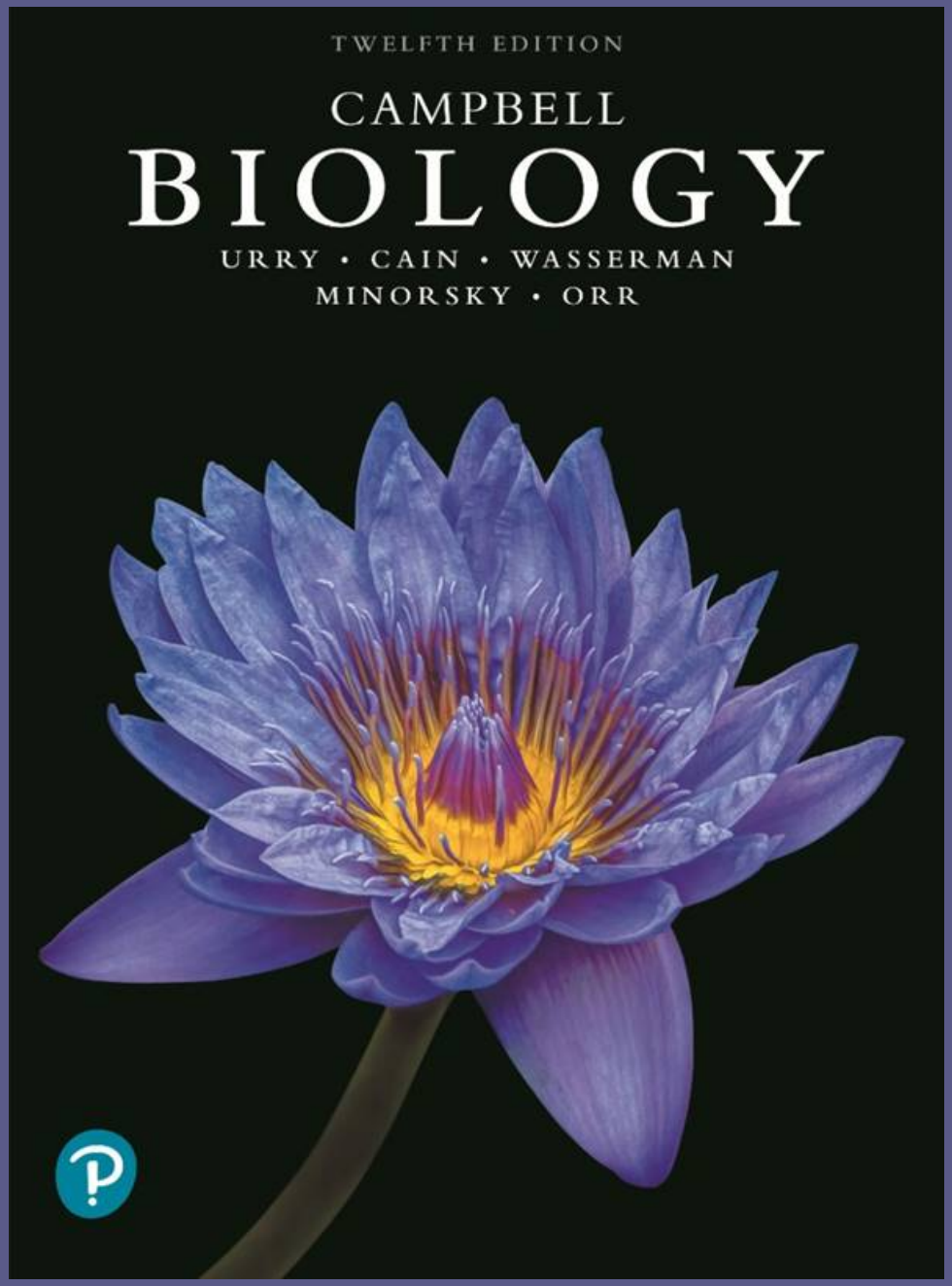


Chapter 21

Genomes and Their Evolution

Lecture Presentations by
Nicole Tunbridge and
Kathleen Fitzpatrick

Figure 21.1a

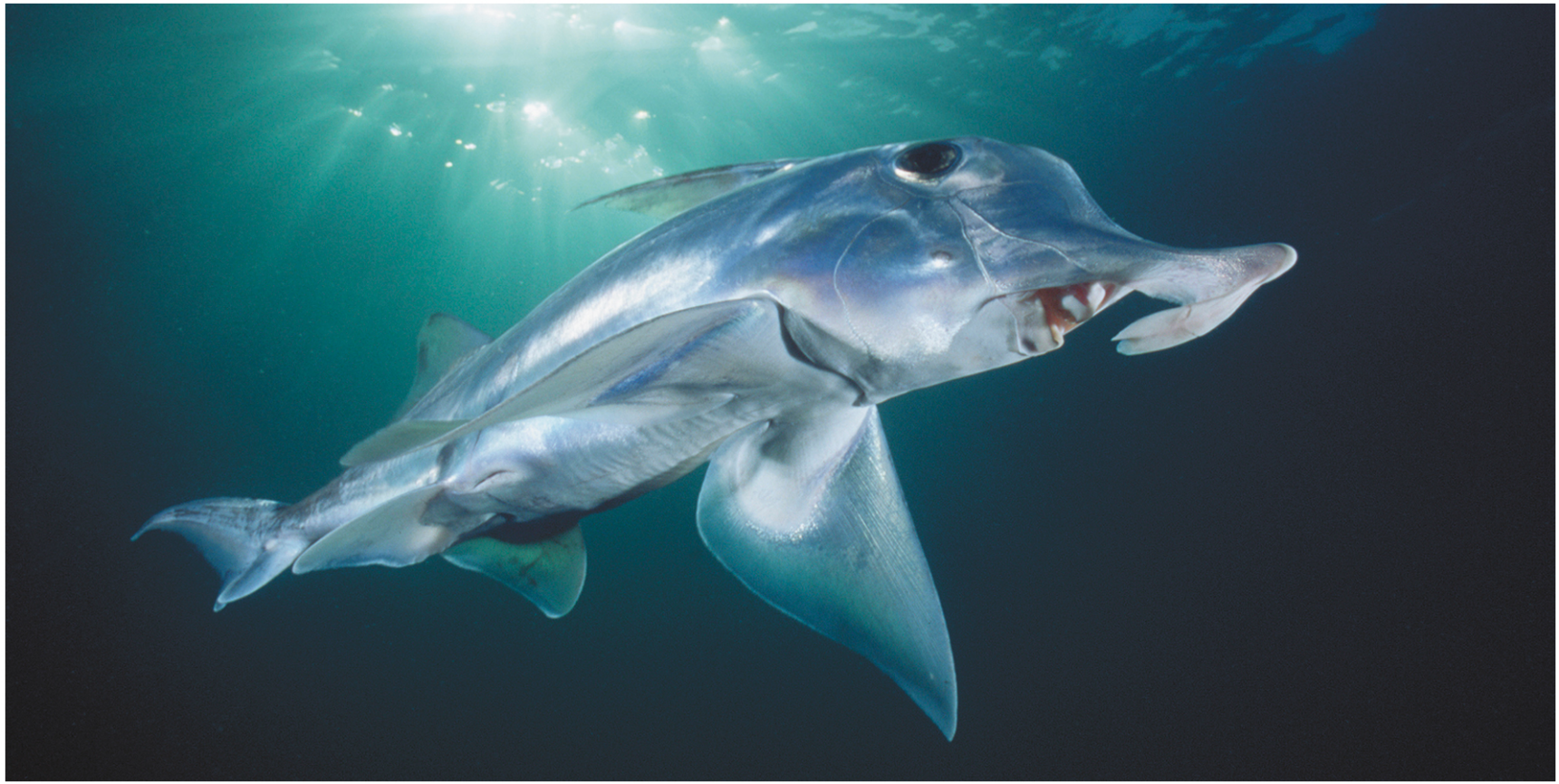
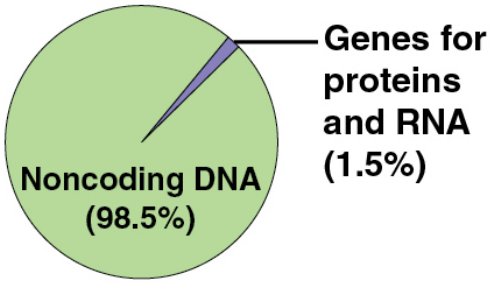


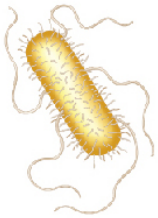
Figure 21.1b

What are some questions that can be explored by sequencing and comparing genomes?

What are the functions of the human genome?



How do genomes differ in number of genes?

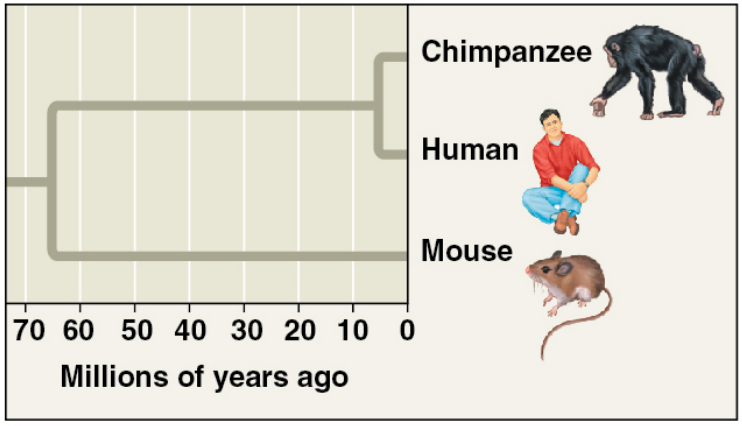


Escherichia coli 4,400 genes

Homo sapiens 21,300 genes

Zea mays (corn) 32,000 genes

What do gene sequences tell us about evolutionary relationships between species?



How do genomes evolve over time?

Elephant shark: slower-evolving genome

Tiger tail sea horse: faster-evolving genome

CONCEPT 21.1: The Human Genome Project fostered development of faster, less expensive sequencing techniques

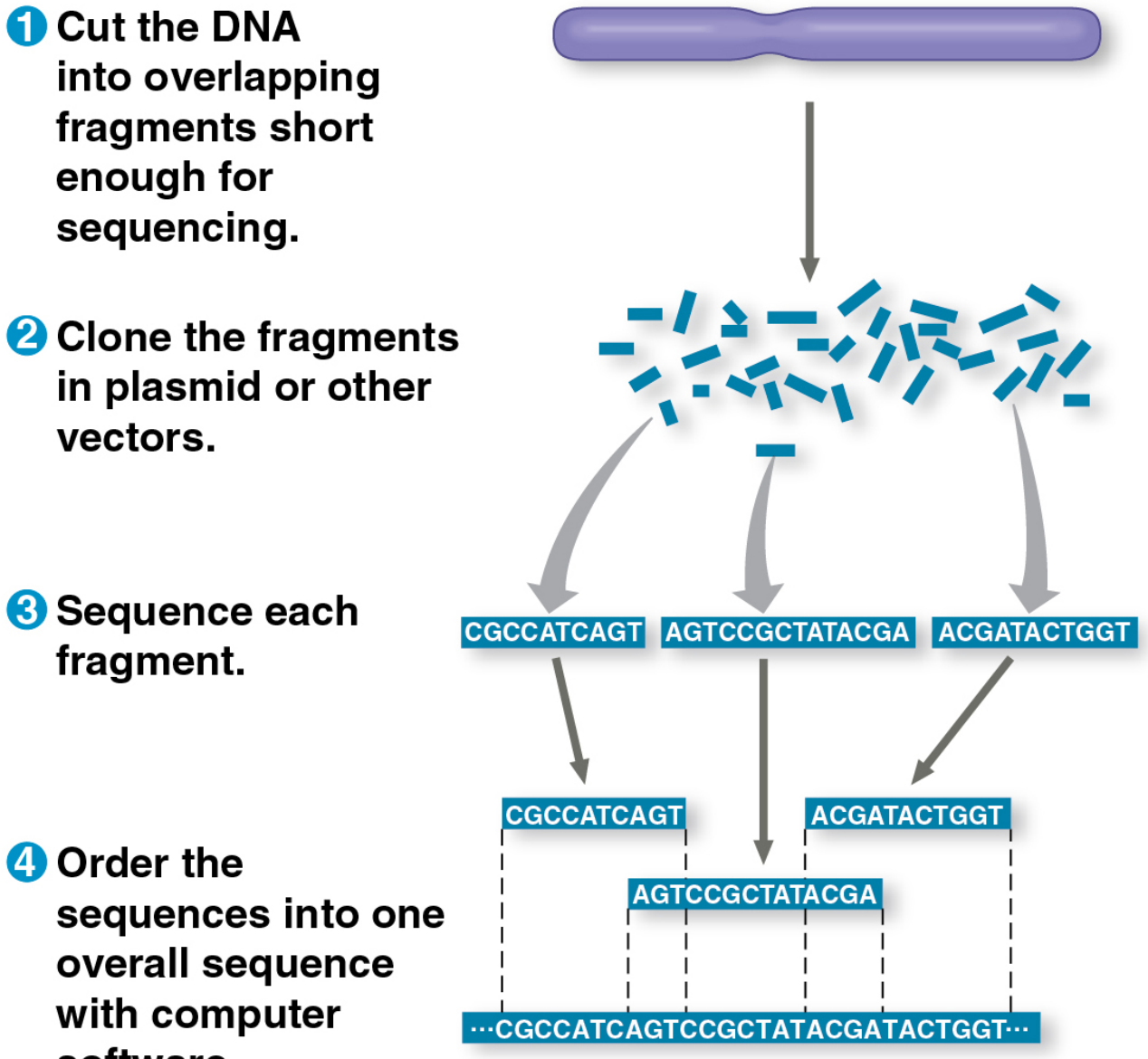
- **Genomics** is the study of whole sets of genes and their interactions
- **Bioinformatics** is the application of computational methods to the storage and analysis of biological data

