21.6 Comparative Genomics Analyzes and Compares Genomes from Different Organisms

- **Comparative genomics** compares the genomes of different organisms in order to answer questions about genetics and other aspects of biology

- Many prokaryotic genomes already sequenced are from organisms causing human diseases such as cholera, tuberculosis, and leprosy

- Bacteria have a single, circular chromosome with substantial variation in chromosome organization and number among species

- Gene density is very high in prokaryotes, and this close packing of DNA demonstrates that a very high proportion of the DNA serves as coding DNA

- Bacterial DNA contains operons

Genome Size and Gene Number in Selected Prokaryotes

	Genome Size (Mb)	Number of Genes
Archaea		
Methanosarcina berkeri	4.84	3680
Archaeoglobus fulgidis	2.17	2437
Methanococcus jannaschii	1.66	1783
Nanoarchaeum equitans	0.49	552
Thermoplasma acidophilium	1.56	1509
Eubacteria		
Pseudomonas aeruginosa	6.30	5570
Rhizobium radiobacter	4.67	5419
Escherichia coli	4.64	4289
Bacillus subtilis	4.21	4779
Haemophilus influenzae	1.83	1738
Aquifex aeolicus	1.55	1749
Rickettsia prowazekii	1.11	834
Mycoplasma pneumonia	0.82	680
Mycoplasma genitalium	0.58	483

Organism (Scientific Name)	Approximate Size of Genome (in million [megabase, Mb] or billion [gigabase, Gb] bases) (Date Completed)	Number of Genes	Appro Genes
Bacterium (*Escherichia coli*)	4.1 Mb (1997)	4403	not de
Chicken (*Gallus gallus*)	1 Gb (2004)	~20,000–23,000	60%
Dog (*Canis familiaris*)	2.5 Gb (2003)	~18,400	75%
Chimpanzee (*Pan troglodytes*)	~3 Gb (2005)	~20,000–24,000	98%
Fruit fly (*Drosophila melanogaster*)	165 Mb (2000)	~13,600	50%
Human (*Homo sapiens*)	3.1 Gb (2004)	~20,000	100%
Mouse (*Mus musculus*)	~2.5 Gb (2002)	~30,000	80%
Rat (*Rattus norvegicus*)	~2.75 Gb (2004)	~22,000	80%
Rhesus macaque (*Macaca mulatta*)	2.87 Gb (2007)	~20,000	93%
Rice (*Oryza sativa*)	389 Mb (2005)	~41,000	not de
Roundworm (*Caenorhabditis elegans*)	97 Mb (1998)	19,099	40%
Sea urchin (*Strongylocentrotus purpuratus*)	814 Mb (2006)	~23,500	60%
Thale cress (plant) (*Arabidopsis thaliana*)	140 Mb (2000)	~27,500	not de
Yeast (*Saccharomyces cerevisiae*)	12 Mb (1996)	~5700	30%

- The basic features of eukaryotic genomes are similar in different species, although genome size in eukaryotes is highly variable

- Eukaryotic genomes have several features not found in prokaryotes

  - Gene density: Varies from chromosome to chromosome
  - Introns: Variation in genomes and in genes
  - Repetitive sequences

- Complete sequences of various organisms show that the number of genes humans share with other species is very high, ranging from about 30% of the genes in yeast to ~80% in mice and ~98% in chimpanzees

- Many mutated genes involved in human diseases are also present in model organisms

- The dog has been used as a model organism and its genome has been sequenced

- Humans share 75 percent of their genes with dogs as well as many genetic disorders
  - Over 400 single-gene disorders
  - Sex-chromosome aneuploidies
  - Multifactorial diseases (e.g., epilepsy)
  - Behavioral conditions (e.g., obsessive compulsive disorder)

- Human and chimpanzee sequences differ by less than 2 percent and share 96 percent of the same genes

- Analysis of human and chimp genome indicates that genome evolution, speciation, and gene expression are interconnected

