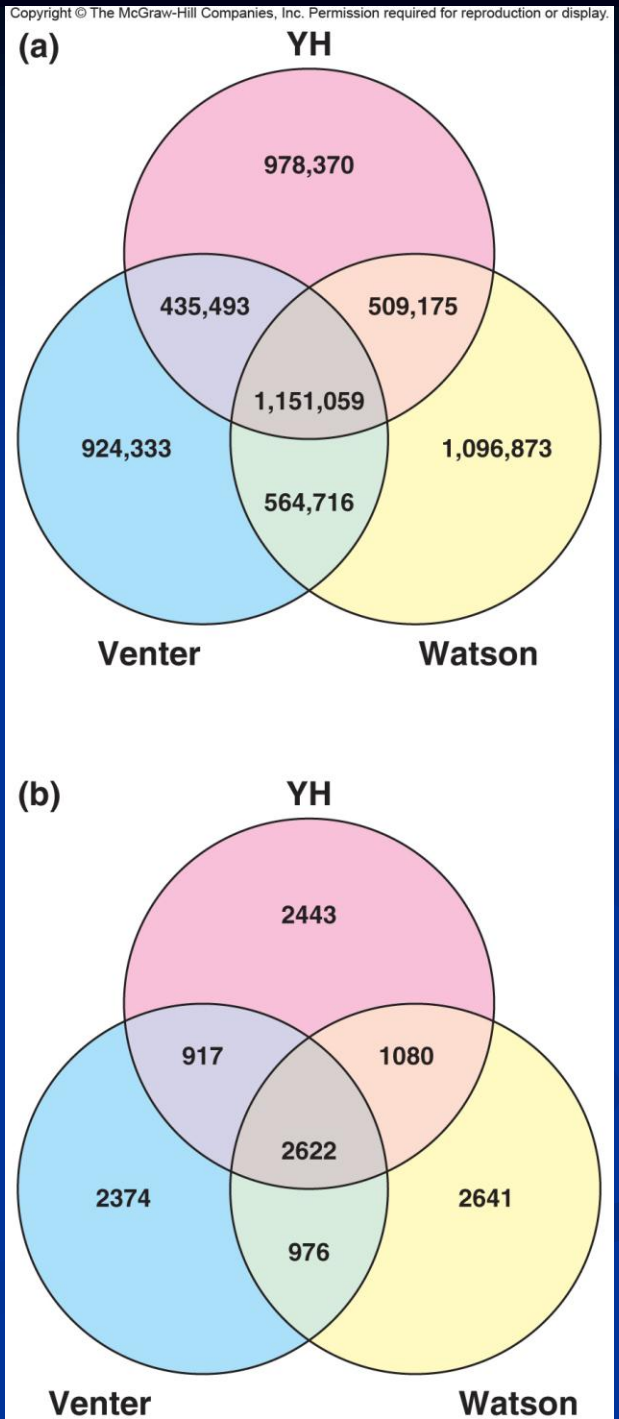
- Three personal genomes: Craig Venter, James Watson, YH (Anonymous Chinese man)
- Single nucleotide substitutions



Differences across genomes

Amino acid changing differences only

Fig. 10.4 p. 344

Genetic mapping

- Based on recombination frequencies
  - The further away two points are on a chromosome, the higher the frequency of recombination between them
  - Determine a relative position for each locus

Genetic markers

- Genetic mapping between positions on chromosomes
  - Positions can be genes
    - Responsible for phenotype like eye color or disease traits
  - Positions can be other physical markers
    - DNA sequence variation

Four classes of DNA polymorphisms or variants

p. 345

Copyright © McGraw-Hill Education. Permission required for reproduction or display.

TABLE 10.1 Categories of Genetic Variants

The right column shows how frequently on average you would find a polymorphism of the indicated class when comparing any two haploid human genomes.

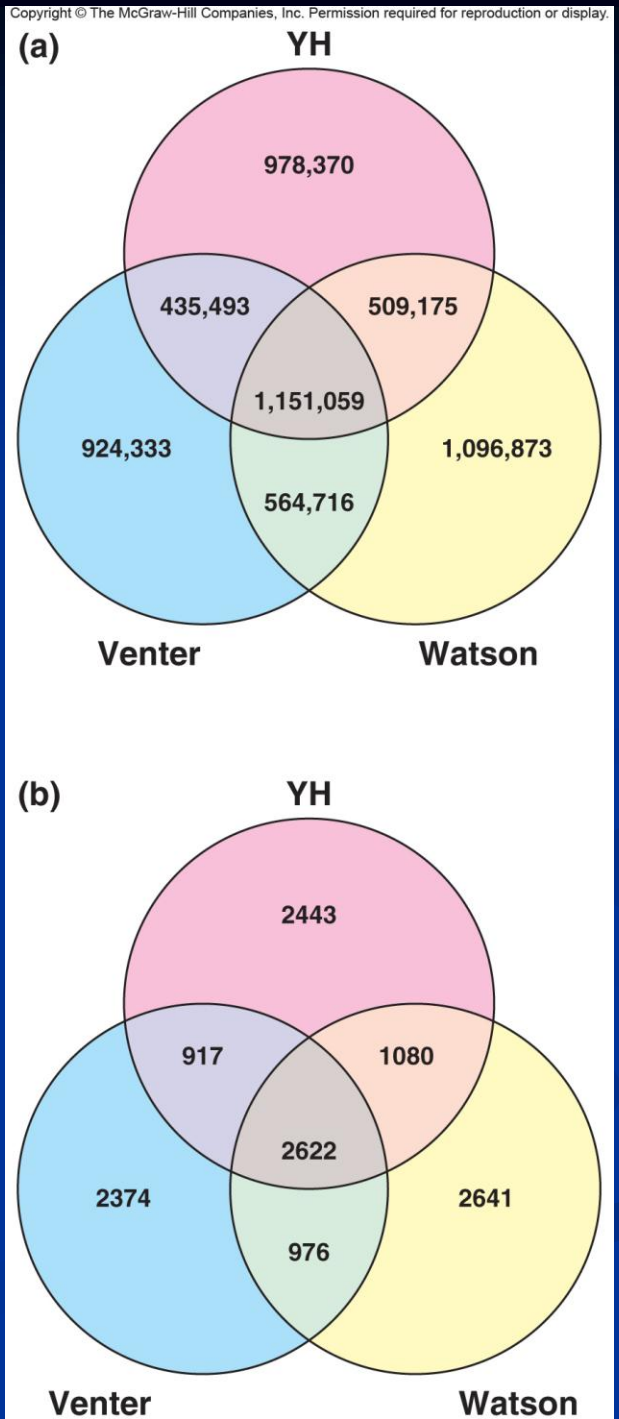
	Size	Frequency (1 per. . . .)
SNP Single nucleotide polymorphism	1 bp	1 kb
DIP or **Indel** Insertion/deletion	1–100 bp	10 kb
SSR or **Microsatellite** Simple sequence repeat	1–10 bp repeat unit	30 kb
CNV Copy number variant	10 bp–1 Mb	3 Mb