- Officially begun as the **Human Genome Project** in 1990, the sequencing of the human genome was published in 2006
- The sequenced DNA was pooled from a few individuals
- Scientists reviewed the results and agreed on a **reference genome**, a full sequence that best represents the genome of a species

- The goal in mapping any genome is to determine the complete nucleotide sequence of each chromosome
- The human genome was completed using sequencing machines and the dideoxy chain termination method
- Two approaches complemented each other in obtaining the complete sequence

- The initial approach ordered each fragment based on earlier mapping of the human genome
- Then, molecular biologist J. Craig Venter set up a company to sequence the entire genome using an alternative **whole-genome shotgun approach**
- This used cloning and sequencing of fragments of randomly cut DNA followed by assembly into a single continuous sequence

Figure 21.2


1 Cut the DNA into overlapping fragments short enough for sequencing.
2 Clone the fragments in plasmid or other vectors.
3 Sequence each fragment.
4 Order the sequences into one overall sequence with computer software.
CGCCATCAGT
AGTCCGCTATACGA
ACGATACTGGT
CGCCATCAGT
ACGATACTGGT
AGTCCGCTATACGA
···CGCCATCAGTCCGCTATACGATACTGGT···

- The whole-genome shotgun approach is widely used today
- A major thrust of the Human Genome Project was the development of technologies for faster sequencing
- These “next-generation” techniques do not require a cloning step

- These techniques have also facilitated a **metagenomics** approach, in which DNA from a group of species in an environmental sample is sequenced
- Making sense of massive amounts of data from many genome sequences has necessitated new analytical approaches

CONCEPT 21.2: Scientists use bioinformatics to analyze genomes and their functions

- The Human Genome Project established databases and refined analytical software to make data available on the Internet

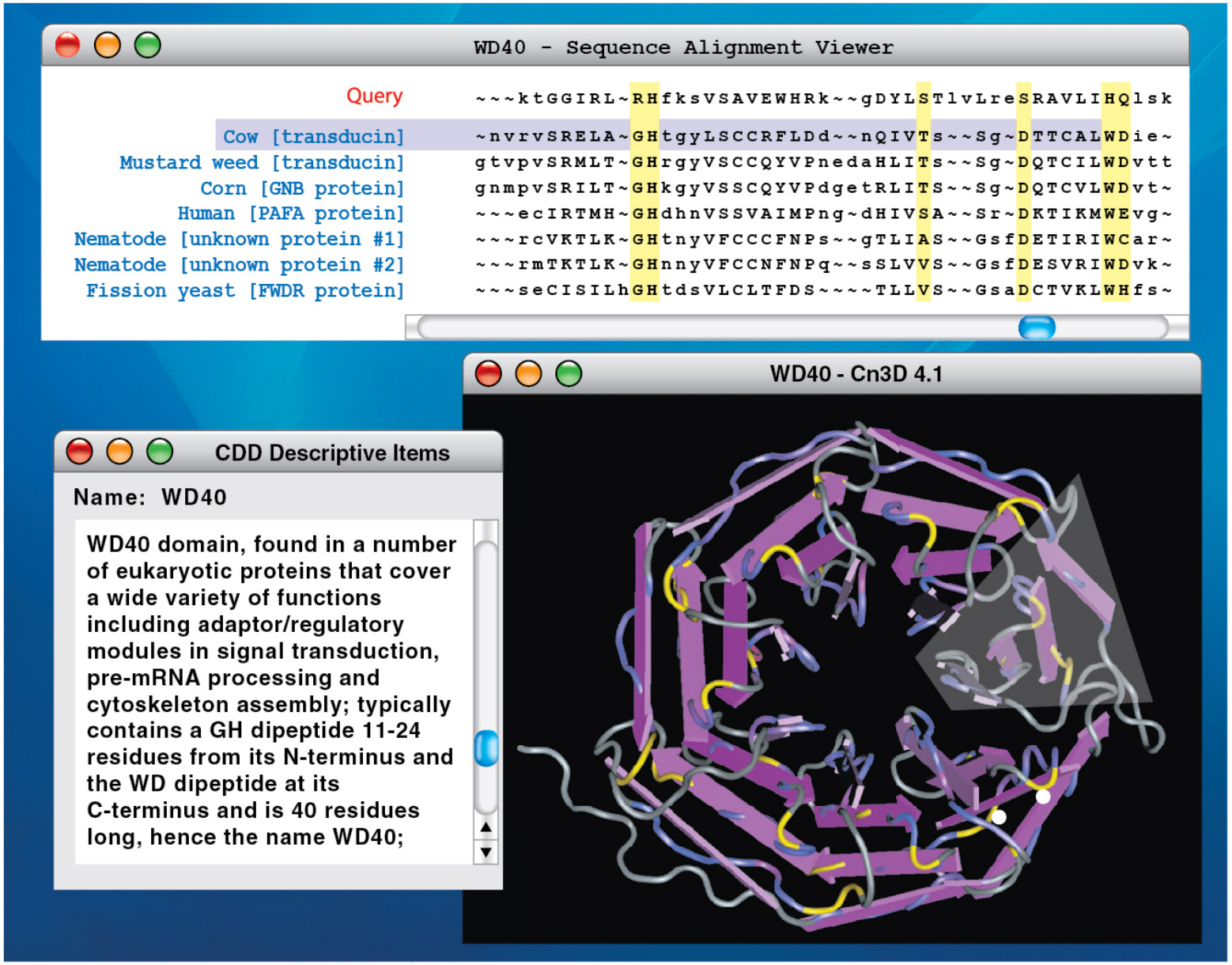
Centralized Resources for Analyzing Genome Sequences

- Bioinformatics resources are provided by a number of sources
  - The National Library of Medicine (NLM) and the National Institutes of Health (NIH) maintain the National Center for Biotechnology Information (NCBI)
  - European Molecular Biology Laboratory
  - DNA Data Bank of Japan
  - BGI in Shenzhen, China

- The NCBI database of sequences is called GenBank
- As of August 2019, it included the sequences of 214 million fragments of genomic DNA, totaling 366 billion base pairs
- A widely used software program on the NCBI website is called BLAST (Basic Local Alignment Search Tool)
- Users of this tool can compare a DNA sequence with every sequence in GenBank

- Another program allows comparison of protein sequences
- A third program can search any protein sequence for conserved (commo) stretches of amino acids (domains) for which a function is known or suspected
- It can show a three-dimensional model of the domain alongside other relevant information

Figure 21.3


WD40 - Sequence Alignment Viewer
Query
~~~ktGGIRL~RHfksVSAVEWHRk~~gDYLSTlvLreSRAVLIHQlsk
Cow [transducin]
~nvrvSRELA~GHtgyLSCCRFLDd~~nQIVTs~~Sg~DTTCALWDie~
Mustard weed [transducin]
gtvpvSRMLT~GHrgyVSCCQYVPnedaHLITs~~Sg~DQTCILWDvtt
Corn [GNB protein]
gnmpvSRILT~GHkgyVSSCQYVPdgetRLITS~~Sg~DQTCVLWDvt~
Human [PAFA protein]
~~~ecIRTMH~GHdhnVSSVAIMPng~dHIVSA~~Sr~DKTIKMWEvg~
Nematode [unknown protein #1]
~~~rcVKTLK~GHtnyVFCCCFNPs~~gTLIAS~~GsfDETIRIWCar~
Nematode [unknown protein #2]
~~~rmTKTLK~GHnnyVFCCNFNPq~~sSLVVS~~GsfDESVRIWDvk~
Fission yeast [FWDR protein]
~~~seCISILhGHtdsVLCLTFDS~~~~TLLVS~~GsaDCTVKLWHfs~
WD40 - Cn3D 4.1
CDD Descriptive Items
Name: WD40
WD40 domain, found in a number of eukaryotic proteins that cover a wide variety of functions including adaptor/regulatory modules in signal transduction, pre-mRNA processing and cytoskeleton assembly; typically contains a GH dipeptide 11-24 residues from its N-terminus and the WD dipeptide at its C-terminus and is 40 residues long, hence the name WD40;
~~~

- Rutgers University and the University of California, San Diego, maintain a world-wide database of all three-dimensional proteins structures that have been determined
- It is called the Protein Data Bank
- There is a vast array of resources available for researchers anywhere in the world to use free of charge

Identifying Protein-Coding Genes and Understanding Their Functions

- Using available DNA sequences, geneticists can study genes directly
- The identification of protein-coding genes within DNA sequences in a database is called **gene annotation**

- Gene annotation uses three lines of evidence to identify a gene
- First computers search for patterns that indicate the presence of genes
- This includes translational start and stop signals, RNA splicing sites and other signs, such as promoter sequences
- The software also looks for short sequences that specify known mRNAs

- The second step is to obtain clues about the identities and functions
- Software is used to compare the sequence of a protein to the products of known genes from other organisms
- The final step is to use RNA-seq or some other method to show that the relevant RNA is actually expressed from the proposed gene

Understanding Genes and Gene Expression at the Systems Level

- Genomics is a rich source of new insights into questions about genome organization, regulation of gene expression, embryonic development, and evolution
- The ENCODE (Encyclopedia of DNA Elements) project ran from 2003 to 2012
- The aim was to learn about the functionally important elements in the human genome

- Besides working to identify enhancers and promoters, investigators also extensively characterized histone and DNA modifications and chromatin structure
- This project allows comparison of results from different projects, yielding a richer picture of the whole genome

- About 75% of the genome is transcribed at some point in at least one cell type studied
- Biochemical functions have been assigned to DNA elements making up at least 80% of the genome
- The ENCODE project analyzed cells in culture, so its potential for clinical applications was limited

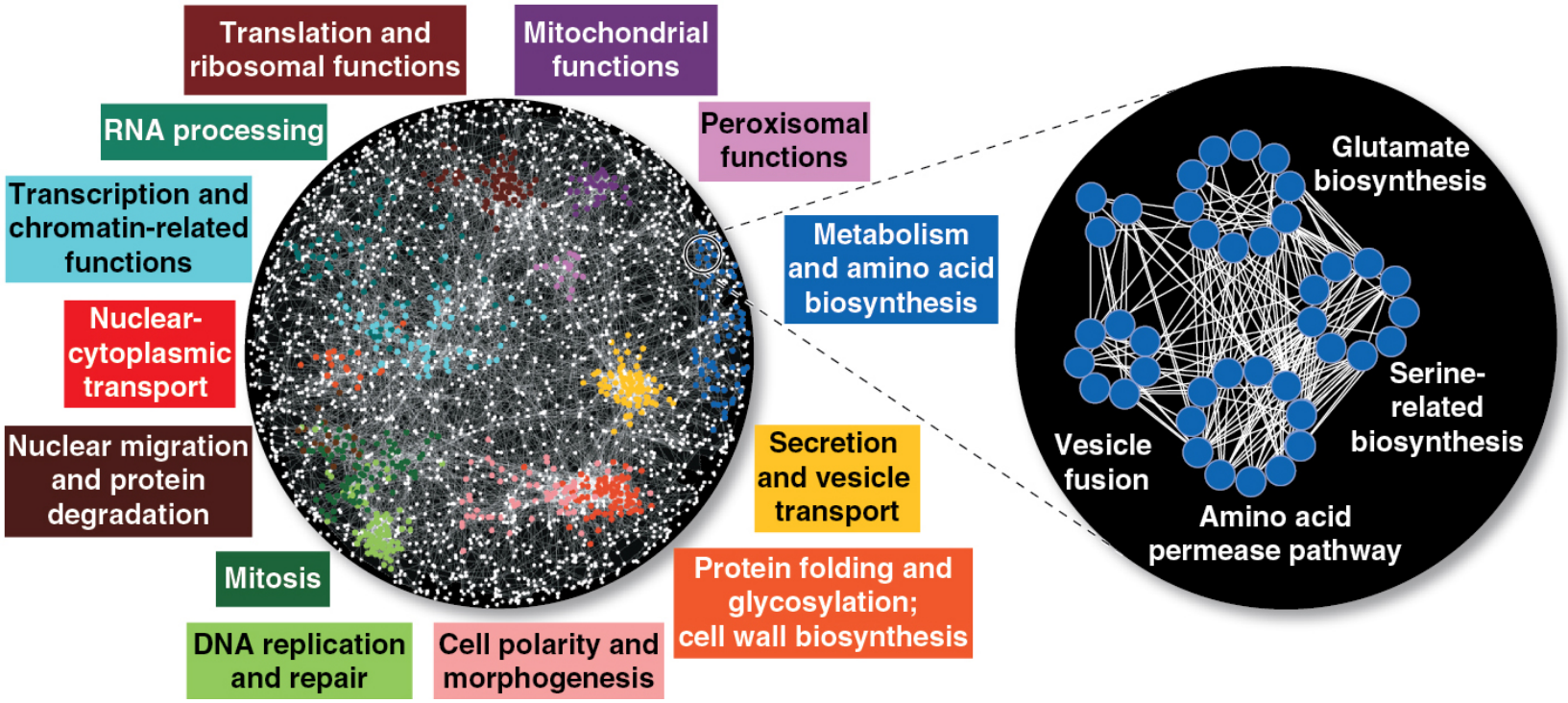
- A related project called the Roadmap Epigenomics Project set out to characterize the epigenetic features of the genome (the epigenome)
- A useful finding was that the original tissue in which a cancer arose can be identified in a secondary tumor based on its epigenomic features