	Human 21	Chimpanzee 22
Size (bp)	33,127,944	32,799,845
%G + C Content	40.94	41.01
CpG Islands	950	885
SINEs (*Alu* elements)	15,137	15,048
Genes	284	272
Pseudogenes	98	89

- The Rhesus macaque monkey (*Macaca mulatta*) is one of the most important model organisms in biomedical research

- It has been central in our understanding of cardiovascular disease, aging, diabetes, cancer, depression, osteoporosis, and many other aspects of human health

- Sea urchins (*Strongylocentrotus purpuratus*) are shallow-water marine invertebrates that have served as important model organisms

- Sequence alignment and homology searches demonstrate that the sea urchin contains many genes with important functions in humans

- A rough draft of the Neanderthal (*Homo neanderthalensis*) genome compasses more than 3 billion bp

- A comparative genomic analysis will help identify areas in the genome where humans have undergone rapid evolution since diverging from Neanderthals

- Genomic studies suggest that interbreeding took place between Neanderthals and modern humans an estimated 45,000 to 80,000 years ago

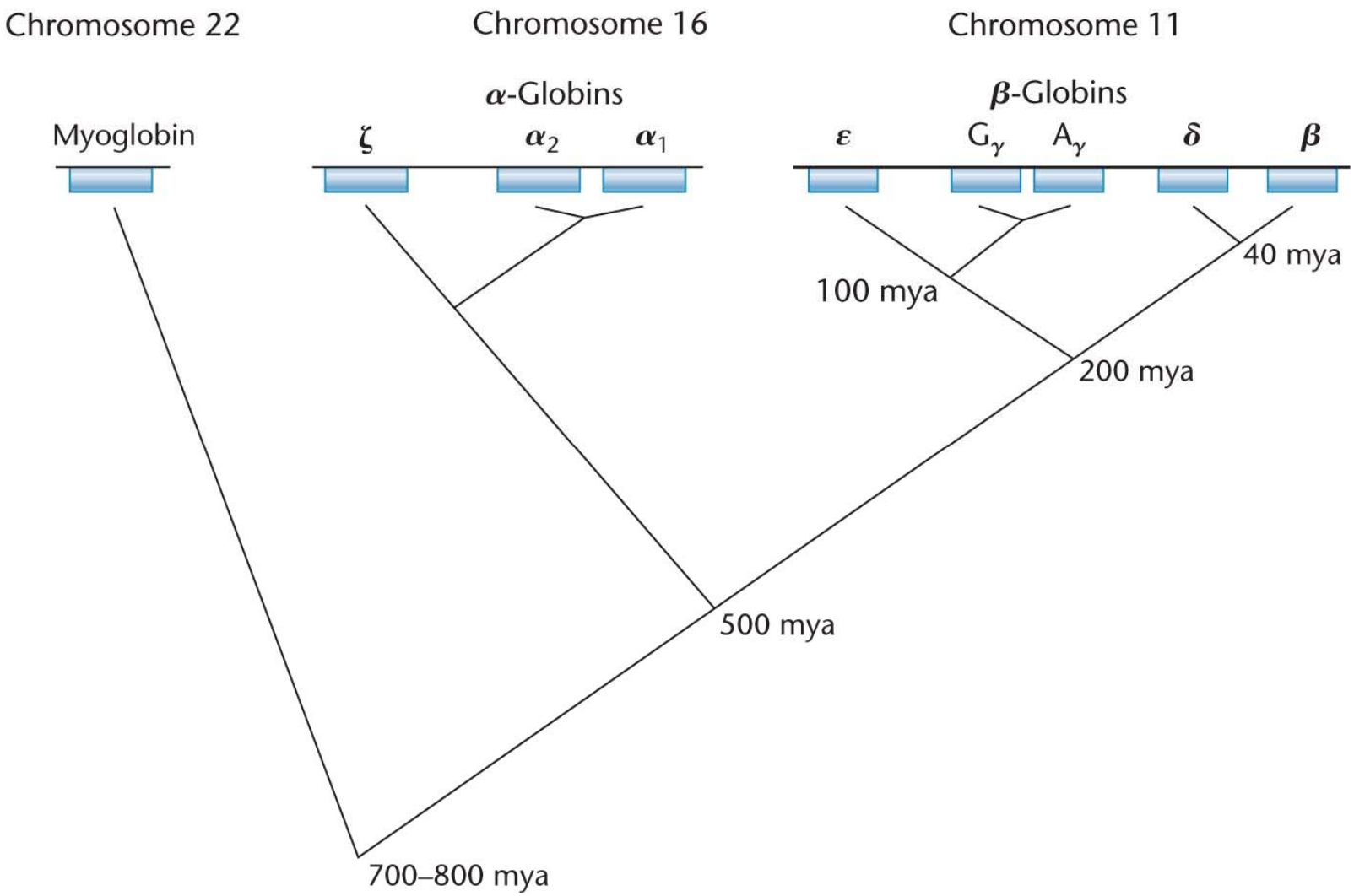
- The genome of non-African *H. sapiens* contains approximately 1–4 percent of sequence inherited from Neanderthals

