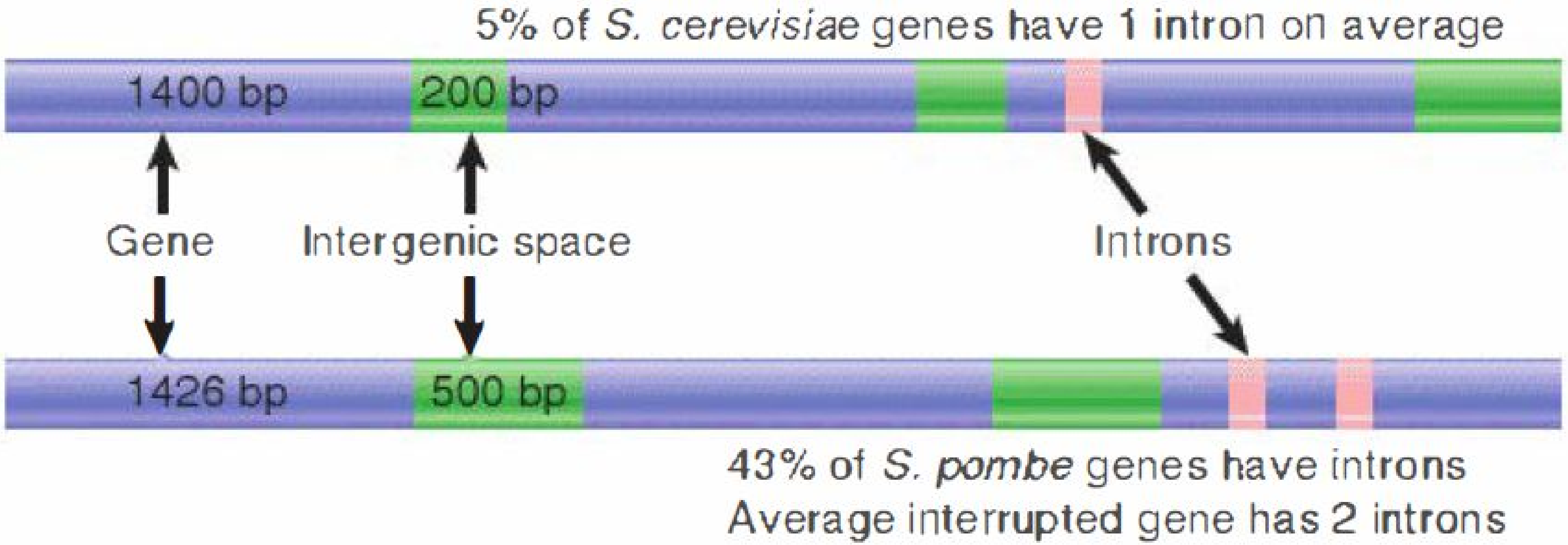


FIGURE 6.5 The *S. cerevisiae* genome of 13.5 Mb has 6000 genes, almost all uninterrupted. The *S. pombe* genome of 12.5 Mb has 5000 genes, almost half having introns. Gene sizes and spacing are fairly similar.

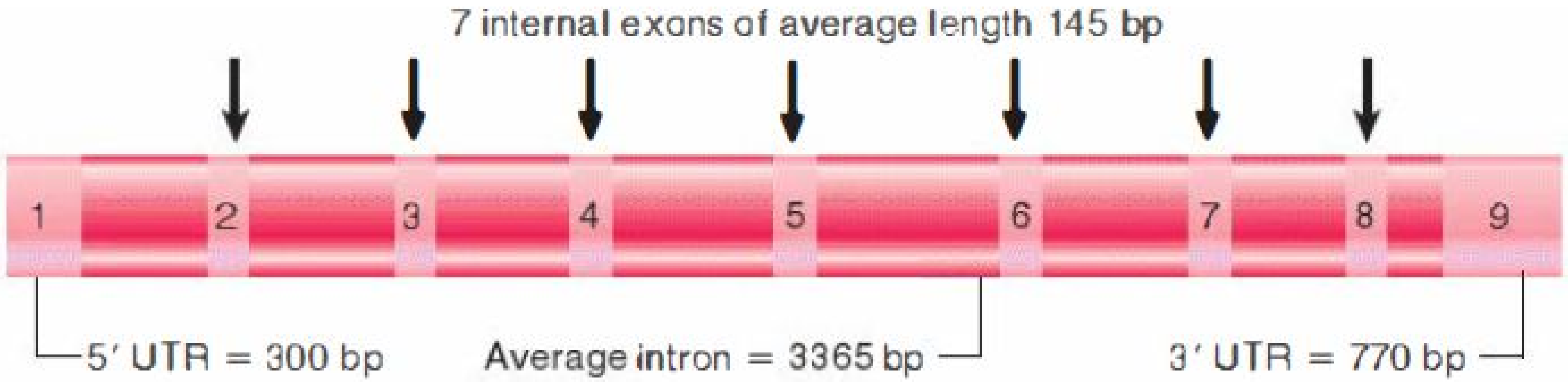
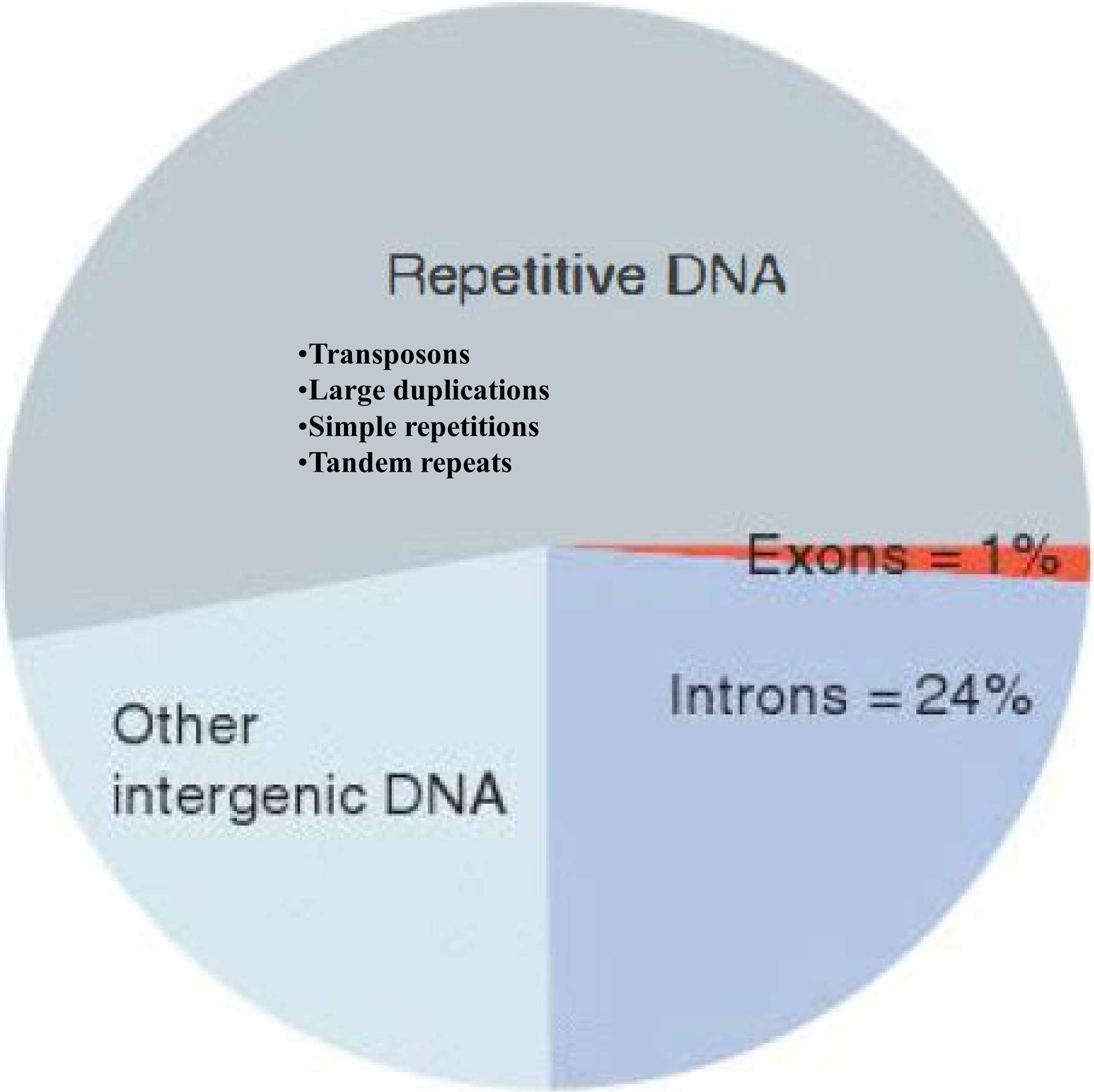


FIGURE 6.12 The average human gene is 27 kb long and has nine exons, usually comprising two longer exons at each end and seven internal exons. The UTRs in the terminal exons are the untranslated (noncoding) regions at each end of the gene. (This is based on the average. Some genes are extremely long, which makes the median length 14 kb with seven exons.)