- **Polymorphisms- two or more alleles that are each present in more than 1% of a species' population**
- **Different alleles of a polymorphic locus are genetic variants**

Single nucleotide polymorphisms (SNP)

- Single base-pair substitutions
- Arise by mutagenic chemicals or mistakes in replication (1 out of about 10^9 gametes)
- Biallelic – only two alleles
- Do not change after initial mutational event
- Make up about 90% of human genetic variation (18 million SNPs)
- Most occur at anonymous loci
- Occur about 1 per kb

- Three personal genomes: Craig Venter, James Watson, YH (Anonymous Chinese man)
- Single nucleotide substitutions

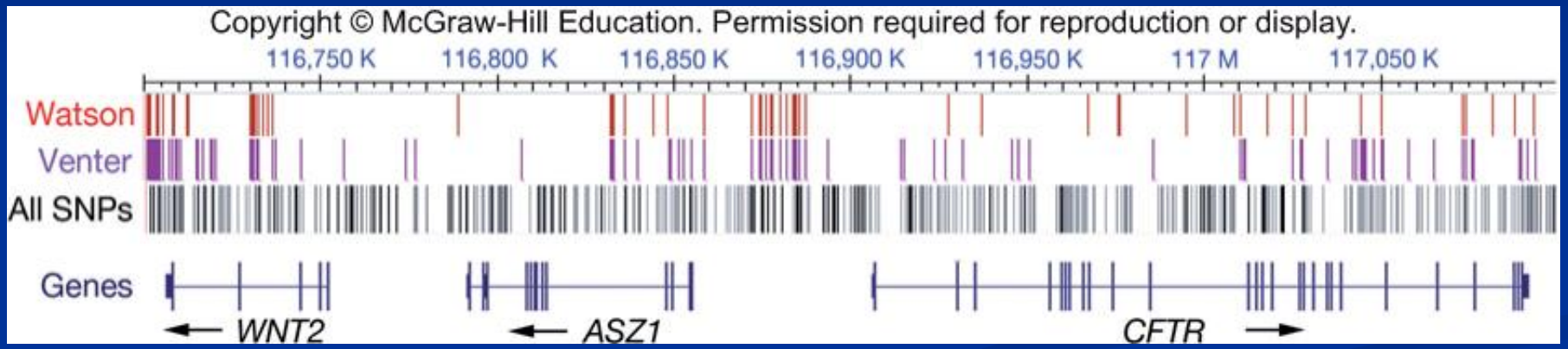


Differences across genomes

Amino acid changing differences only

Fig. 10.4 p. 344

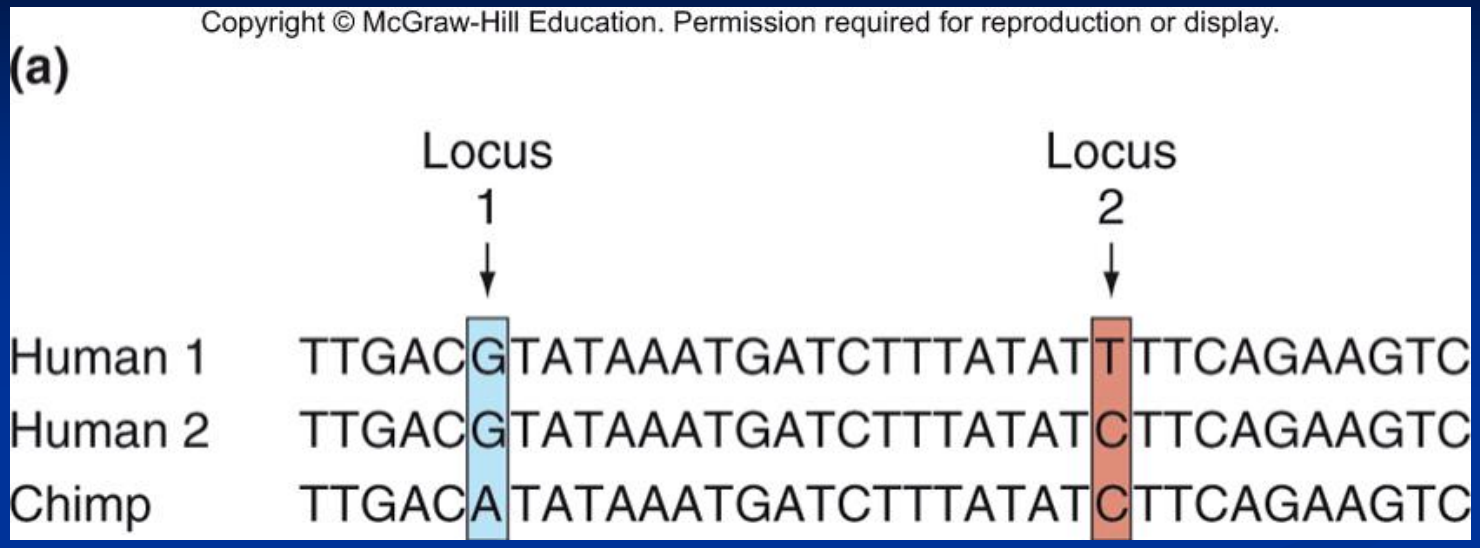
Single nucleotide polymorphisms (SNPs)



- Vertical marks indicate locations for SNPs different from the human RefSeq
- All SNPs in database as of 2009

Fig. 10.5 p. 344

Single nucleotide polymorphisms (SNPs)