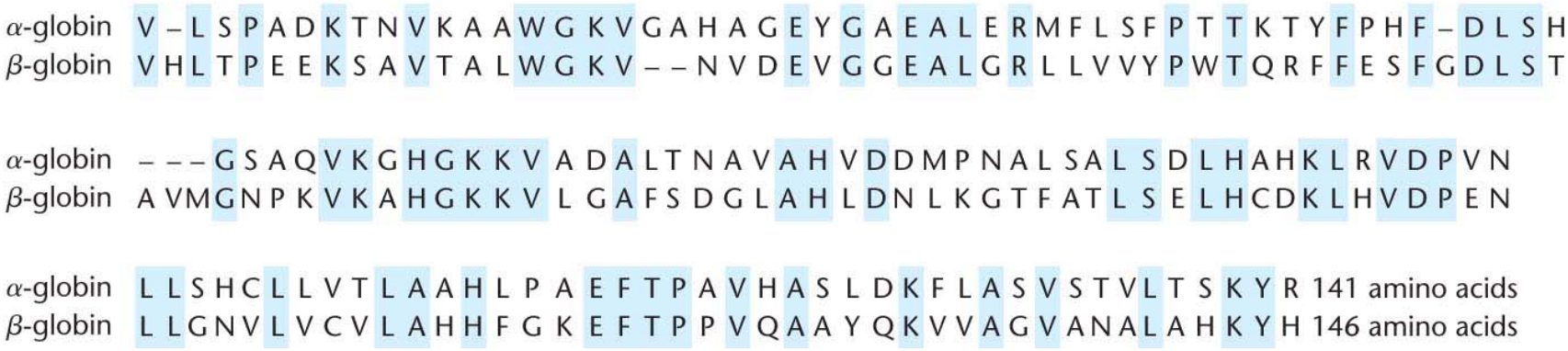
- The ancestors of modern humans encountered Neanderthals tens of thousands of years ago, soon after they migrated out of Africa. That would explain why modern people of African descent have little to no Neanderthal DNA.