- Transposons
- Large duplications
- Simple repetitions
- Tandem repeats

FIGURE 6.11 Genes occupy 25% of the human genome, but protein-coding sequences are only a small part of this fraction.

FIGURE 6.6 Functions of *Drosophila* genes based on comparative genomics of 12 species. The functions of about a quarter of the genes of *Drosophila* are unknown. Adapted from *Drosophila* 12 Genomes Consortium, "Evolution of genes and genomes on the *Drosophila* phylogeny," *Nature* **450** (2007): 203–218.

How many genes are expressed?

- About 10 000 genes are expressed in all cells
- Some genes are expressed only in some cells

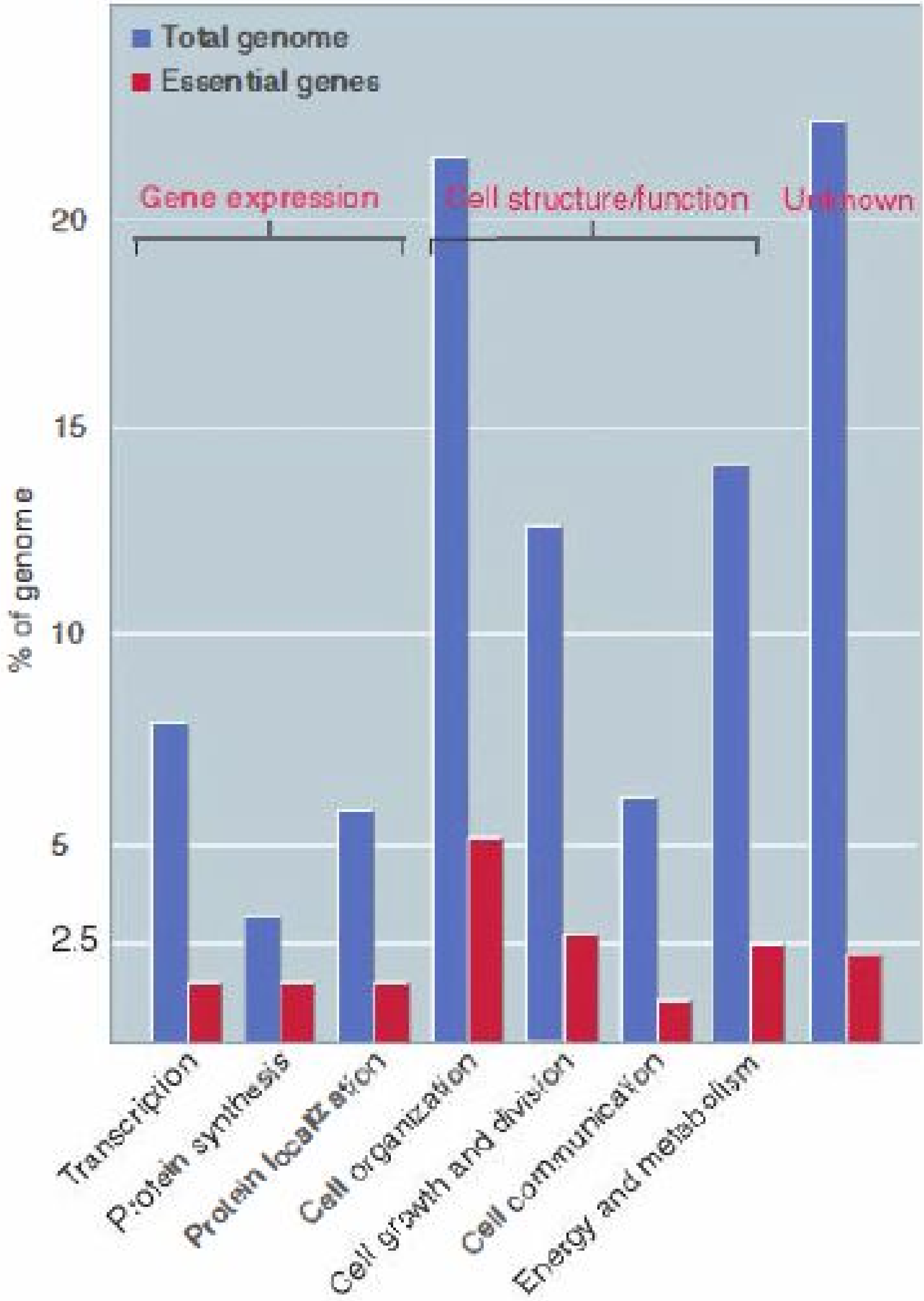
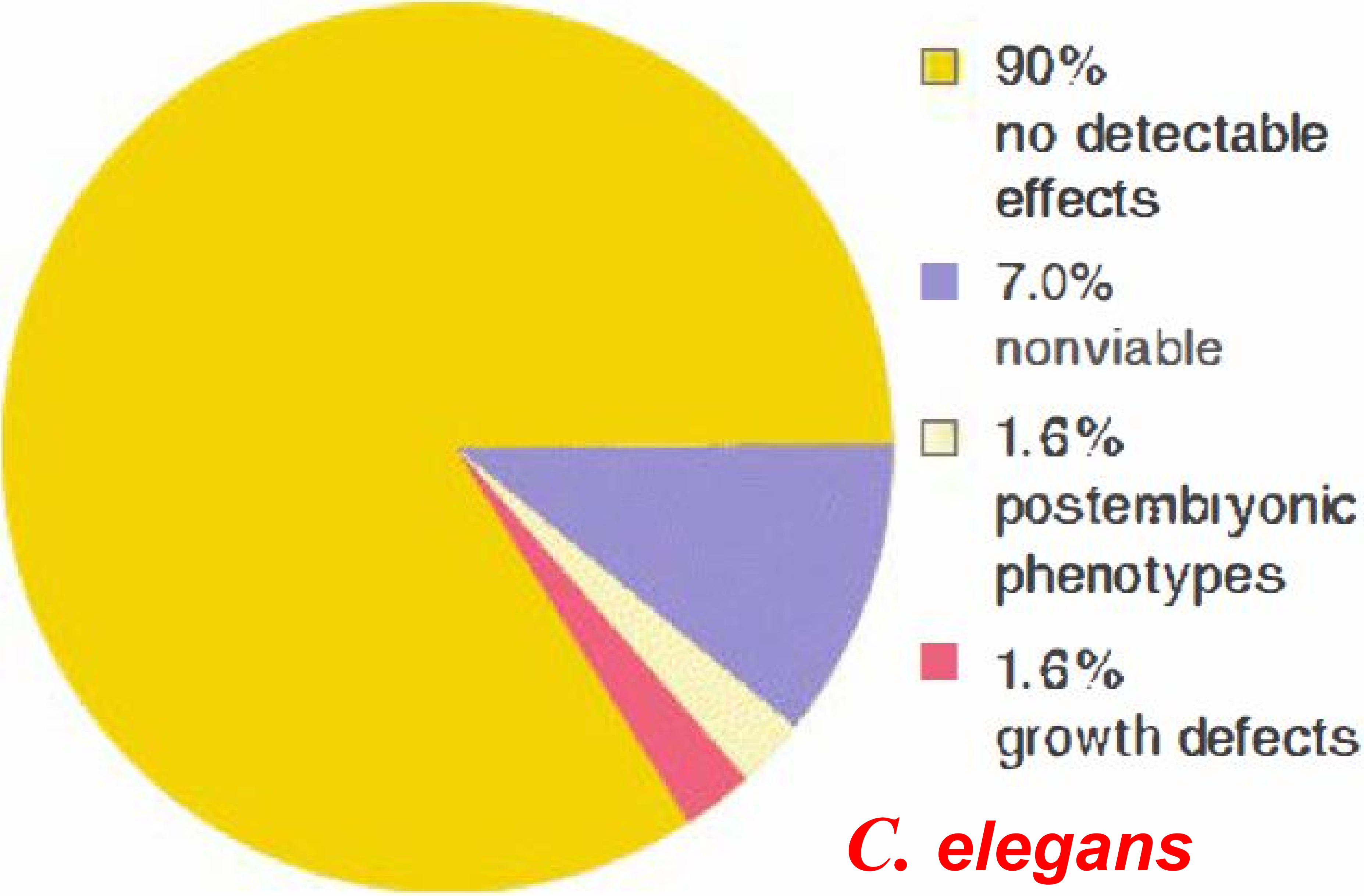


FIGURE 6.16 Essential yeast genes are found in all classes. Blue bars show the total proportion of each class of genes, and red bars show those that are essential.