Inferring evolutionary history of SNPs

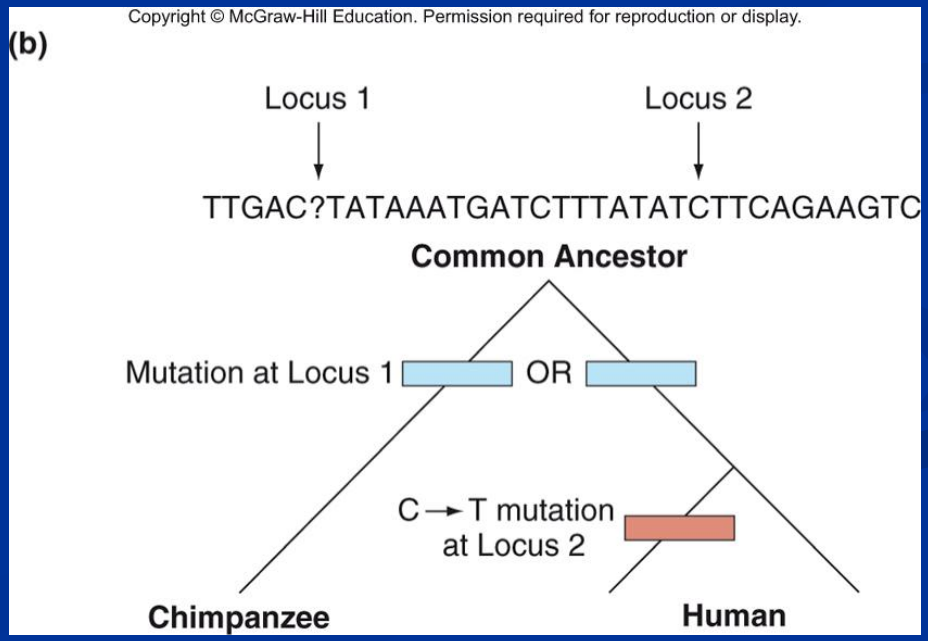
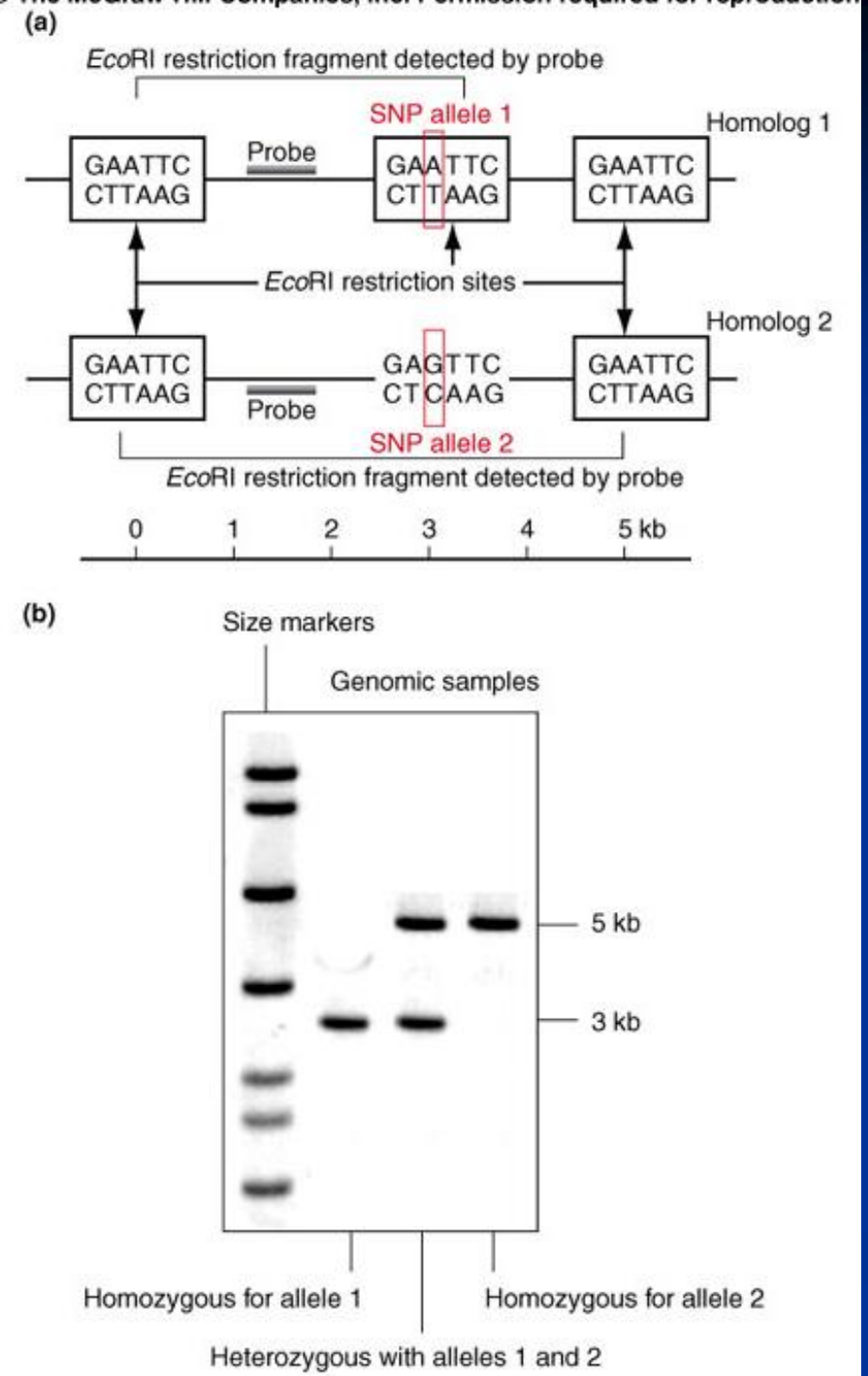