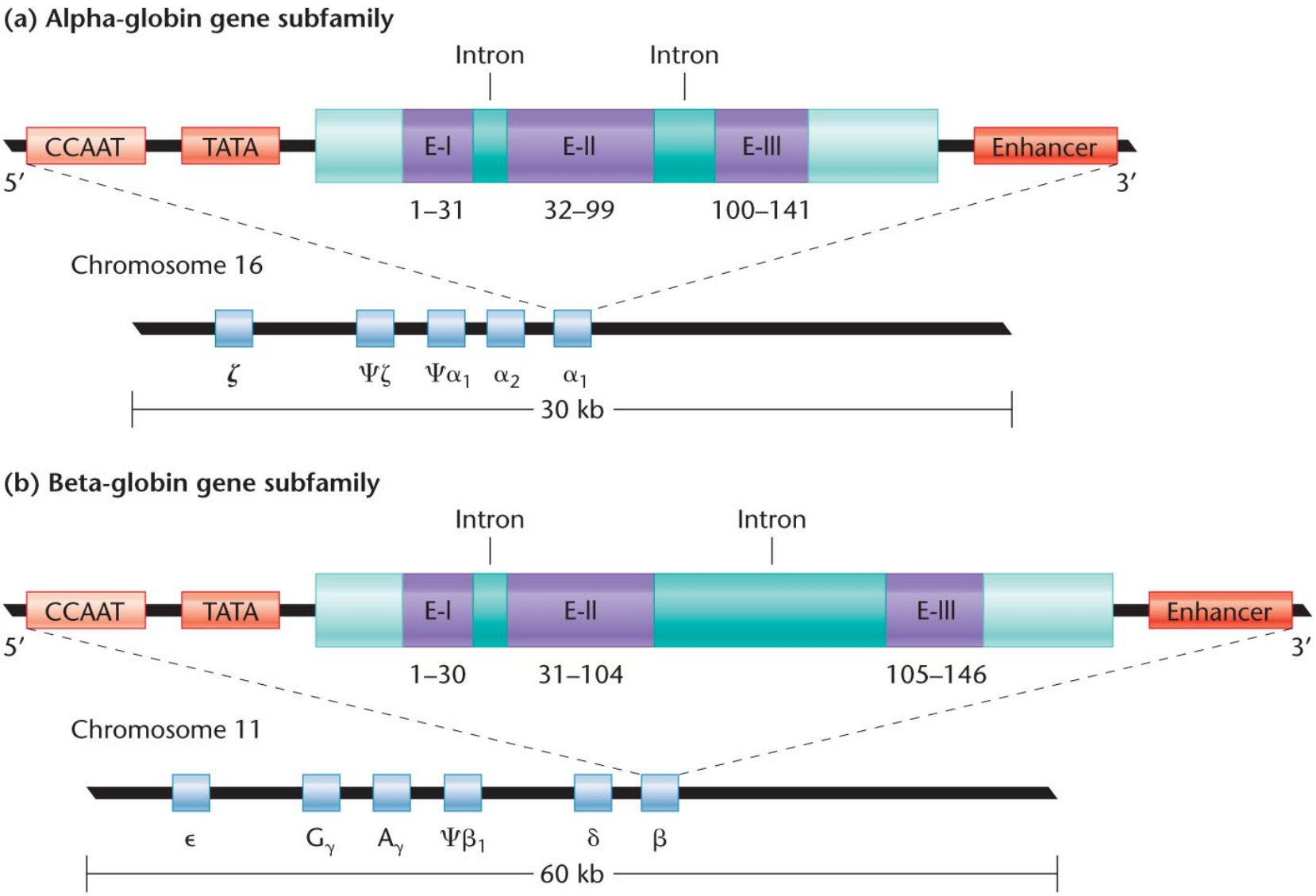
- Comparative genomics has been valuable in identifying members of **multigene families** including nonfunctional **pseudogenes**

- A group of related multigene families is called a **superfamily**

- One of the best-studied examples of gene family evolution is the **globin family gene**


Chromosome 22
Chromosome 16
Chromosome 11
α-Globins
β-Globins
Myoglobin
ζ
α2
α1
ε
Gγ
Aγ
δ
β
100 mya
40 mya
200 mya
500 mya
700–800 mya


α-globin V – L S P A D K T N V K A A W G K V G A H A G E Y G A E A L E R M F L S F P T T K T Y F P H F – D L S H
β-globin V H L T P E E K S A V T A L W G K V – – N V D E V G G E A L G R L L V V Y P W T Q R F F E S F G D L S T
α-globin – – – G S A Q V K G H G K K V A D A L T N A V A H V D D M P N A L S A L S D L H A H K L R V D P V N
β-globin A V M G N P K V K A H G K K V L G A F S D G L A H L D N L K G T F A T L S E L H C D K L H V D P E N
α-globin L L S H C L L V T L A A H L P A E F T P A V H A S L D K F L A S V S T V L T S K Y R 141 amino acids
β-globin L L G N V L V C V L A H H F G K E F T P P V Q A A Y Q K V V A G V A N A L A H K Y H 146 amino acids


