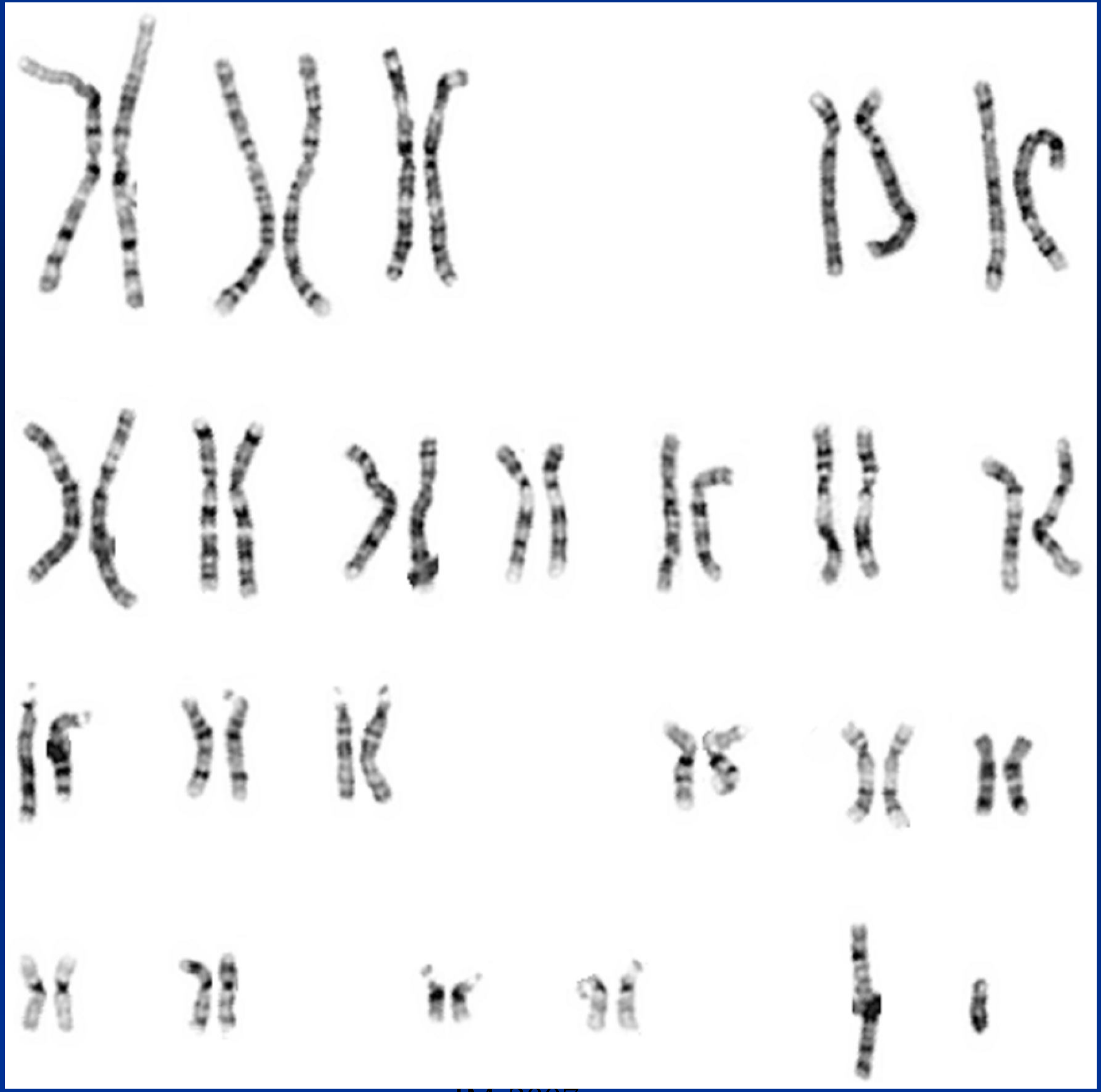
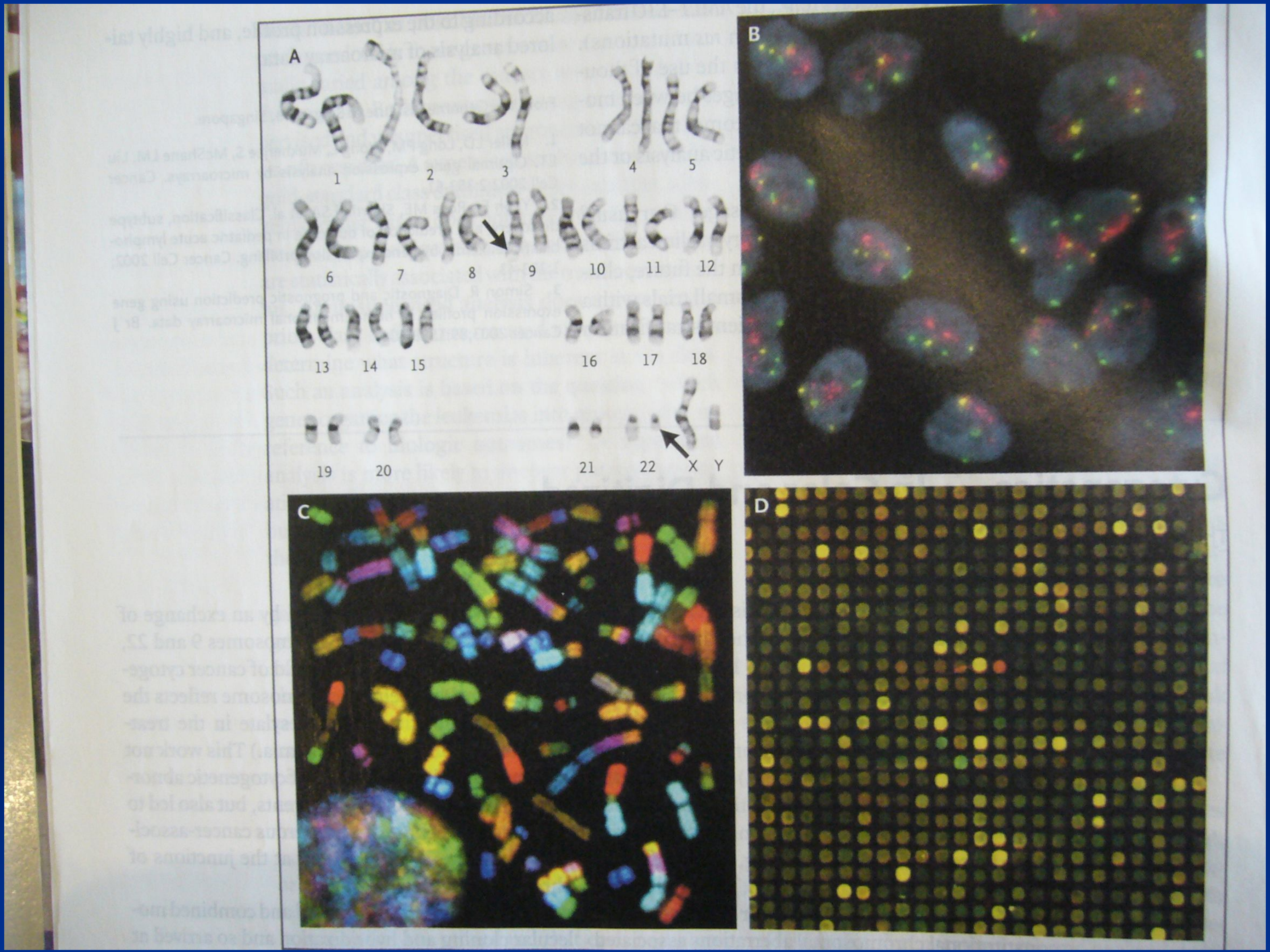
PROYECTO GENOMA HUMANO

Medicina Molecular 2009
Maestría en Biología Molecular Médica- UBA

CARIOTIPO NORMAL





N. Engl. J. Med 350, 1598, 2004 JM-2007

JM.2009

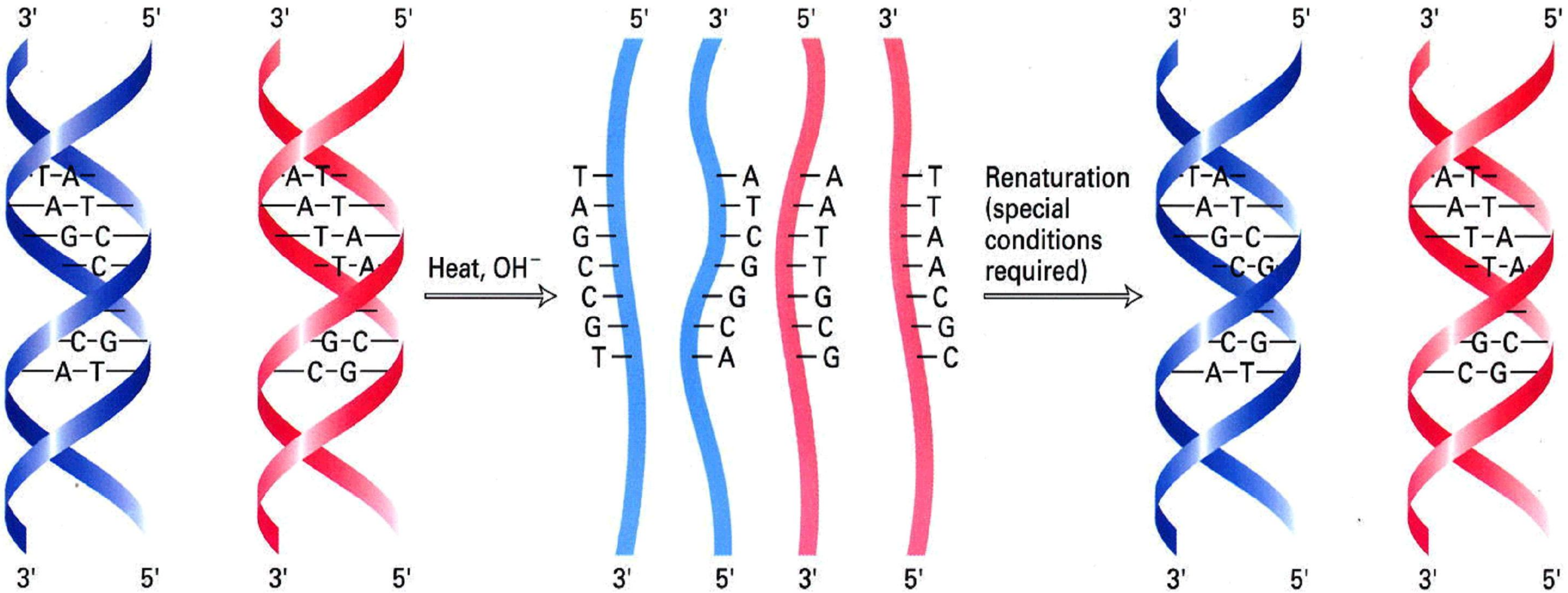
Some landmarks on the way to the genome sequence

- 1940s Recognition that DNA is the hereditary material
- 1953 Double-helix structure described
- 1966 Genetic code cracked
- 1972 Recombinant DNA technology developed
- 1975-77 DNA sequencing technology developed

DNA Base Pairing

A G C G A T C T G G
T C G C T A G A C C

Double helix consists of 2 complementary strands of DNA.

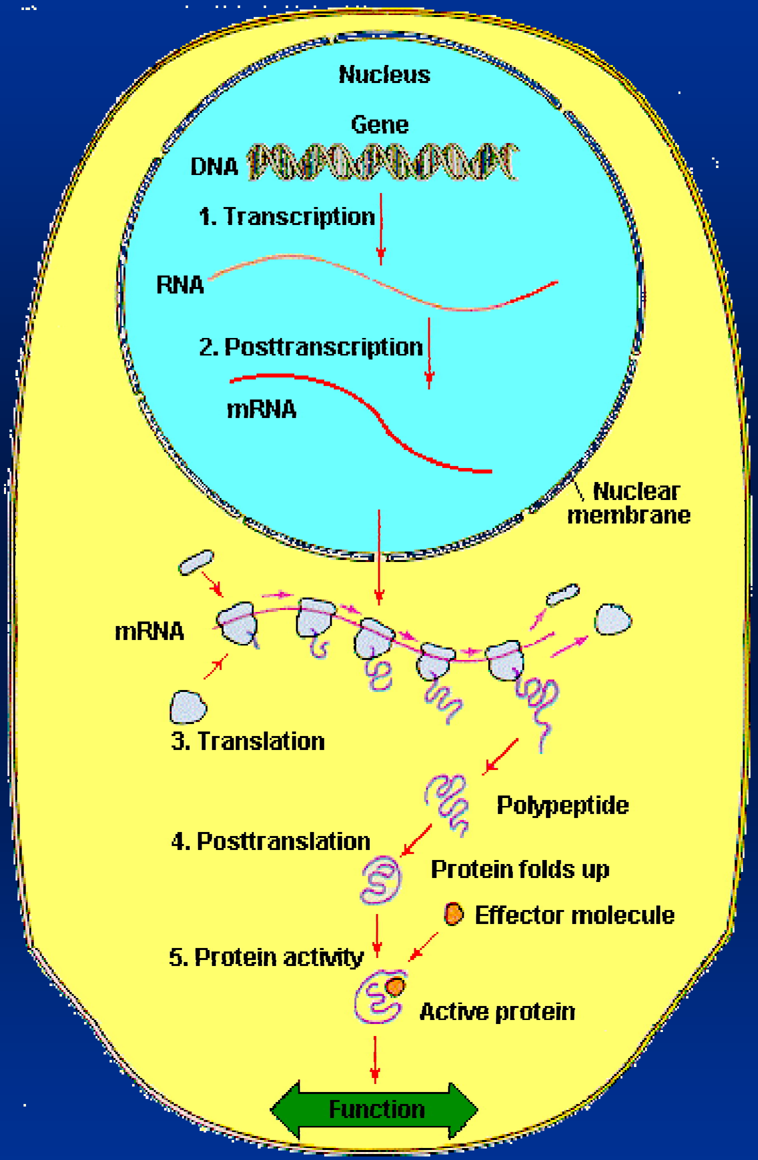


Native state

Single-stranded denatured state

Renatured state

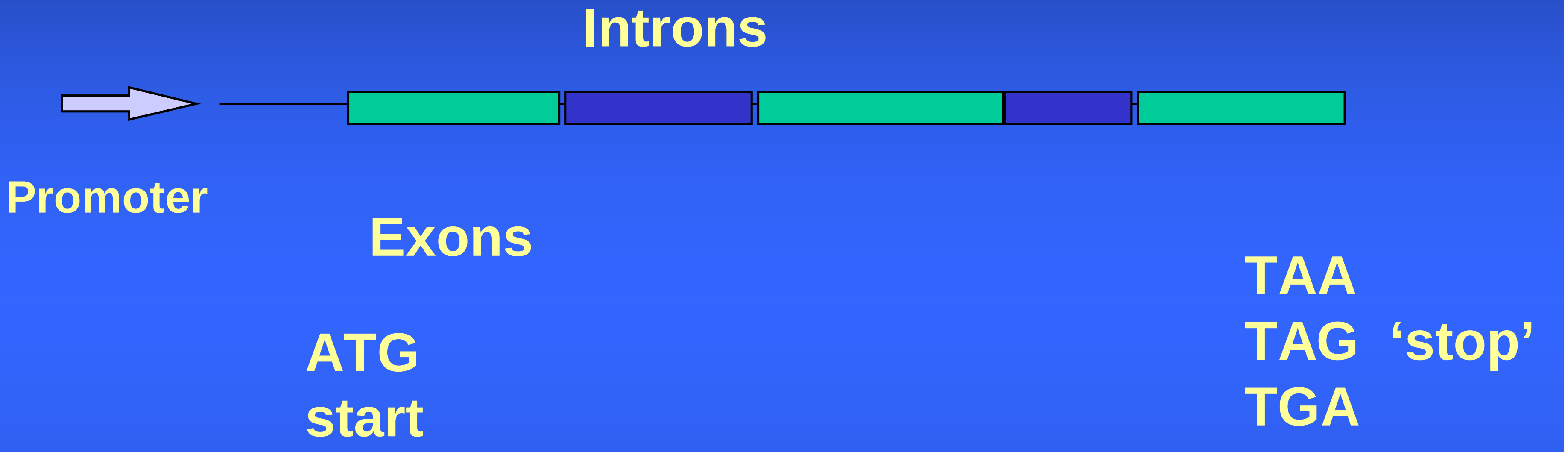
Figure 4-8
Lodish et al. MOLECULAR CELL BIOLOGY, Fourth Edition
Copyright © by W. H. Freeman and Company


Nucleus
Gene
DNA
1. Transcription
RNA
2. Posttranscription
mRNA
Nuclear membrane
mRNA
3. Translation
Polypeptide
4. Posttranslation
Protein folds up
Effector molecule
5. Protein activity
Active protein
Function

JM-2007

TRANSCRIPCION

Gene Structure



Exon = coding sequence

Intron= intervening sequence (non-coding)

Protein Synthesis - Transcription

- Each gene codes for a protein
- DNA sense strand acts as template and is ‘transcribed’ into messenger RNA (mirror image of the DNA but Uracil instead of Thymine)