Fig. 10.6 p. 345

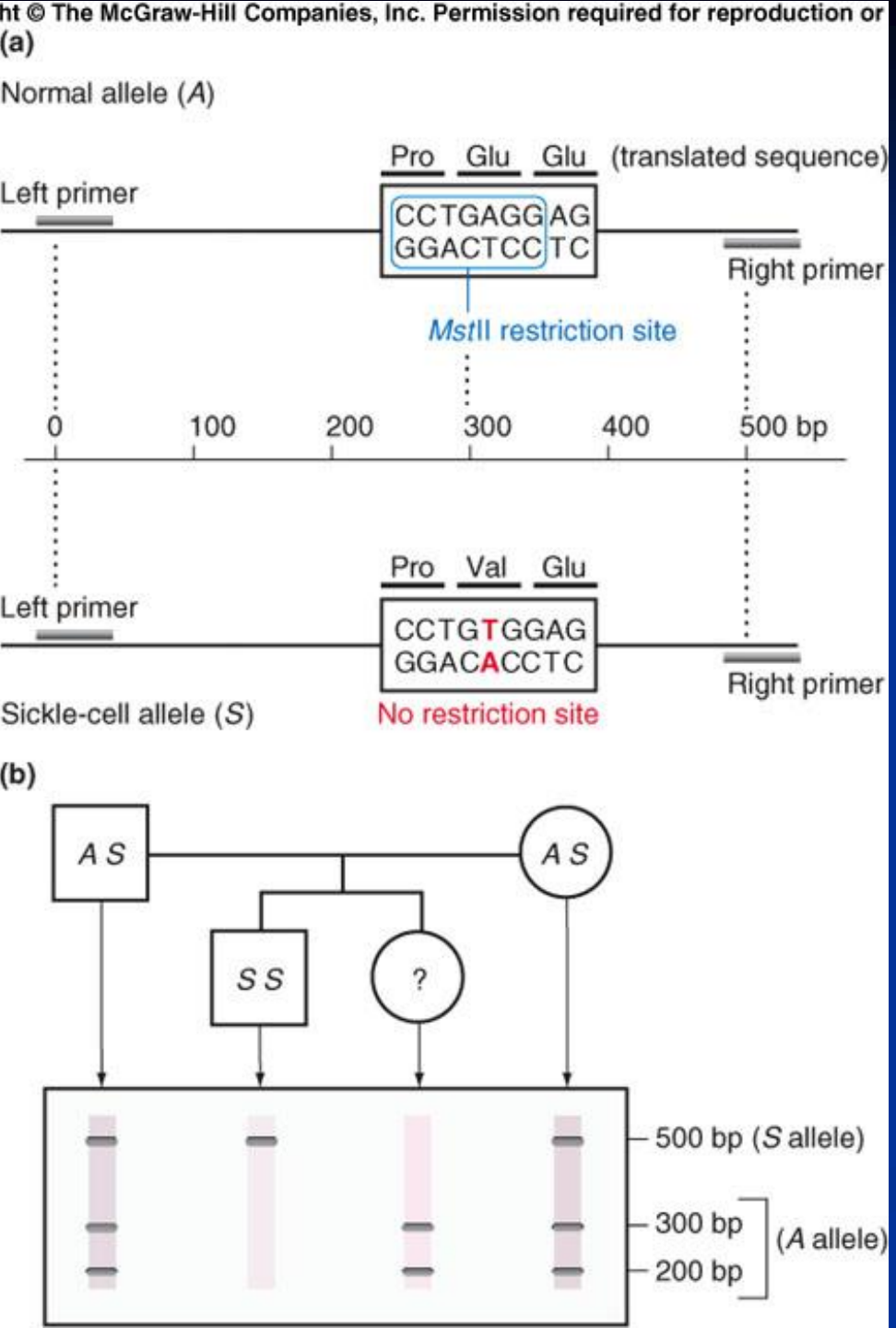
© The McGraw-Hill Companies, Inc. Permission required for reproduction



SNP detection using Southern blots

- RFLPs (restriction fragment length polymorphisms)
- Changes in fragment lengths due to the loss or gain of a restriction site

ht © The McGraw-Hill Companies, Inc. Permission required for reproduction or



SNP detection by PCR

- Must know sequence on either side of polymorphism
  - Amplify fragment
  - Expose to restriction enzyme
  - Gel electrophoresis