(a) Alpha-globin gene subfamily
Intron
Intron
CCAAT
TATA
E-I
E-II
E-III
Enhancer
5′
3′
1–31
32–99
100–141
Chromosome 16
ζ
Ψζ
Ψα1
α2
α1
30 kb
(b) Beta-globin gene subfamily
Intron
Intron
CCAAT
TATA
E-I
E-II
E-III
Enhancer
5′
3′
1–30
31–104
105–146
Chromosome 11
ϵ
Gγ
Aγ
Ψβ1
δ
β
60 kb

21.7 Metagenomics Applies Genomics Techniques to Environmental Samples

- **Metagenomics**, also called **environmental genomics**, is the use of whole-genome shotgun approaches to sequence genomes from entire communities of microbes in environmental samples of water, air, and soil

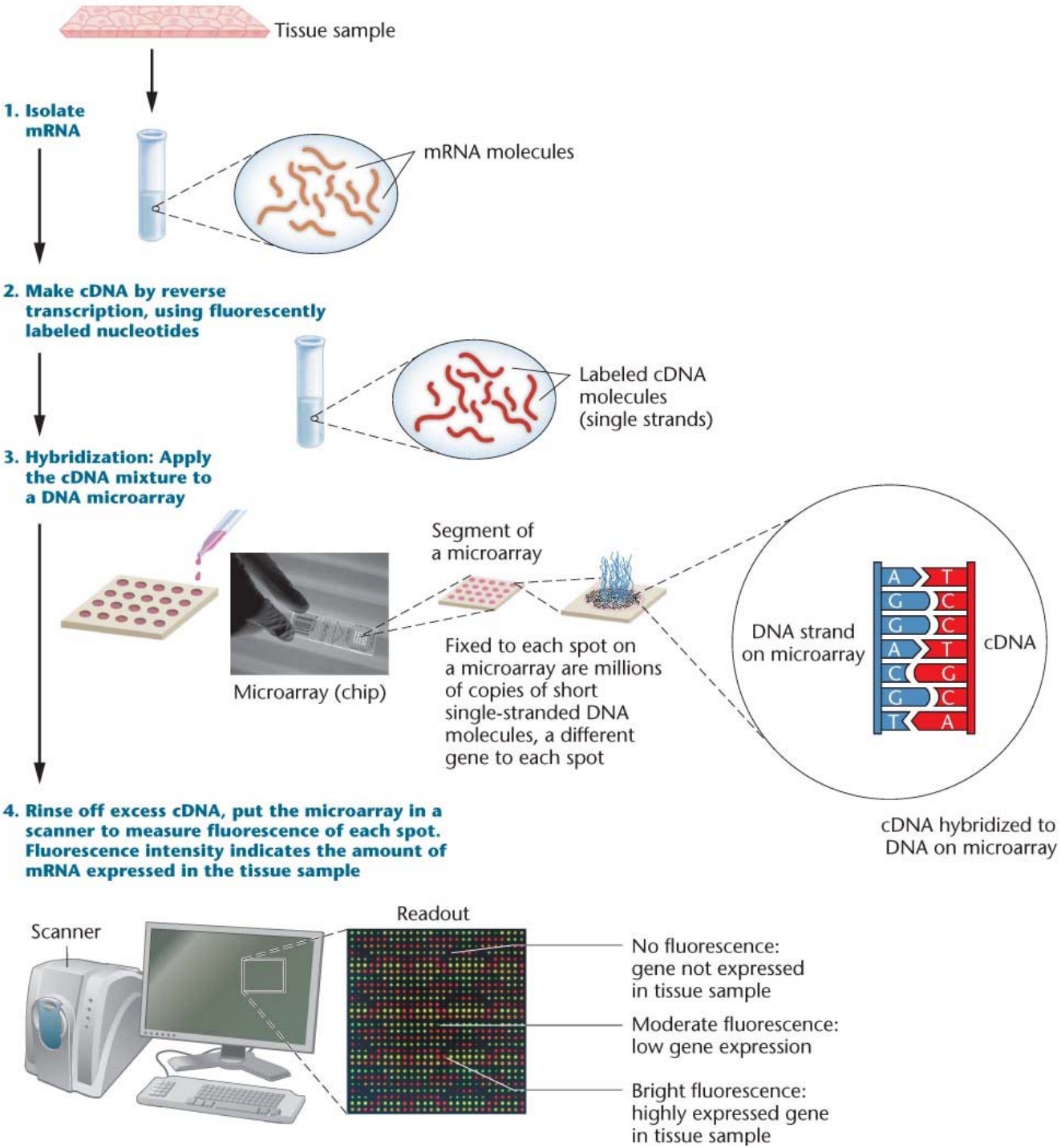
- The general method for metagenomics is to sequence genomes for all microbes in a given environment
  - This will teach us more about millions of species of bacteria as well as viruses, particularly bacteriophages
  - Has a great potential for identifying genes with novel functions, some of which may have valuable applications in medicine and biotechnology