C. elegans

FIGURE 6.17 A systematic analysis of loss of function for 86% of worm genes shows that only 10% have detectable effects on the phenotype.

Most of yeast genes are <u>not essential genes</u>

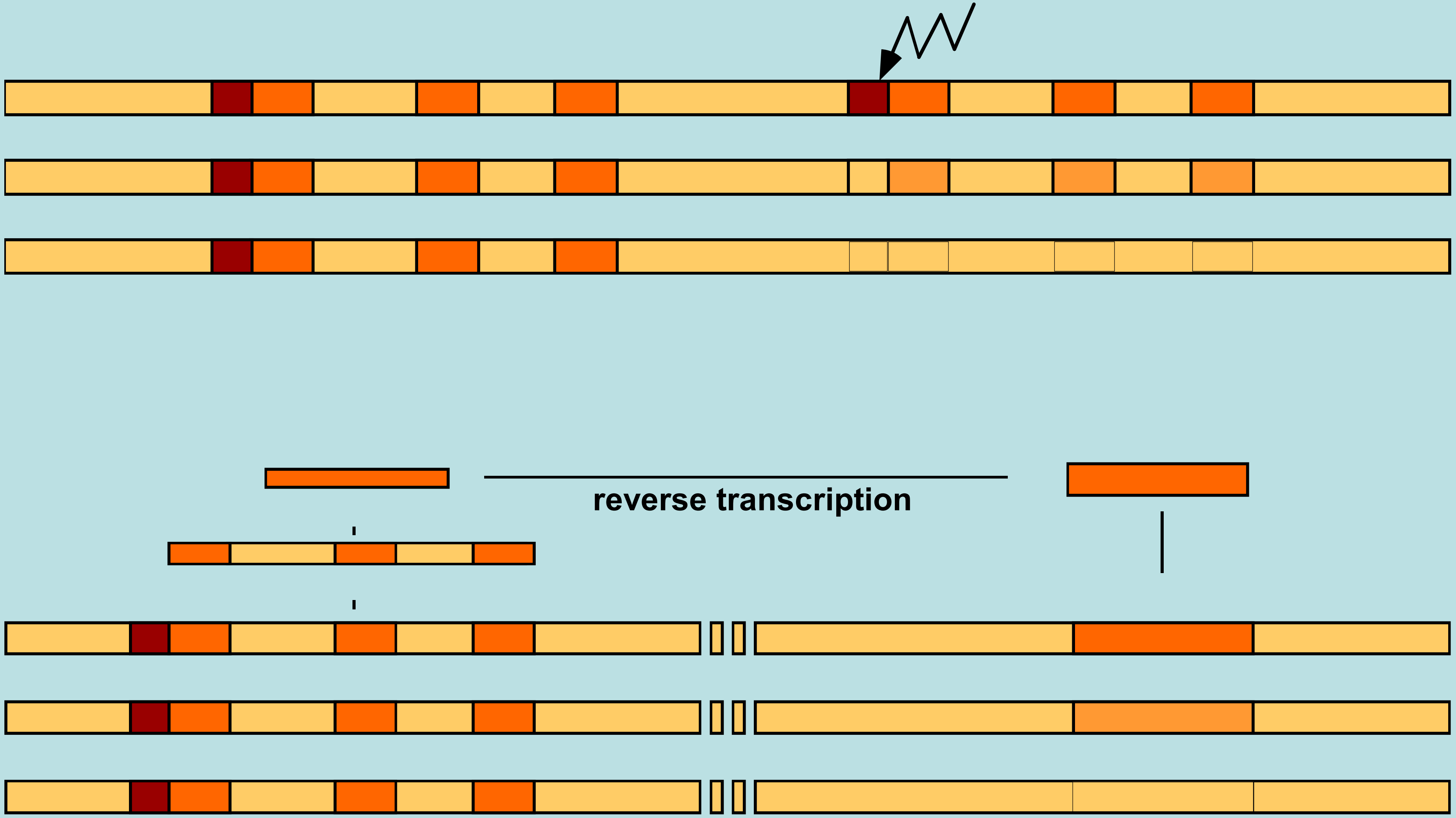
- Redundant genes (gene families)
- Genes with ancillary functions

PSEUDOGENES

Nonfunctional gene copies

- Non processed (with introns)
- Processed (without introns)

Origin of pseudogenes (inactive copies of a coding gene)



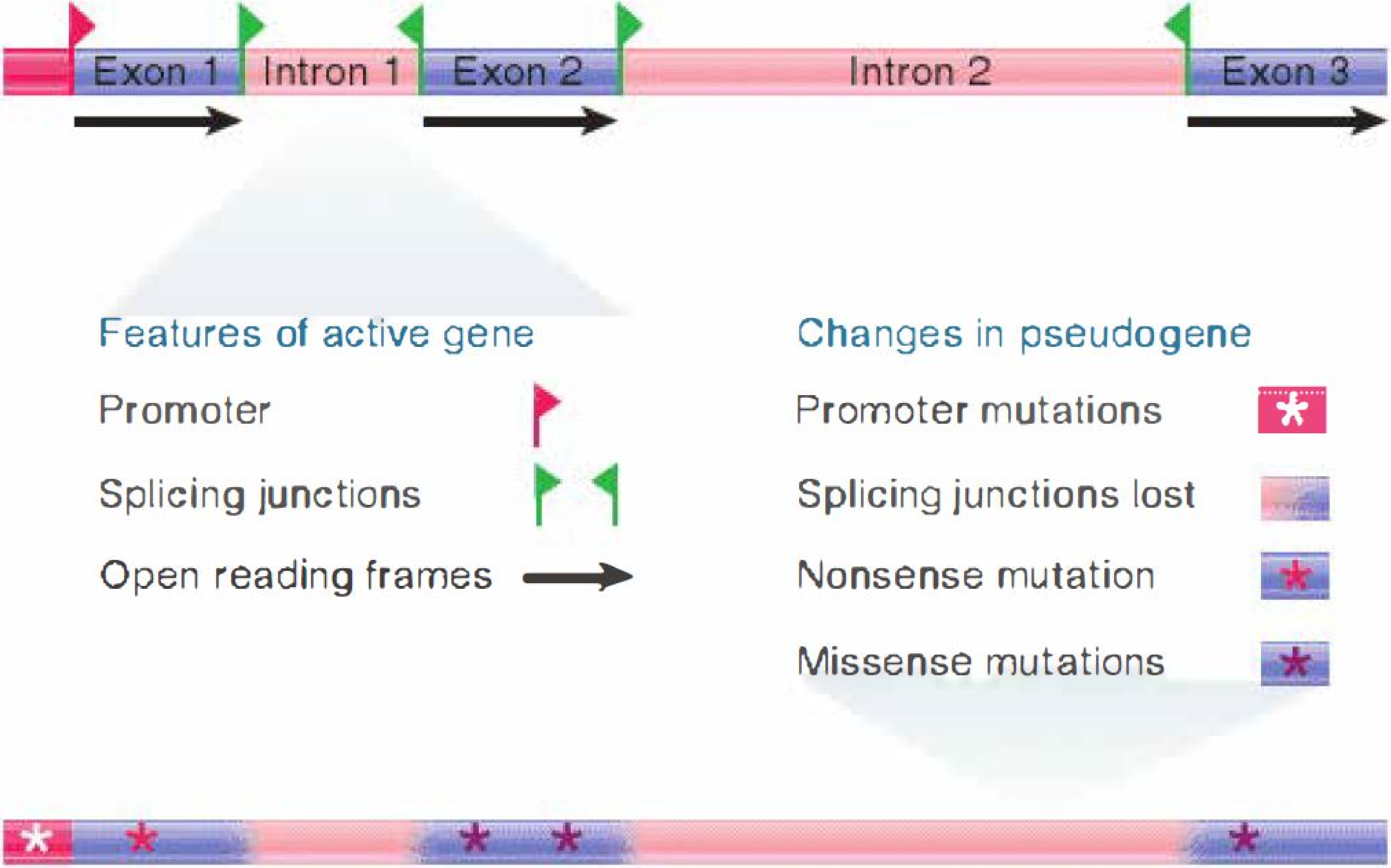


FIGURE 8.23 Many changes have occurred in a β-globin gene since it became a pseudogene.