Systems Biology

- **Proteomics** is an approach to studying large sets of proteins and their properties
- A **proteome** is the entire set of proteins expressed by a cell or group of cells
- Biologists have begun to compile catalogs of genes and proteins and have begun to focus on their functional integration in biological systems
- This approach is called **systems biology**

- Researchers working on the yeast *Saccharomyces cerevisiae* used sophisticated techniques to disable pairs of genes one pair at a time, creating double mutants
- Computer software then mapped genes to produce a network-like “functional map” of their interactions
- The systems biology approach is possible because of advances in bioinformatics

Figure 21.4


Translation and ribosomal functions
Mitochondrial functions
RNA processing
Peroxisomal functions
Transcription and chromatin-related functions
Metabolism and amino acid biosynthesis
Nuclear-cytoplasmic transport
Nuclear migration and protein degradation
Secretion and vesicle transport
Mitosis
Protein folding and glycosylation; cell wall biosynthesis
DNA replication and repair
Cell polarity and morphogenesis
Glutamate biosynthesis
Serine-related biosynthesis
Vesicle fusion
Amino acid permease pathway

Application of Systems Biology to Medicine

- The Cancer Genome Atlas project culminated in 2018 with publications called the Pan-Cancer Atlas
- In this project, many interacting genes and gene products were analyzed together as a group
- High-throughput techniques are increasingly being applied to the problem of cancer
- Overall, the Pan-Cancer Atlas contributed significantly to understanding how, where, and why tumors arise

- DNA microarrays on glass or silicon chips and, increasingly, RNA-seq are used to analyze gene expression patterns in patients with cancers or other diseases
- Analyzing which genes are overexpressed or underexpressed in a cancer allows physicians to tailor the treatment to unique genetic makeup of the patient and the cancer

Figure 21.5


P/N: 520019
Lot #: SAMPLES
Exp. Date: 00/00/00
For Research Use Only
Human Genome U133 Plus 2.0
GeneChip
Affymetrix

CONCEPT 21.3: Genomes vary in size, number of genes, and gene density

- The sequences of thousands of genomes have been completed
- Tens of thousands of genomes are either in progress or are considered permanent drafts
- Among the sequences in progress are roughly 22,000 metagenomes

Genome Size

- Genomes of most bacteria and archaea range from 1 to 6 million base pairs (Mb)
- Eukaryotic genomes tend to be larger
- Most plants and animals have genomes greater than 100 Mb; humans have 3,000 Mb
- Within each domain, there is no systematic relationship between genome size and phenotype

Table 21.1 Genome Sizes and Estimated Numbers of Genes*

Organism	Haploid Genome Size (Mb)	Number of Genes	Genes per Mb
Bacteria			
Haemophilus influenzae	1.8	1,700	940
Escherichia coli	4.6	4,400	950
Archaea			
Archaeoglobus fulgidus	2.2	2,500	1,130
Methanosarcina barkeri	4.8	3,600	750
Eukaryotes			
Saccharomyces cerevisiae (yeast, a fungus)	12	6,300	525
Utricularia gibba (floating bladderwort)	82	28,500	348
Caenorhabditis elegans (nematode)	100	20,100	200
Arabidopsis thaliana (mustard family plant)	120	27,000	225
Drosophila melanogaster (fruit fly)	165	14,000	85
Daphnia pulex (water flea)	200	31,000	155
Zea mays (corn)	2,300	32,000	14
Ailuropoda melanoleuca (giant panda)	2,400	21,000	9
Homo sapiens (human)	3,000	21,300	7
Paris japonica (Japanese canopy plant)	149,000	ND	ND

*Some values given here are likely to be revised as genome analysis continues. Mb = million base pairs; the haploid number is used because it represents a complete set of genetic information. ND = not determined.

Number of Genes

- Free-living bacteria and archaea have 1,500 to 7,500 genes
- Unicellular fungi have about 5,000 genes and multicellular eukaryotes up to at least 40,000 genes
- Number of genes is not correlated to genome size

- It is estimated that the nematode *C. elegans* has 100 Mb and 20,100 genes, while *Drosophila melanogaster* has 165 Mb and 14,000 genes
- Researchers predicted the human genome would contain about 50,000 to 100,000 genes; however, the number is around 21,300
- Vertebrate genomes can produce more than one polypeptide per gene because of alternative splicing of RNA transcripts

Gene Density and Noncoding DNA

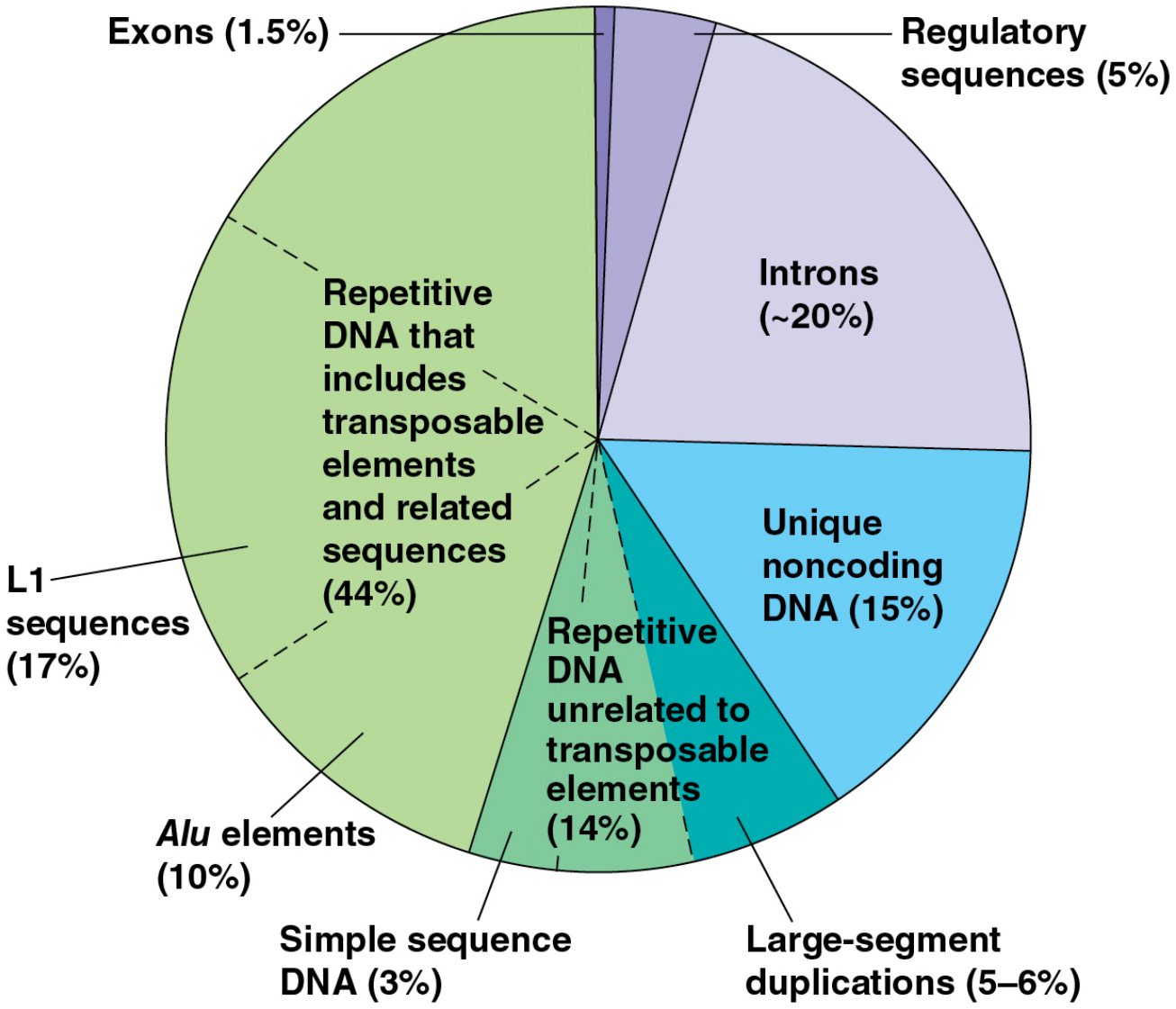
- Humans and other mammals have the lowest gene density, or number of genes in a given length of DNA
- Multicellular eukaryotes have many introns within genes and a large amount of noncoding DNA between genes

CONCEPT 21.4: Multicellular eukaryotes have a lot of noncoding DNA and many multigene families

- Sequencing of the human genome revealed that 98.5% does not code for proteins, rRNAs, or tRNAs
- Gene regulatory sequences and introns account for 5% and 20%, respectively, of the human genome

- Noncoding DNA, found between genes, includes:
  - **Pseudogenes**, former genes that have accumulated mutations and are nonfunctional
  - **Repetitive DNA**, present in multiple copies in the genome
- A high level of sequence conservation in some noncoding DNA among humans, rats, and mice suggests that these regions have important functions

Figure 21.6


Exons (1.5%)
Regulatory sequences (5%)
Introns (~20%)
Repetitive DNA that includes transposable elements and related sequences (44%)
L1 sequences (17%)
Unique noncoding DNA (15%)
Repetitive DNA unrelated to transposable elements (14%)
Alu elements (10%)
Simple sequence DNA (3%)
Large-segment duplications (5–6%)

Transposable Elements and Related Sequences

- Prokaryotes and eukaryotes have stretches of DNA that can move from one location to another within the genome, called **transposable elements**
- About 75% of human repetitive DNA is made up of transposable elements and the sequences related to them
- The first evidence of these mobile elements came from Barbara McClintock's breeding experiments with Indian corn (maize)

Figure 21.7

Movement of Transposons and Retrotransposons

- Eukaryotic transposable elements are of two types
  - **Transposons** move by means of a DNA intermediate and require a transposase enzyme
  - **Retrotransposons** move by means of an RNA intermediate using a reverse transcriptase