21.8 Transcriptome Analysis Reveals Profiles of Expressed Genes in Cells and Tissues

- **Transcriptome analysis** or **global analysis of gene expression** studies the expression of genes by a genome

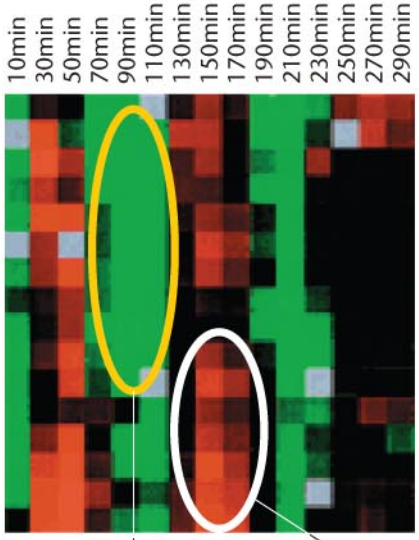
  - both qualitatively by identifying which genes are expressed and which are not
  - quantitatively by measuring the varying levels of expression of different genes

- **Microarray analysis** enables researchers to analyze all of a sample's expressed genes simultaneously
  - Microarrays (**gene chips)**, consist of glass microscope slides onto which single-stranded DNA molecules are attached using a high-speed robotic arm called an arrayer
  - A single microarray can have over 20,000 different spots of DNA with entire genomes available on microarrays
- Computerized microarray data analysis programs are essential for organizing gene-expression profile data from microarrays
  - Cluster algorithm programs


Tissue sample
1. Isolate mRNA
mRNA molecules
2. Make cDNA by reverse transcription, using fluorescently labeled nucleotides
Labeled cDNA molecules (single strands)
3. Hybridization: Apply the cDNA mixture to a DNA microarray
Segment of a microarray
Microarray (chip)
Fixed to each spot on a microarray are millions of copies of short single-stranded DNA molecules, a different gene to each spot
DNA strand on microarray
cDNA
cDNA hybridized to DNA on microarray
4. Rinse off excess cDNA, put the microarray in a scanner to measure fluorescence of each spot. Fluorescence intensity indicates the amount of mRNA expressed in the tissue sample
Scanner
Readout
No fluorescence: gene not expressed in tissue sample
Moderate fluorescence: low gene expression
Bright fluorescence: highly expressed gene in tissue sample

(a)