SNP detection by sequencing PCR

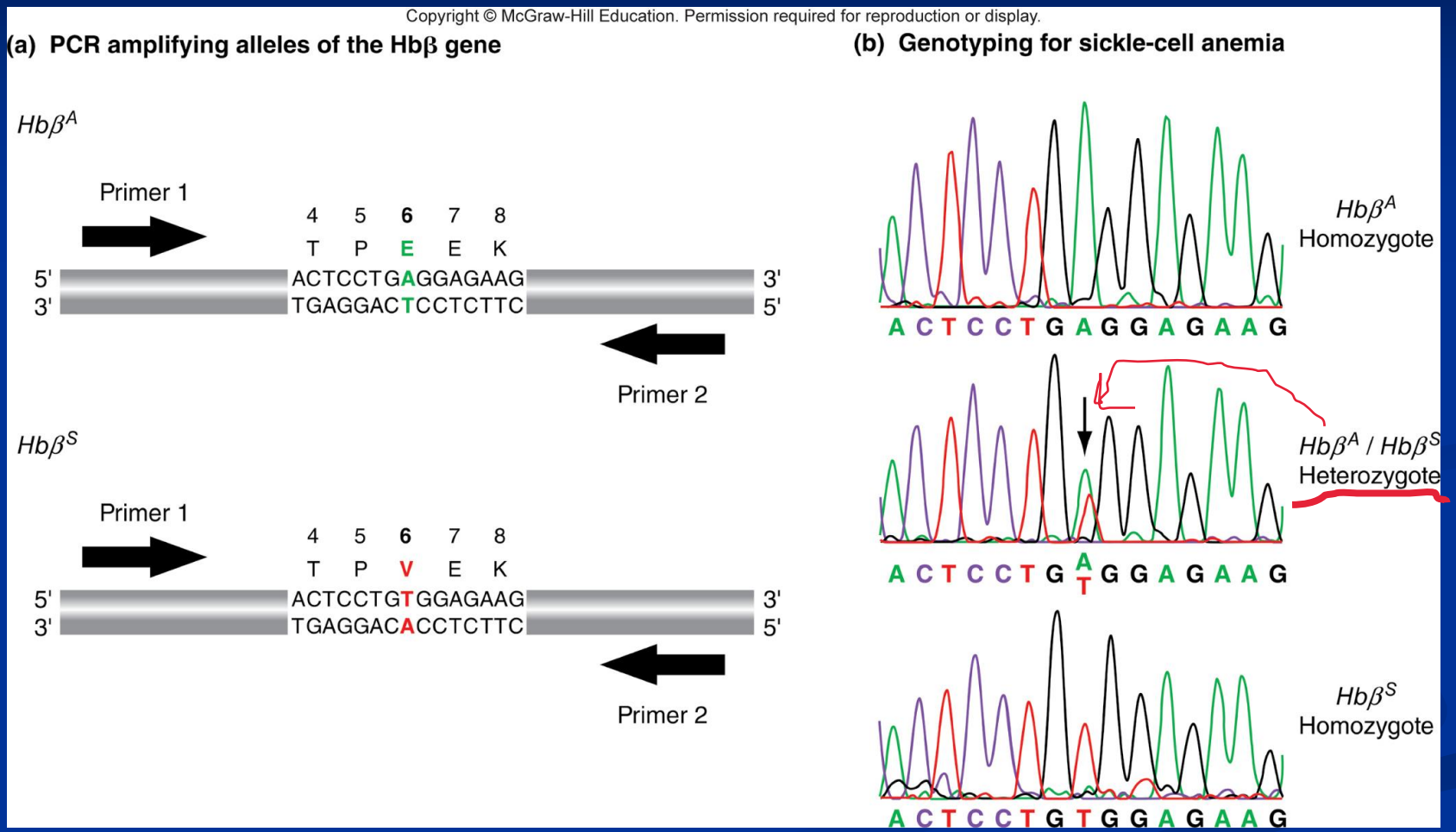
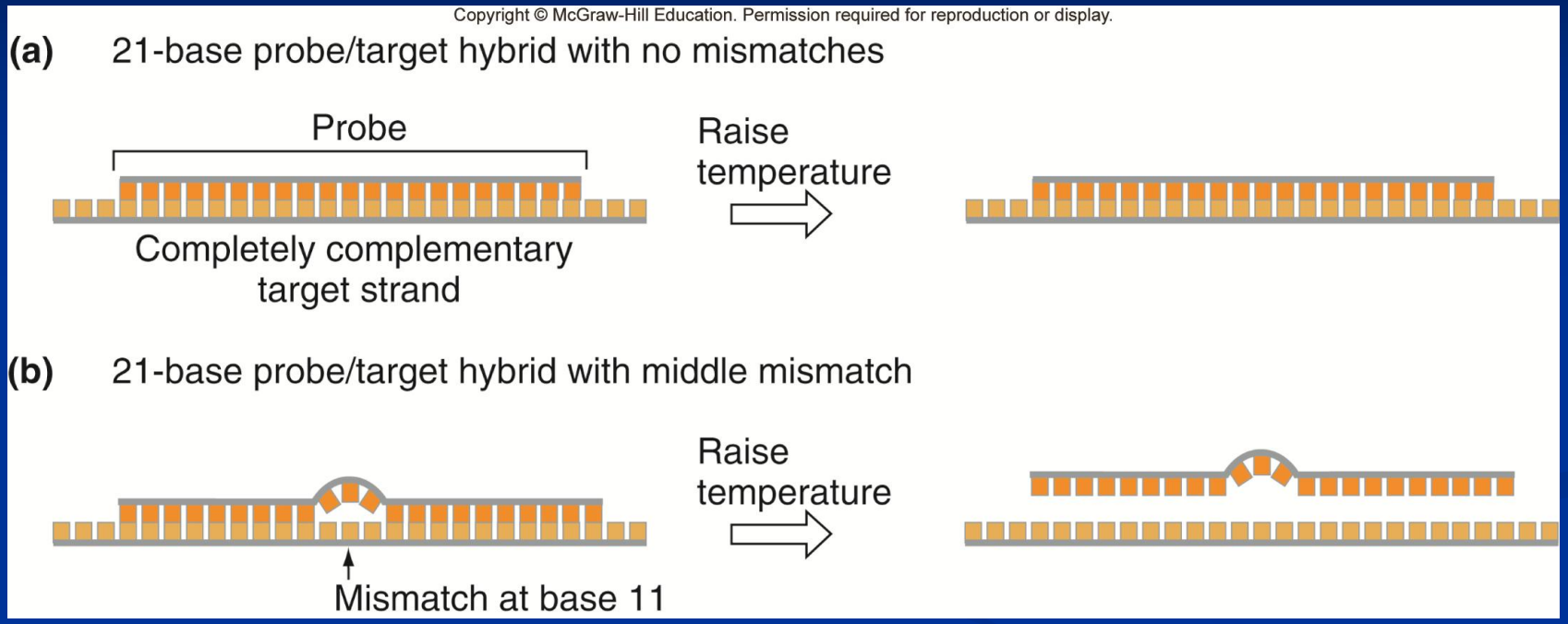