DNA

A T C G G

U A G C C

mRNA

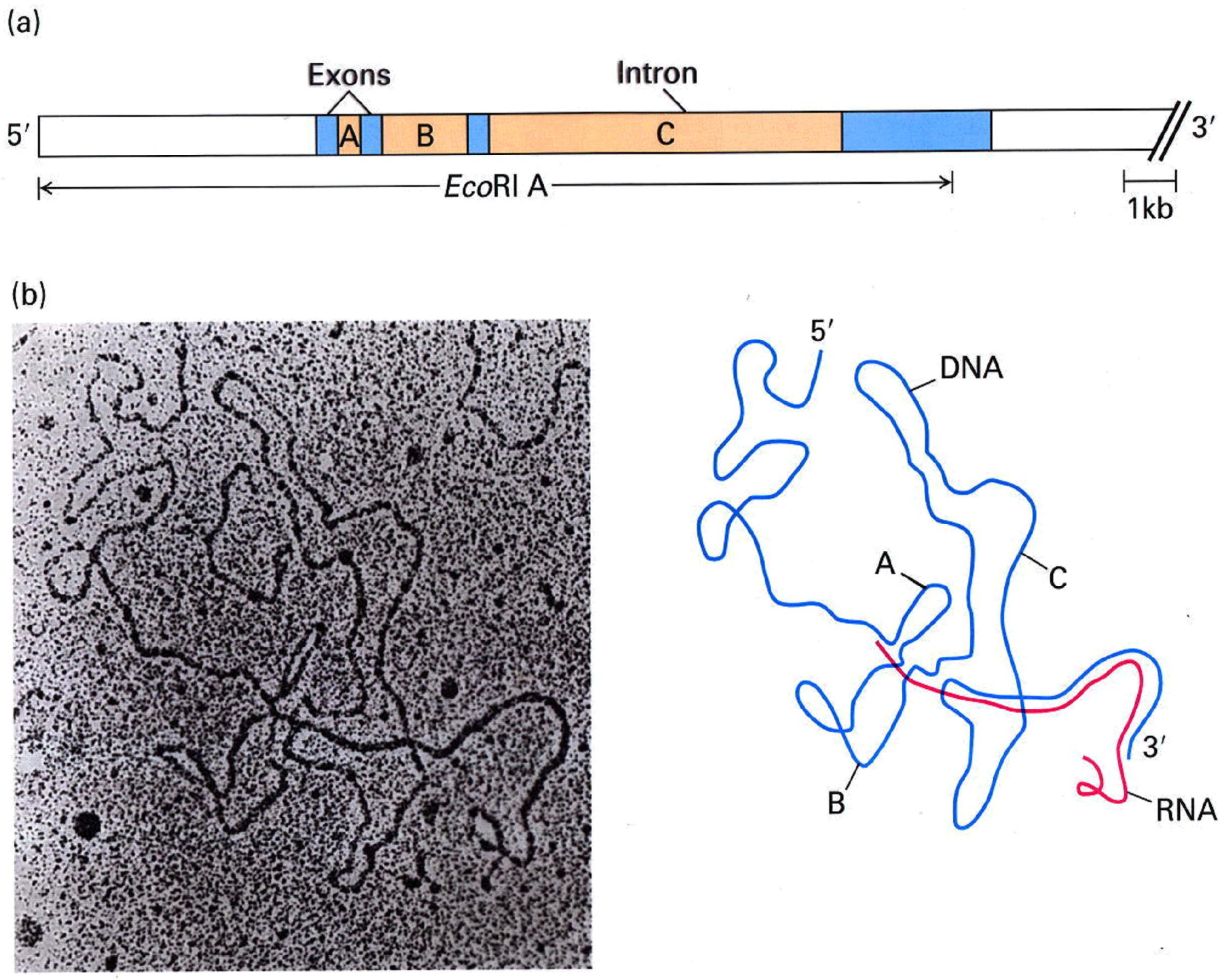
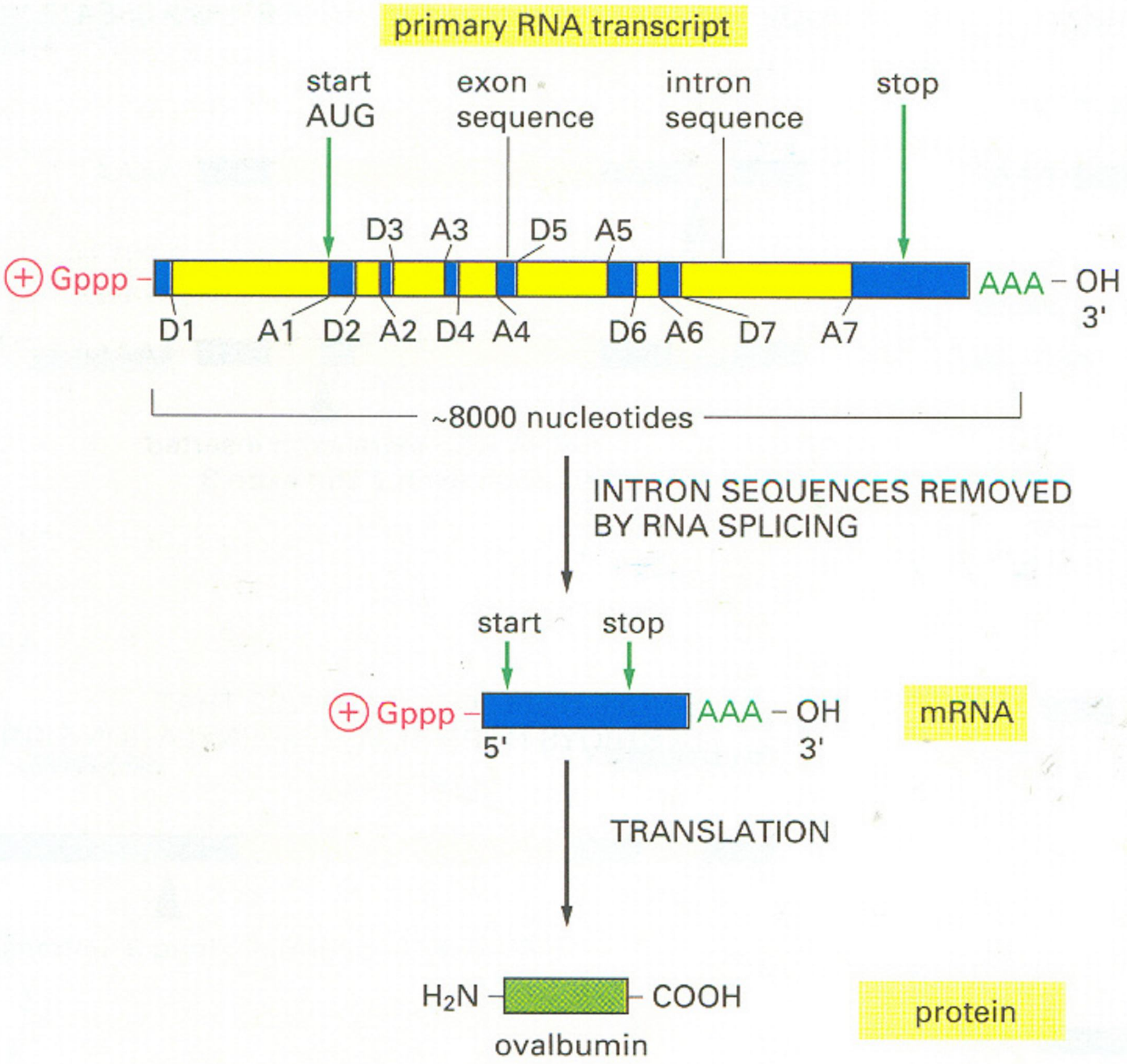
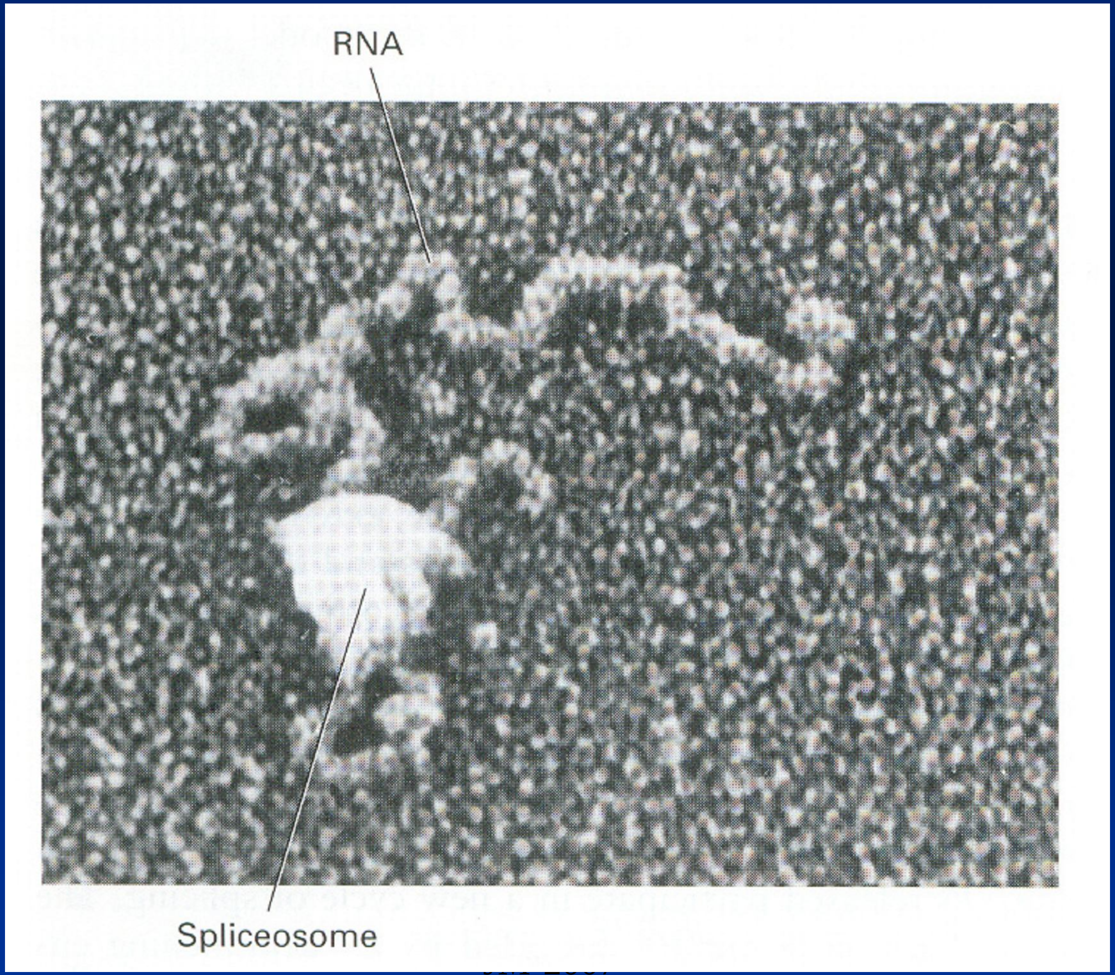


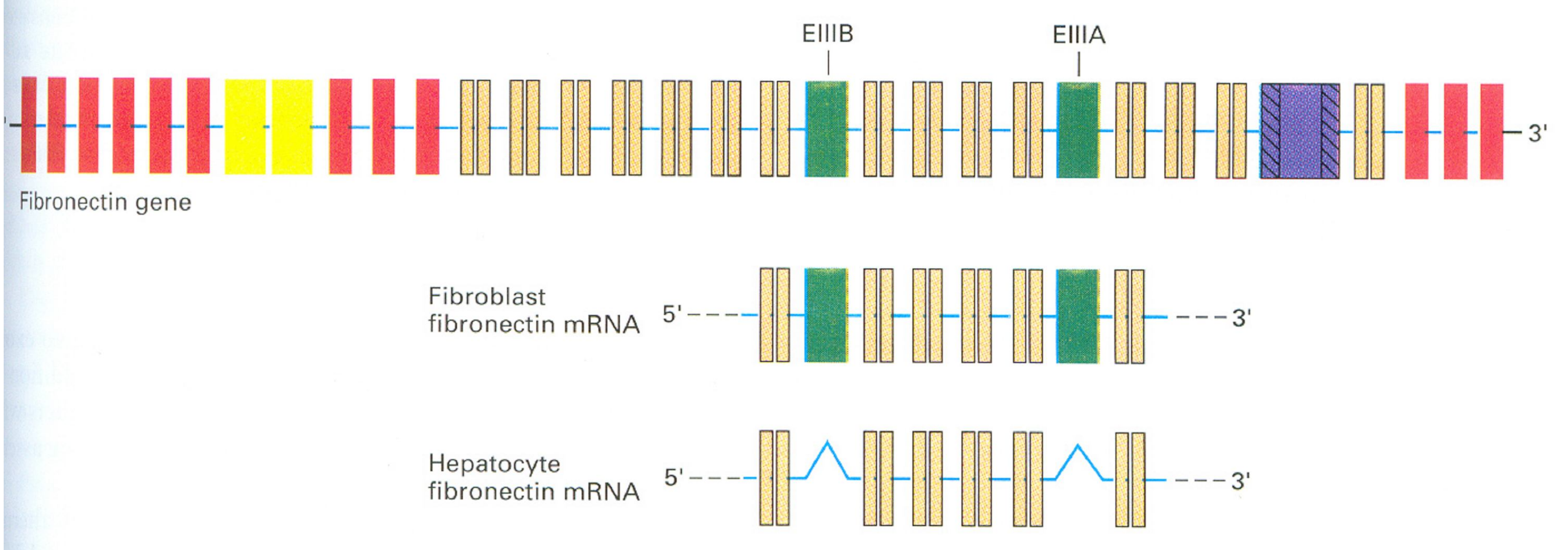
Figure 11-13
Lodish et al. MOLECULAR CELL BIOLOGY, Fourth Edition


primary RNA transcript
start
AUG
exon
sequence
intron
sequence
stop
D3
A3
D5
A5
⊕ Gppp –
AAA – OH
3'
D1
A1
D2
A2
D4
A4
D6
A6
D7
A7
~8000 nucleotides
INTRON SEQUENCES REMOVED
BY RNA SPLICING
start
stop
⊕ Gppp –
AAA – OH
5'
3'
mRNA
TRANSLATION
H_2N –
– COOH
ovalbumin
protein

SPLICEOSOMA



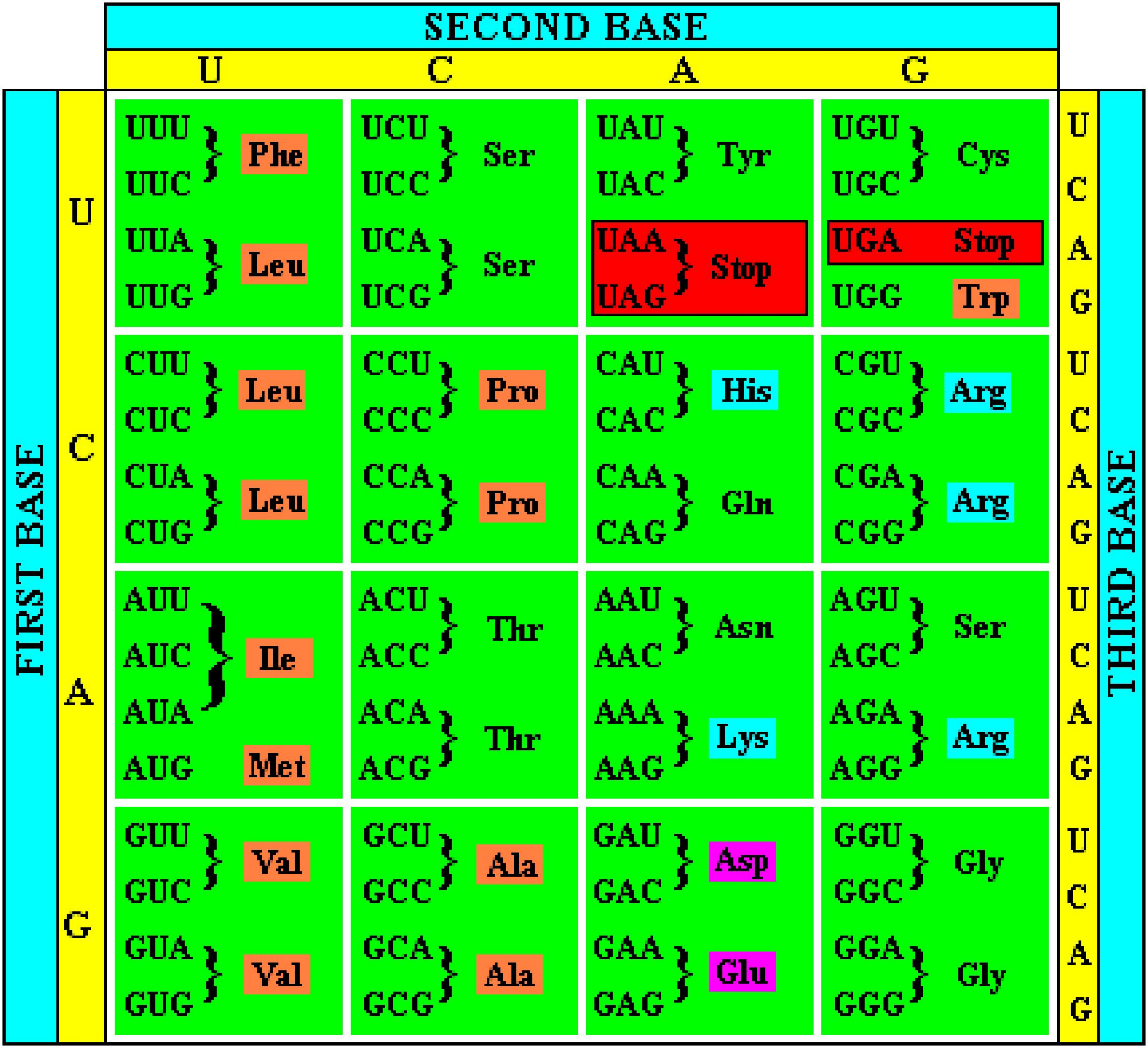
SPLICING DIFERENCIAL EN DISTINTOS TEJIDOS



Some landmarks on the way to the genome sequence

- 1940s Recognition that DNA is the hereditary material
- 1953 Double-helix structure described
- 1966 Genetic code cracked
- 1972 Recombinant DNA technology developed
- 1975-77 DNA sequencing technology developed

T



	SECOND BASE				
FIRST BASE	U	C	A	G	THIRD BASE
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC	UCC	UAC	UGC	C
	UUA Leu	UCA Ser	UAA Stop	UGA Stop	A
	UUG	UCG	UAG	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC	CCC	CAC	CGC	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG	CCG	CAG	CGG	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC	ACC	AAC	AGC	C
	AUA	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG	AAG	AGG	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC	GCC	GAC	GGC	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG	GCG	GAG	GGG	G

TRADUCCION

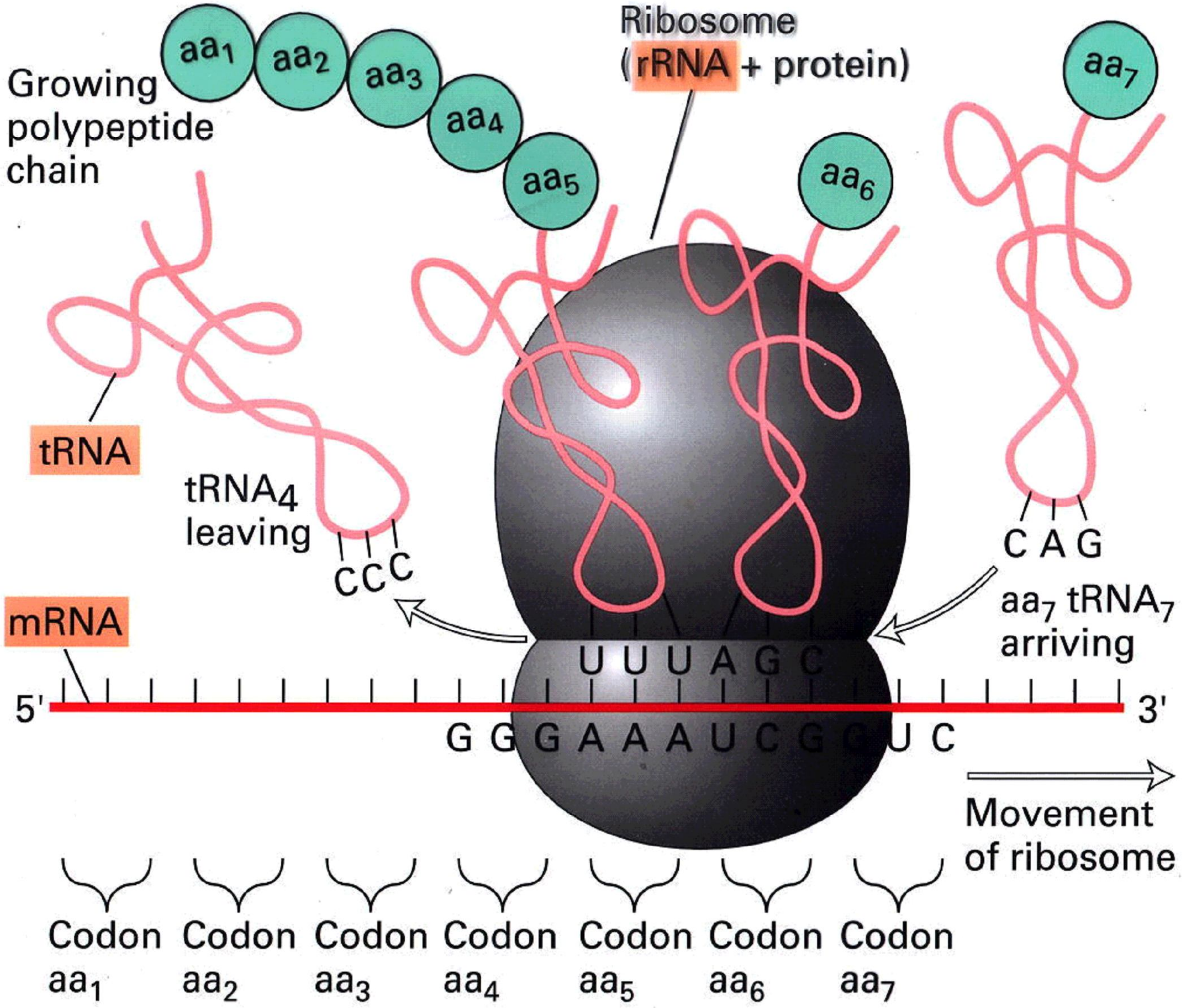


Figure 4-20
Lodish et al. MOLECULAR CELL BIOLOGY, Fourth Edition

Some landmarks on the way to the genome sequence

- 1940s Recognition that DNA is the hereditary material
- 1953 Double-helix structure described
- 1966 Genetic code cracked
- 1972 Recombinant DNA technology developed
- 1975-77 DNA sequencing technology developed