Figure 21.8

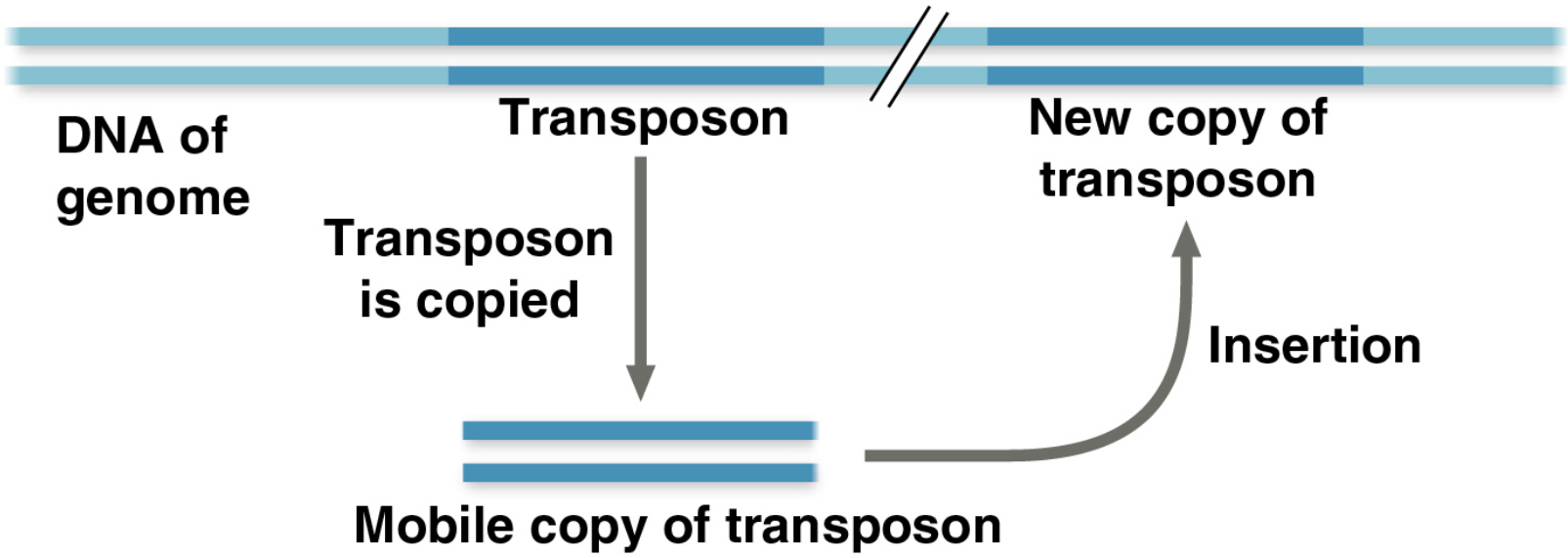
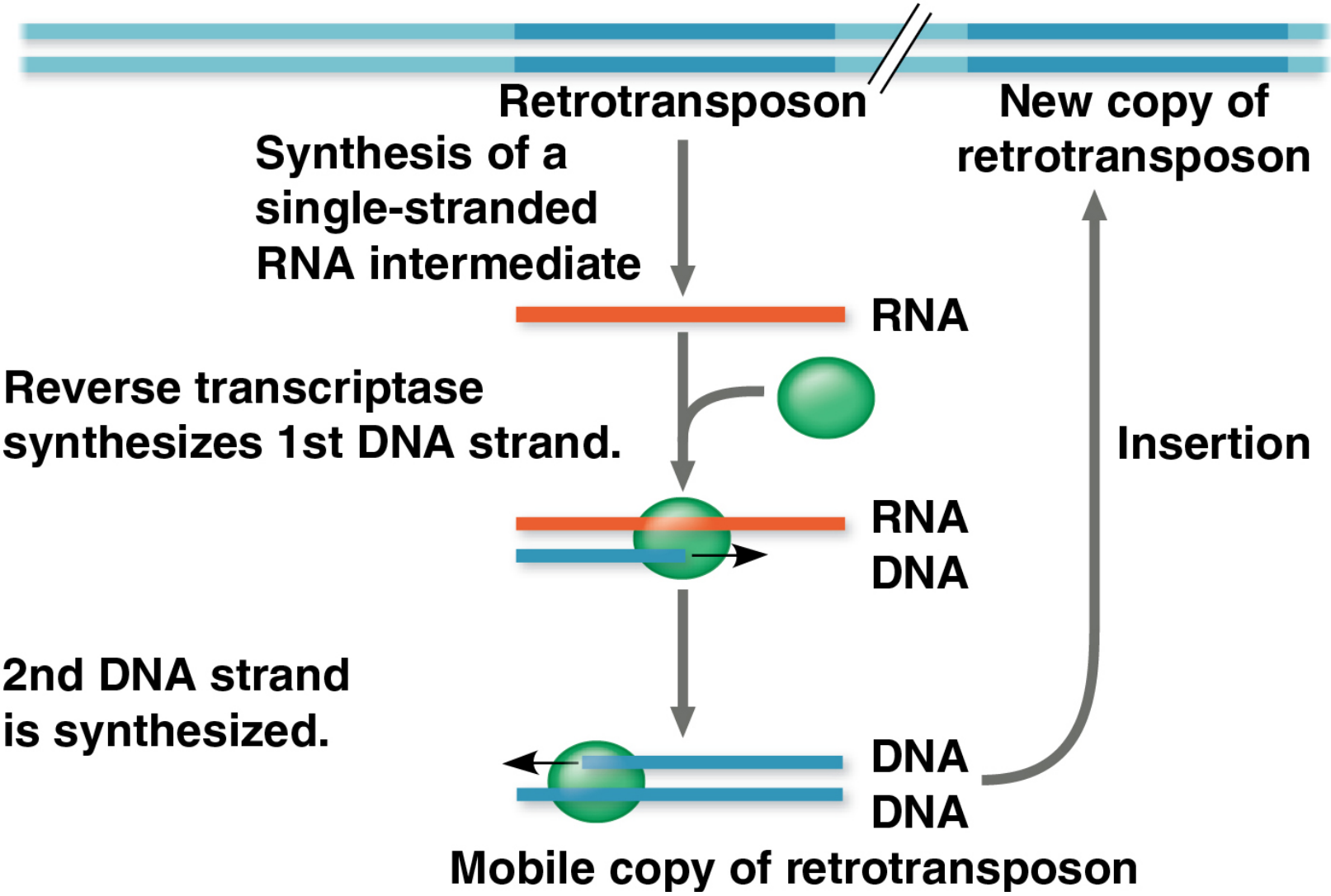

DNA of genome
Transposon
New copy of transposon
Transposon is copied
Insertion
Mobile copy of transposon

Figure 21.9


Retrotransposon
New copy of retrotransposon
Synthesis of a single-stranded RNA intermediate
RNA
Reverse transcriptase synthesizes 1st DNA strand.
Insertion
RNA
DNA
2nd DNA strand is synthesized.
DNA
DNA
Mobile copy of retrotransposon

Sequences Related to Transposable Elements

- Multiple copies of transposable elements and related sequences are scattered throughout eukaryotic genomes
- In humans and other primates, a large portion of transposable element–related DNA consists of a family of similar sequences called *Alu* elements
- Many *Alu* elements are transcribed into RNA molecules; some are thought to help regulate gene expression

- The human genome also contains many sequences of a type of retrotransposon called LINE-1 (L1)
- L1 sequences have a low rate of transposition and may affect chromatin structure
- Transposable elements are included in the “noncoding” DNA category, along with other repetitive sequences

Other Repetitive DNA, Including Simple Sequence DNA

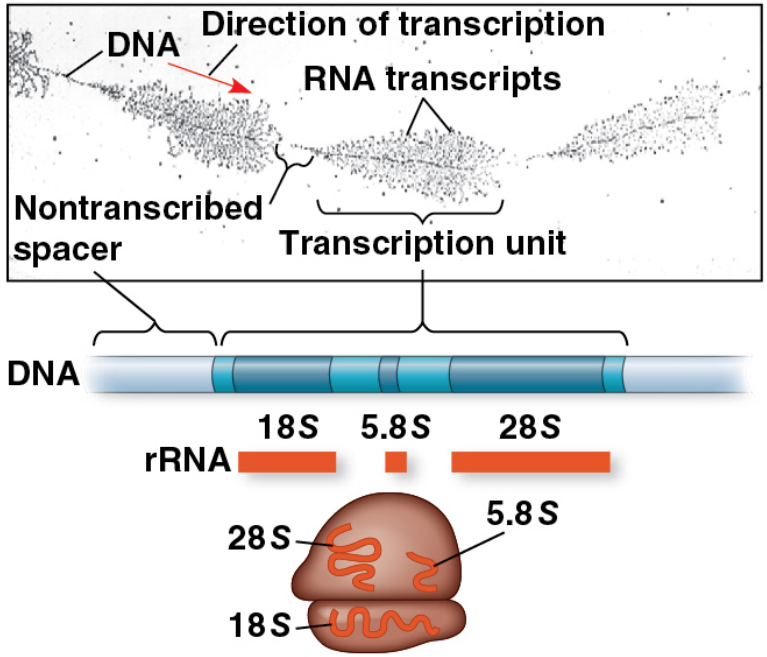
- Repetitive DNA accounts for about 14% of the human genome
- About 5–6% of the human genome consists of duplication of long sequences of DNA from one location to another
- In contrast, **simple sequence DNA** contains many copies of tandemly repeated short sequences

- A series of repeating units of 2 to 5 nucleotides is called a **short tandem repeat (STR)**
- The repeat number for STRs can vary among sites (within a genome) or individuals
- Simple sequence DNA makes up 3% of the human genome
- It is common in centromeres and telomeres, where it probably plays structural roles in the chromosome

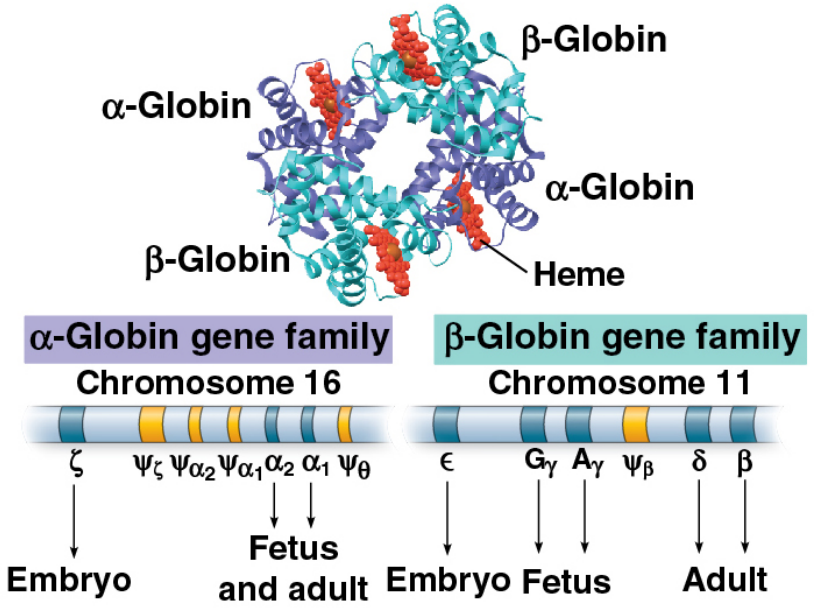
Genes and Multigene Families

- Many eukaryotic genes are present in one copy per haploid set of chromosomes
- The rest of the genes occur in **multigene families**, collections of two or more identical or very similar genes
- Some multigene families consist of identical DNA sequences, usually clustered tandemly, such as those that code for rRNA products

Figure 21.10



(a) Part of the ribosomal RNA gene family



(b) The human α-globin and β-globin gene families

- The classic examples of multigene families of nonidentical genes are two related families of genes that encode globins
- α-globins and β-globins are polypeptides of hemoglobin coded by genes on different human chromosomes and are expressed at different times in development

CONCEPT 21.5: Duplication, rearrangement, and mutation of DNA contribute to genome evolution

- The basis of change at the genomic level is mutation, which underlies much of genome evolution
- The earliest forms of life likely had only those genes necessary for survival and reproduction
- The size of genomes has increased over evolutionary time, with the extra genetic material providing raw material for gene diversification

Duplication of Entire Chromosome Sets

- Accidents in meiosis can lead to one or more extra sets of chromosomes, a condition known as polyploidy
- The genes in one or more of the extra sets can diverge by accumulating mutations
- These variations may persist if the organism carrying them survives and reproduces
- In this way, genes with novel functions can evolve

Alterations of Chromosome Structure

- Humans have 23 pairs of chromosomes, while chimpanzees have 24 pairs
- Following the divergence of humans and chimpanzees from a common ancestor, two ancestral chromosomes fused in the human line
- Large blocks of genes on human chromosome 16 are found on four mouse chromosomes
- This indicates that the genes in each block stayed together in both the human and mouse lineages