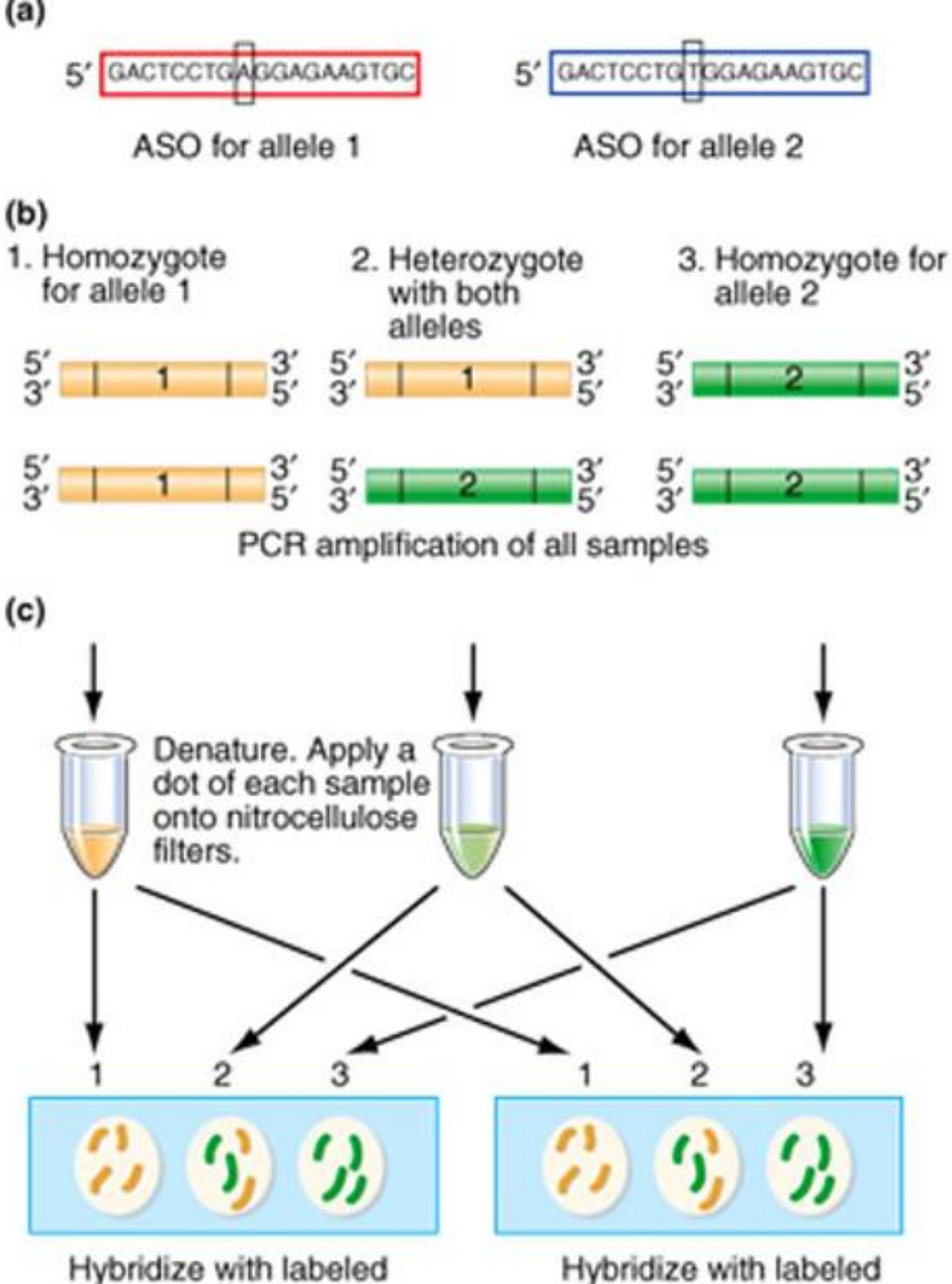
Fig. 10.11 p. 350

SNP detection by ASO



- **Very short probes (<21 bp) that hybridize to one allele or other**
- **Such probes are allele-specific oligonucleotides (ASOs)**