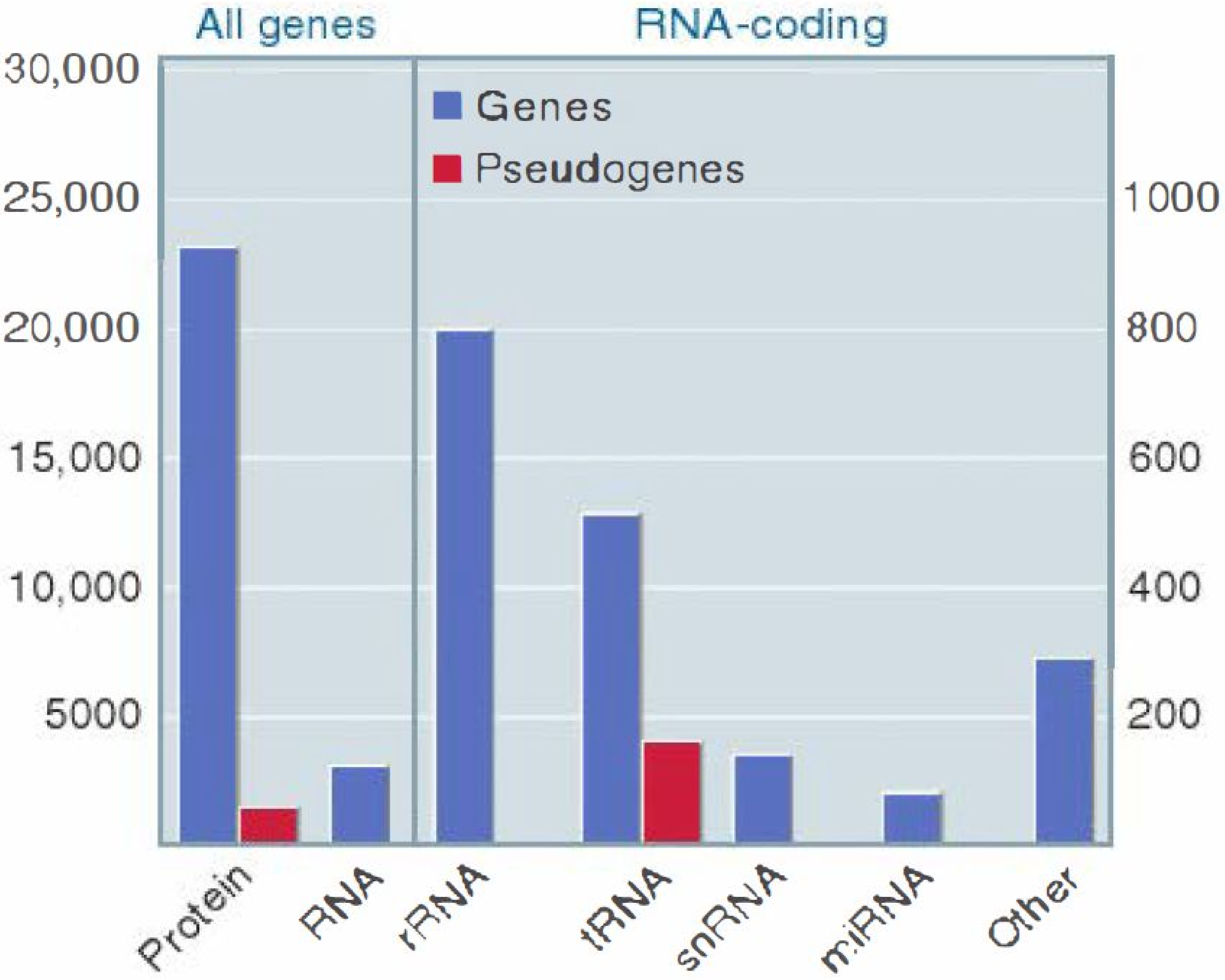


FIGURE 6.10 The mouse genome has ~23,000 protein-coding genes, which have ~1200 pseudogenes. There are ~3000 RNA-coding genes.

To summarize……

Pseudogenes = non functional genes

- With introns
- Processed

Gene Families

Gene family → a group of genes that encode related or identical products as a result of gene-duplication events from a common ancestor.

Superfamily → genes that are more distantly related but that still can be recognized as having common ancestry and are coding for proteins that exerts very different roles.

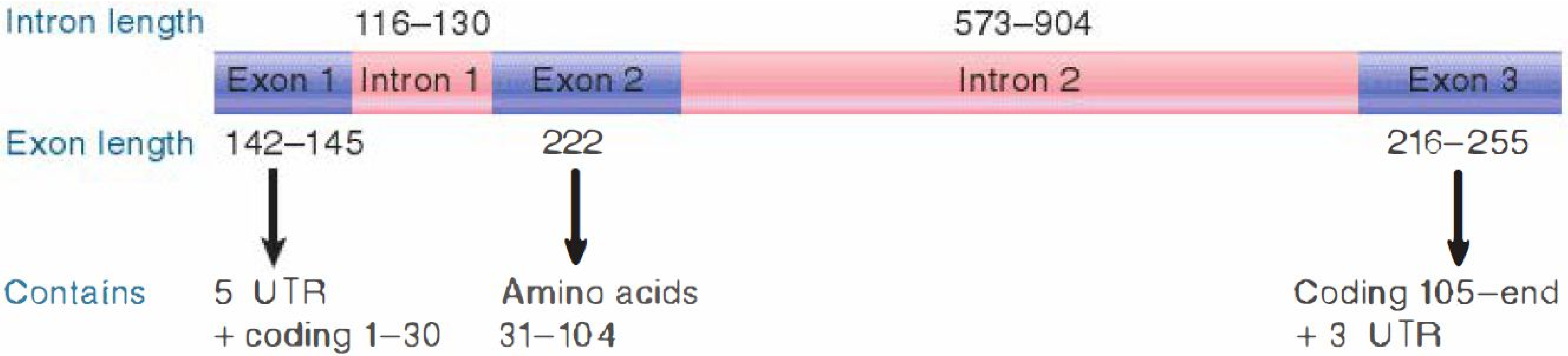


FIGURE 4.4 All functional globin genes have an interrupted structure with three exons. The lengths indicated in the figure apply to the mammalian β-globin genes.

	Unique genes	Families with 2–4 members	Families with >4 members
H. influenzae	89%	10%	1%
S. cerevisiae	72%	19%	9%
D. melanogaster	72%	14%	14%
C. elegans	55%	20%	26%
A. thaliana	35%	24%	41%

FIGURE 6.8 The proportion of genes that are present in multiple copies increases with genome size in multicellular eukaryotes.