ENZIMAS DE RESTRICCION

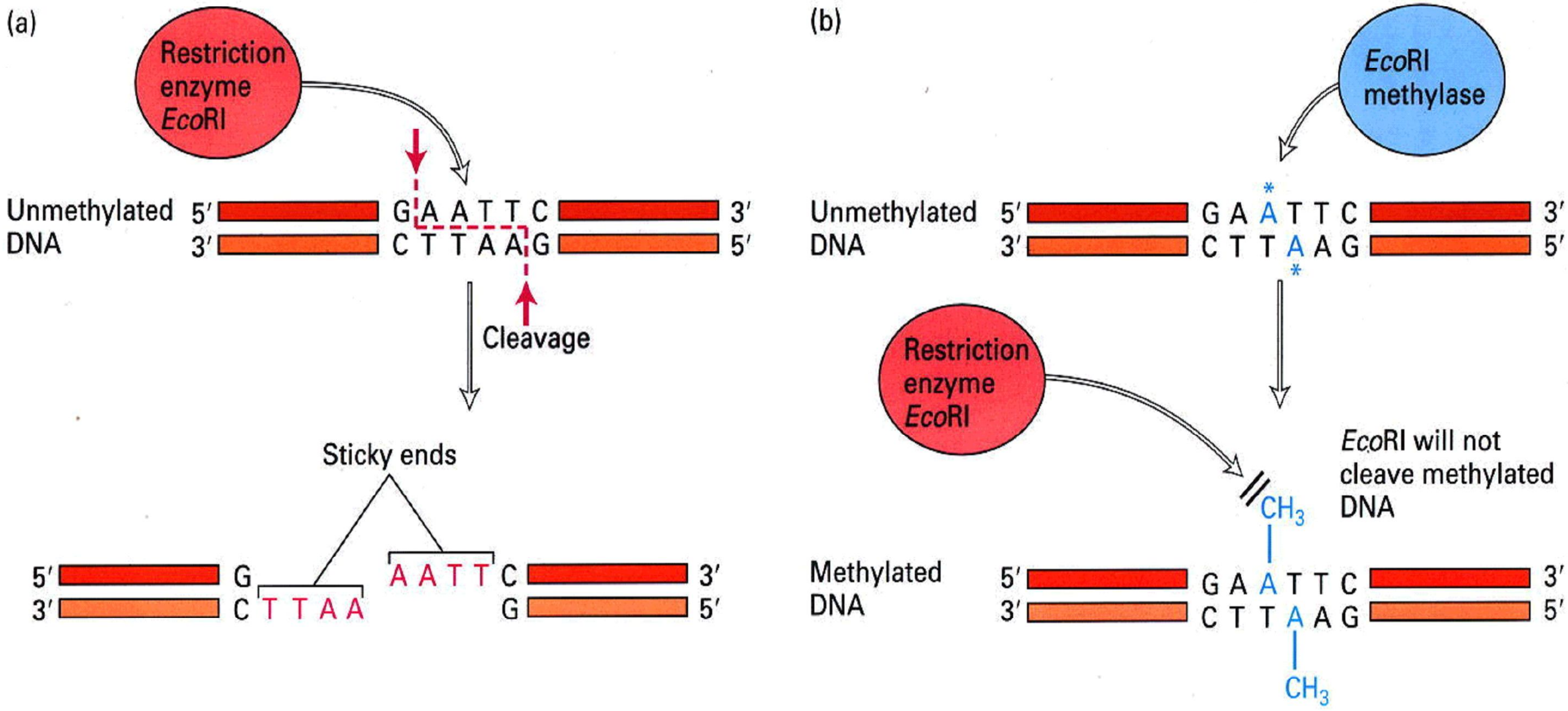
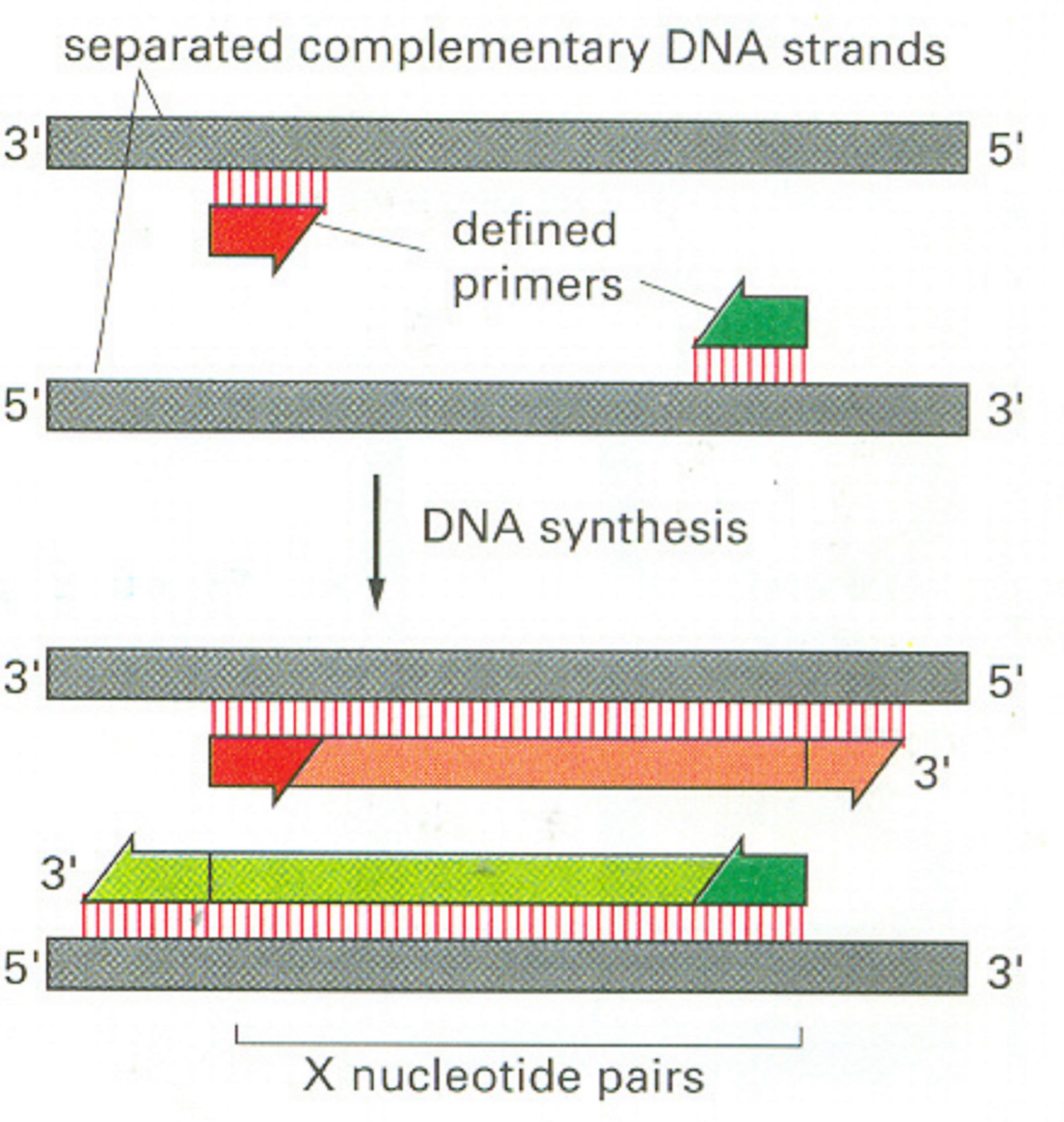
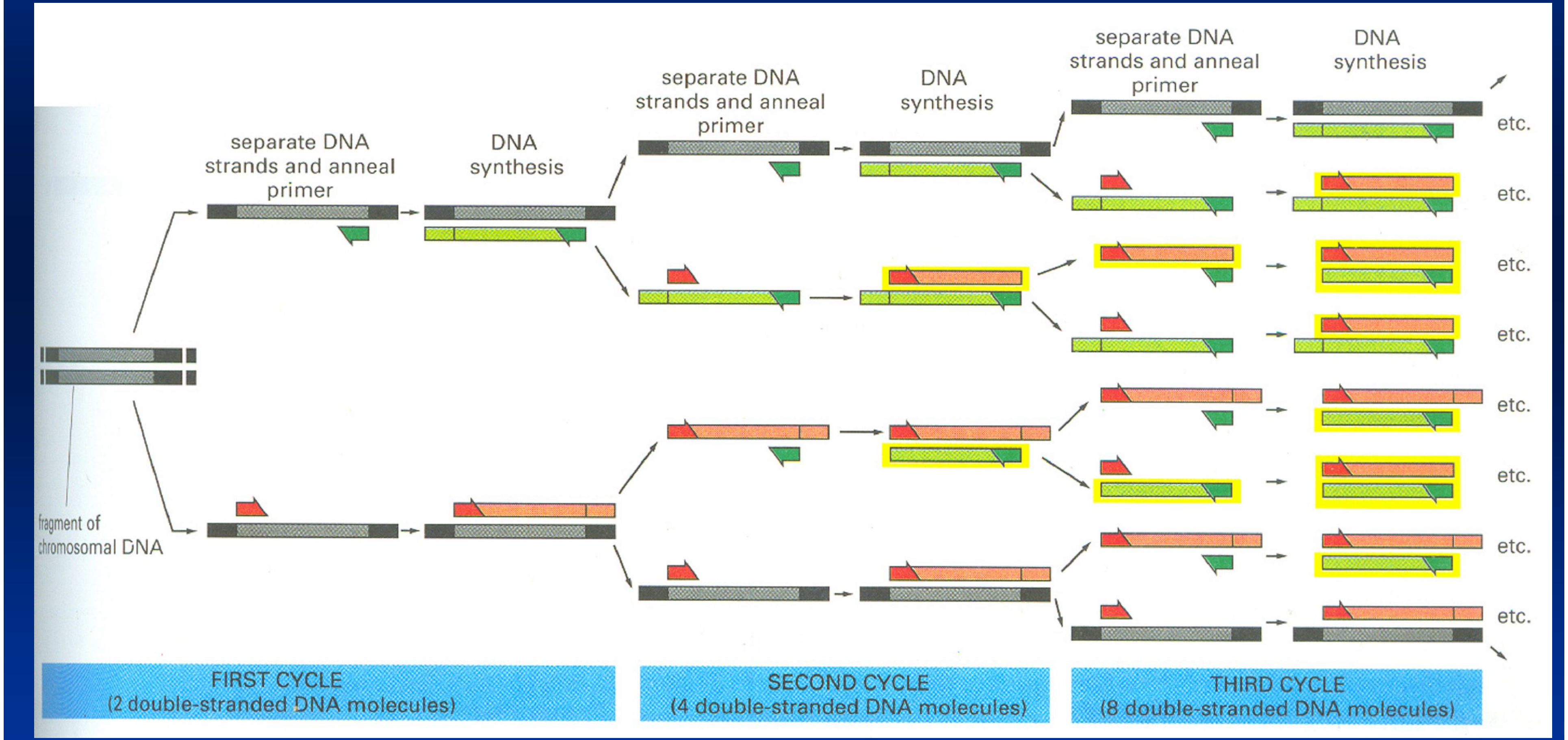


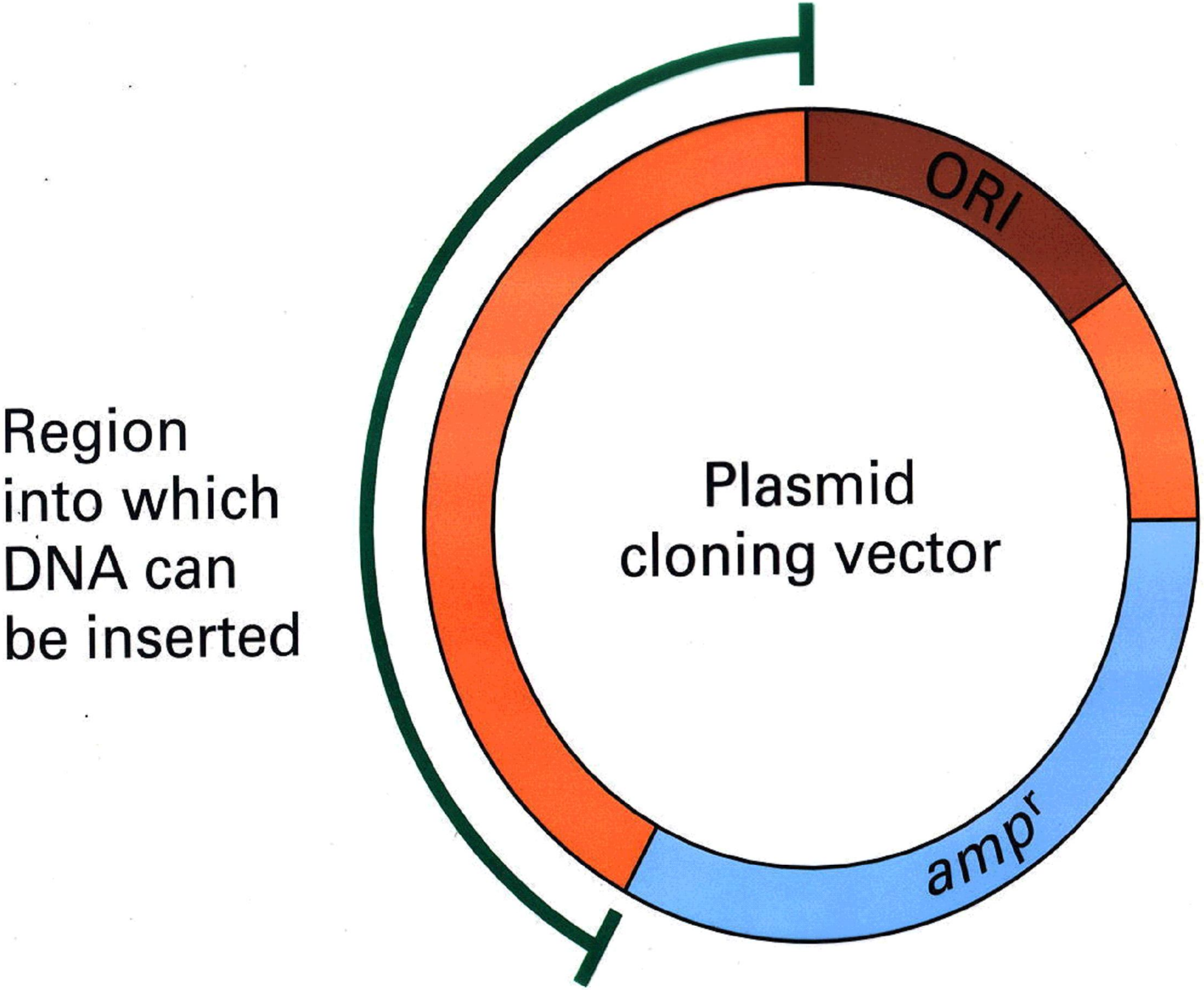
Figure 7-5
Lodish et al. MOLECULAR CELL BIOLOGY, Fourth Edition

POLYMERASE CHAIN REACTION


separated complementary DNA strands
3'
5'
defined primers
5'
3'
DNA synthesis
3'
5'
3'
3'
5'
3'
X nucleotide pairs

PCR

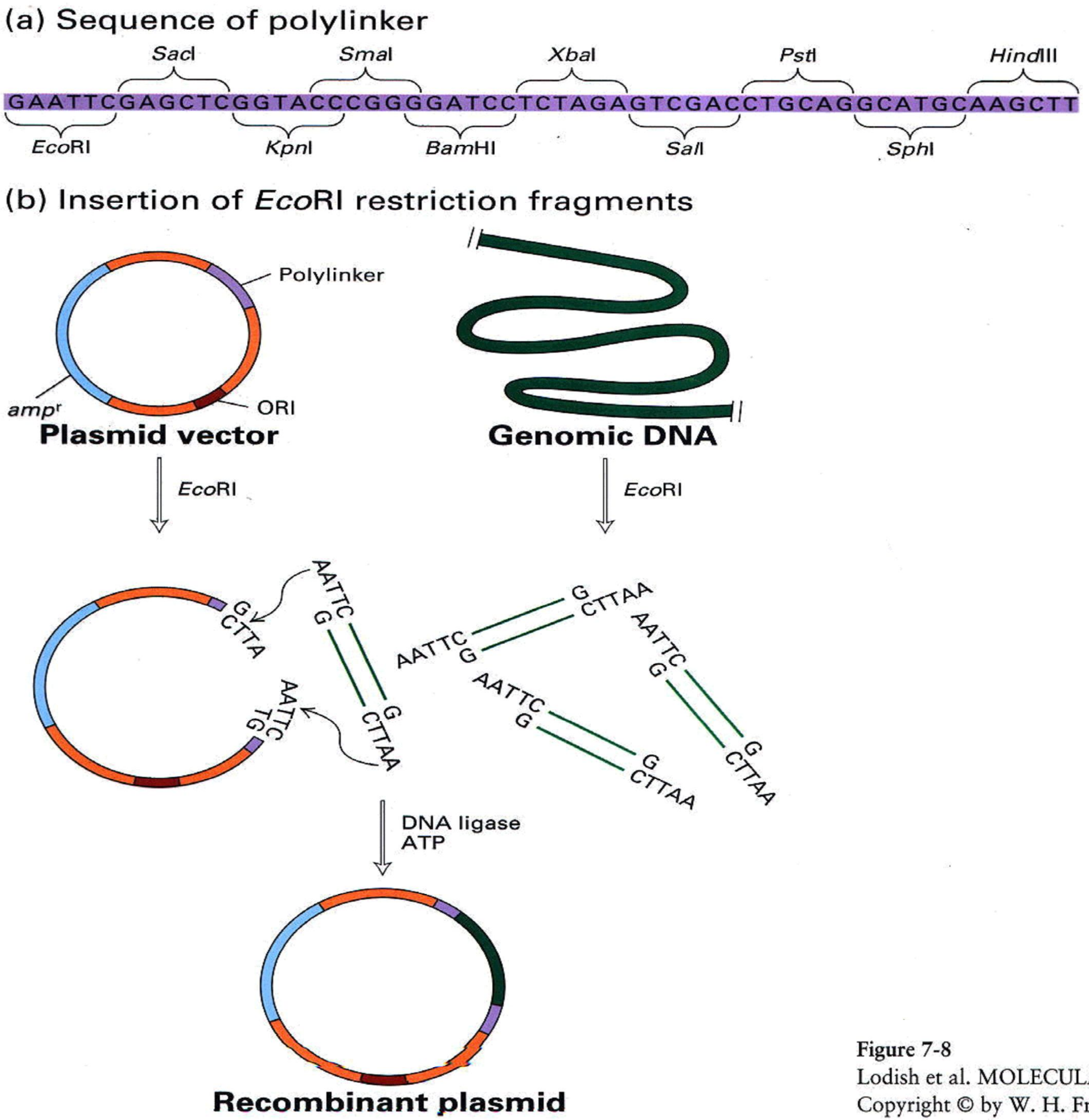

separate DNA strands and anneal primer
DNA synthesis
separate DNA strands and anneal primer
DNA synthesis
separate DNA strands and anneal primer
DNA synthesis
etc.
etc.
etc.
etc.
etc.
etc.
etc.
etc.
fragment of chromosomal DNA
FIRST CYCLE
(2 double-stranded DNA molecules)
SECOND CYCLE
(4 double-stranded DNA molecules)
THIRD CYCLE
(8 double-stranded DNA molecules)