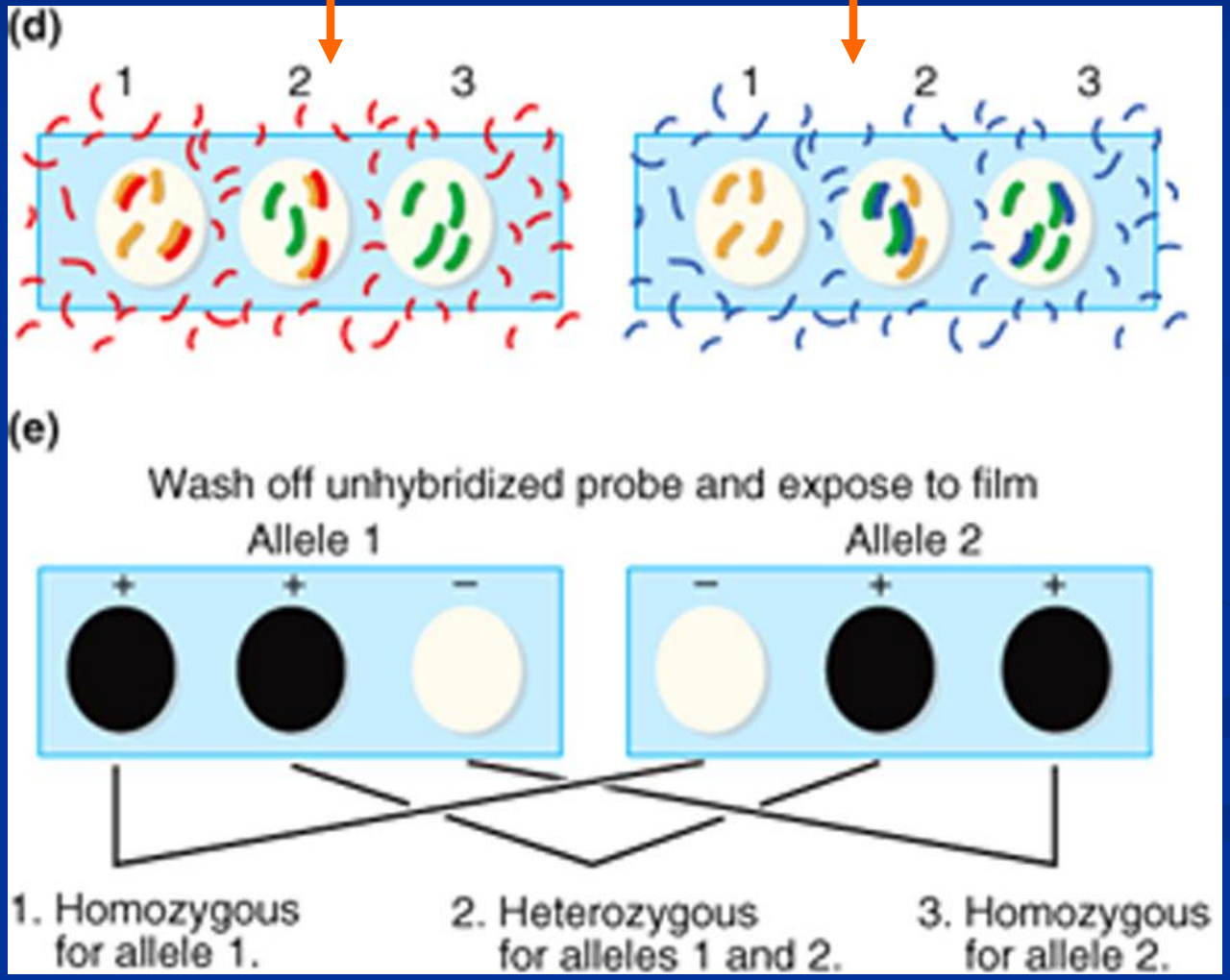
Fig. 10.16 p. 355



ASOs can determine genotype at any SNP locus

Hybridized and labeled with ASO for allele 1

Hybridized and labeled with ASO for allele 2



Preimplantation embryo diagnosis of CF using ASO analysis

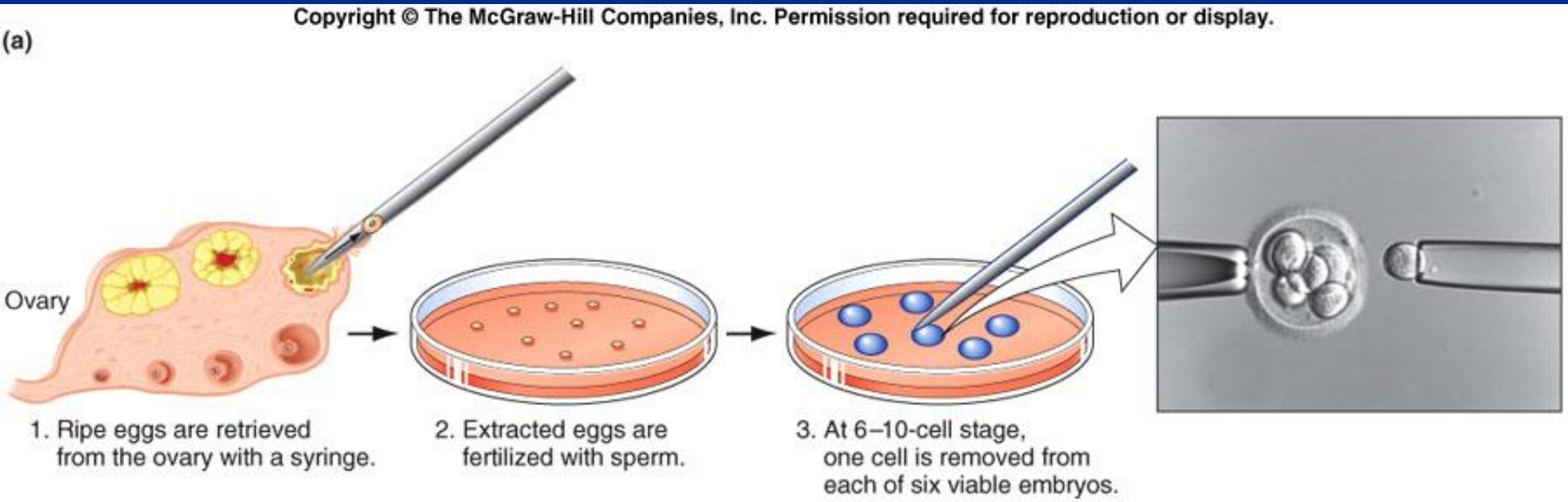


Fig. 10.14 p. 353

Preimplantation embryo diagnosis of CF using ASO analysis

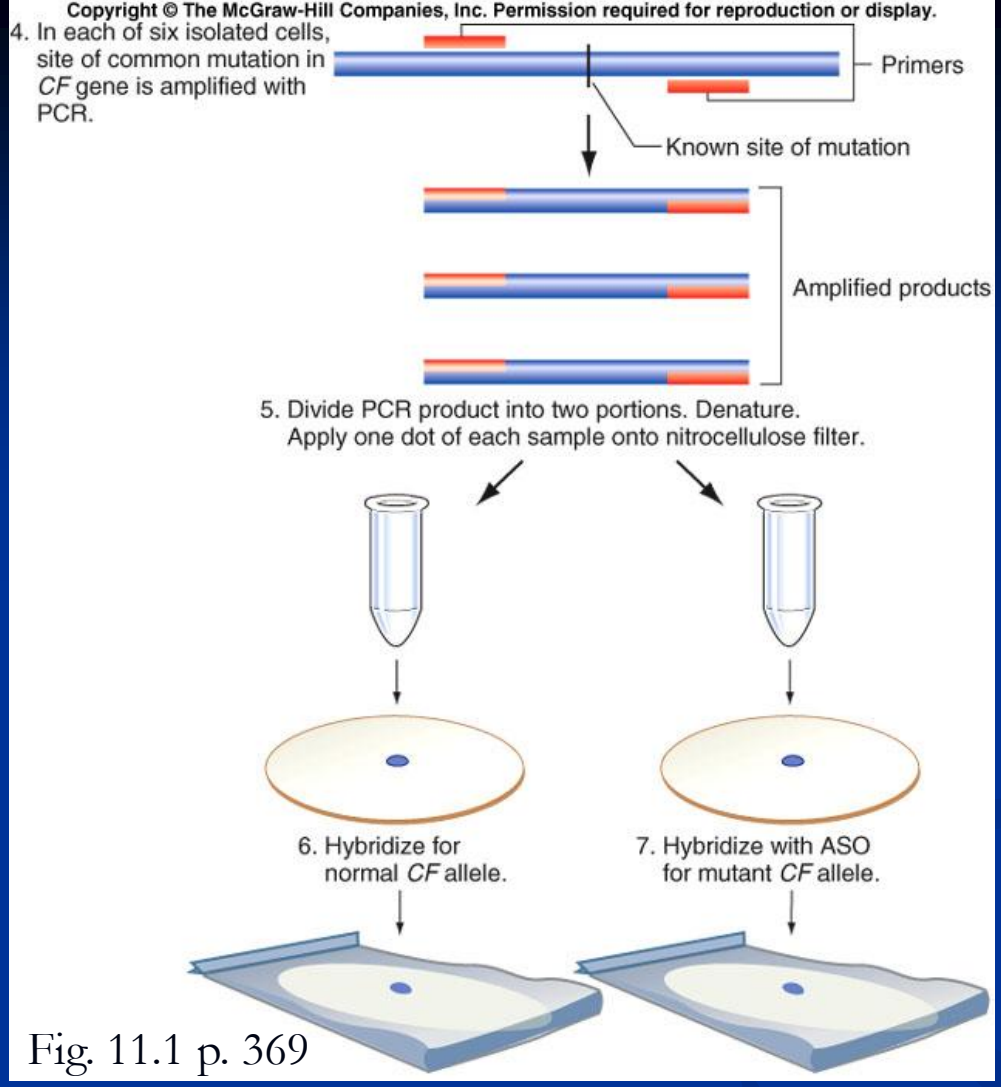


Fig. 11.1 p. 369

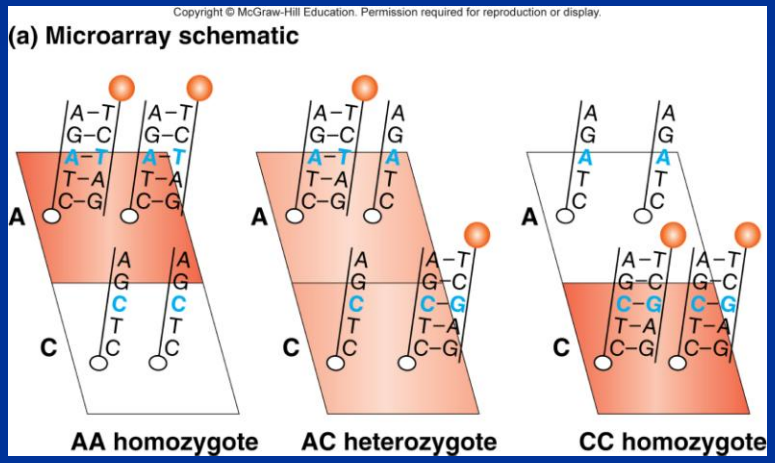
Copyright © The McGraw-Hill Companies, Inc. Permission required for reproduction or display.