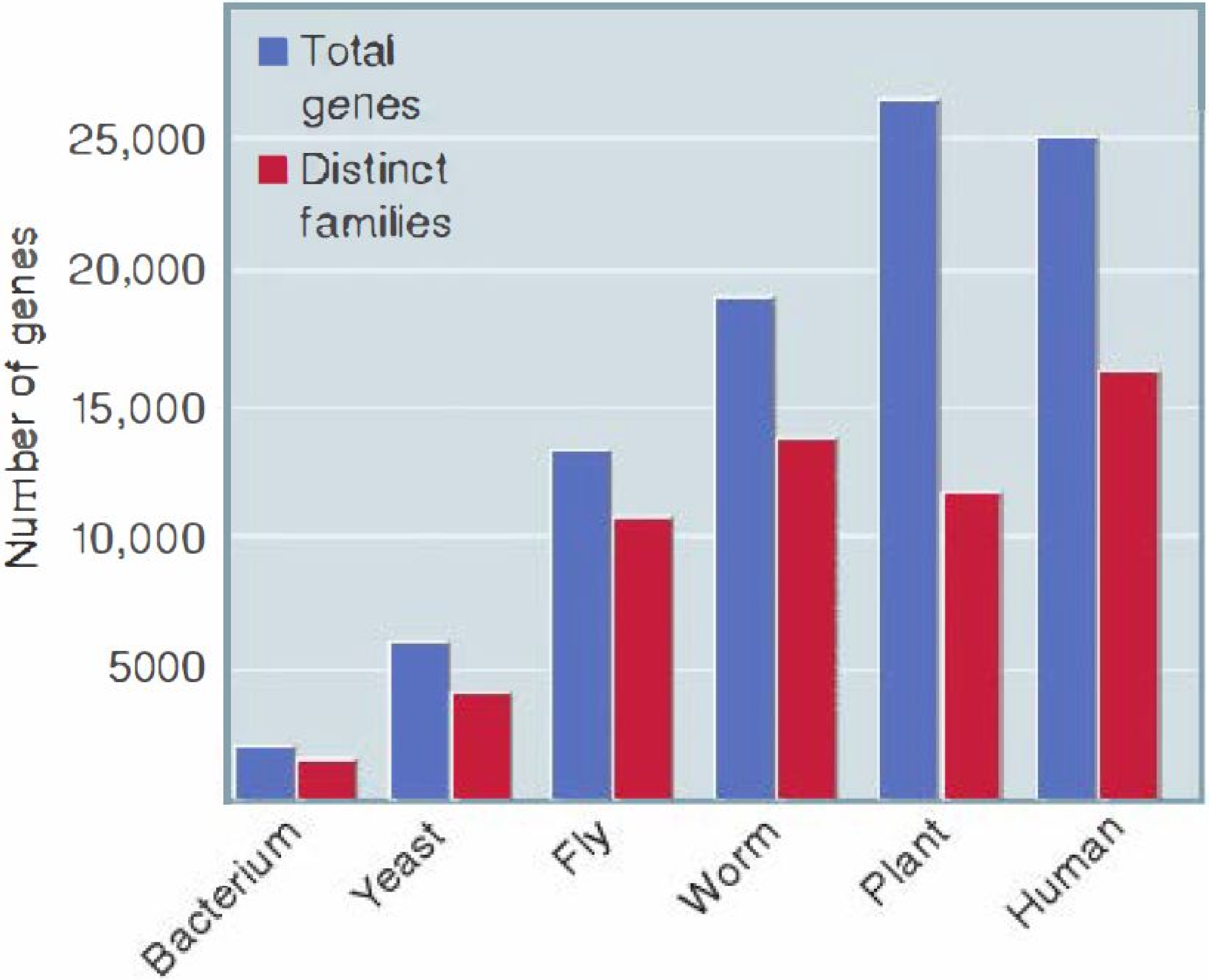
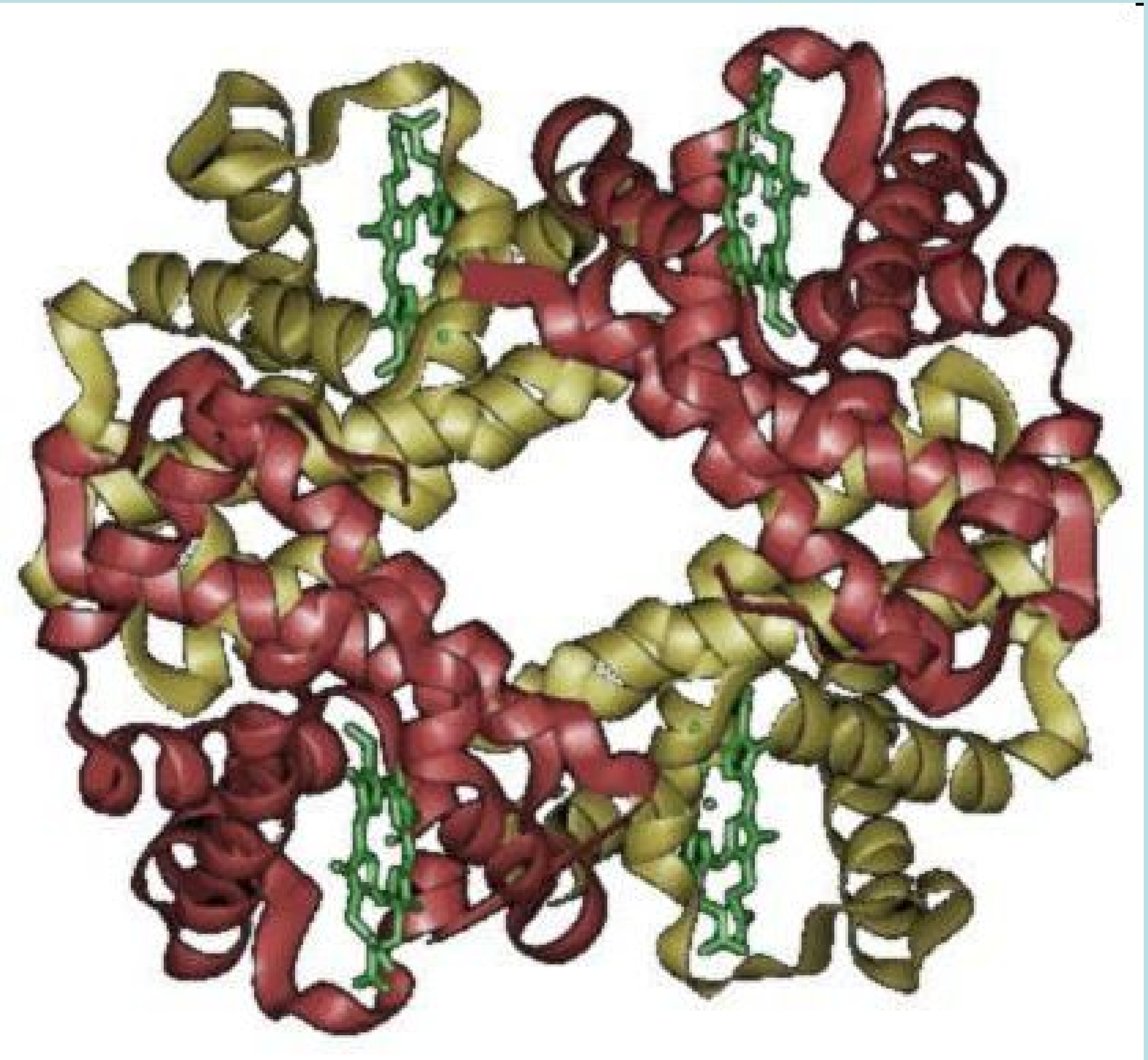
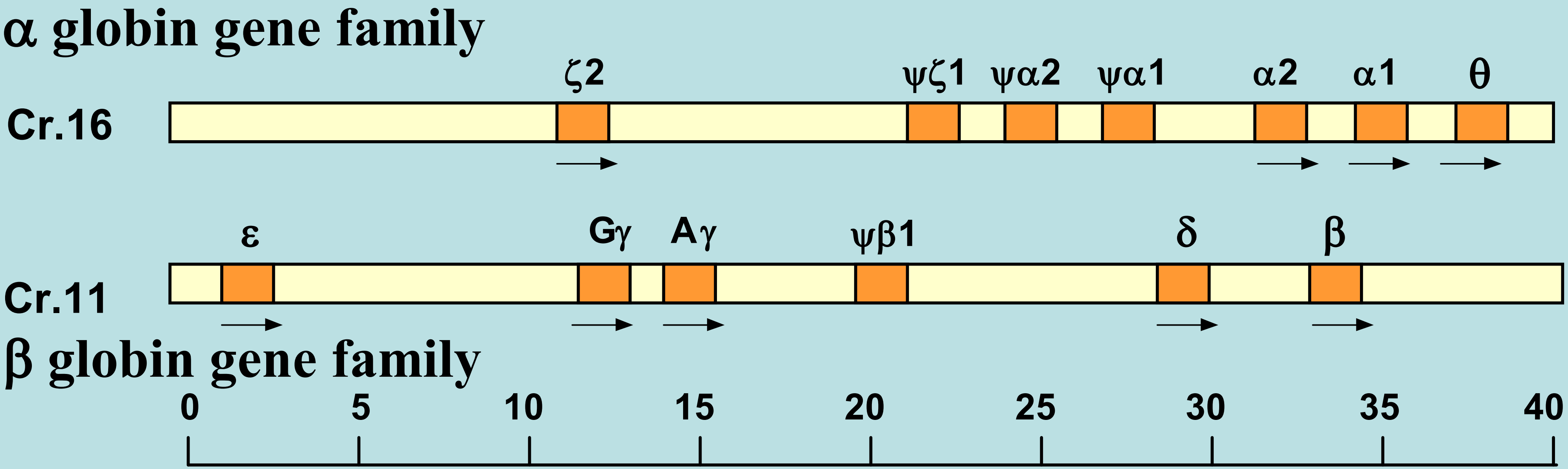


FIGURE 6.7 Many genes are duplicated, and as a result the number of different gene families is much smaller than the total number of genes. This histogram compares the total number of genes with the number of distinct gene families.

Human hemoglobins



Globin gene family in humans and other primates



Human hemoglobins

SUBUNITS	EMBRYO	FETUS	ADULT
α type	ζ	α	α
β type	ε	Gγ, Aγ	β, γ
Tetramers	$\zeta_2\varepsilon_2$ (Gower 1) $\alpha_2\varepsilon_2$ (Gower 2) $\zeta_2\gamma_2$ (Portland)	$\alpha_2\gamma_2$ (HbF)	$\alpha_2\beta_2$ (HbA) $\alpha_2\gamma_2$ (HbA2)