Region into which DNA can be inserted

Plasmid cloning vector

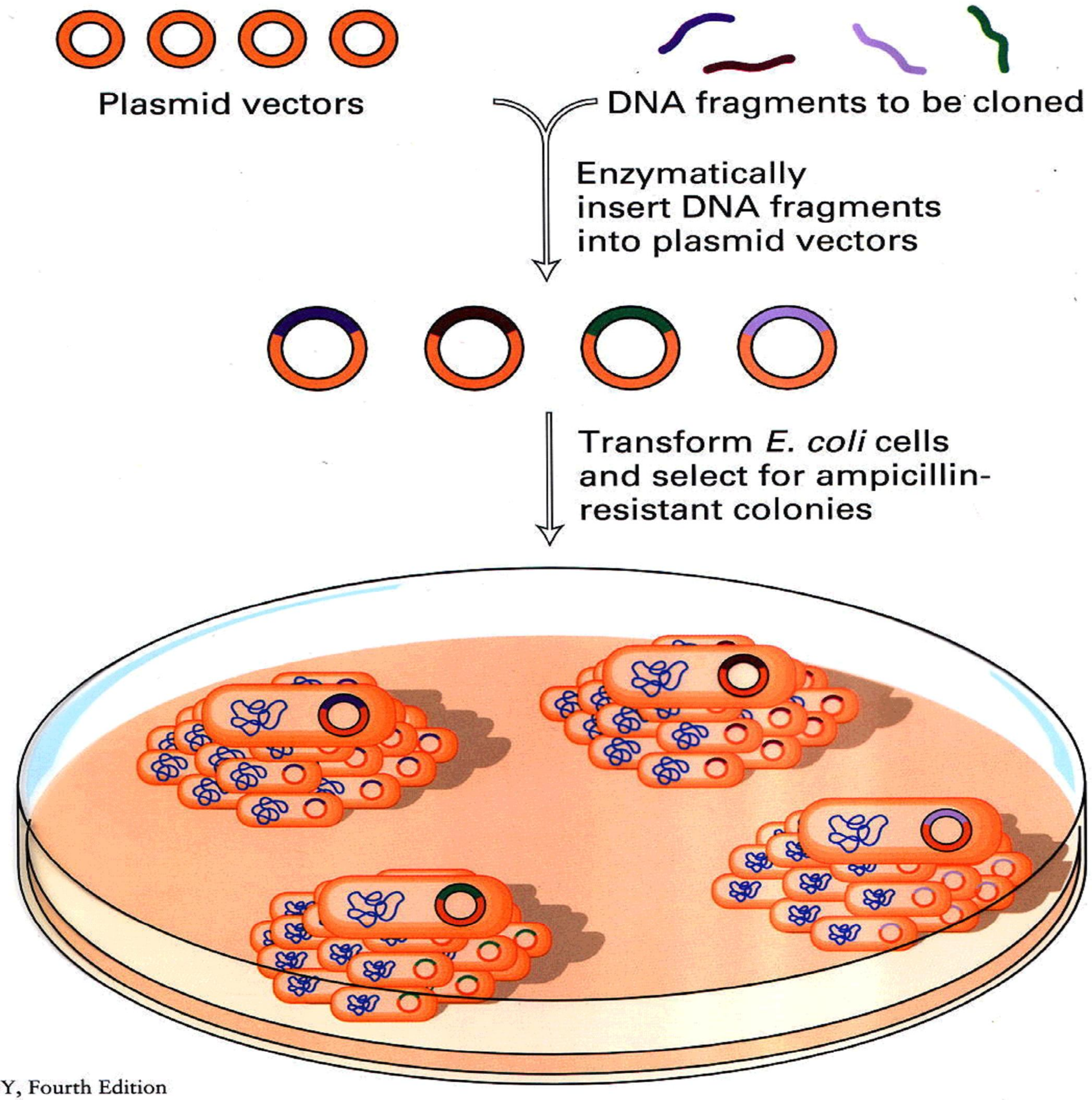
Figure 7-1
Lodish et al. MOLECULAR CELL BIOLOGY, Fourth Edition



(a) Sequence of polylinker

(b) Insertion of *Eco*RI restriction fragments

Figure 7-8
Lodish et al. MOLECULAR CELL BIOLOGY, Fourth Edition
Copyright © by W. H. Freeman and Company



Plasmid vectors

DNA fragments to be cloned

Enzymatically insert DNA fragments into plasmid vectors

Transform *E. coli* cells and select for ampicillin-resistant colonies

Figure 7-4
Lodish et al. MOLECULAR CELL BIOLOGY, Fourth Edition
Copyright © by W. H. Freeman and Company

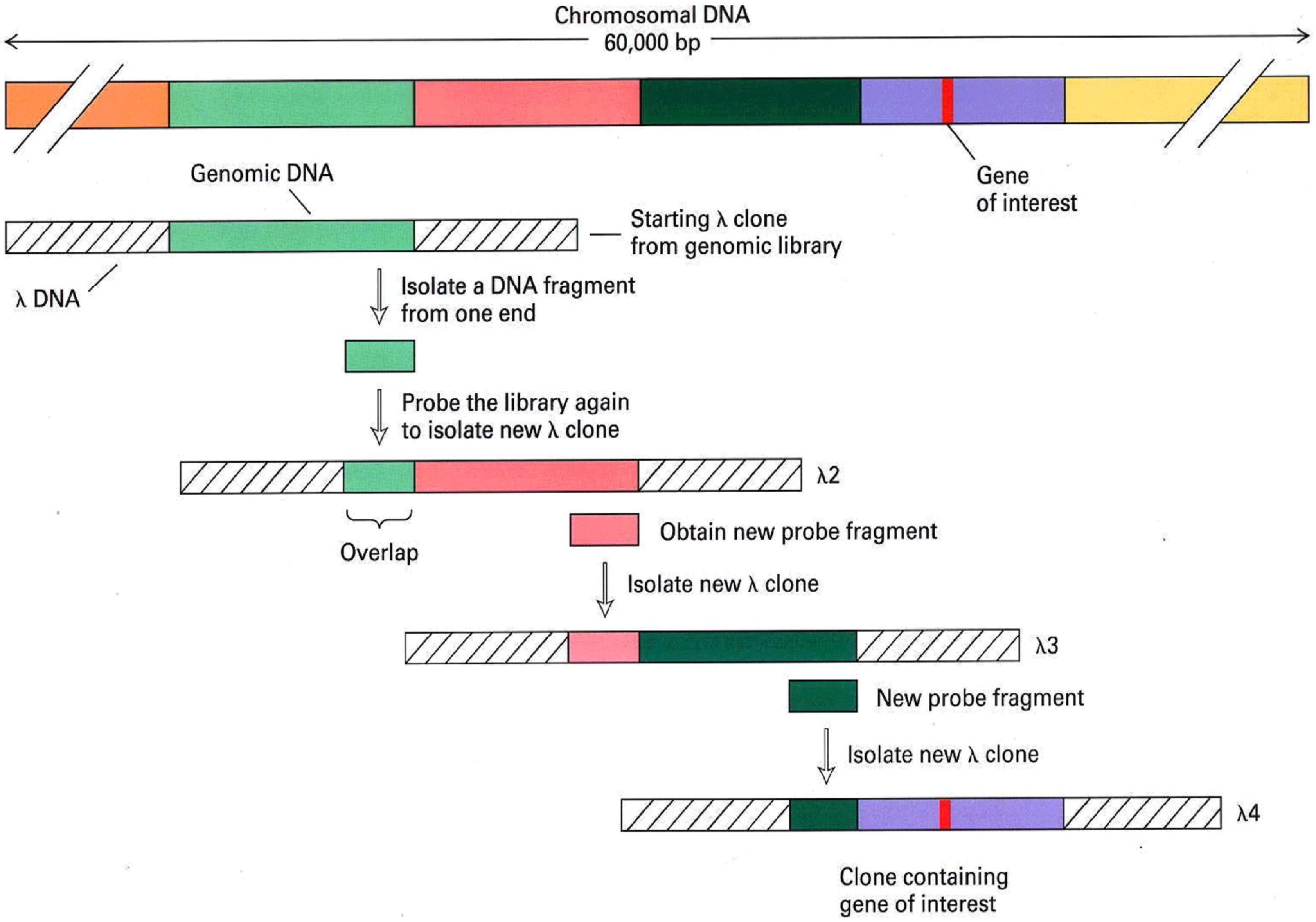


Figure 8-24
Lodish et al. MOLECULAR CELL BIOLOGY, Fourth Edition

Some landmarks on the way to the genome sequence

- 1940s Recognition that DNA is the hereditary material
- 1953 Double-helix structure described
- 1966 Genetic code cracked
- 1972 Recombinant DNA technology developed
- 1975-77 DNA sequencing technology developed

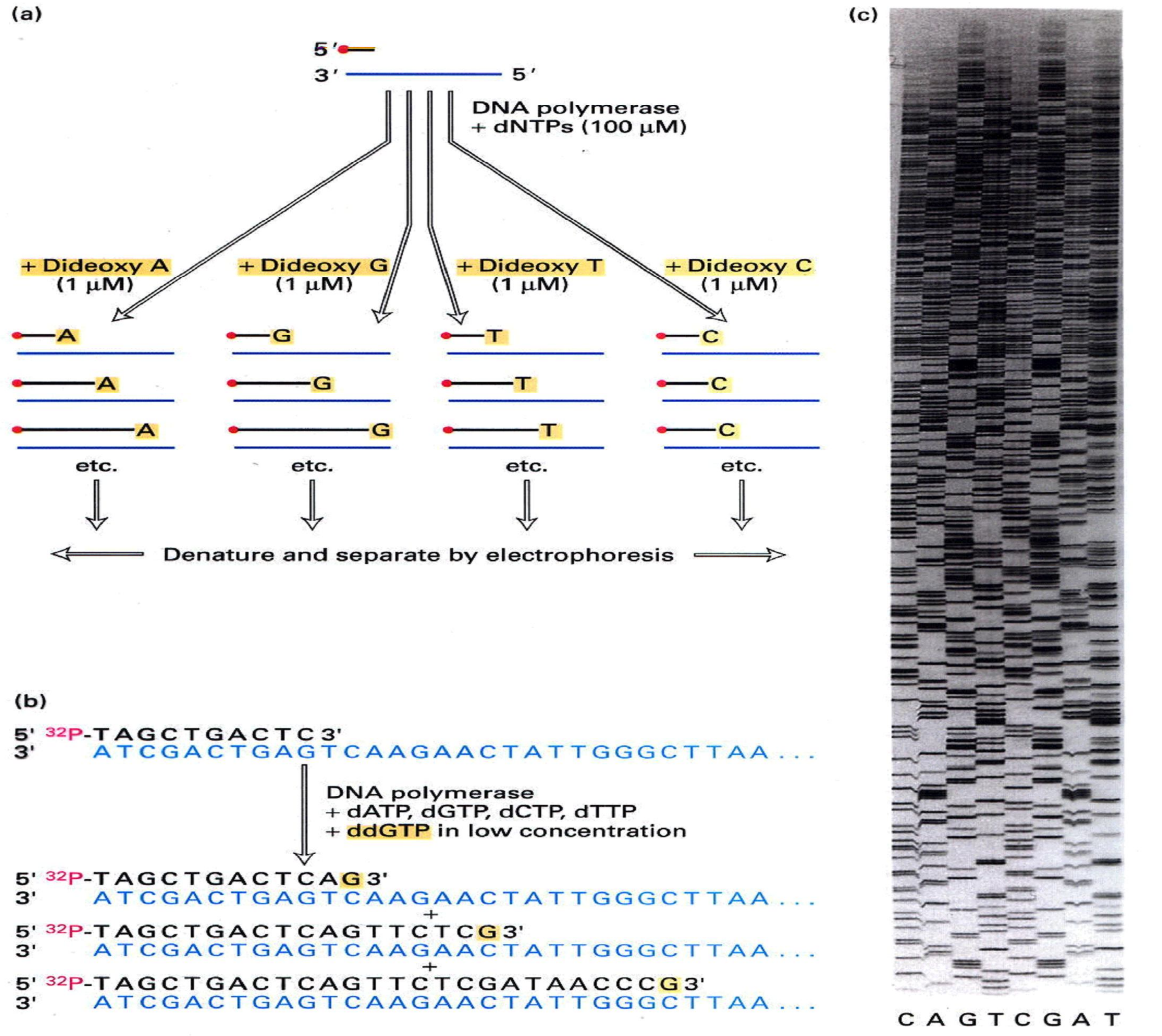
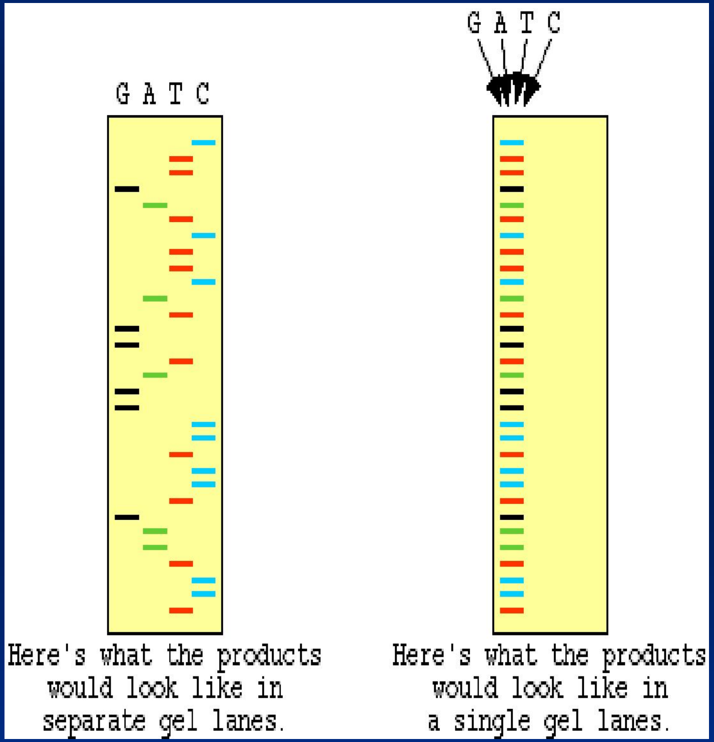
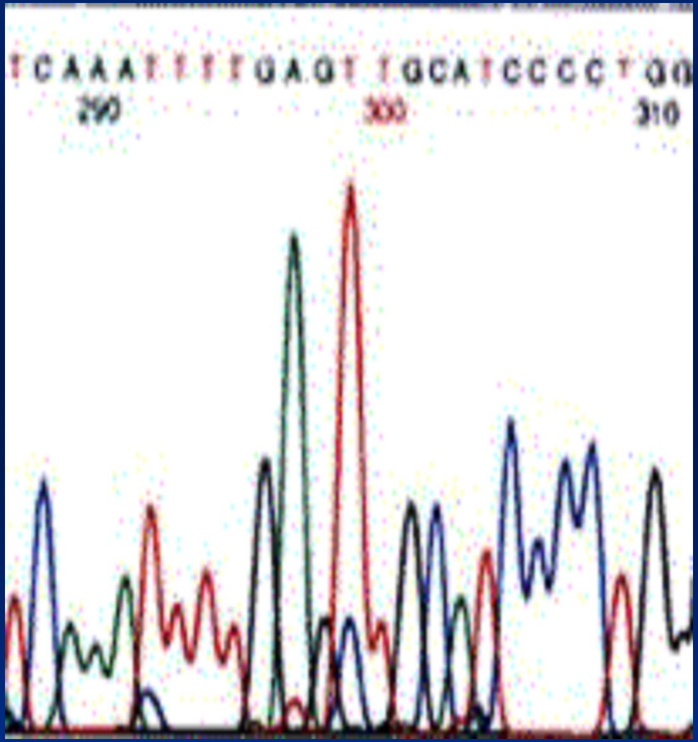

(a)
5′
3′
5′
DNA polymerase
+ dNTPs (100 μM)
+ Dideoxy A
(1 μM)
+ Dideoxy G
(1 μM)
+ Dideoxy T
(1 μM)
+ Dideoxy C
(1 μM)
A
A
A
G
G
G
T
T
T
C
C
C
etc.
etc.
etc.
etc.
Denature and separate by electrophoresis
(b)
5' 32P-TAGCTGACTC 3'
3' ATCGACTGAGTCAAGAACTATTGGGCTTAA . . .
DNA polymerase
+ dATP, dGTP, dCTP, dTTP
+ ddGTP in low concentration
5' 32P-TAGCTGACTCAG 3'
3' ATCGACTGAGTCAAGAACTATTGGGCTTAA . . .
+
5' 32P-TAGCTGACTCAGTTCTCG 3'
3' ATCGACTGAGTCAAGAACTATTGGGCTTAA . . .
+
5' 32P-TAGCTGACTCAGTTCTCGATAACCCG 3'
3' ATCGACTGAGTCAAGAACTATTGGGCTTAA . . .
(c)
C A G T C G A T

Figure 7-29
Lodish et al. MOLECULAR CELL BIOLOGY, Fourth Edition

T-47


G A T C
G A T C
Here's what the products would look like in separate gel lanes.
Here's what the products would look like in a single gel lanes.



Figures taken from http b572/L8/L8.htm
://www.escience.ws/

Landmarks

- 1983 First human disease gene mapped (Huntington’s disease). Gene finally isolated in 1993
- 1990 Human Genome Project launched
- 1995 Human Physical map completed
- 1996 Sequencing begins
- 1999 Full-scale sequencing begins
- 2001 ‘Working draft’ produced
- 2003 Final sequence published

Human Genome Project Goals

- identify all the approximate 30,000 genes in human DNA,
- determine the sequences of the 3 billion chemical base pairs that make up human DNA
- store this information in databases,
- improve tools for data analysis,
- transfer related technologies to the private sector, and
- address the ethical, legal, and social issues (ELSI) that may arise from the project.

Why?

- virtually all disease states arise through complex interplay between genes & environment
- almost all progress to date is on single-gene disorders
- Cancers, heart disease, hypertension etc, all have genetic component,
  - many genes involved
  - hard to get at by traditional techniques

Proyecto Genoma Humano

Tres instancias de análisis del GH y sus productos:

- **GENOMA HUMANO**: análisis de la secuencia completa
- **TRANSCRIPTOMA HUMANO**: análisis de las secuencias transcriptas en mRNA (genes)
- **PROTEOMA HUMANO**: análisis de las proteínas presentes en el hombre

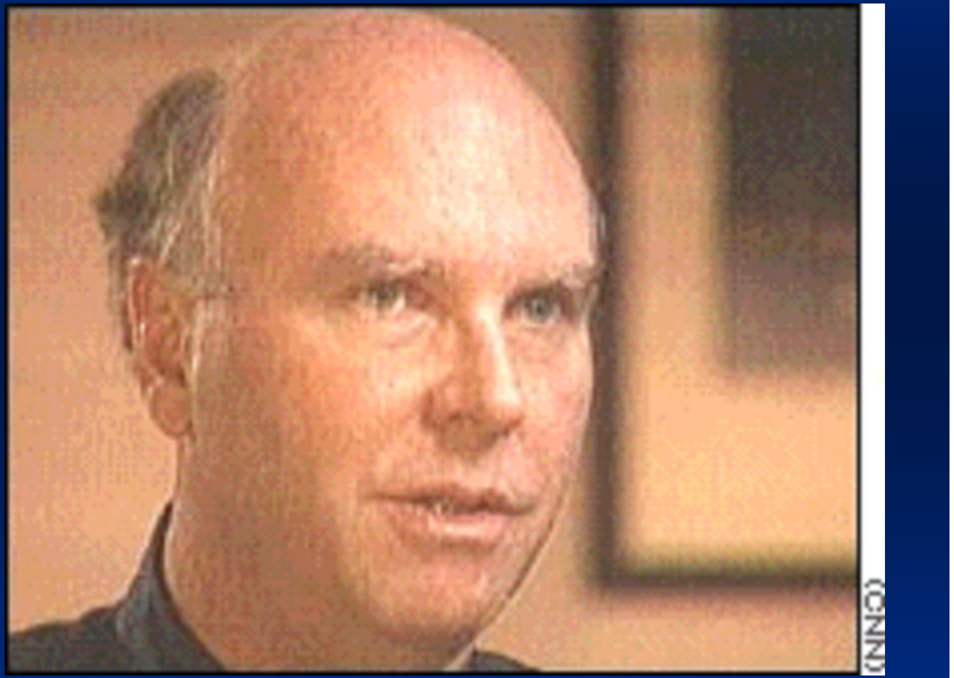
Proyecto Genoma Humano

Tabla 1. **Tamaño comparativo de los genomas de diversas especies**

Especie	**Genoma Haploide** **(pares de Bases, bp)**	**Genes** **(n)**
Homo sapiens	3.000.000.000.	30.000 - 35.000
Mus musculus	3.000.000.000.	30.000-35.000
Rattus norvegicus	3.000.000.000.	30.000-35.000
Drosophila (mosca de la fruta)	165.000.000.	15.000 - 25.000
C. elegans	100.000.000.	19.000
Levadura y hongos	14.000.000.	8.355 - 8.947
E. Coli	4.670.000.	3.237
H.Influenzae	1.800.000.	
M. Genitalium	580.000.	