Figure 21.11

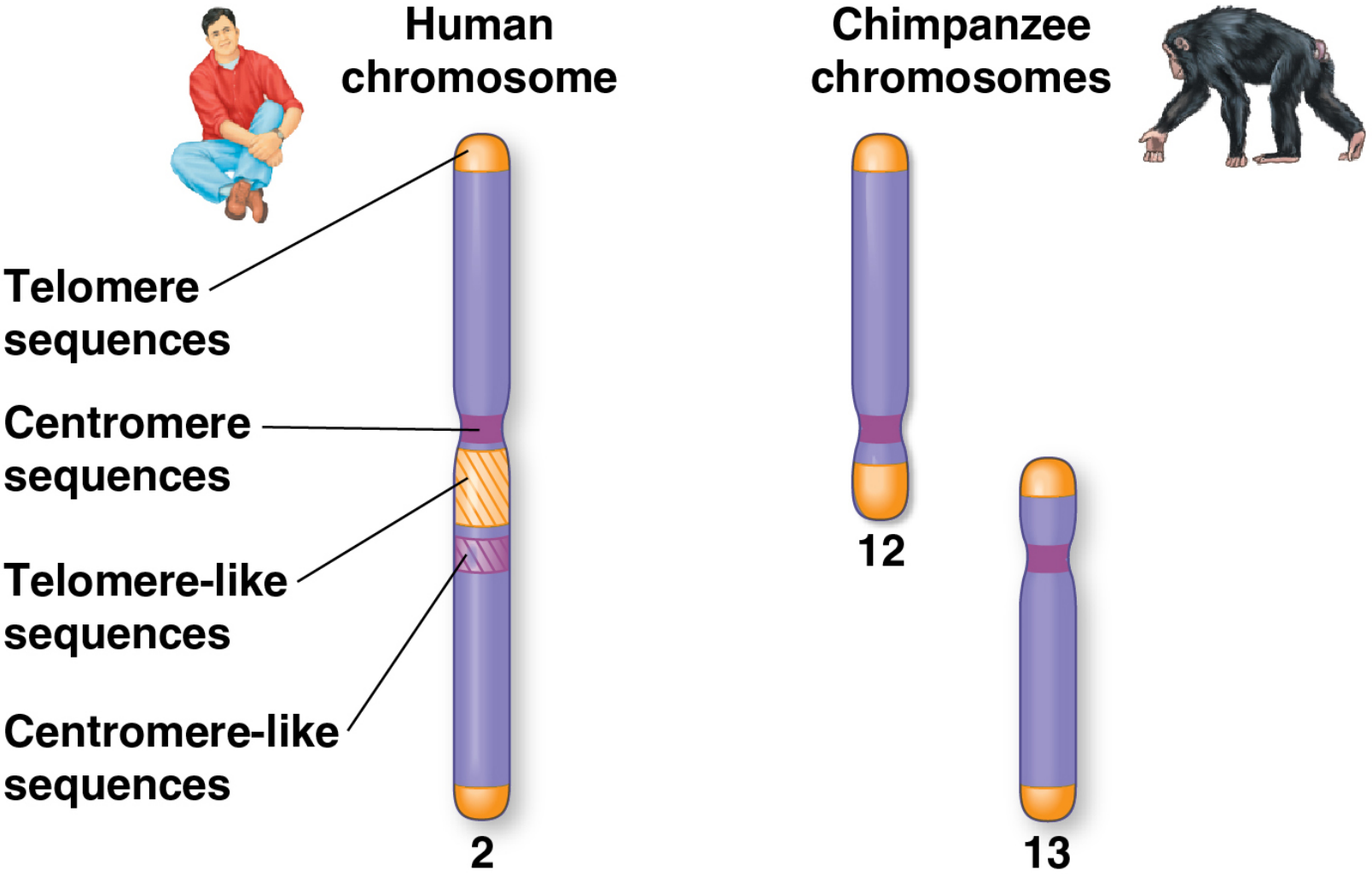
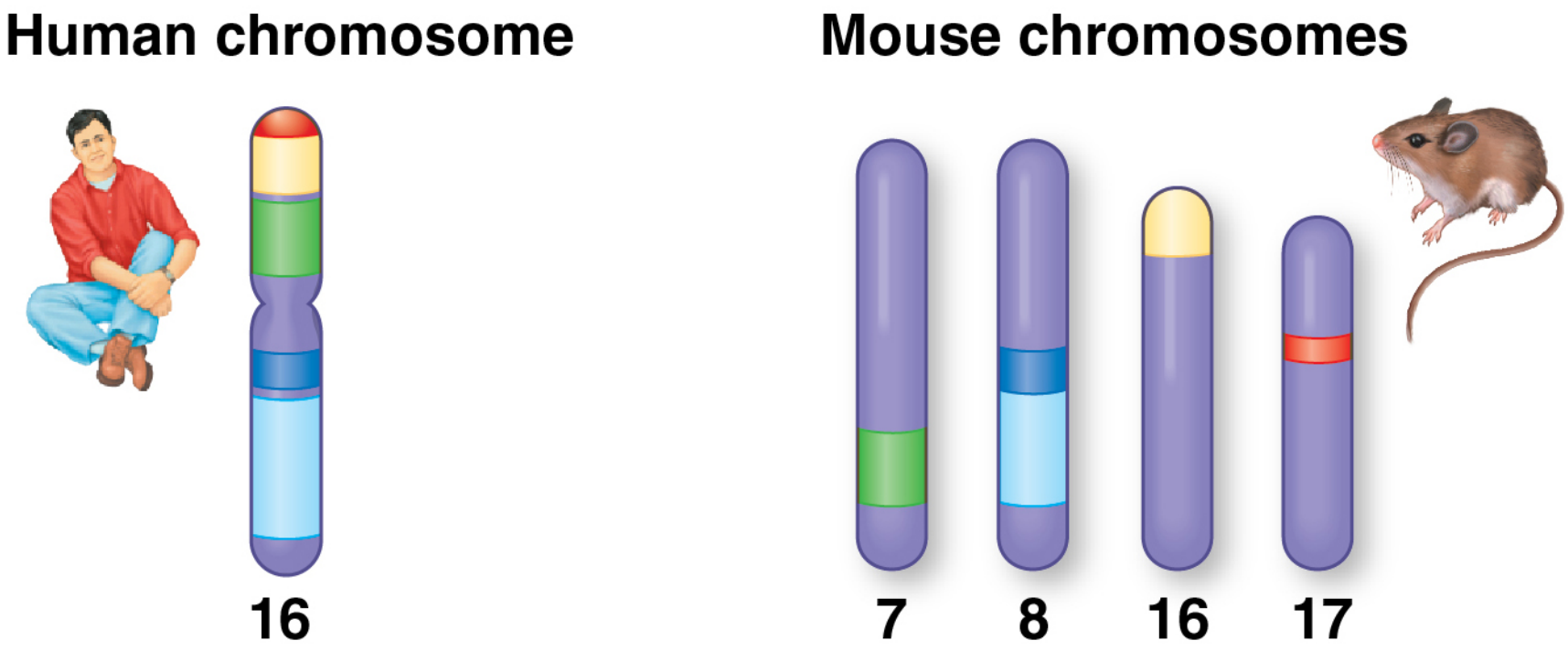

Human chromosome
Chimpanzee chromosomes
Telomere sequences
Centromere sequences
Telomere-like sequences
Centromere-like sequences
2
12
13

Figure 21.12

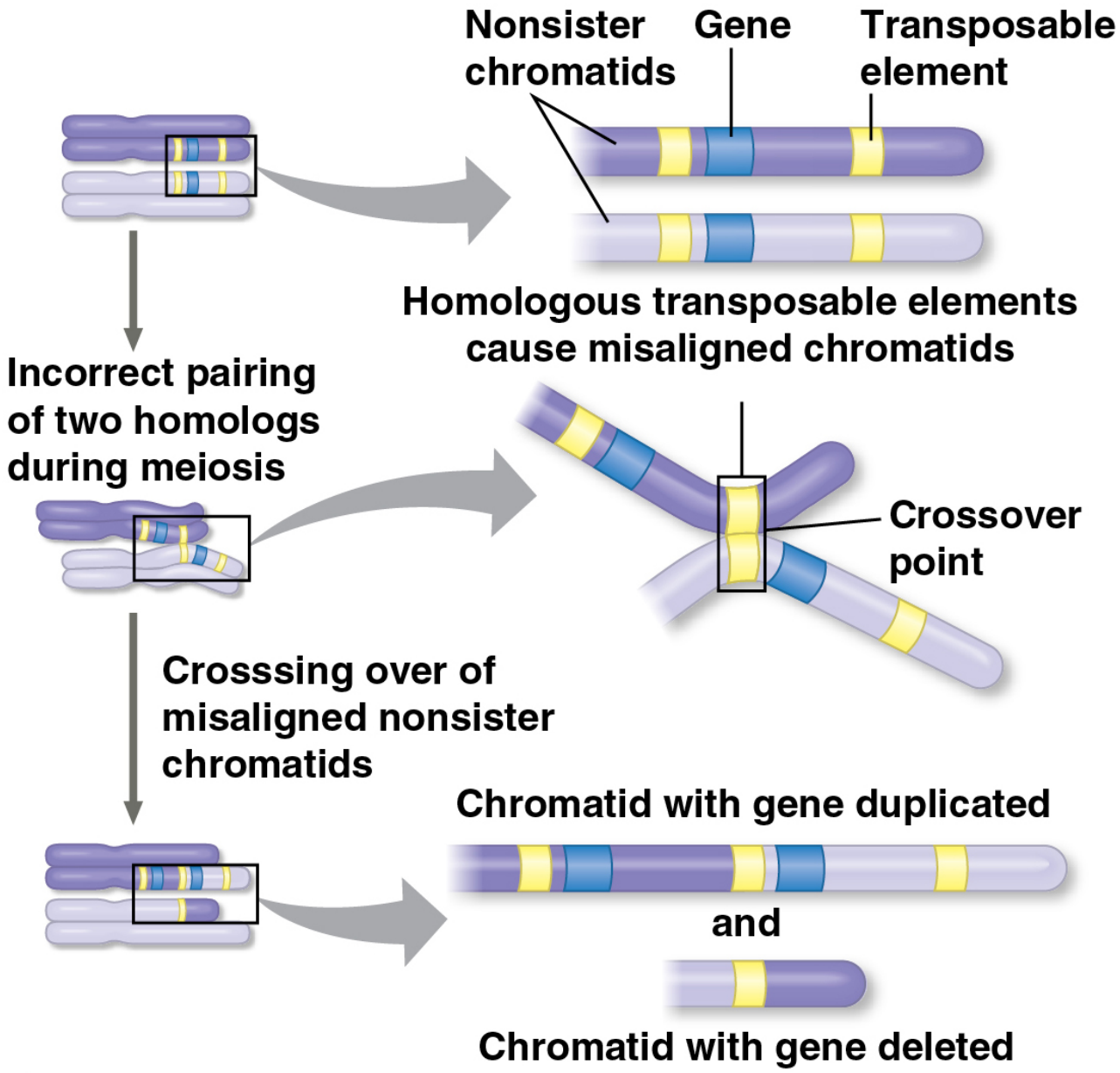


- Comparative analysis between chromosomes of humans and six other mammalian species paints a hypothetical chromosomal evolutionary history
- The rate of duplications and inversions seems to have accelerated about 100 million years ago
- This coincides with when large dinosaurs went extinct and mammals diversified
- Chromosomal rearrangements are thought to contribute to the generation of new species

Duplication and Divergence of Gene-Sized Regions of DNA

- Unequal crossing over during prophase I of meiosis can result in one chromosome with a deletion and another with a duplication of a particular region
- Transposable elements can provide sites for crossover between nonsister chromatids
- Also, slippage can occur during DNA replication so that a part of the template is either skipped, or replicated twice

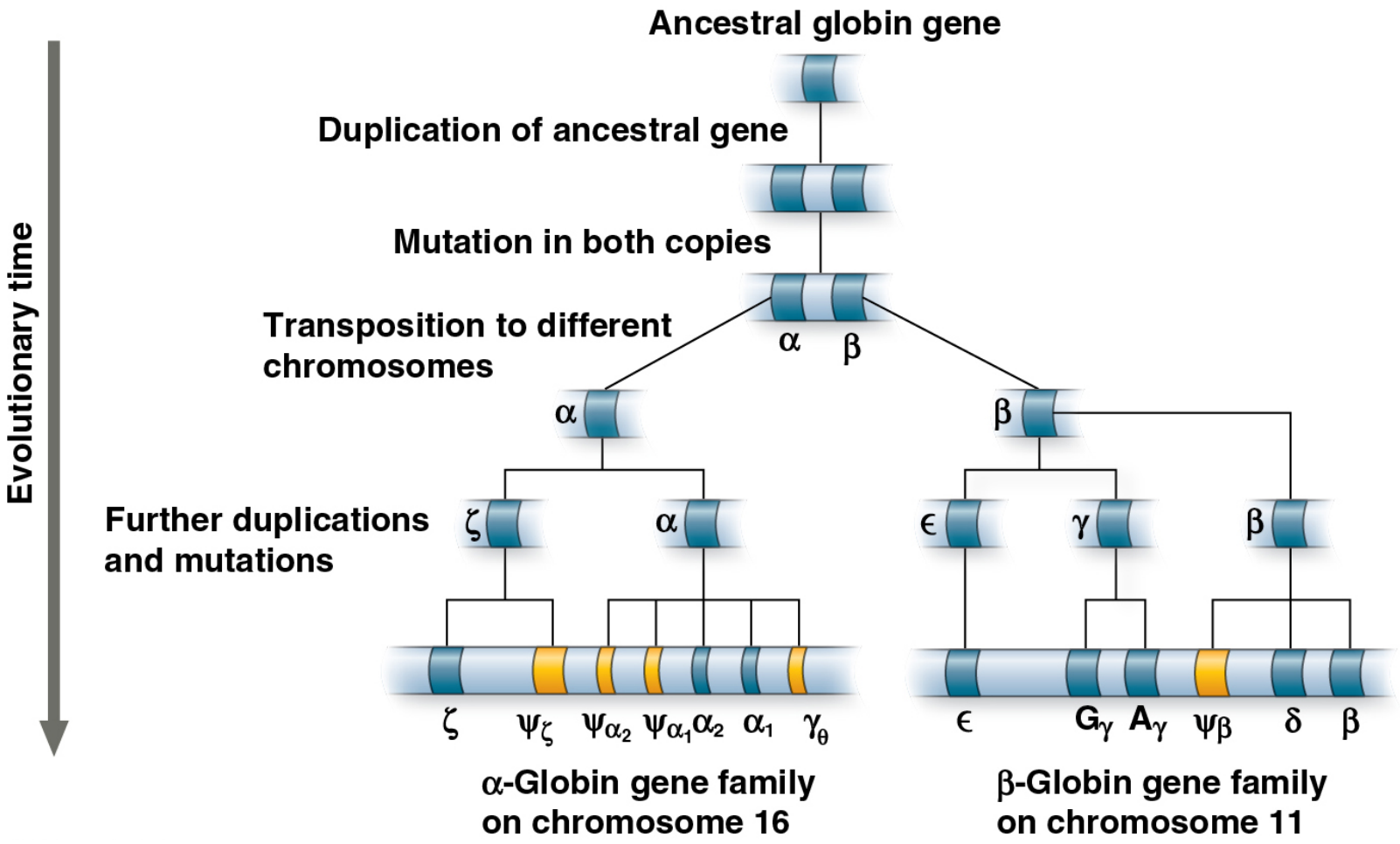
Figure 21.13


Nonsister chromatids
Gene
Transposable element
Homologous transposable elements cause misaligned chromatids
Incorrect pairing of two homologs during meiosis
Crossover point
Crosssing over of misaligned nonsister chromatids
Chromatid with gene duplicated
and
Chromatid with gene deleted

Evolution of Genes with Related Functions: The Human Globin Genes

- Evidence suggests that the genes encoding the globin proteins evolved from one common ancestral globin gene, which duplicated and diverged about 450–500 million years ago
- After the duplication events, differences between the genes in the globin family arose from the accumulation of mutations