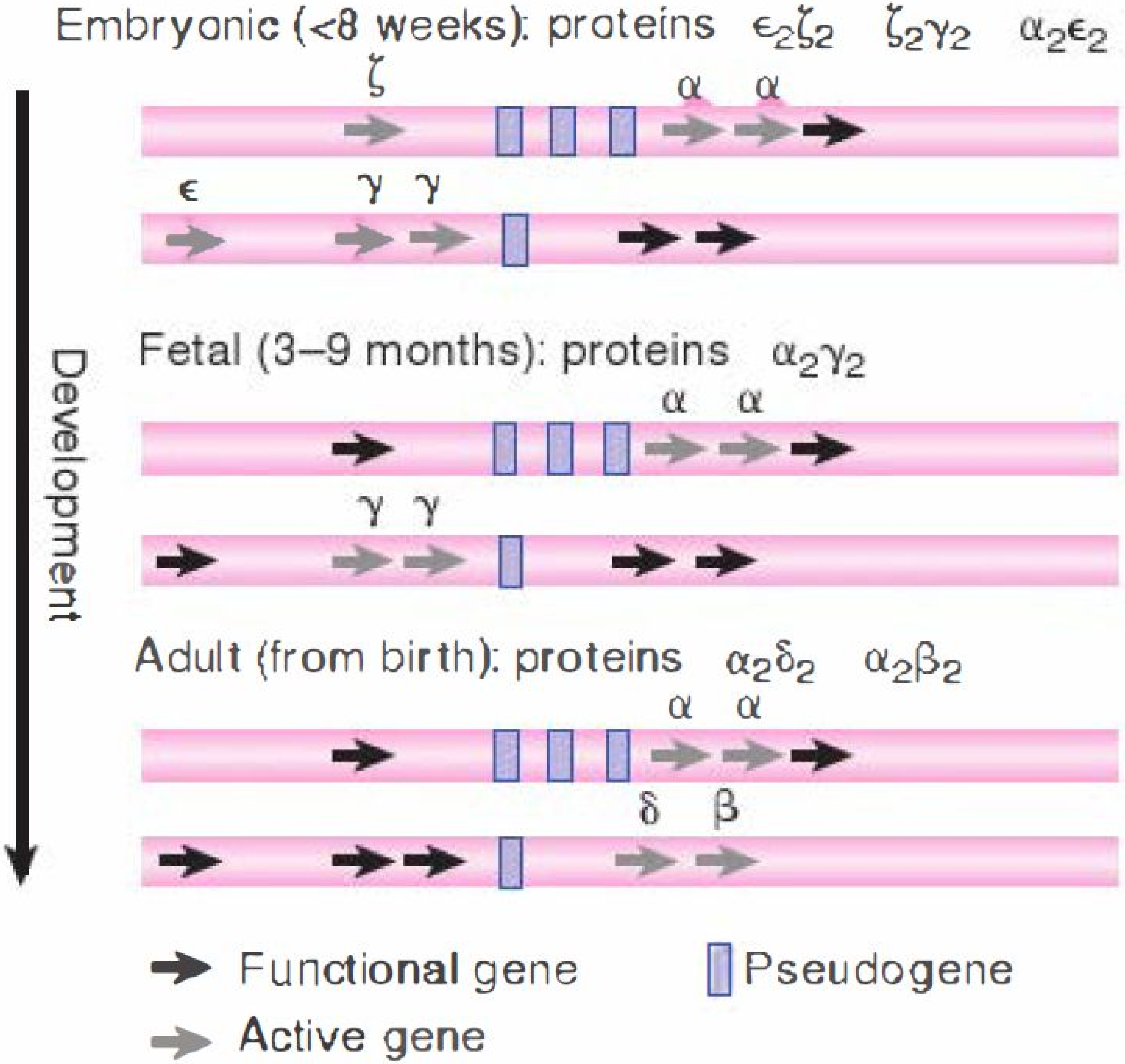


FIGURE 8.21 Different hemoglobin genes are expressed during embryonic, fetal, and adult periods of human development.

Gene duplication

- Gene duplication contributes to genome evolution

- Immediately following a gene duplication, changes might accumulate more rapidly in one of the copies, eventually leading to a new function.

- Duplicated genes may diverge to generate different genes, or one copy may become an inactive pseudogene.

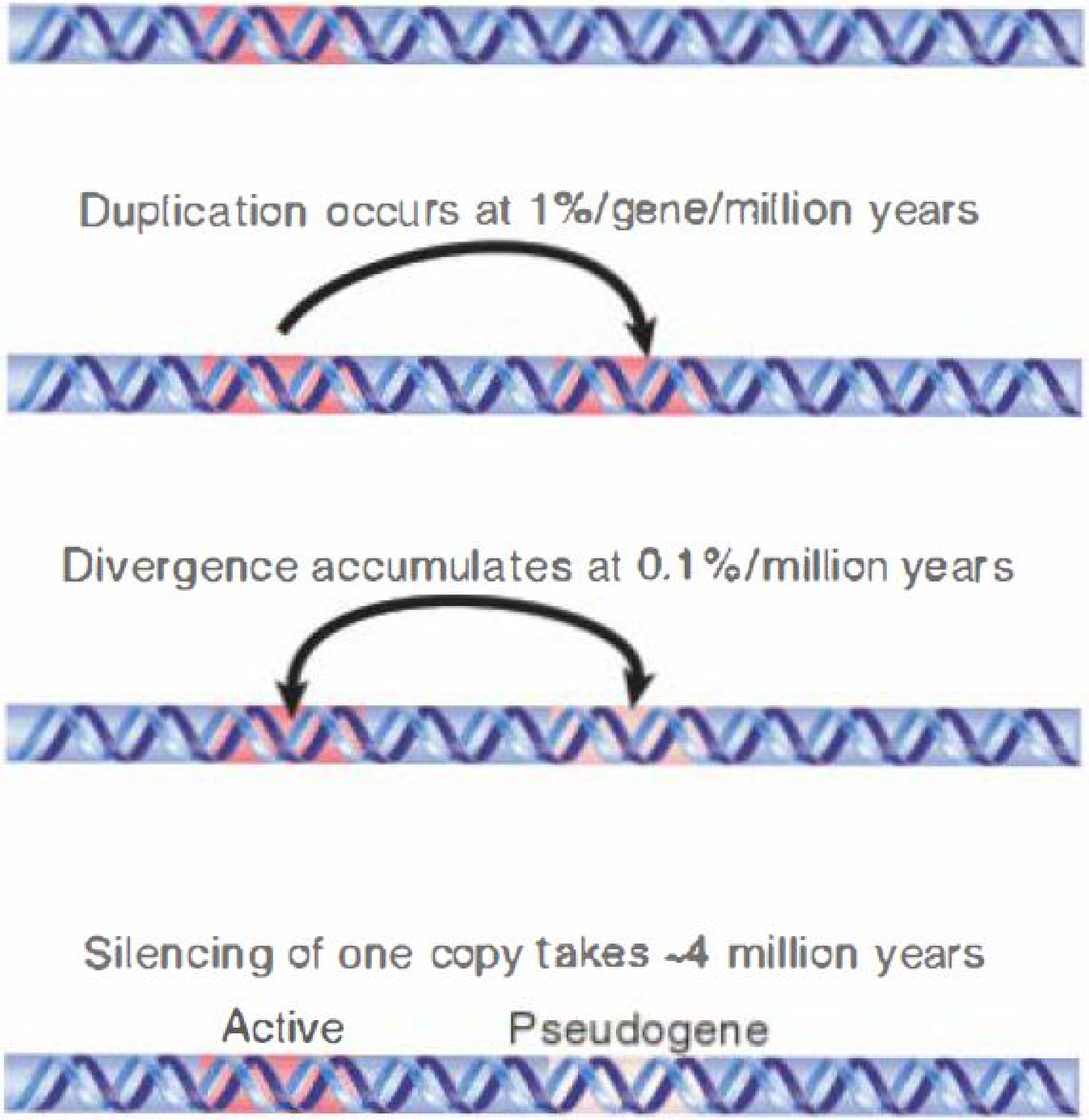
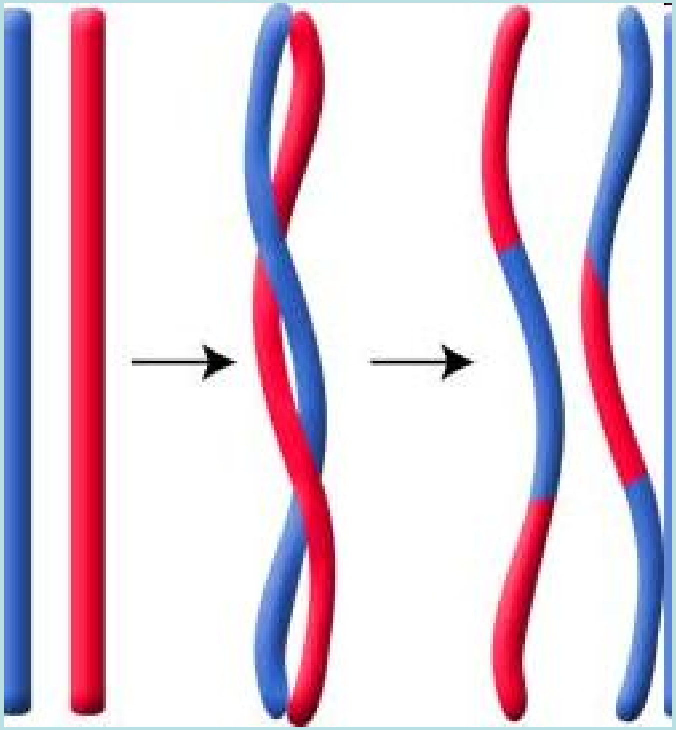
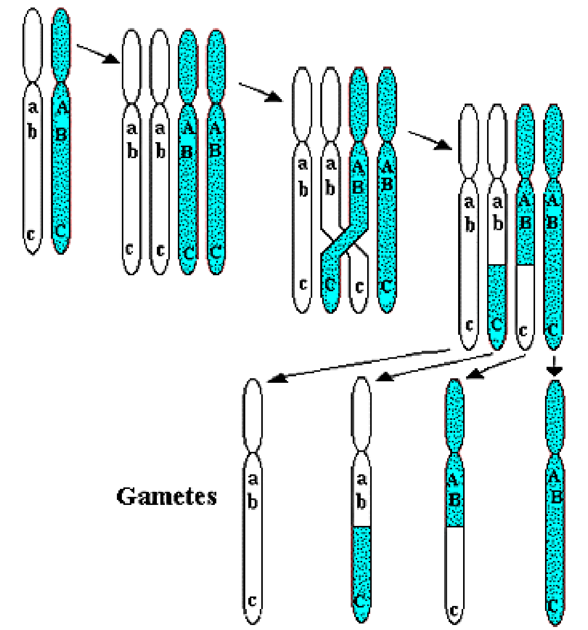