The Human Genome Race

Collins vs. Venter



Francis Collins

Craig Venter

Collins

- Francis Collins, a physician, is director of the National Human Genome Research Institute.
- His research laboratory was responsible for identifying the genes responsible for **Cystic Fibrosis, Neurofibromatosis**, and **Huntington's disease**.

Venter

Venter founded the nonprofit Institute for Genomic Research in 1992. Before that he was section chief and a laboratory chief at the National Institutes of Neurological Disorders and the National Institutes of Health. Celera Genomics is part of the PE Corporation.

Venter

Celera is a for-profit organization whose motto is "Discovery Can't Wait"

Intro to Sequencing

- To read the DNA, the chromosomes are cut into tiny pieces, each of which is read individually.
- When all the segments have been read they are assembled in the correct order. Link these fragments to self-replicating forms of DNA = vectors.

Intro to Sequencing

- Two approaches have been used to sequence the genome.
- They differ in the methods they use to cut up the DNA, assemble it in the correct order, and **whether** they map the chromosomes before decoding the sequence.

Intro to Sequencing: BAC to BAC

- The BAC-to-BAC method:
  - the first to be employed in human genome studies
  - slow but sure
  - also called the “map based method”

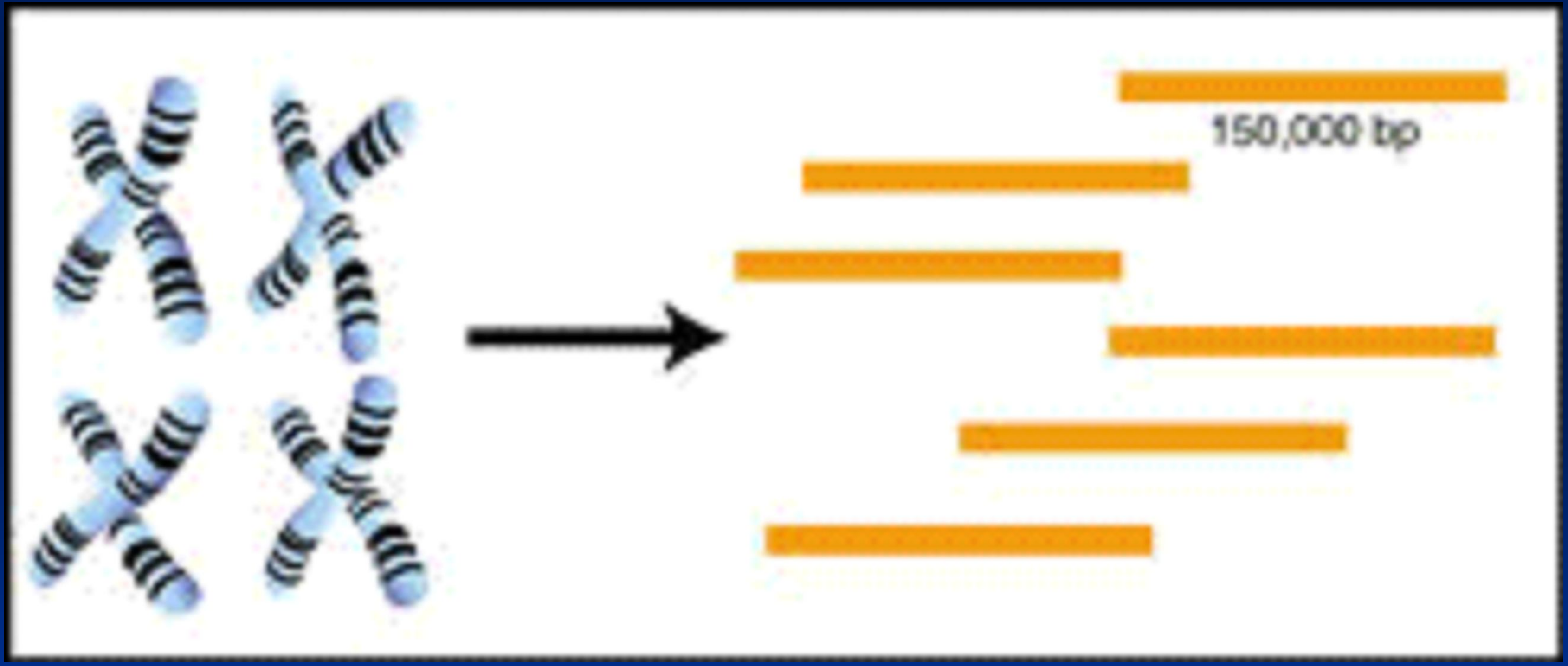
BAC to BAC - 1

- First create a rough physical map of the whole genome before sequencing the DNA
  - requires cutting the chromosomes into large pieces and then figuring out the order of these big chunks of DNA before taking a closer look and sequencing all the fragments.

BAC to BAC - 2

II. Several copies of the genome are randomly cut into pieces that are about 150,000 base pairs (bp) long.

BAC to BAC - 2



BAC to BAC - 3

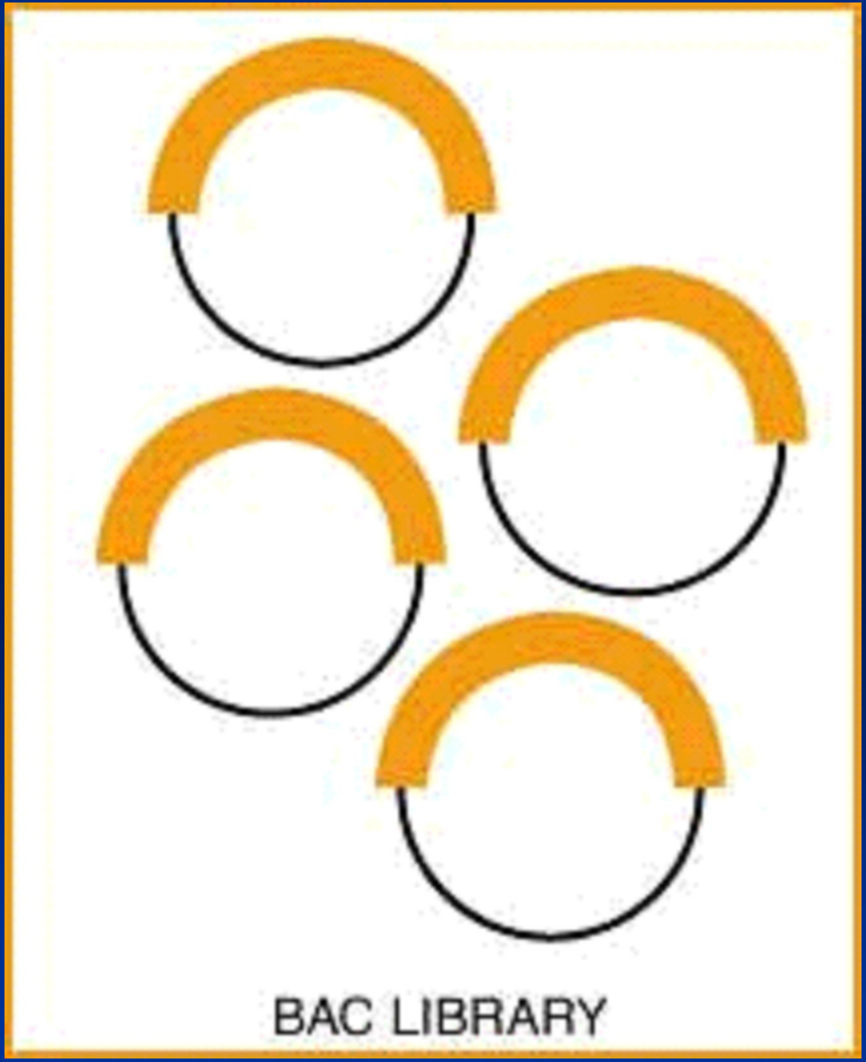
Each of these 150,000 bp fragment

is inserted into a BAC

♦ A BAC is a man made piece of DNA that can replicate inside a bacterial cell.

♦ The collection of BACs containing the entire human genome is called **“a BAC library”.**

BAC to BAC - 3



BAC to BAC - 4

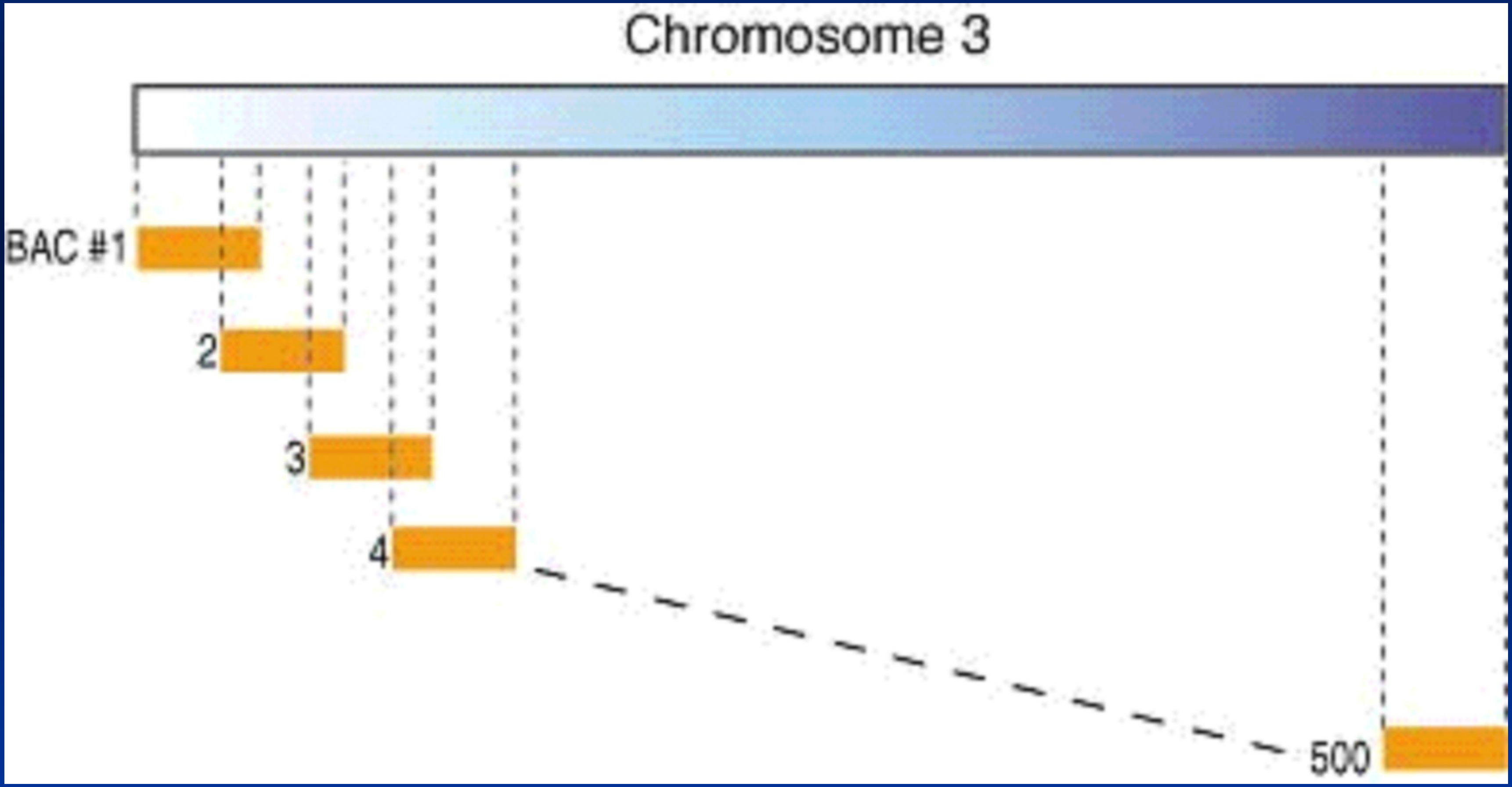
These pieces are fingerprinted to give each piece a unique identification tag that determines the order of the fragments.

Cutting each BAC fragment with a single enzyme and finding common sequence landmarks in overlapping fragments that determine the location of each BAC along the chromosome.

BAC to BAC - 4

Then overlapping BACs with markers every 100,000 bp form a map of each chromosome

BAC to BAC - 4



BAC to BAC - 5

1 Each BAC is then broken randomly into 1,500 bp pieces and placed in another artificial piece of DNA called M13. This collection is known as an M13 library.

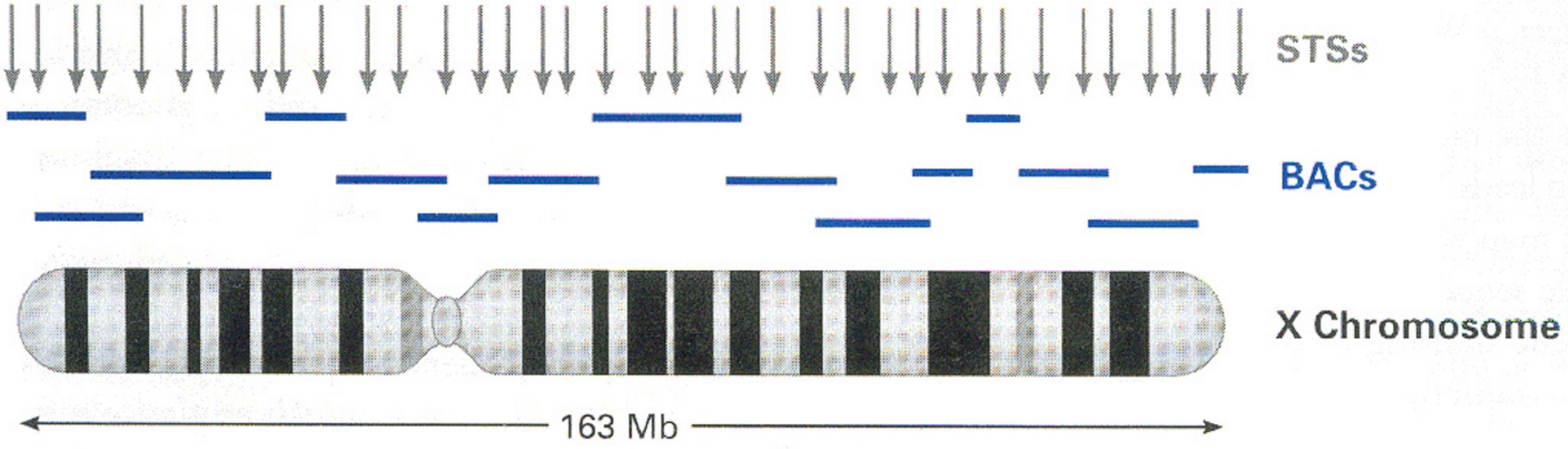
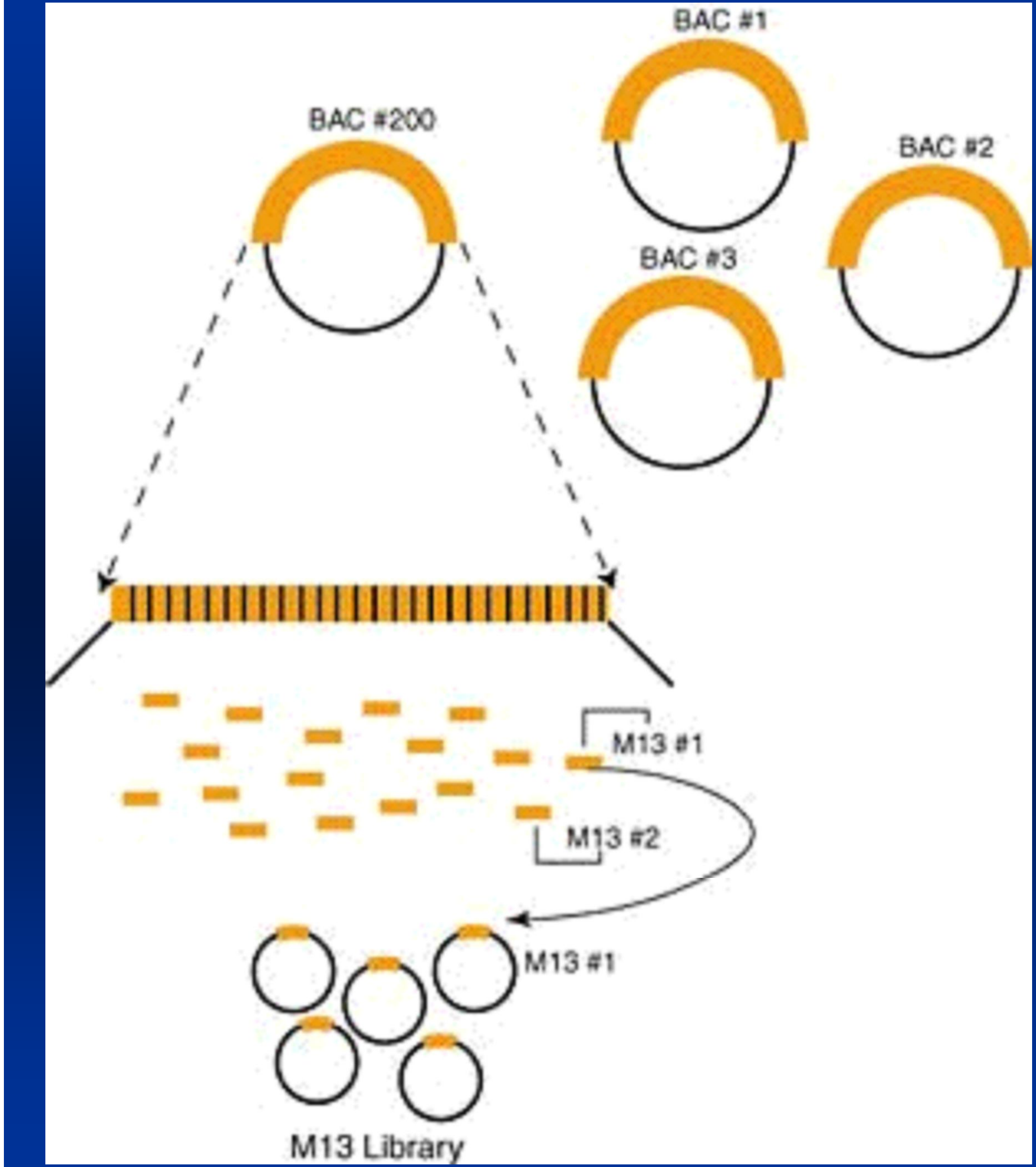


FIGURE 1.3 • Relationships of chromosomes to genome sequencing markers. The X chromosome is about 163 Mb in length. In this diagram, there are 16 overlapping BAC clones that span the entire length. In reality, 1,408 BACs were needed to span the X chromosome. Arrows (top) mark STSs scattered throughout the chromosome and on overlapping BACs.