Experiments 1–15



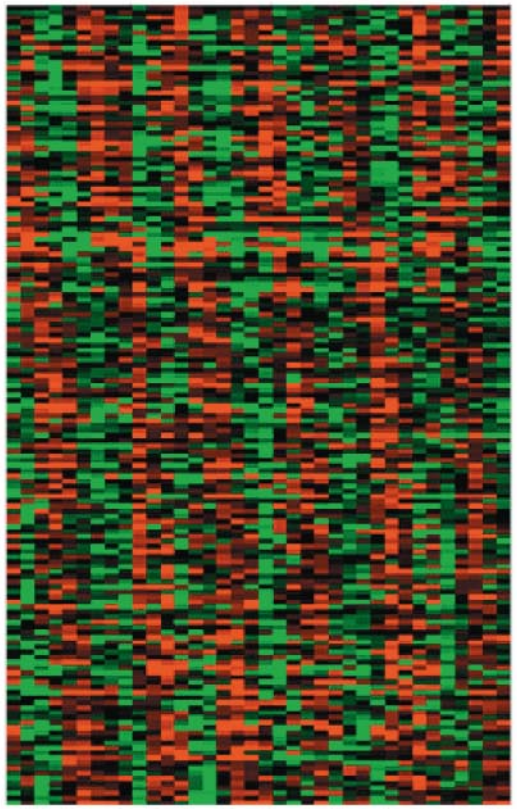
Genes

	unknown
SCW11	cell wall biogenesis glucanase (putative)
CLB6	cell cycle B-type cyclin; S phase
KCC4	bud growth protein kinase
MCD1	mitosis, sister chromatid cohesion unknown
POL30	DNA replication DNA polymerase processivity factor
	unknown
RFA1	DNA replication replication factor A, 69 kD subunit
	unknown
STP3	tRNA splicing
TOS6	unknown; similar to Mid2p
ISR1	staurosporine resistance protein kinase
SRO4	bud site selection, plasma membrane protein
CSI2	cell wall biogenesis chitin synthase 3 subunit
SVS1	vanadate resistance
CLN2	cell cycle G1/S cyclin

Cluster of down-regulated genes

Cluster of up-regulated genes

(b)



Time (h) →

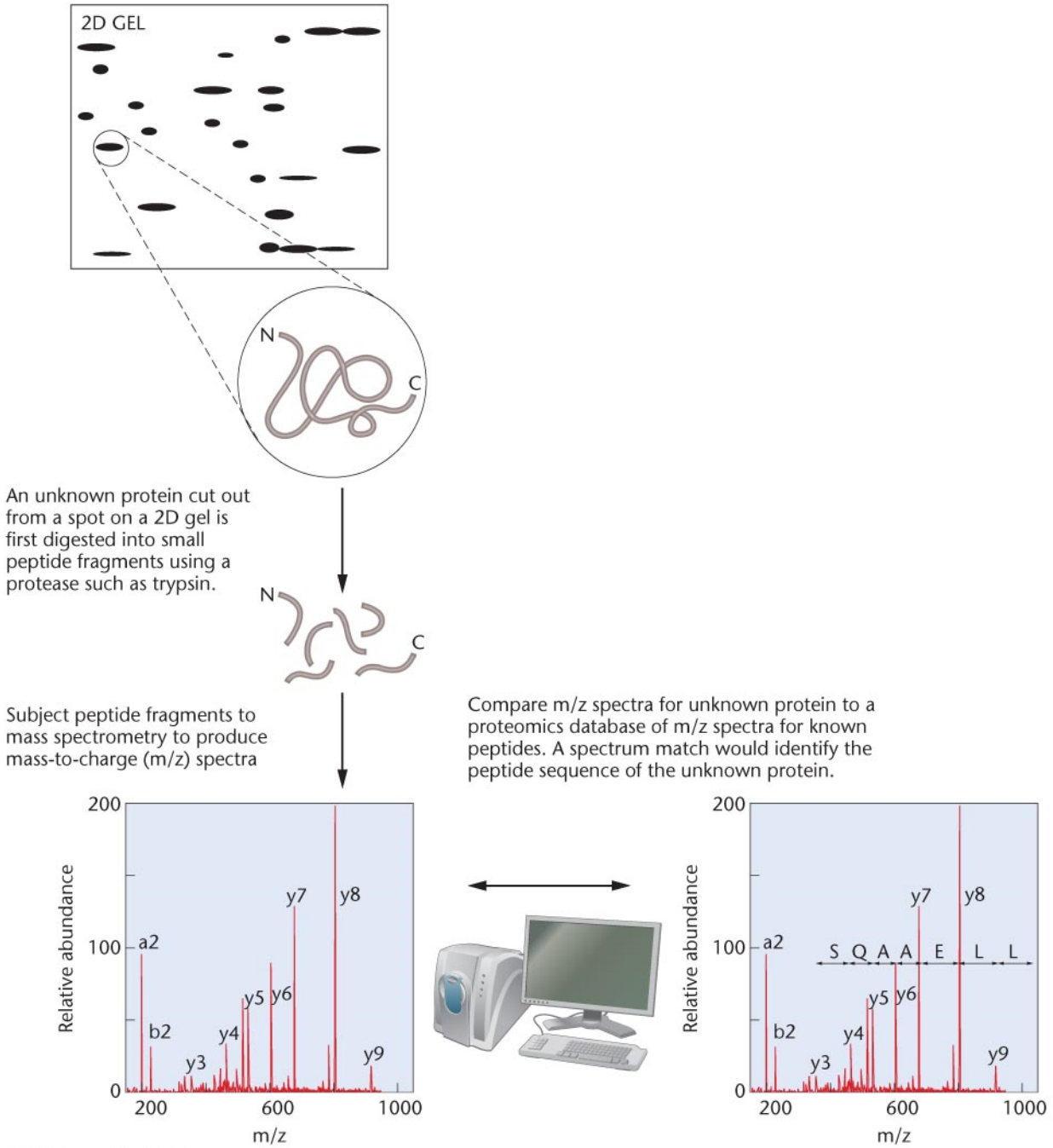
21.9 Proteomics Identifies and Analyzes the Protein Composition of Cells