Figure 21.14

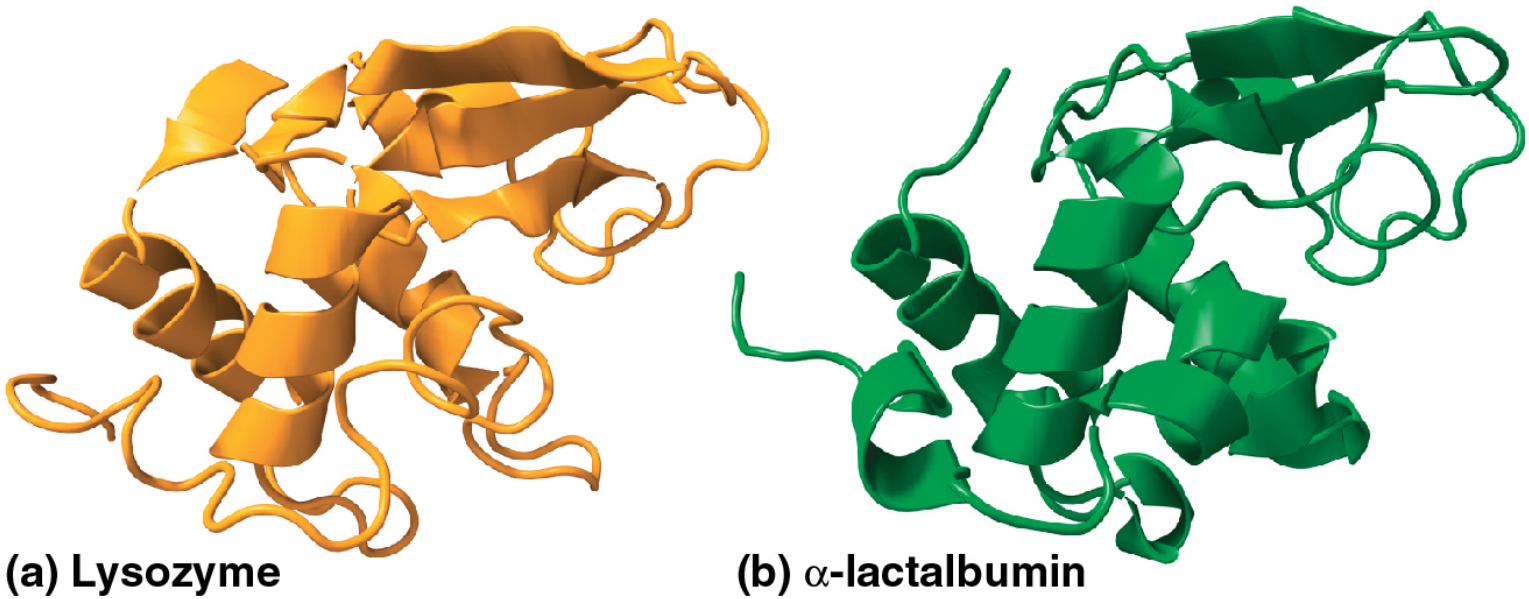

Ancestral globin gene
Duplication of ancestral gene
Mutation in both copies
Transposition to different chromosomes
α β
α
β
Further duplications and mutations
ζ α
ϵ γ β
ζ ψζ ψα2 ψα1α2 α1 γθ
ϵ Gγ Aγ ψβ δ β
α-Globin gene family on chromosome 16
β-Globin gene family on chromosome 11
Evolutionary time

- Subsequent duplications of these genes and random mutations gave rise to the present globin genes, which code for oxygen-binding proteins
- The similarity in the amino acid sequences of the various globin proteins supports this model of gene duplication and mutation

Evolution of Genes with Novel Functions

- One copy of a duplicated gene can undergo alterations that lead to a completely new function for the protein product
- For example, the lysozyme gene was duplicated and evolved into the gene that encodes α-lactalbumin in mammals
- Lysozyme is an enzyme that helps protect animals against bacterial infection
- α-lactalbumin is a nonenzymatic protein that plays a role in milk production in mammals

Figure 21.15



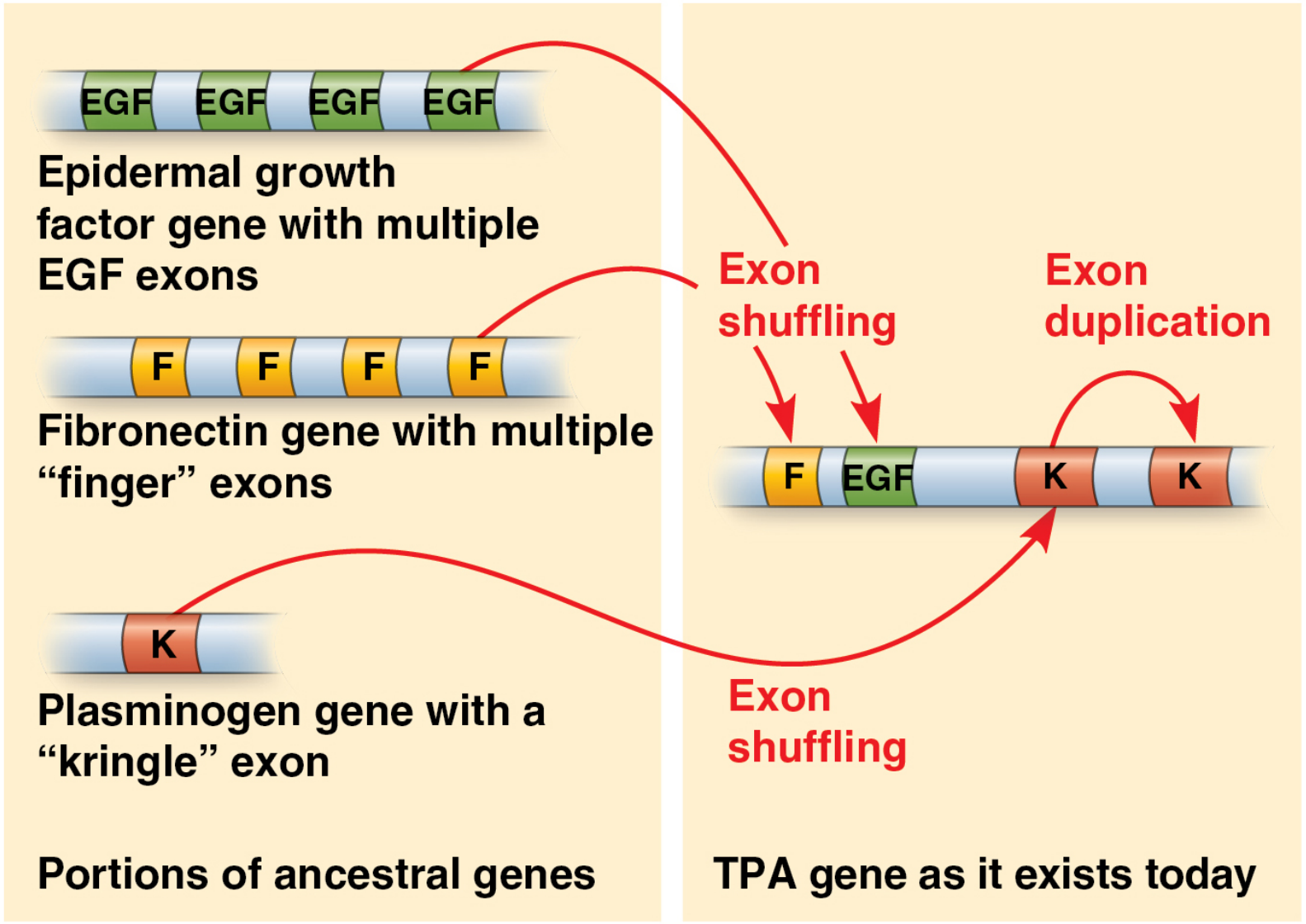
Lysozyme	1	KVFERCELARTLKRLGMDGYRGISLANWMCLAKWESGYNTRATN
α-lactalbumin	1	KQFTKCELSQLLK--DIDGYGGIALPELICTMFHTSGYDTQAIV
Lysozyme	51	STDYGIFQINSRYWCNDGKTPGAVNACHLSCSALLQDNIADAVA
α-lactalbumin	51	STEYGLFQISNKLWCKSSQVPQSRNICDISCDKFLDDDITDDIM
Lysozyme	101	RDPQGIRAWVAWRNRCQ-NRDVRQYVQGCGV
α-lactalbumin	101	D-IKGIDYWLAHKALCT--EKLEQWLCEKL-

(c) Amino acid sequence alignments of lysozyme and α-lactalbumin

Rearrangements of Parts of Genes: Exon Duplication and Exon Shuffling

- Errors in meiosis can result in an exon being duplicated on one chromosome and deleted from the homologous chromosome
- In exon shuffling, errors in meiotic recombination lead to some mixing and matching of exons, either within a gene or between two nonallelic genes
- The current version of the gene for tissue plasminogen activator (TPA) is thought to have arisen by several instances of exon shuffling and subsequent duplication

Figure 21.16


EGF
EGF
EGF
EGF
Epidermal growth factor gene with multiple EGF exons
F
F
F
F
Fibronectin gene with multiple "finger" exons
K
Plasminogen gene with a "kringle" exon
Portions of ancestral genes
Exon shuffling
Exon duplication
Exon shuffling
F
EGF
K
K
TPA gene as it exists today

How Transposable Elements Contribute to Genome Evolution

- Multiple copies of similar transposable elements facilitate recombination, or crossing over, between different chromosomes
- Insertion of transposable elements within a protein-coding sequence may block protein production
- Insertion of transposable elements within a regulatory sequence may increase or decrease protein production

- Transposable elements may carry a gene or groups of genes to a new position
- Transposable elements may also create new sites for alternative splicing in an RNA transcript
- In all cases, changes are usually detrimental but may on occasion prove advantageous to an organism

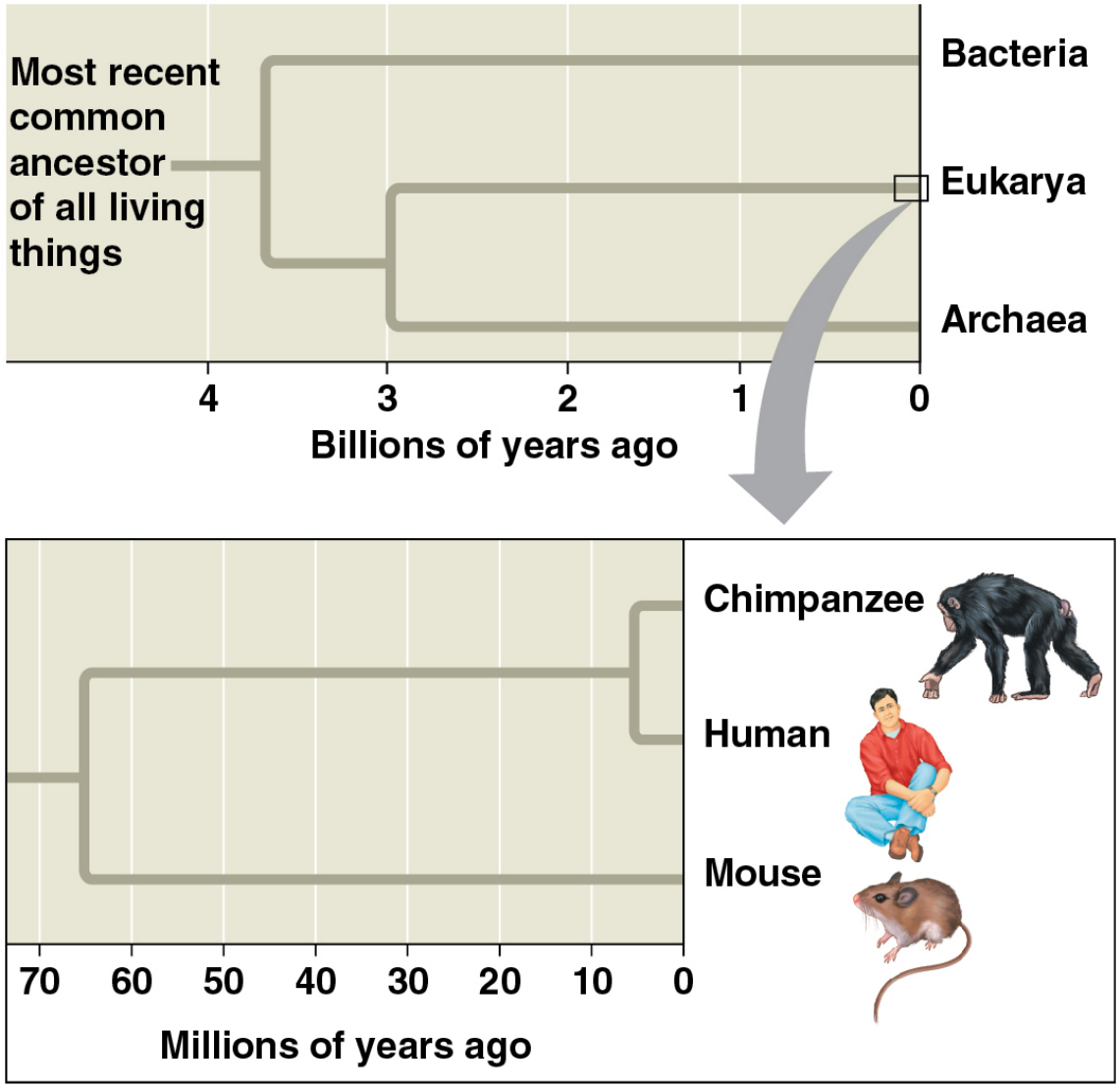
CONCEPT 21.6: Comparing genome sequences provides clues to evolution and development

- Comparisons of genome sequences from different species reveal much about the evolutionary history of life
- Comparative studies of genetic programs that affect embryonic development are beginning to clarify the mechanisms that generated the diversity of life-forms present today

Comparing Genomes

- Genome comparisons of closely related species help shed light on recent evolutionary events
- Comparing genomes of very distantly related species helps us understand ancient evolutionary history
- Relationships among species can be represented by a tree-shaped diagram