BAC to BAC - 5



BAC to BAC - 6

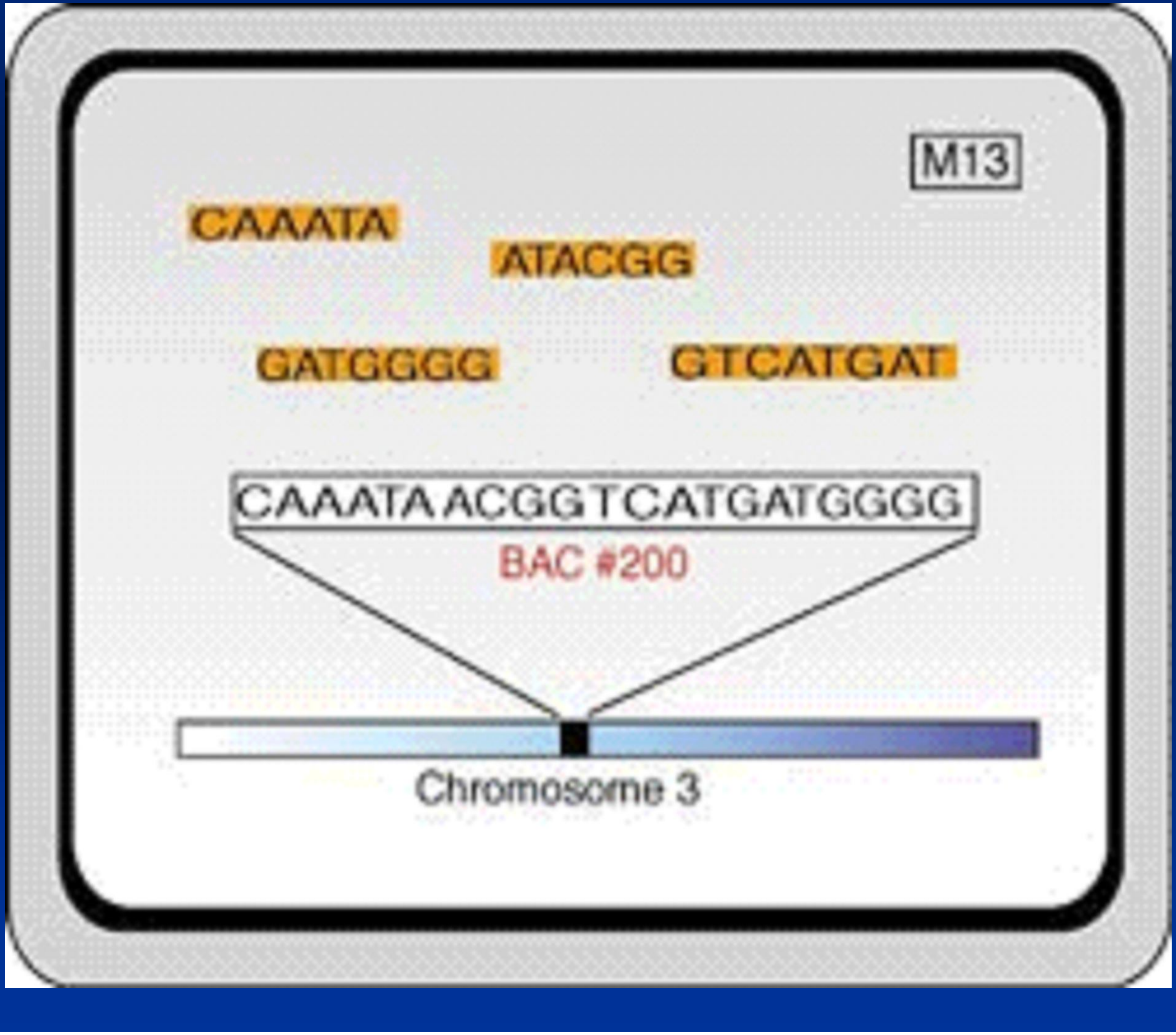
All the M13 libraries are sequenced.

500 bp from one end of the fragment are sequenced generating millions of sequences

BAC to BAC - 7

These sequences are fed into a computer program called PHRAP that looks for common sequences that join two fragments together.

BAC to BAC - 7


M13
CAAATA
ATACGG
GATGGGG
GTCATGAT
CAAATA ACGGTCATGATGGGG
BAC #200
Chromosome 3

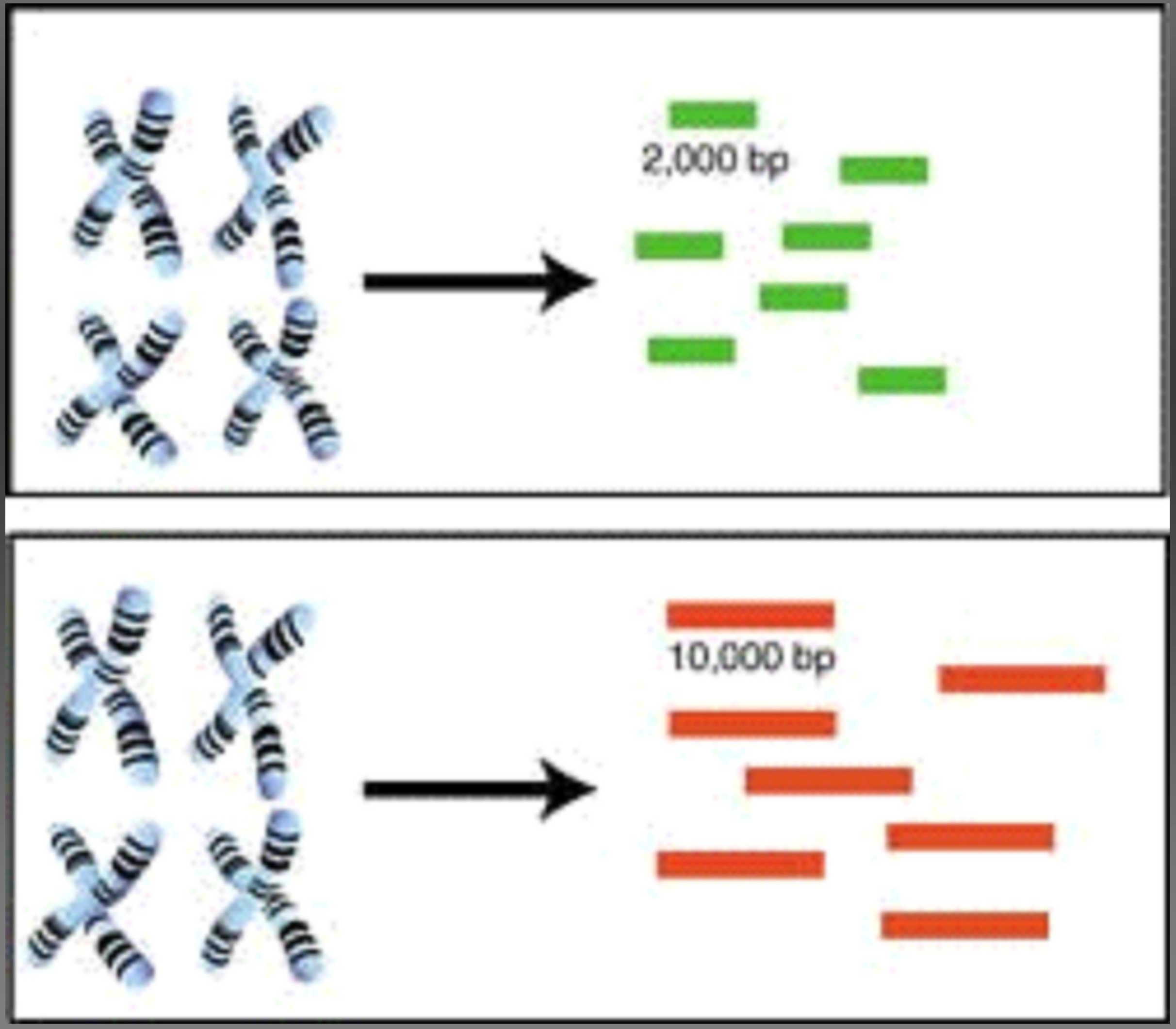
Whole Genome Shotgun - 1

1 The shotgun sequencing method goes straight to the job of decoding, bypassing the need for a physical map.

Whole Genome Shotgun - 2

1 Multiple copies of the genome are randomly shredded into pieces that are 10,000 bp long by squeezing the DNA through a pressurized syringe. This is done a second time to generate pieces that are 2,000 bp long.

Whole Genome Shotgun - 2



Whole Genome Shotgun - 3

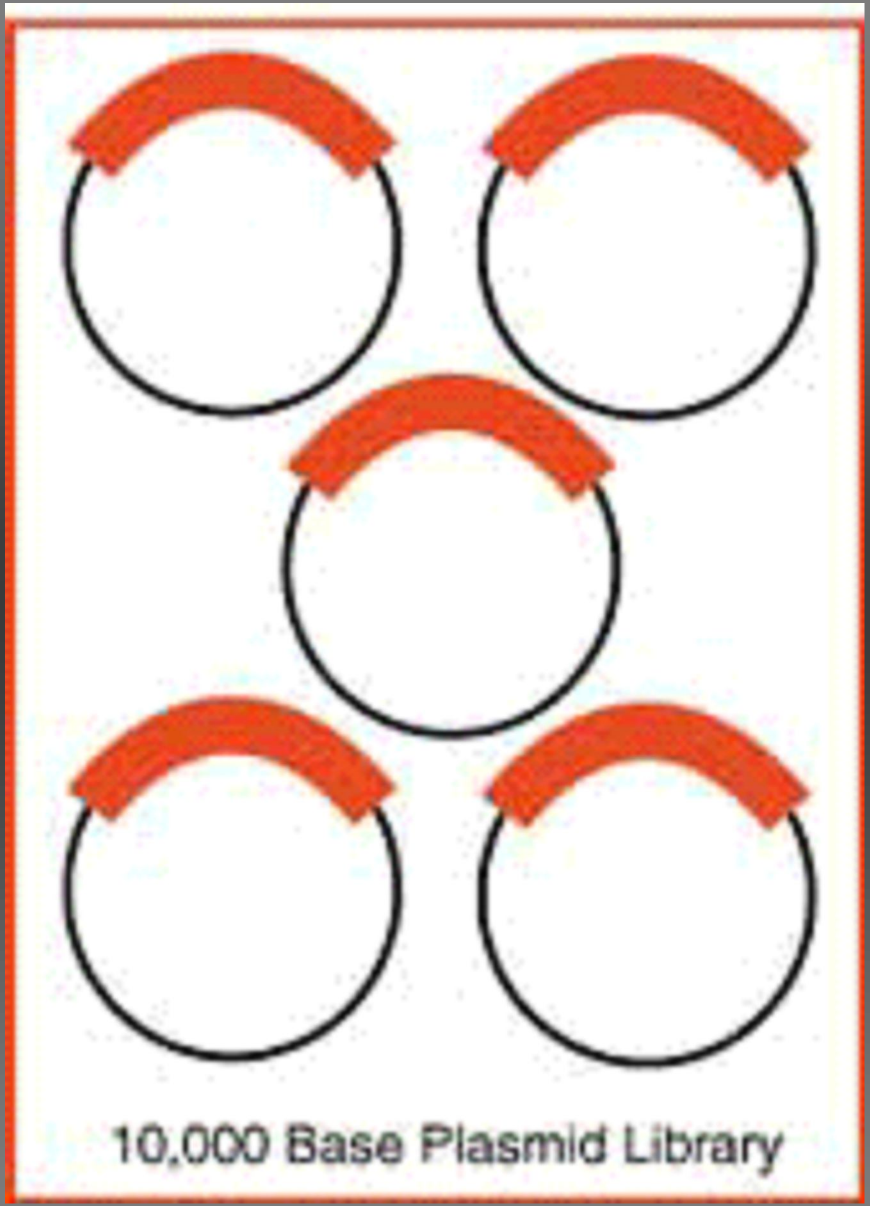
1 Each 2,000 and 10,000 bp fragment is inserted into a plasmid, which is a piece of DNA that can replicate in bacteria.

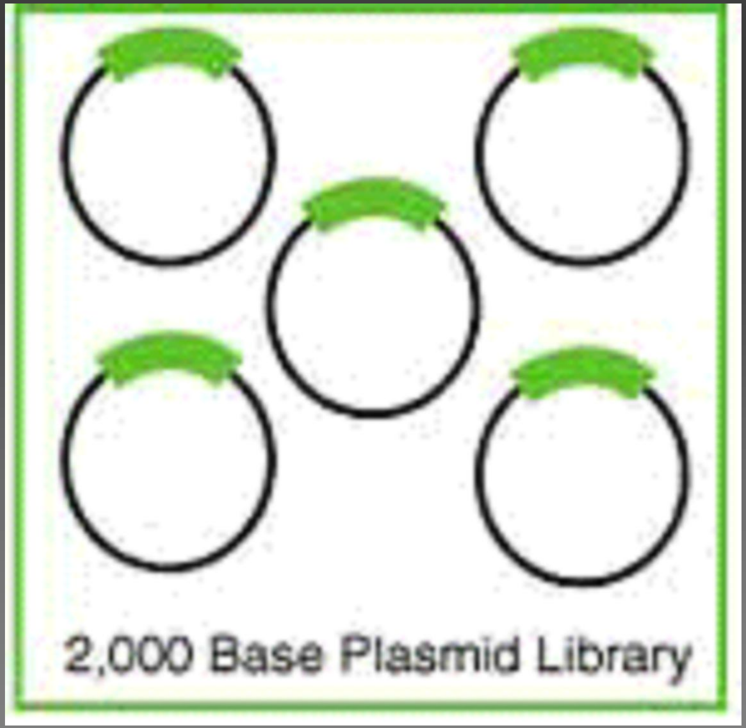
  1 The two collections of plasmids containing 2,000 and 10,000 bp chunks of human DNA are known as plasmid libraries.

Intro to Sequencing: Whole Genome Shotgun

- Whole Genome Shotgun Method brings speed into the picture, enabling researchers to do the job in months to a year.
- Developed by Celera president Craig Venter in 1996 when he was at the Institute for Genomic Research.

Whole Genome Shotgun - 3



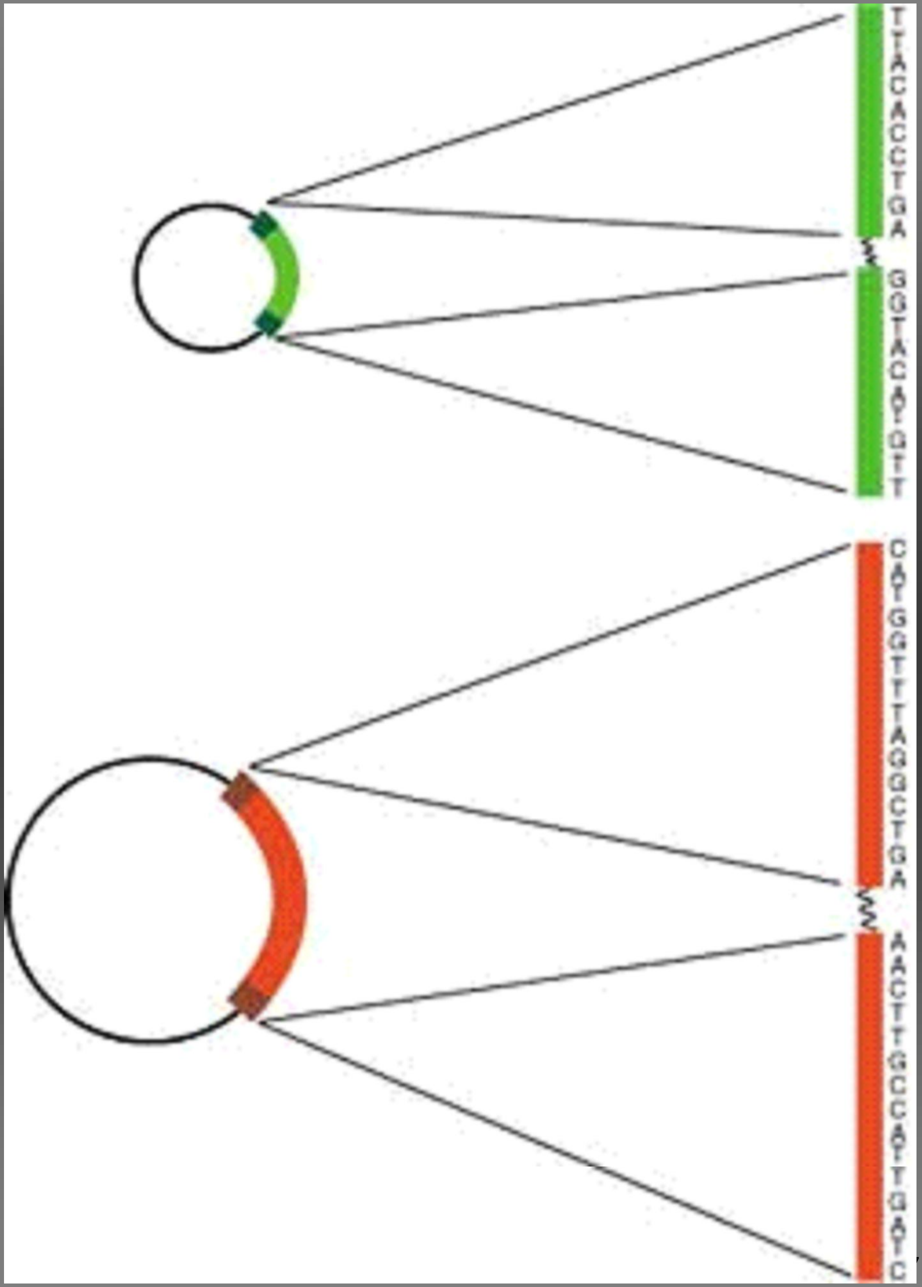


JM-2007

Whole Genome Shotgun - 4

Both plasmid libraries are sequenced.

500 bp from each end of each fragment are decoded generating millions of sequences. Sequencing both ends of each insert is critical for assembling the entire chromosome.