FIGURE 8.19 After a globin gene has been duplicated, differences may accumulate between the copies. The genes may acquire different functions or one of the copies may become nonfunctional.

Crossing Over

Exchange of corresponding segments between the homologs recombinant chromosomes that are different from the parental chromosomes can be generated. Recombination results from a physical exchange of chromosomal material.

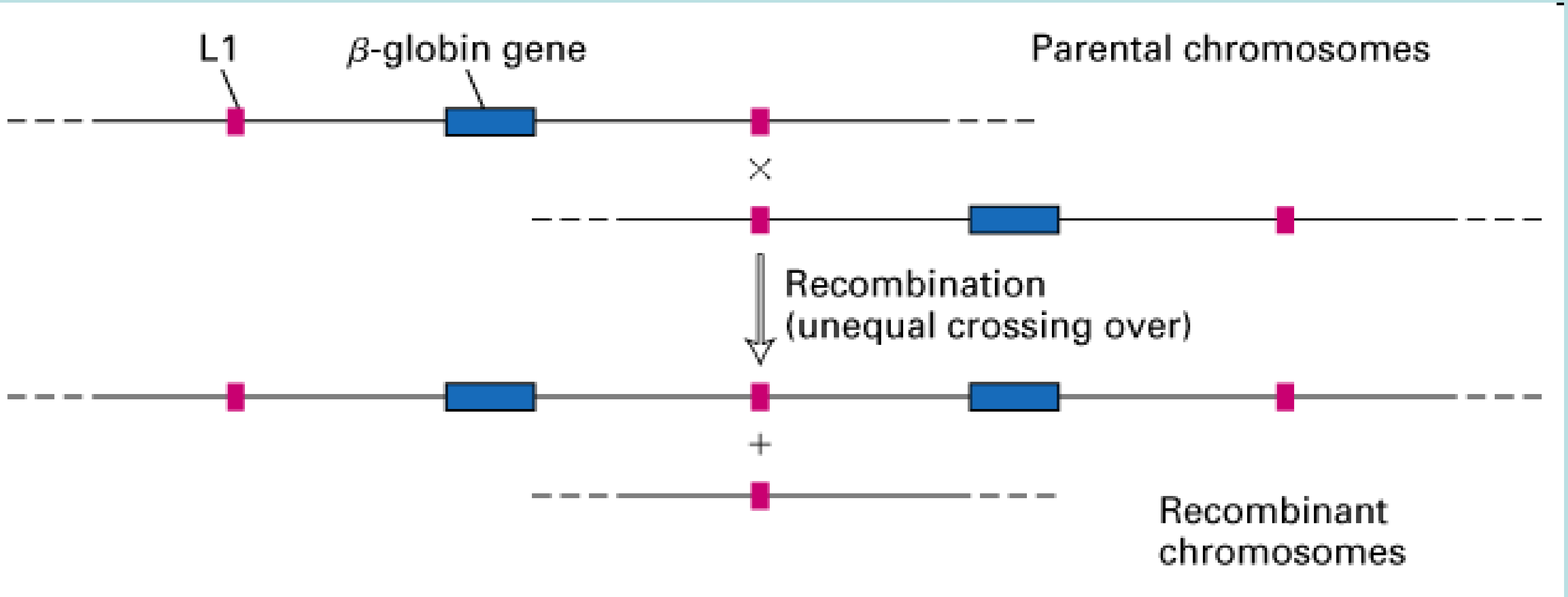




Crossing-over and recombination during meiosis

Crossing Over Ineguale

Results from pairing between nonequivalent repeats in regions of DNA consisting of repeating units.



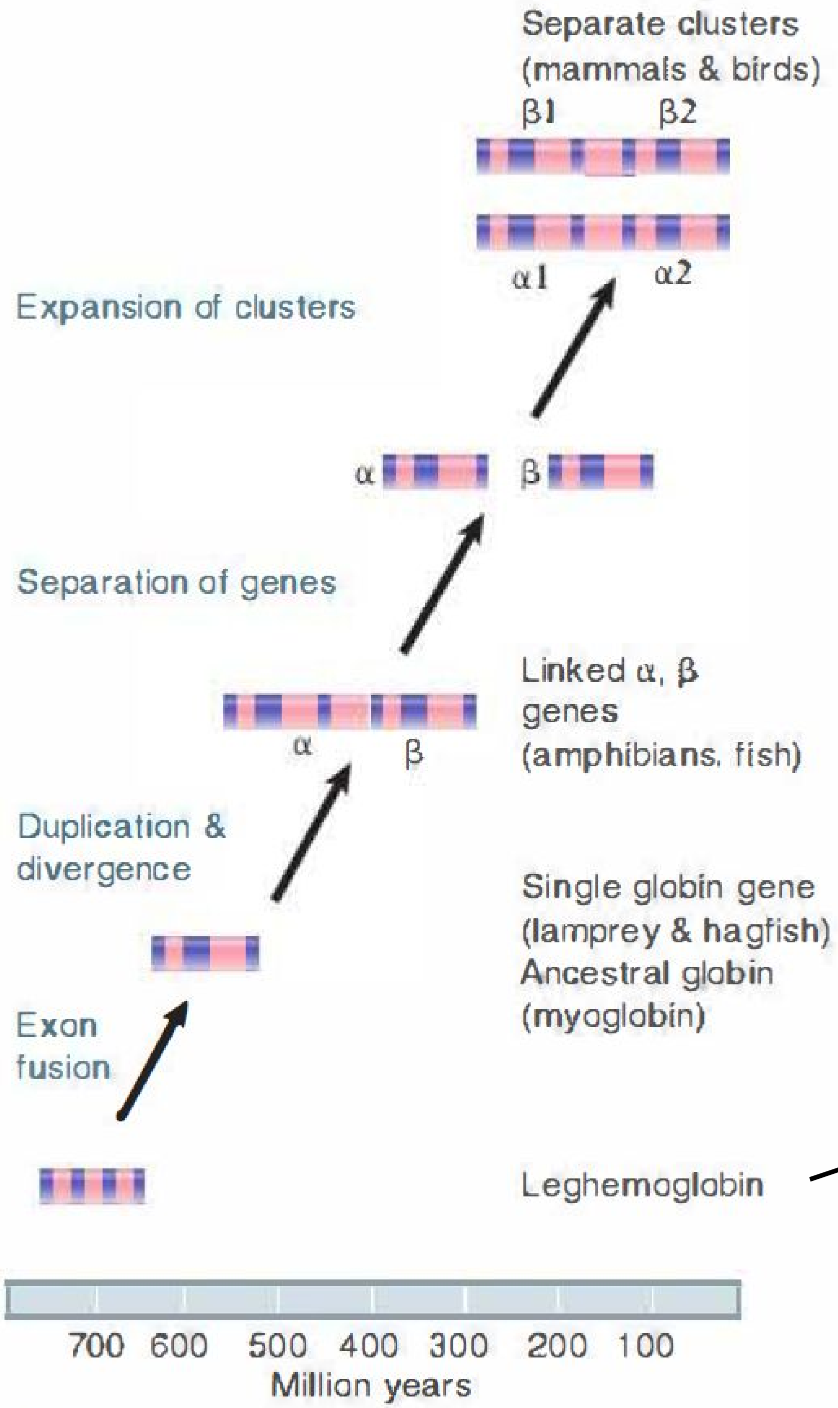


FIGURE 8.9 All globin genes have evolved by a series of duplications, transpositions, and mutations from a single ancestral gene.

Ancestral precursor

Divergence → percentage difference between correlated nucleotide sequences.