Figure 21.17


Most recent common ancestor of all living things
Bacteria
Eukarya
Archaea
4
3
2
1
0
Billions of years ago
Chimpanzee
Human
Mouse
70
60
50
40
30
20
10
0
Millions of years ago

Comparing Distantly Related Species

- Highly conserved genes have changed very little over time
- These help clarify relationships among species that diverged from each other long ago
- Bacteria, archaea, and eukaryotes diverged from each other between 2 and 4 billion years ago
- Very ancient genes can still be surprisingly similar in disparate species

Comparing Closely Related Species

- Genomes of closely related species are likely to be organized similarly
- For example, comparison of the human genome with that of other mammals gives us clues about what it takes to make a mammal
- Analysis of the human and chimpanzee genomes reveals some general differences that underlie the differences between the two organisms

- Human and chimpanzee genomes differ by 1.2% at single base pairs and by 2.7% because of insertions and deletions
- Sequencing of the bonobo genome in 2012 revealed that in some regions there is greater similarity between human and bonobo or chimpanzee sequences than between chimpanzee and bonobo
- We don’t know how the genetic differences revealed by genome sequencing account for the distinct characteristics of each species

- A number of genes are apparently evolving faster in the human than in the chimpanzee or mouse
- Among them are genes involved in defense against malaria and tuberculosis and one that regulates brain size
- The *FOXP2* gene shows evidence of rapid change in the human lineage compared to other primates
- The gene may be related to human speech; there are two amino acids found only in the human protein sequence

- In 2014, a high-quality sequence of the Neanderthal (*Homo neanderthalensis*) genome was achieved
- Shortly after that, the DNA of the Denisovan member of the species *Homo* was also sequenced
- DNA sequences from both species encode the same two amino acids (in the *FOXP2* gene) so these are not human-specific
- Further research shows no evidence for selection for these two amino acids in the human lineage during the time frame relevant to language acquisition

- However, several lines of evidence suggest that *FOXP2* regulates genes that function in vocalization in vertebrates
- Mutation in this gene leads to severe speech and language impairment in humans
- It is expressed in the brains of zebra finches and canaries at the time when they are learning their songs
- Mice with *FOXP2* knocked out have malformed brains and fail to emit normal ultrasonic vocalizations

- When *FOXP2* knockout mice have the human form of the gene introduced, the mice are generally healthy, and can vocalize
- However, their vocalizations are subtly different from normal mice
- The *FOXP2* story is an excellent example of how different approaches can complement each other in leading to understanding of biological phenomena

Figure 21.18

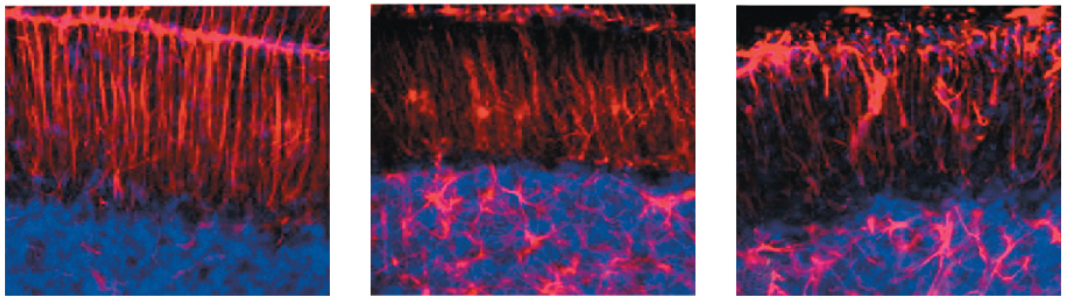
Wild type: two normal copies of *FOXP2*

Heterozygote: one copy of *FOXP2* disrupted

Homozygote: both copies of *FOXP2* disrupted

Experiment 1: Researchers cut thin sections of brain and stained them with reagents that allow visualization of brain anatomy in a UV fluorescence microscope.

Results

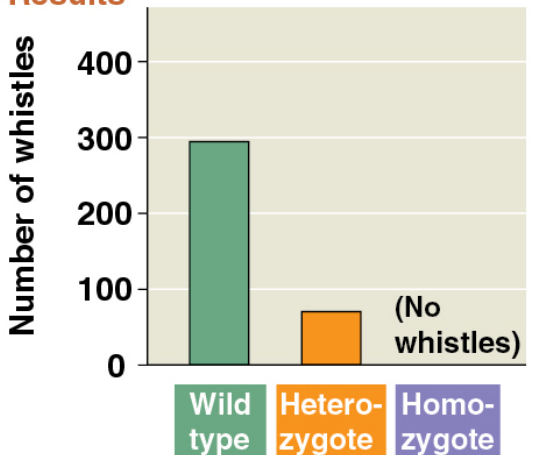


Wild type | Heterozygote | Homozygote

Data from W. Shu et al., Altered ultrasonic vocalization in mice with a disruption in the *Foxp2* gene, *Proceedings of the National Academy of Sciences USA* 102:9643–9648 (2005).

Experiment 2: To induce stress, researchers separated each newborn mouse pup from its mother and recorded the number of ultrasonic whistles produced by the pup.

Results



Comparing Genomes Within a Species

- As a species, humans have only existed for about 200,000 years and have low within-species genetic variation
- Variation within humans is due to single nucleotide polymorphisms (SNPs), inversions, deletions, and duplications
- Several million SNPs have been identified in the human genome, stored in databases around the world

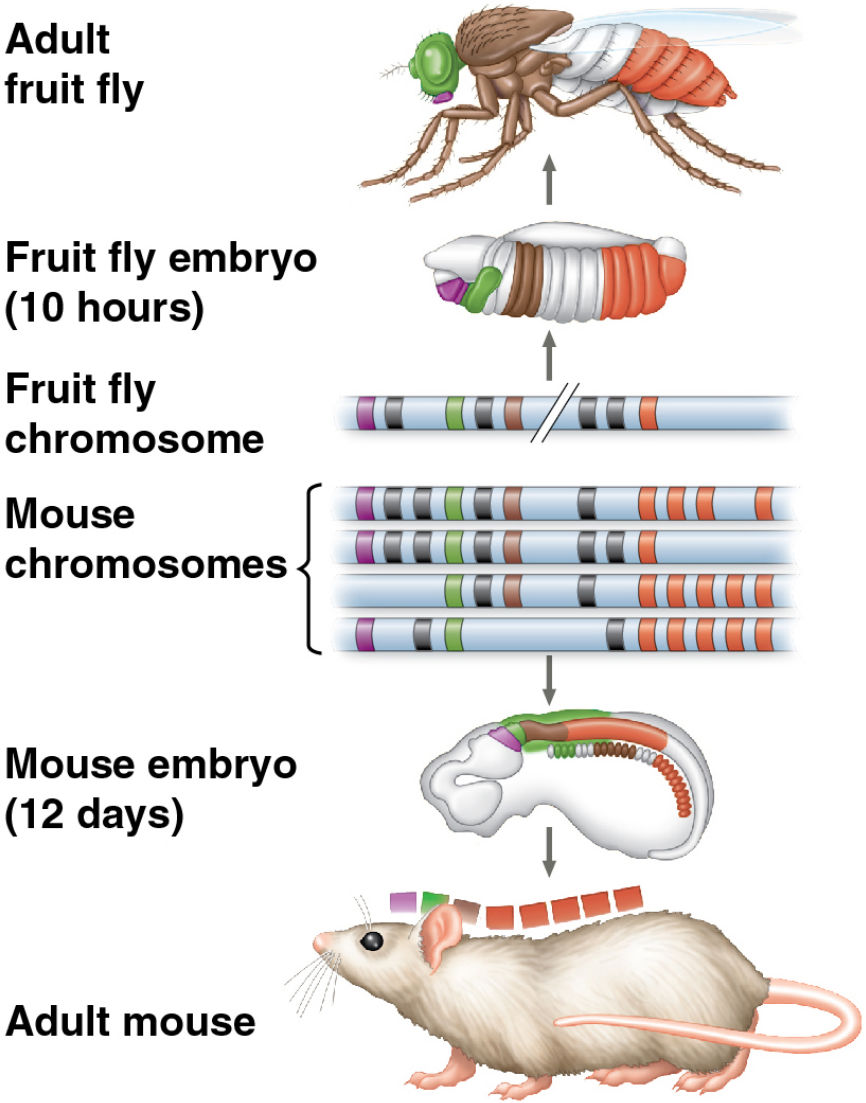
- Most surprising is the large number of copy-number variants
- These variations are likely to play a role in complex diseases and disorders
- Copy-number variants, SNPs, and other polymorphisms are useful genetic markers for studying human evolution
- African genomes have higher genetic diversity than other genomes, suggesting these populations have been evolving longer than non-African populations

Widespread Conservation of Developmental Genes Among Animals

- Evolutionary developmental biology, or **evo-devo**, compares developmental processes of different multicellular organisms
- Genomic information shows that minor differences in gene sequence or regulation can result in striking differences in form

- Homeotic genes in *Drosophila melanogaster* encode genes that specify identity of body segments of the fly
- All of these genes contain a 180-nucleotide sequence called a **homeobox**
- An identical or very similar nucleotide sequence has been discovered in the homeotic genes of both vertebrates and invertebrates

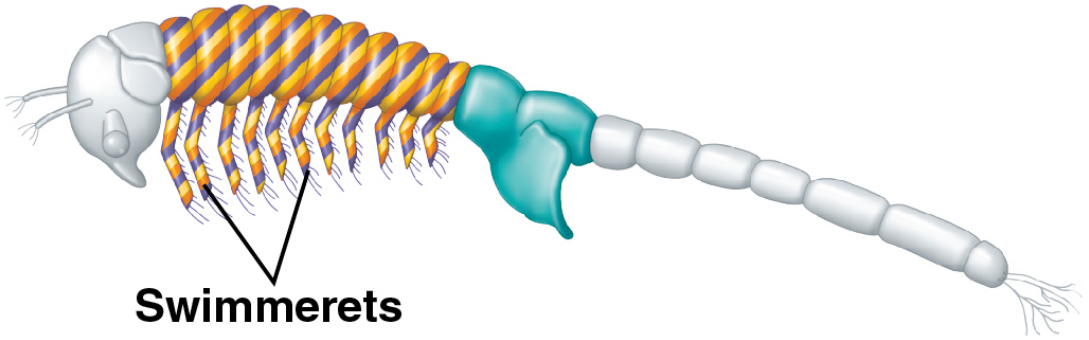
Figure 21.19


Adult
fruit fly
Fruit fly embryo
(10 hours)
Fruit fly
chromosome
Mouse
chromosomes
Mouse embryo
(12 days)
Adult mouse

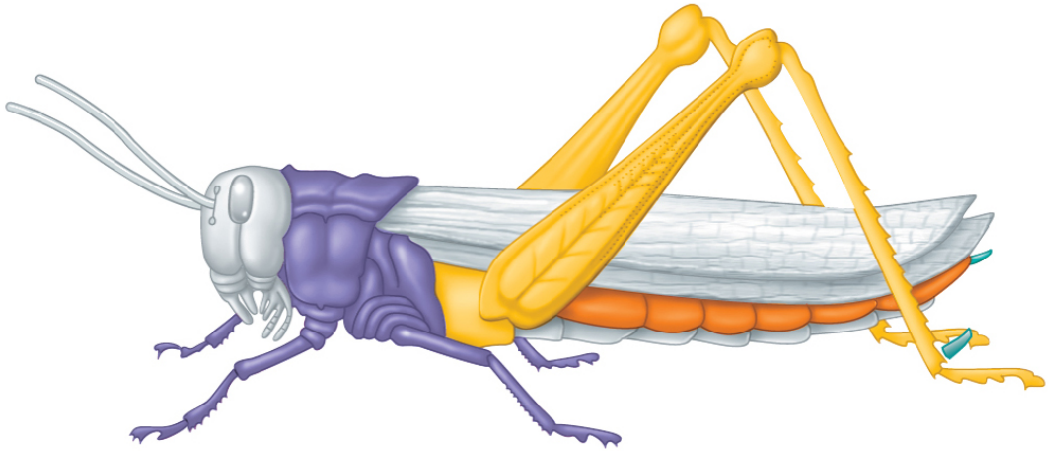
- Related homeobox sequences have been found in regulatory genes of yeasts and plants
- Homeotic genes in animals are called *Hox* genes
- The homeodomain is the part of the protein that binds to the DNA, where the protein functions as a transcription factor
- In addition to homeotic genes, many other developmental genes are highly conserved from species to species

- Sometimes small changes in regulatory sequences of certain genes lead to major changes in body form
- For example, variation in *Hox* gene expression controls variation in leg-bearing segments of crustaceans and insects
- In other cases, genes with conserved sequences play different roles in different species