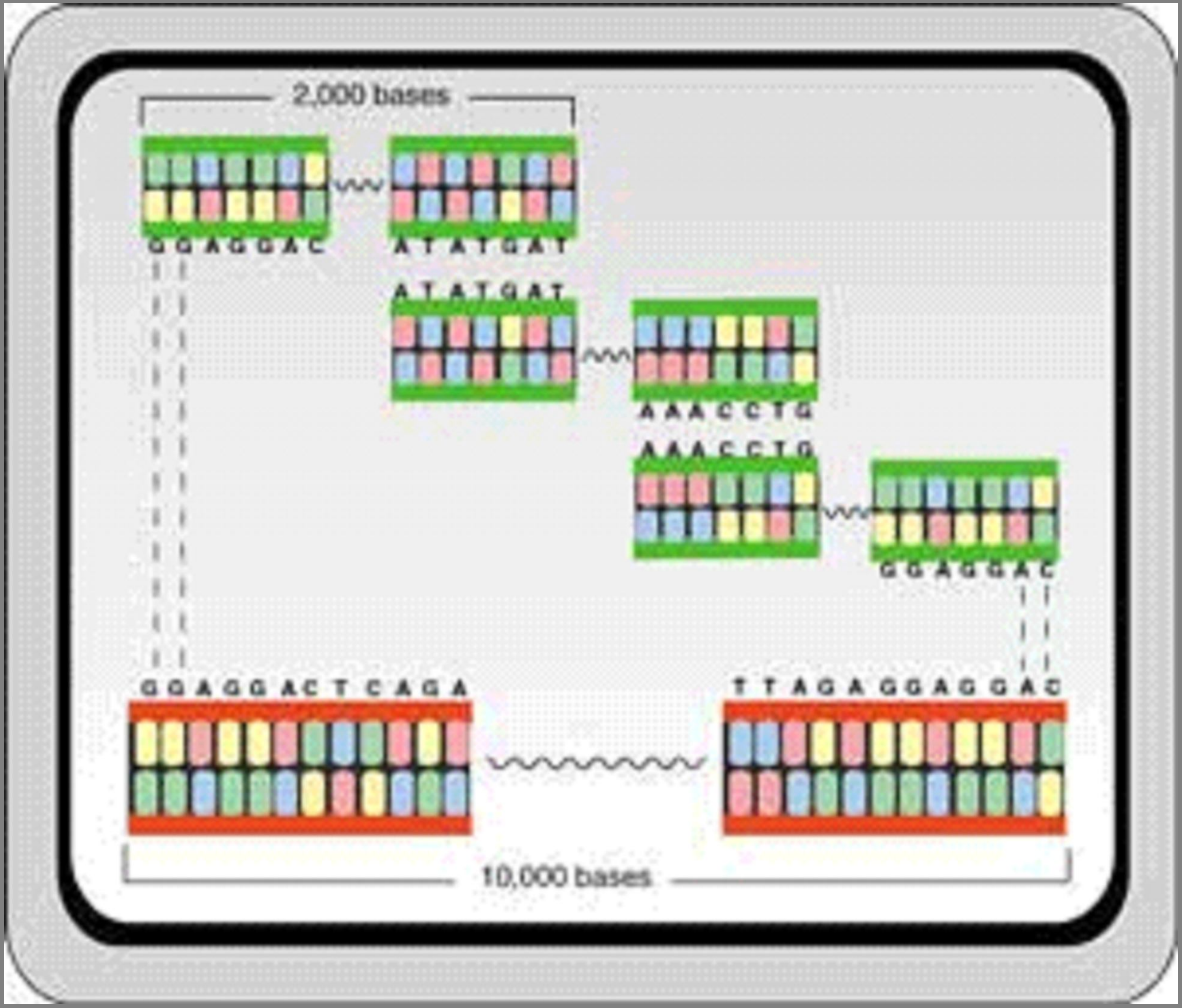


Whole Genome Shotgun - 4

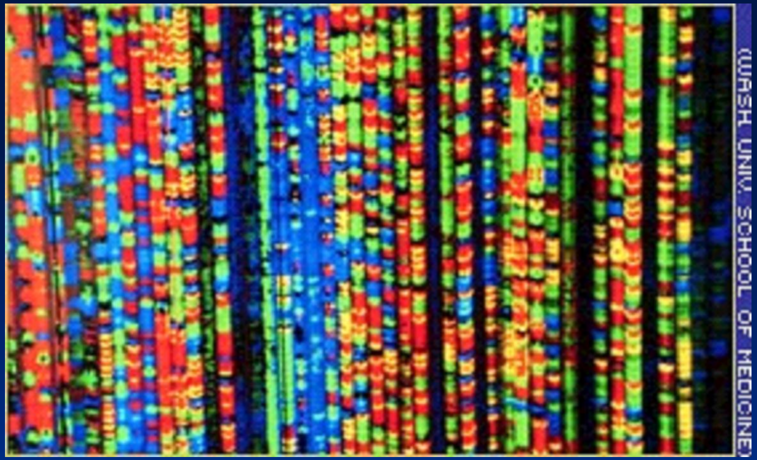
Whole Genome Shotgun - 5

1 Computer algorithms assemble the millions of sequenced fragments into a continuous stretch resembling each chromosome.

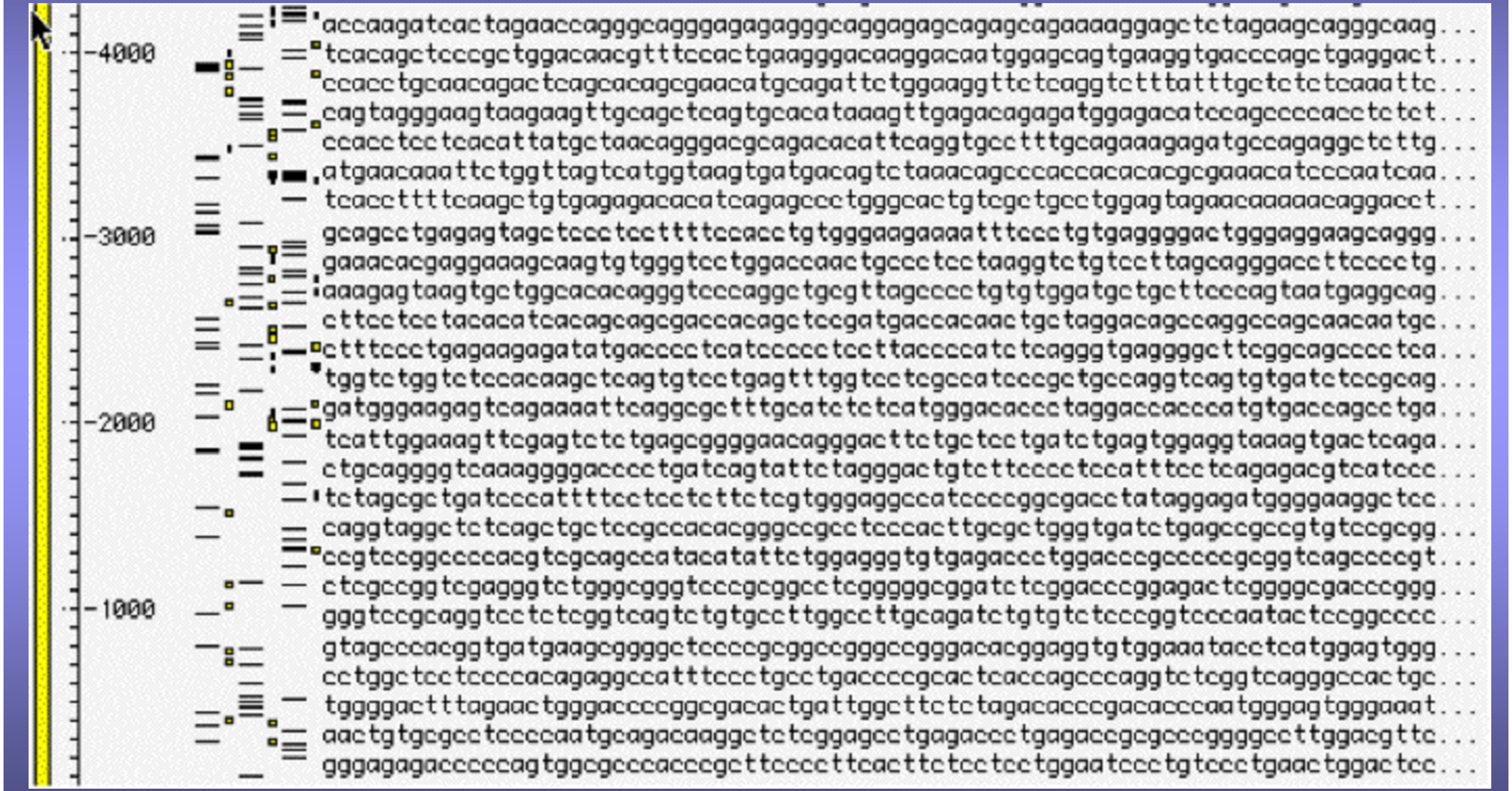
Whole Genome Shotgun - 5



- Genome sequencing factories churn out raw sequence data at an ever increasing rate
- Fewer scientists are involved in generating data and more are involved in data analysis



Raw Genome Data:



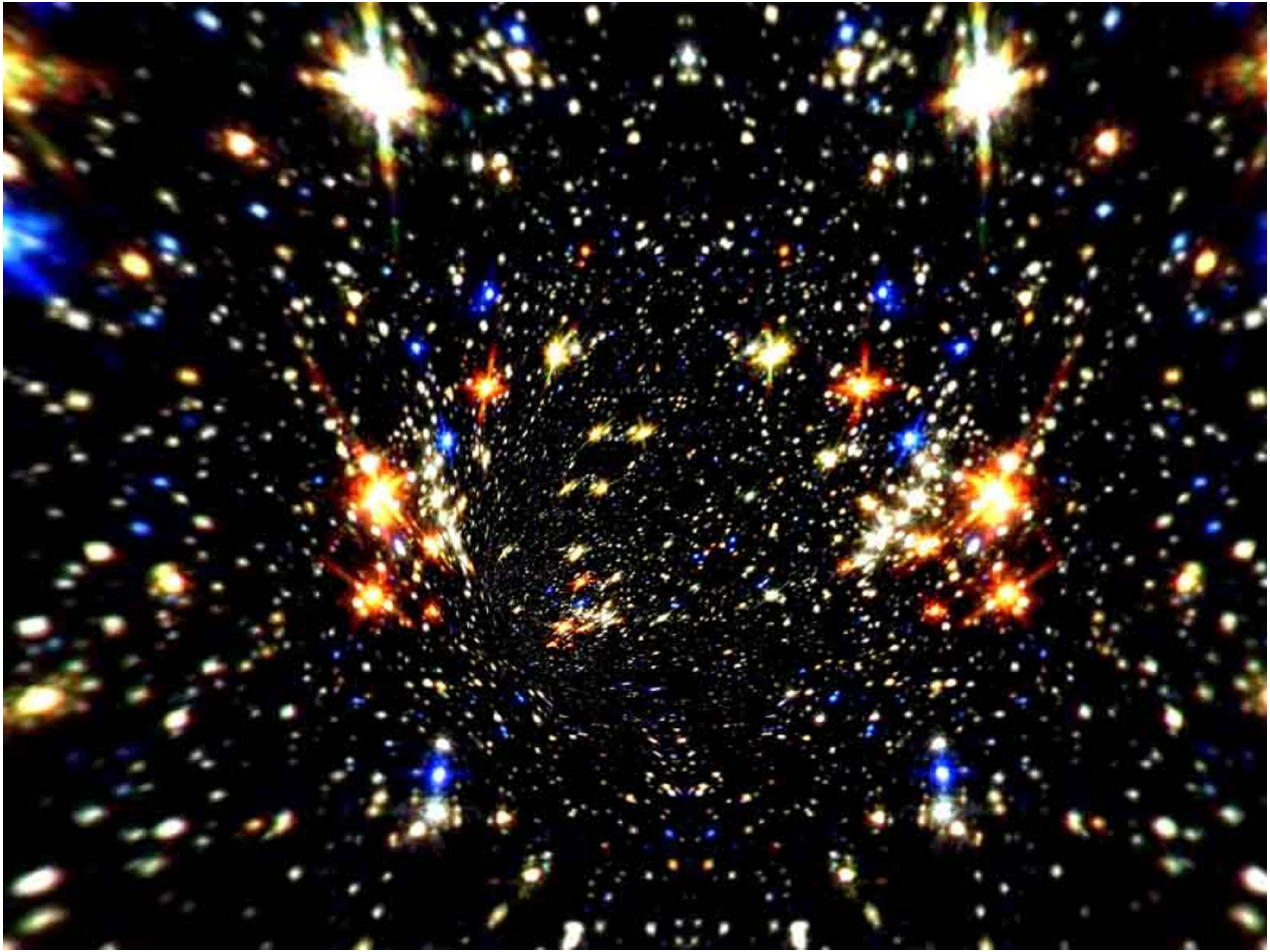
```
accaagatcactagaaccagggcagggagagagggcaggagagcagagcagaaaaggagctctagaagcagggcaag...
tcacagctcccgctggacaacgtttccactgaagggacaaggacaatggagcagtgaaggtgacccagctgaggact...
ccacctgcaacagactcagcacagcgaacatgcagattctggaaggttctcaggtctttatttgctctctcaaattc...
cagtagggaagtaagaagttgcagctcagtgcacataaagttgagacagagatggagacatccagccccacctctct...
ccacctcctcacattatgctaacagggacgcagacacattcaggtgcctttgcagaaagagatgccagaggctcttg...
atgaacaaattctggttagtcatggtaagtgatgacagtctaaacagcccaccacacacgcgaaacatcccaatcaa...
tcacctttttcaagctgtgagagacacatcagagccctgggcactgtcgctgcctggagtagaacaaaaacaggacct...
gcagcctgagagtagctccctcctttttccacctgtgggaagaaaatttccctgtgaggggactgggaggaagcaggg...
gaaacacgaggaaagcaagtgtgggtcctggaccaactgccctcctaaggtctgtccttagcagggaccttcccctg...
aaagagtaagtgctggcacacagggtcccaggctgcgttagcccctgtgtggatgctgcttcccagtaatgaggcag...
cttcctcctacacatcacagcagcgaccacagctccgatgaccacaactgctaggacagccaggccagcaacaatgc...
ctttccctgagaagagatatgacccctcatcccccctccttaccccatctcagggtgaggggcttcggcagccccctca...
tggtctggtctccacaagctcagtgtcctgagtttggtcctcgccatcccgctgccaggtcagtgtgatctccgcag...
gatgggaagagtcagaaaattcaggcgctttgcatctctcatgggacaccctaggaccacccatgtgaccagcctga...
tcattggaaagttcgagtctctgagcggggaacagggacttctgctcctgatctgagtggaggtaaagtgactcaga...
ctgcaggggtcaaaggggacccctgatcagtattctagggactgtcttcccctccatttcctcagagacgtcatccc...
tctagcgctgatcccatttttcctcctcttctcgtgggaggccatccccggcgacctataggagatggggaaggctcc...
caggtaggctctcagctgctccgccacacgggccgcctcccacttgcgctgggtgatctgagccgccgtgtccgcgg...
ccgtccggccccacgtcgcagccatacatattctggagggtgtgagaccctggacccgcccccgcggtcagccccgt...
ctcgccggtcgagggtctgggcgggtcccgcggcctcgggggcggatctcggacccggagactcggggcgacccggg...
gggtccgcaggtcctctcggtcagtctgtgccttggccttgcagatctgtgtctcccggtcccaatactccggcccc...
gtagcccacggtgatgaagcggggctccccgcggccgggccgggacacggaggtgtggaaatacctcatggagtggg...
cctggctcctccccacagaggccatttccctgcctgaccccgcactcaccagcccaggtctcggtcagggccactgc...
tggggactttagaactgggaccccggcgacactgattggcttctctagacacccgacacccaatgggagtgggaaat...
aactgtgcgcctccccaatgcagacaaggctctcggagcctgagacccctgagaccgcgcccggggccttggacgttc...
gggagagacccccagtggcgcccacccgcttccccttcacttctcctcctggaatccctgtccctgaactggactcc...
```

Finding genes in genome sequence is not easy

- About 1% of human DNA encodes functional genes.

- Genes are interspersed among long stretches of non-coding DNA.

- Repeats, pseudo-genes, and introns confound matters



RFLP

(Restriction Fragment Length Polymorphisms)

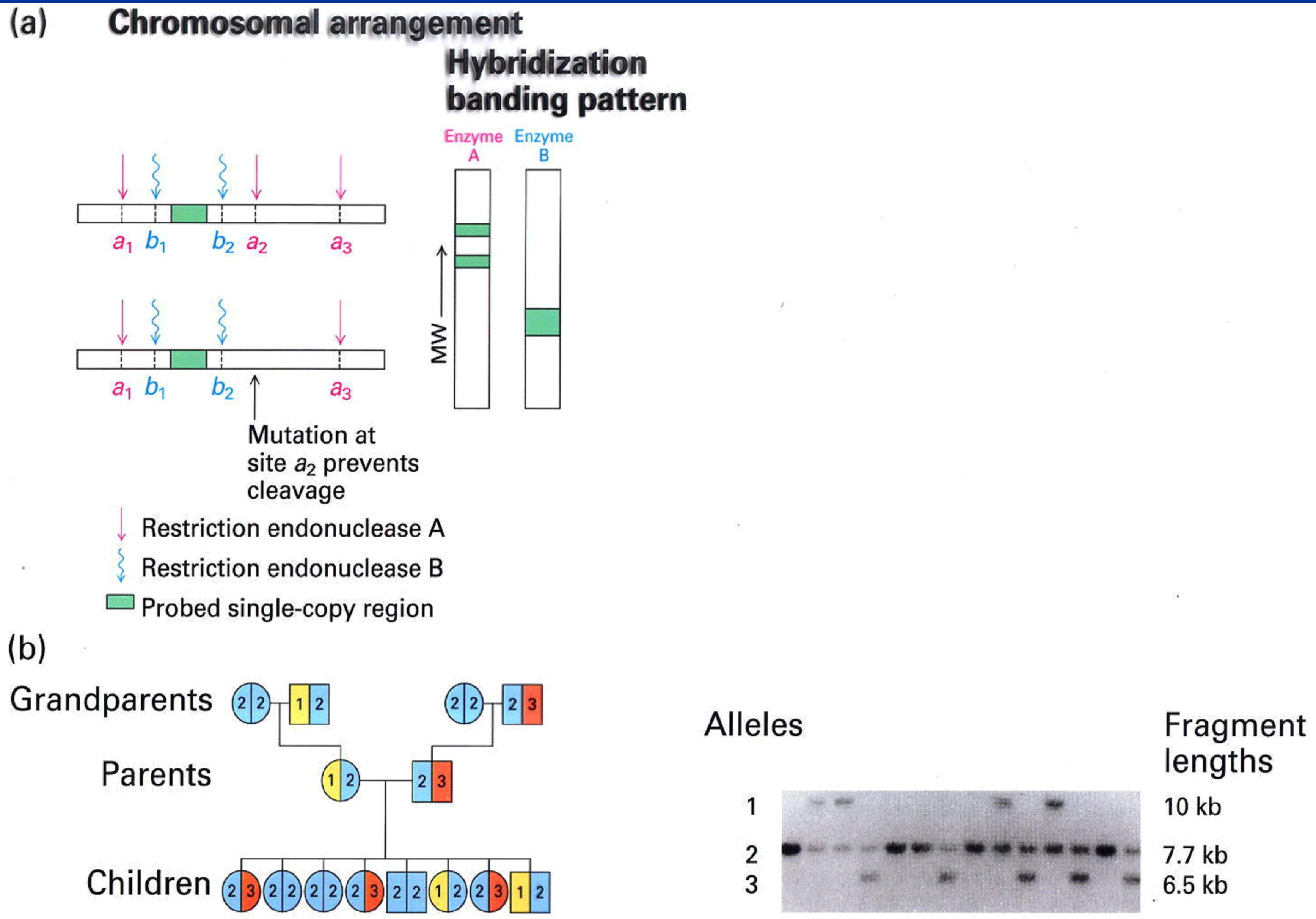


Figure 8-20
Lodish et al. MOLECULAR CELL BIOLOGY, Fourth Edition

SNPs are Very Common

- SNPs are very common in the human population.
- Between any two people, there is an average of one SNP every ~1250 bases.
- Most of these have no phenotypic effect
  - Venter et al. estimate that only <1% of all human SNPs impact protein function (non-coding regions)
  - Selection against mis-sense mutations
- Some are alleles of genes.

Genome Sequencing finds SNPS

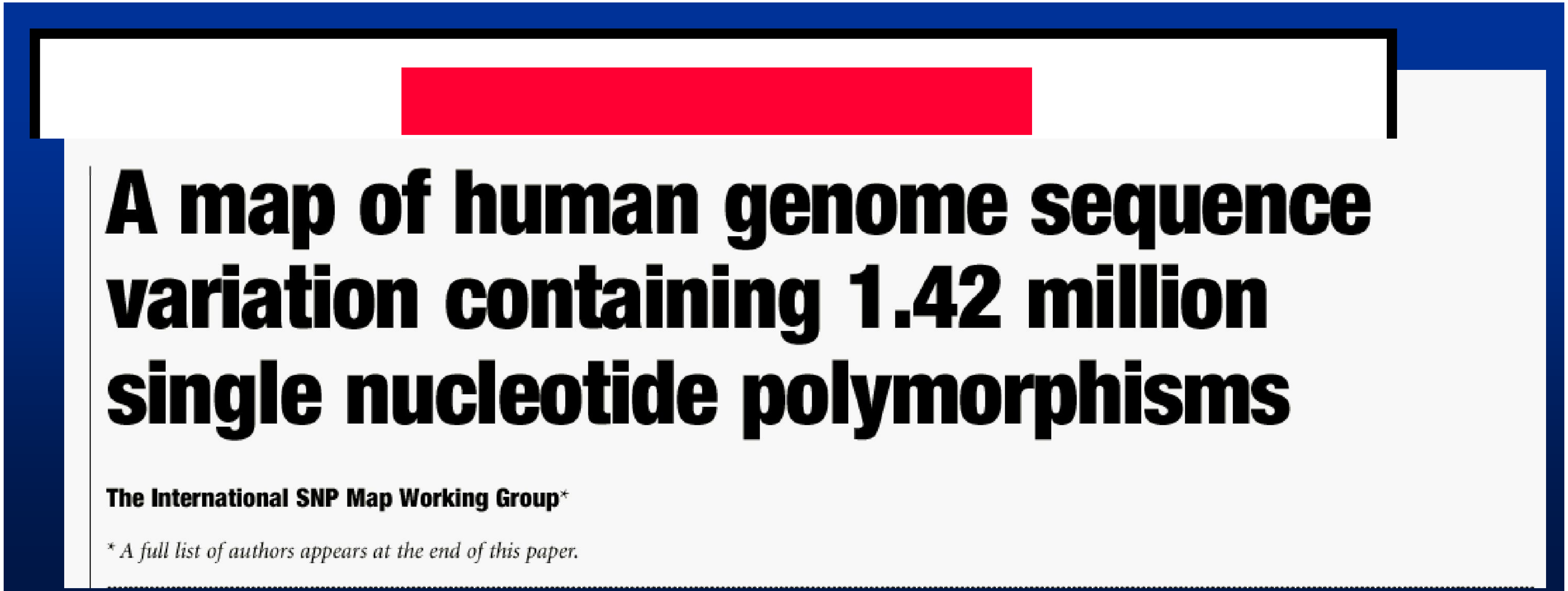
- The Human Genome Project involves sequencing DNA cloned from a number of different people.