Mutations can occur in <u>synonymous sites or in nonsynonymous sites.</u>

Nonsynonymous and synonymous sites

.Val.Cys.Leu.Ala.Pro.

.GTG.TGT.CTA.GCT.CCA.

.GTG.TGT.CGA.GCT.CCA.

.Val.Cys.Arg.Ala.Pro.

.Val.Cys.Leu.Ala.Pro.

.GTG.TGT.CTA.GCT.CCA.

.GTG.TGT.CTT.GCT.CCA.

.Val.Cys.Leu.Ala.Pro.

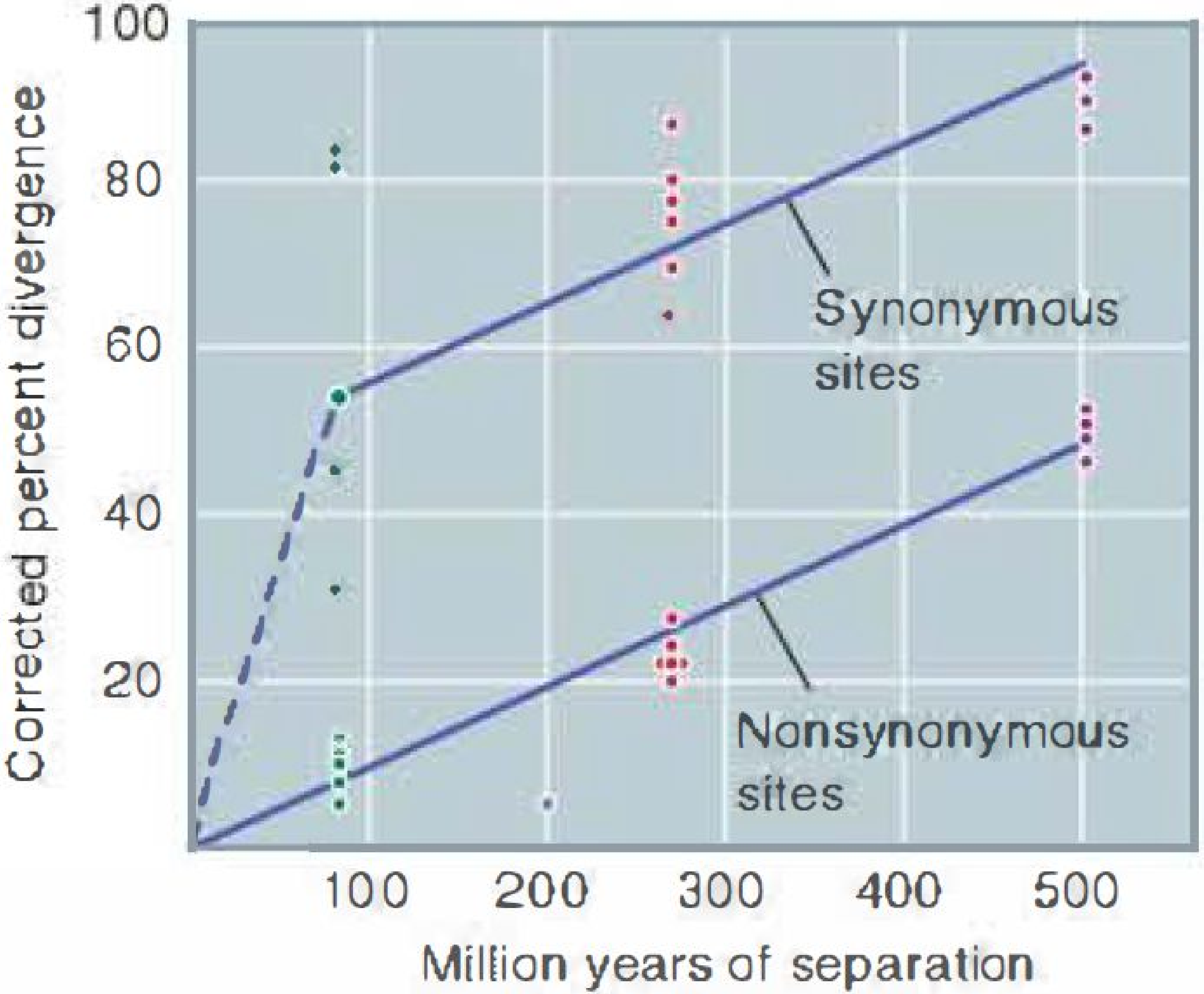


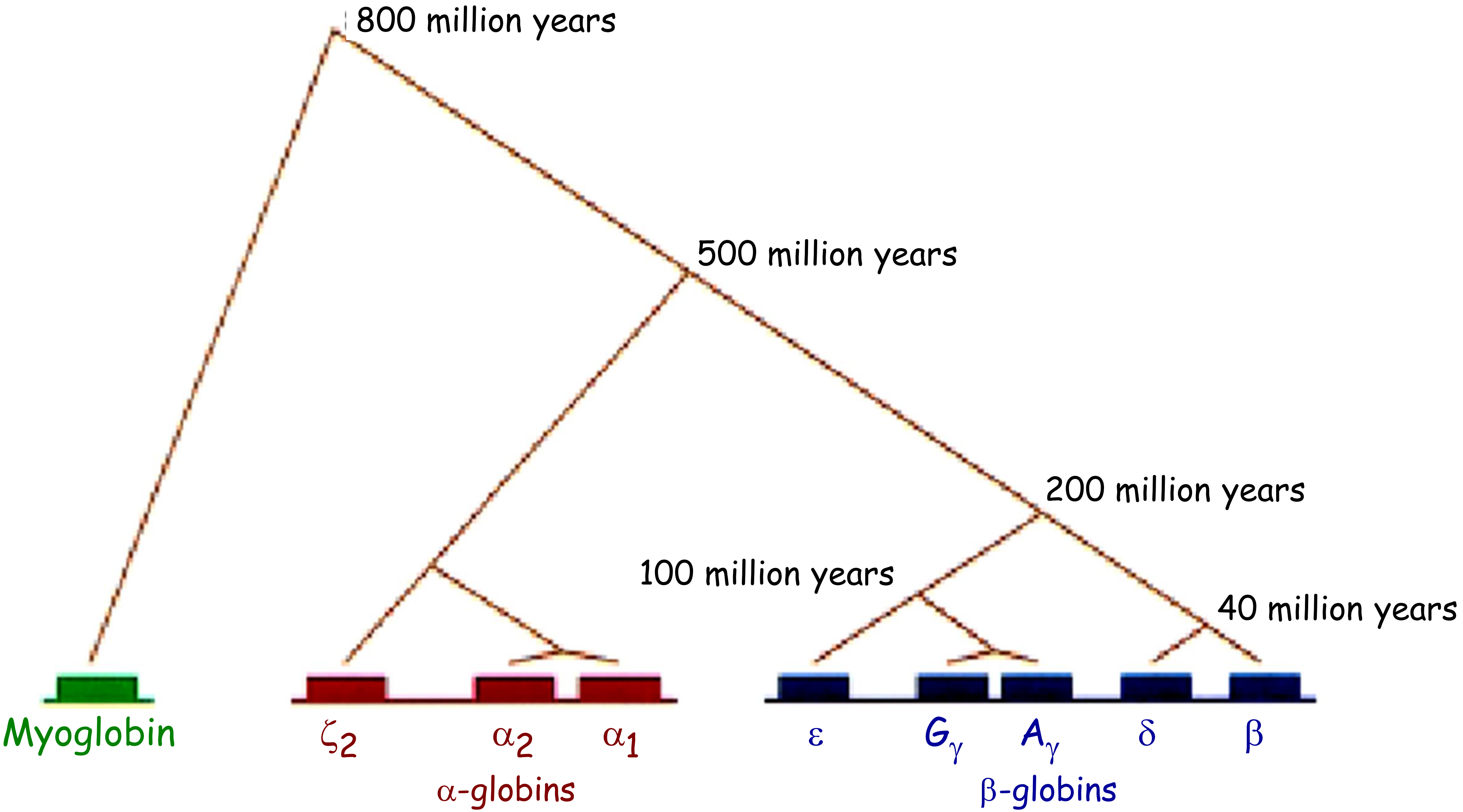
FIGURE 8.8 Divergence of DNA sequences depends on evolutionary separation. Each point on the graph represents a pairwise comparison.

The divergence between any pair of sequences is (more or less) proportional to the time since they separated.

From their divergence, we can calculate how much time has passed since the duplication that generated them.

Gene duplications during the evolution of human globin gene family

The comparison of nucleotide sequences allows to establish the evolutive relationship among globin genes



There may be more members of a gene family, both functional and nonfunctional, than we would suspect on the basis of protein analysis. (proteome).

The extra functional genes may represent duplicates that encode identical polypeptides, or they may be related to-but different from-known proteins (and presumably expressed only briefly or in low amounts).