- **Proteomics** is the identification and characterization and quantitative analysis of all proteins (**proteome**) encoded by the genome of a cell, tissue, or organism

  - Can be used to reconcile differences between the number of genes in a genome and the number of different proteins produced
  - Allows comparison of proteins in normal and diseased tissue

- The **Protein Structure Initiative** (**PSI**) is a 10-year project designed to analyze the structures of more than 4000 protein families

- **Mass spectrometry** (**MS**) techniques analyze ionized samples in gaseous form and measure the **mass-to-charge (m/z)** ratio of different ions in a sample
  - **Matrix-assisted laser desorption ionization (MALDI)** is used for proteomic analysis of tissue samples treated under different conditions


2D GEL
N
C
An unknown protein cut out from a spot on a 2D gel is first digested into small peptide fragments using a protease such as trypsin.
N
C
Subject peptide fragments to mass spectrometry to produce mass-to-charge (m/z) spectra
Compare m/z spectra for unknown protein to a proteomics database of m/z spectra for known peptides. A spectrum match would identify the peptide sequence of the unknown protein.
Relative abundance
200
100
0
a2
b2
y3
y4
y5
y6
y7
y8
y9
200
600
1000
m/z
S Q A A E L L

- Mass spectrometry analysis of bone tissue from a *T. rex* skeleton estimated at over 68 million years old demonstrated fossilization does not fully destroy all protein in well-preserved fossils

- Similar results were obtains from 160,000- to 600,000-year-old mastodon (*Mammut americanum*) peptides that showed matches to collagen from extant species, including collagen isoforms from humans, chimps, dogs, cows, chicken, elephants, and mice

21.10 Systems Biology Is an Integrated Approach to Studying Interactions of All Components of an Organism's Cells

- **Systems biology** incorporates data from genomics, transcriptomics, proteomics, and other areas of biology to further elucidate components of interacting pathways and the interrelationships of molecules. This is referred to as the **interactome**

- A **network map** is a sketch showing the interacting proteins, genes, and other molecules
  - Helps scientists model intricate potential interactions of molecules involved in normal and disease processes

- Systems biology is becoming increasingly important in the drug discovery and development process

- Many databases are being developed to model interactomes for human disease, including breast cancer, diabetes, asthma, and cardiovascular disease

- Systems biology is also being used to create biofuels and design genetically modified organisms for cleaning up the environment