  [The Celera sequence comes from 5 people]

- Even in a library made from one person’s DNA, the homologous chromosomes have SNPs

- This inevitably leads to the discovery of SNPs - any single base sequence difference

- These SNPs can be valuable as the basis for diagnostic tests



A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms

The International SNP Map Working Group*

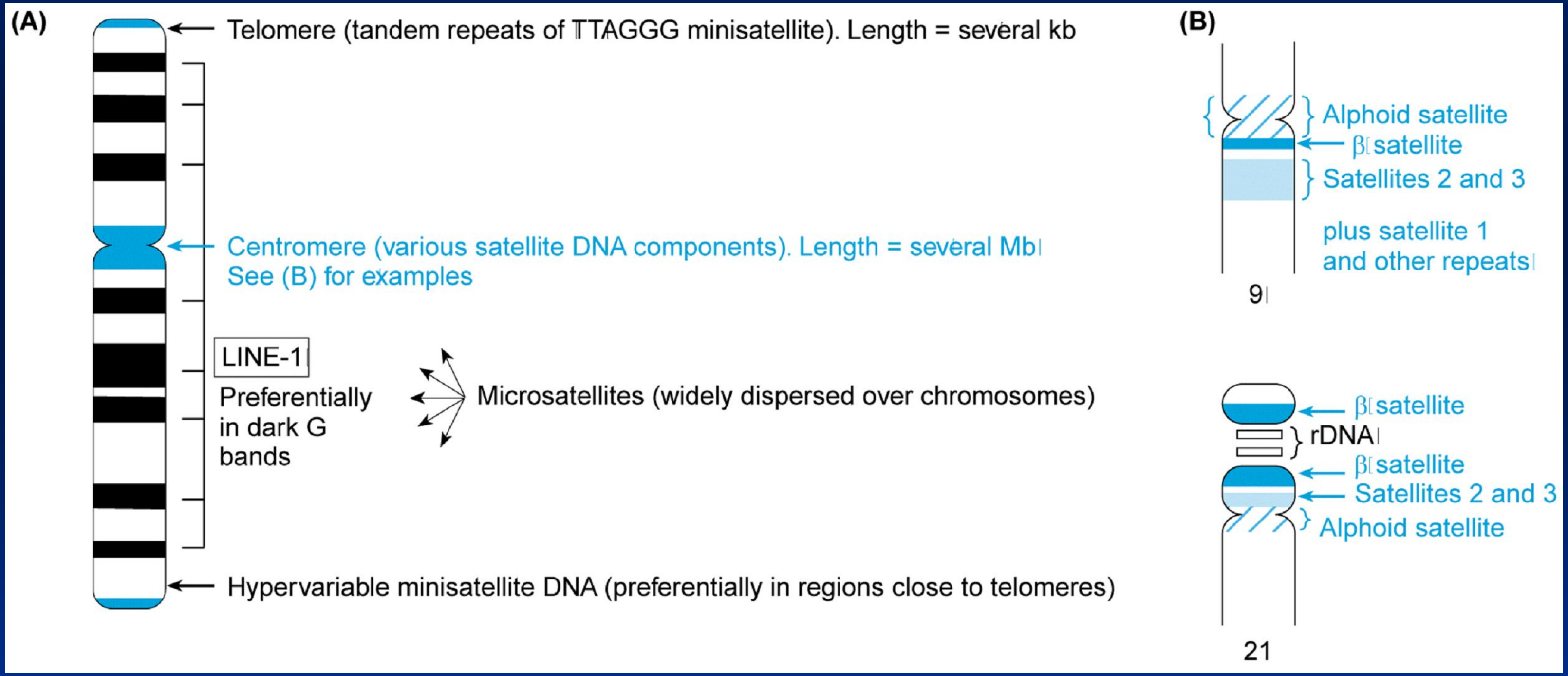
** A full list of authors appears at the end of this paper.*

We describe a map of 1.42 million single nucleotide polymorphisms (SNPs) distributed throughout the human genome, providing an average density on available sequence of one SNP every 1.9 kilobases. These SNPs were primarily discovered by two projects: The SNP Consortium and the analysis of clone overlaps by the International Human Genome Sequencing Consortium. The map integrates all publicly available SNPs with described genes and other genomic features. We estimate that 60,000 SNPs fall within exon (coding and untranslated regions), and 85% of exons are within 5 kb of the nearest SNP. Nucleotide diversity varies greatly across the genome, in a manner broadly consistent with a standard population genetic model of human history. This high-density SNP map provides a public resource for defining haplotype variation across the genome, and should help to identify biomedically important genes for diagnosis and therapy.

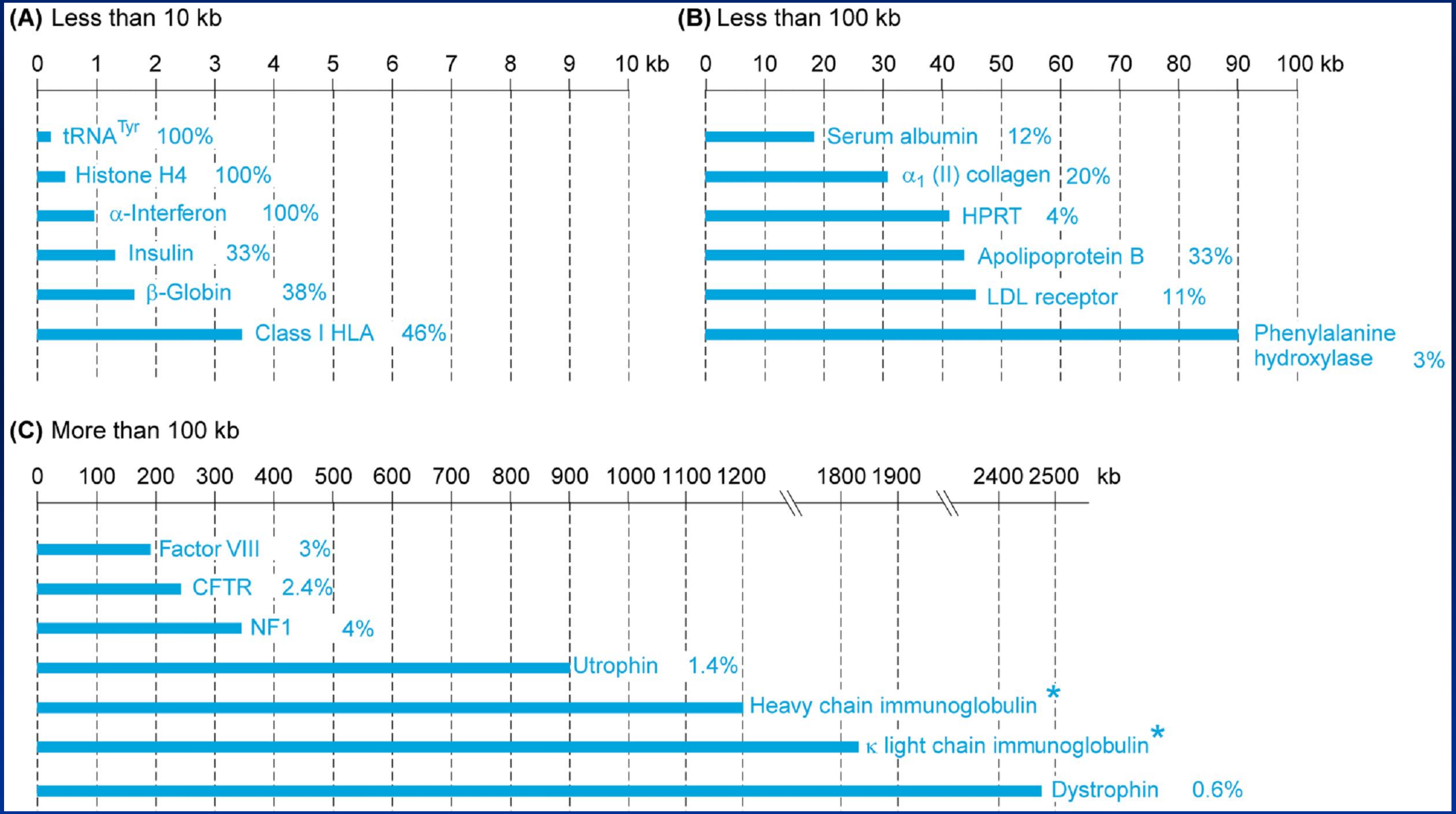
Gene families

- Members may exhibit high sequence homology
- sometimes contain a highly conserved domain (e.g. SOX box)
- sometimes contain a very short conserved "motif" (e.g. DEAD box, asp-glu-ala-asp RNA helicases)
- superfamilies (e.g. Ig superfamily)
- sometimes clustered (e.g. globin genes)
- Often associated with truncated and non processed pseudogenes

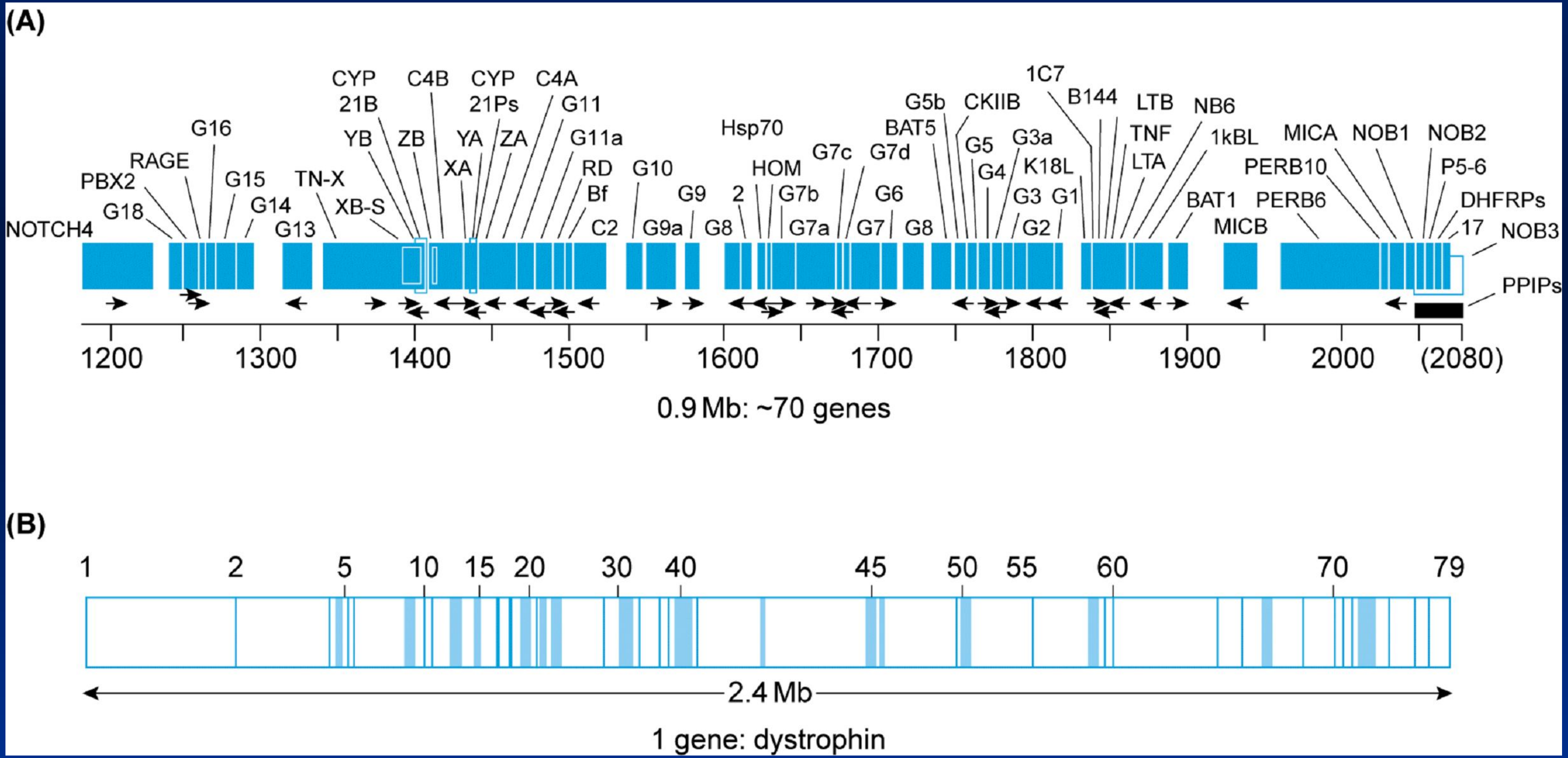
Chromosomal location of repetitive DNA


(A)
Telomere (tandem repeats of TTAGGG minisatellite). Length = several kb
Centromere (various satellite DNA components). Length = several Mb
See (B) for examples
LINE-1
Preferentially in dark G bands
Microsatellites (widely dispersed over chromosomes)
Hypervariable minisatellite DNA (preferentially in regions close to telomeres)
(B)
Alphoid satellite
β satellite
Satellites 2 and 3
plus satellite 1 and other repeats
9
β satellite
rDNA
β satellite
Satellites 2 and 3
Alphoid satellite
21

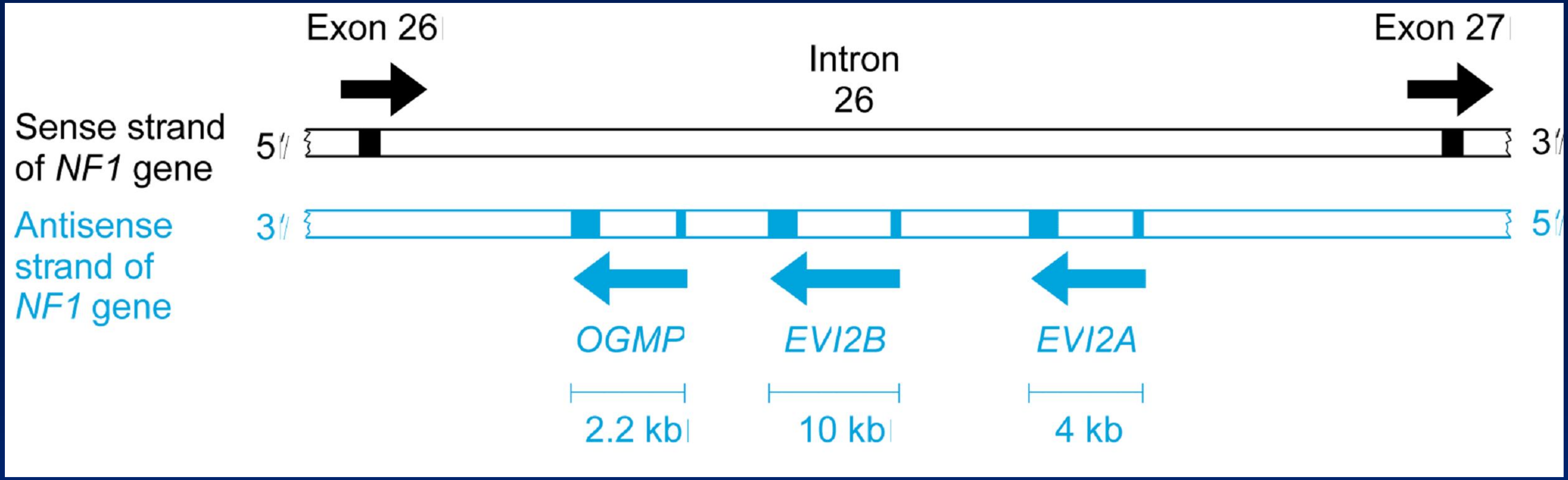
Human genes vary enormously in size and exon content



GENE DENSITIES

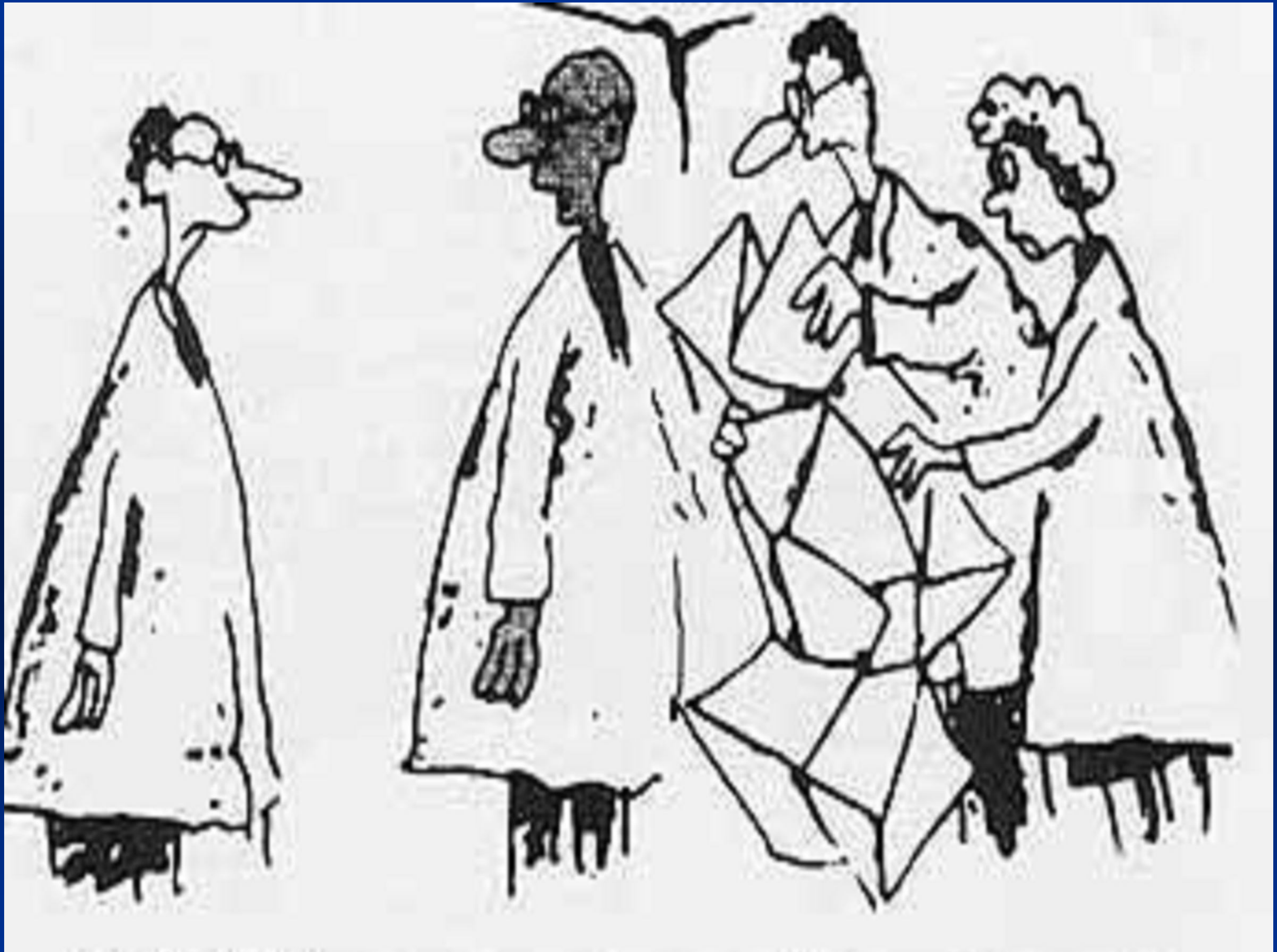


GENES WITHIN GENES



Proteomics

- Identify all of the proteins in an organism
  - Potentially many more than genes due to alternative splicing and post-translational modifications
- Quantitate in different cell types and in response to metabolic/environmental factors
- Protein-protein interactions



"We finished the genomic map, now we can't figure out how to fold it."