Figure 21.20

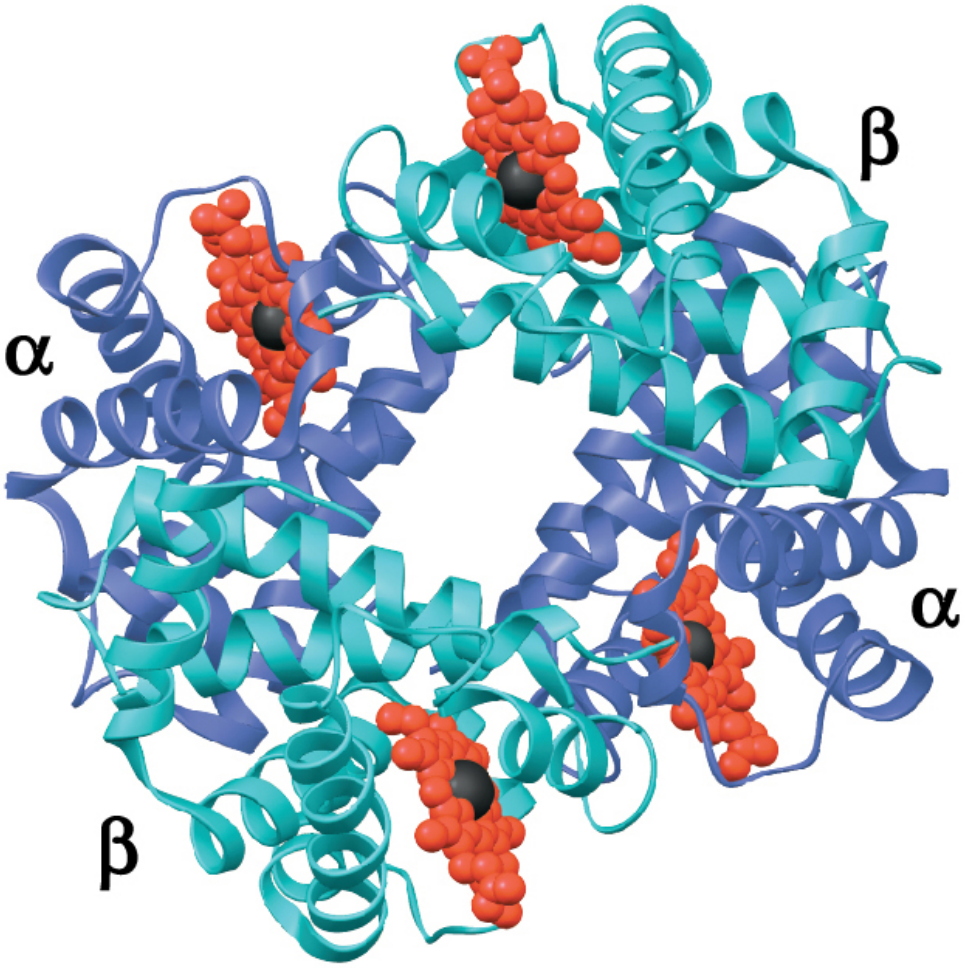


(a) Expression of four *Hox* genes in the brine shrimp *Artemia*



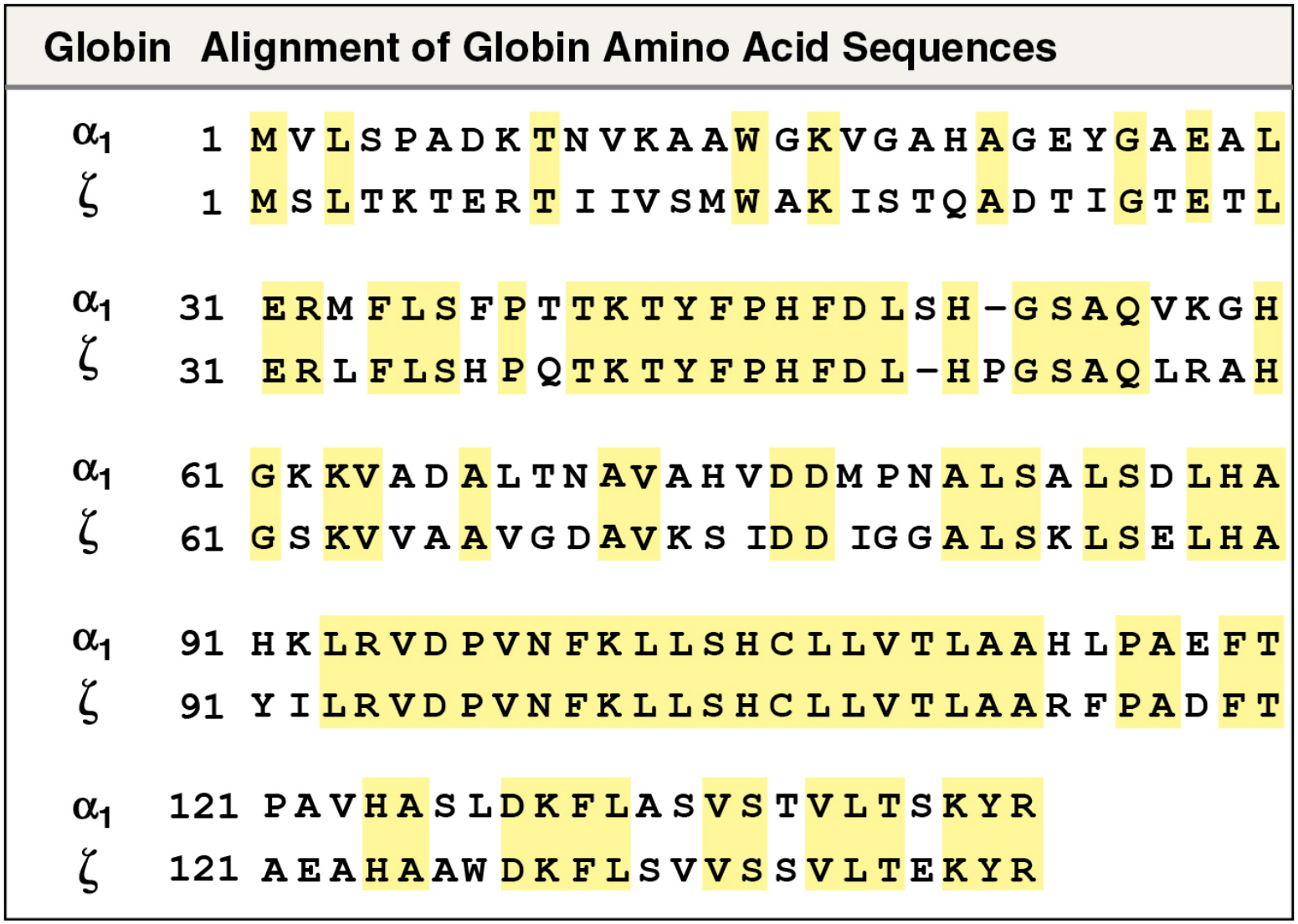
(b) Expression of the grasshopper versions of the same four *Hox* genes

Figure 21.UN01



Hemoglobin

Figure 21.UN01a



Further Reading R. C. Hardison, Globin genes on the move, *Journal of Biology* 7:35.1–35.5 (2008).

Figure 21.UN01b

Amino Acid Identity Table									
		α Family			β Family				
		α_1 (alpha 1)	α_2 (alpha 2)	ζ (zeta)	β (beta)	δ (delta)	ϵ (epsilon)	A_γ (gamma A)	G_γ (gamma G)
α Family	α_1	-----	100	60	45	44	39	42	42
	α_2		-----	60	45	44	39	42	42
	ζ			-----	38	40	41	41	41
β Family	β				-----	93	76	73	73
	δ					-----	73	71	72
	ϵ						-----	80	80
	A_γ							-----	99
	G_γ								-----

Compiled using data from the National Center for Biotechnology Information (NCBI).
Further Reading R. C. Hardison, Globin genes on the move, *Journal of Biology* 7:35.1–35.5 (2008).

Figure 21.UN02

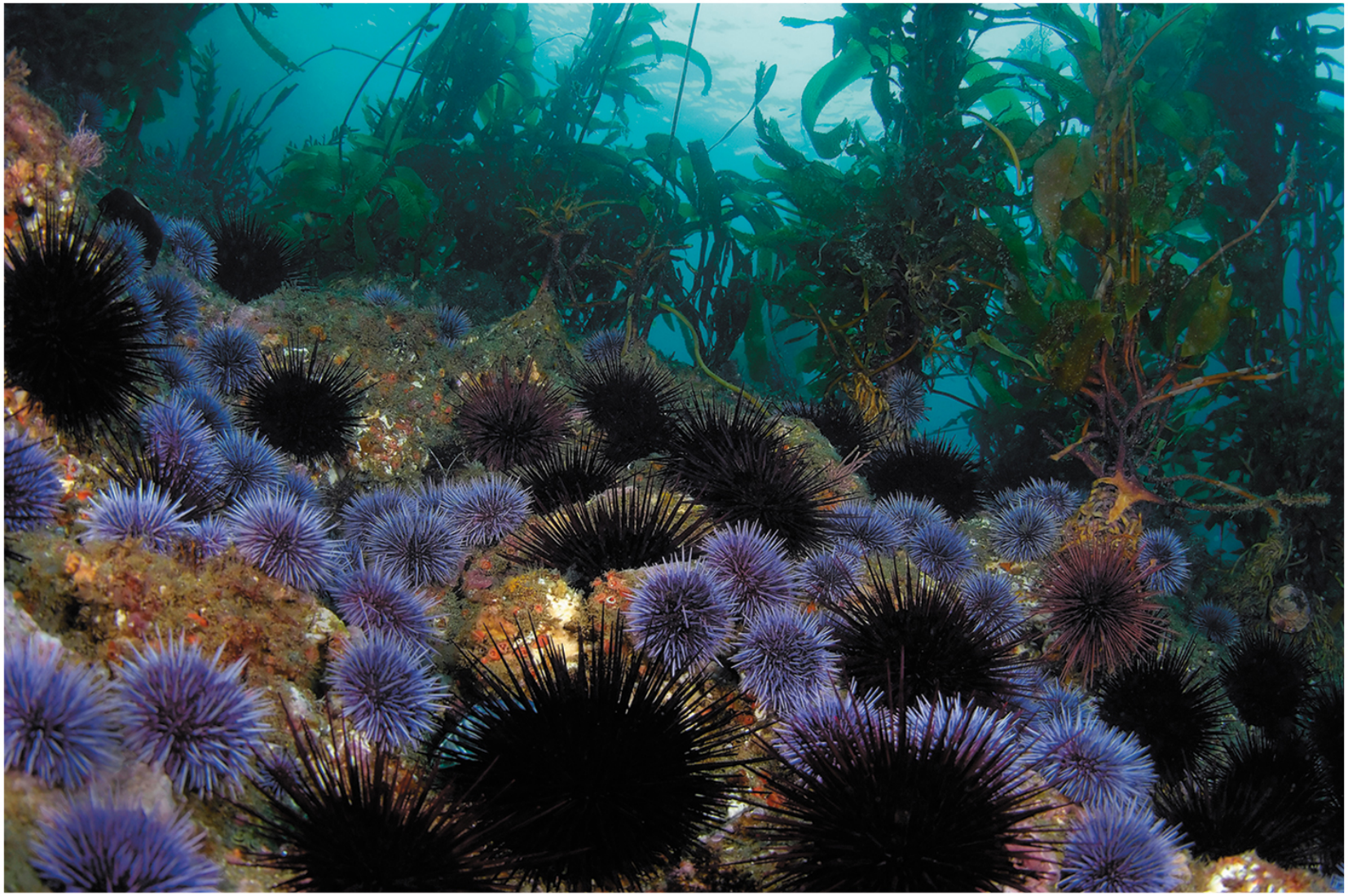


Figure 21.UN04

	Bacteria	Archaea	Eukarya
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		Most are 5,000–45,000
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	Present in most genes of multicellular eukaryotes, but only in some genes of unicellular eukaryotes
Other noncoding DNA	Very little		Can exist in large amounts; generally more repetitive noncoding DNA in multicellular eukaryotes

Figure 21.UN05

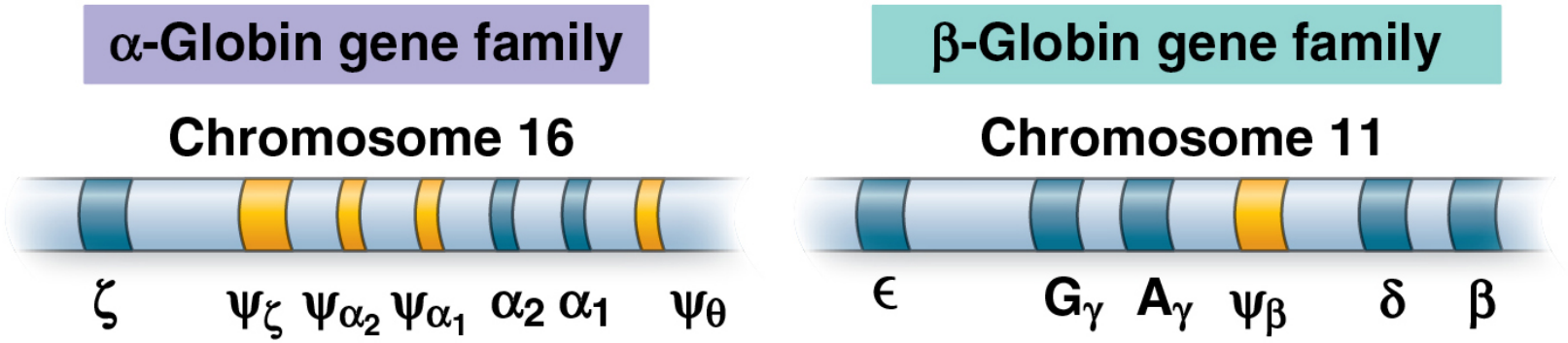


Figure 21.UN06

Figure 21.UN07