(b) Diagnosis

	Cell 1	Cell 2*	Cell 3	Cell 4*	Cell 5	Cell 6
Normal ASO	○	●	○	●	●	○
Mutant ASO	●	○	●	●	○	●

* Cells from embryos later transplanted into uterus

Oligonucleotide arrays



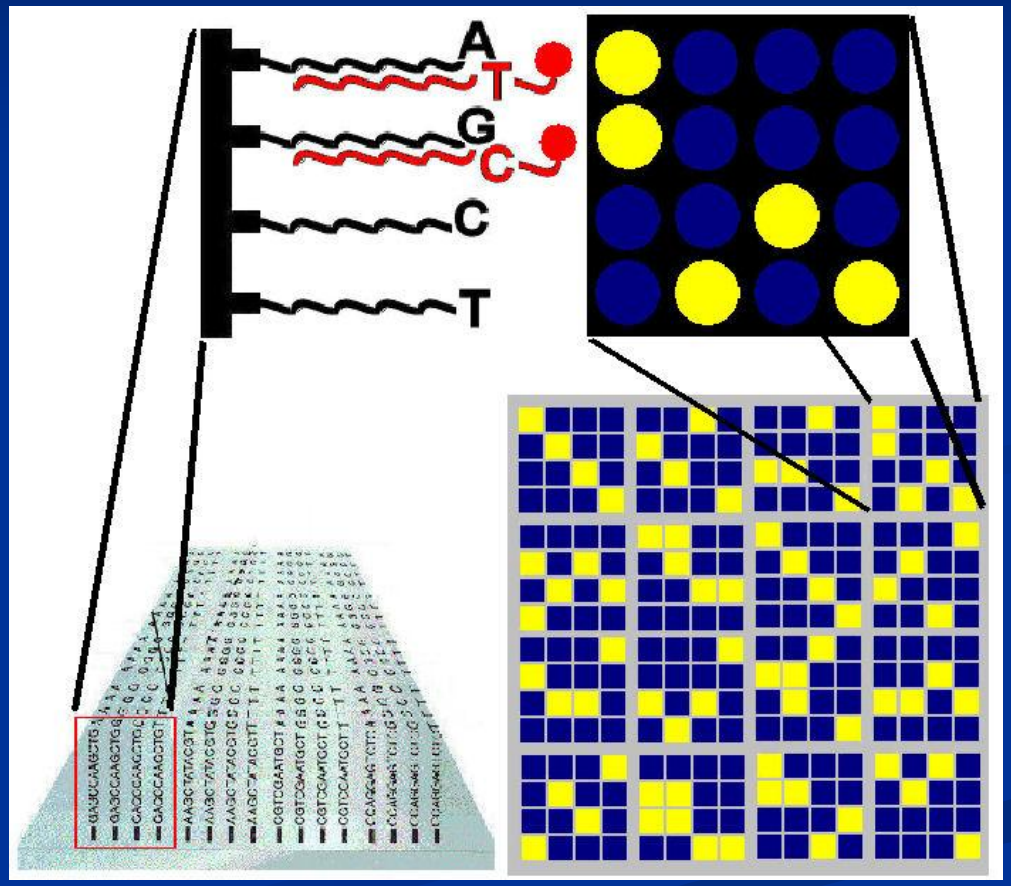
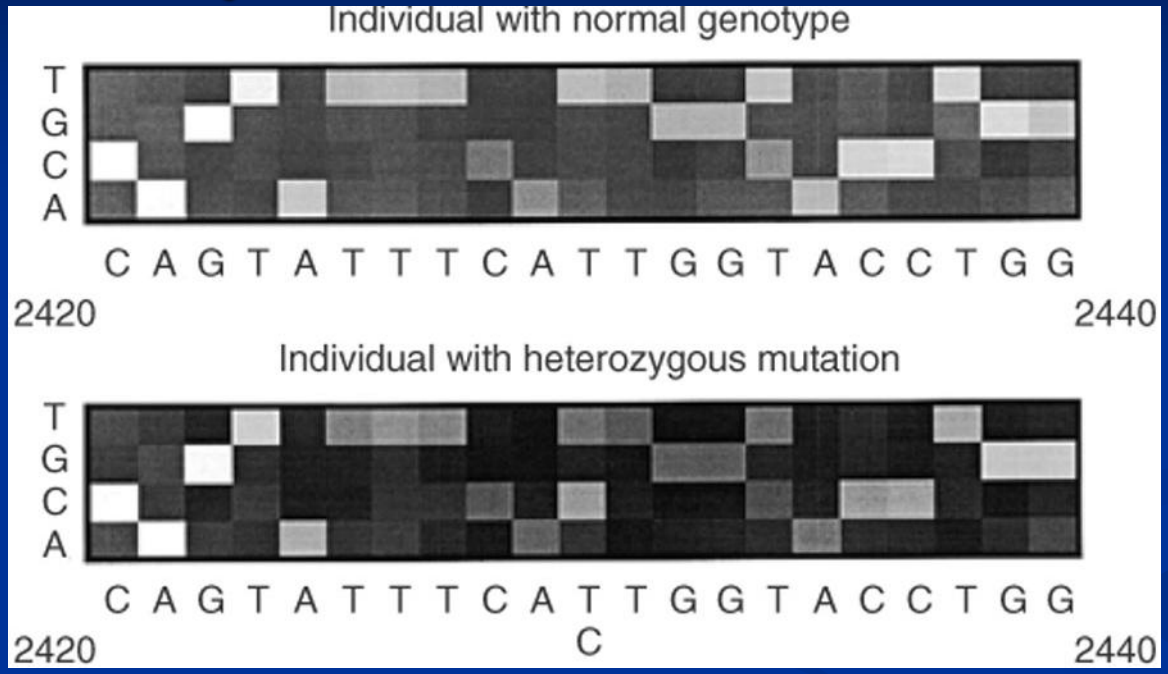


Fig. 10.17a p. 356

http://www.mun.ca/biology/scarr/DNA_chip_62Kbp_array.html

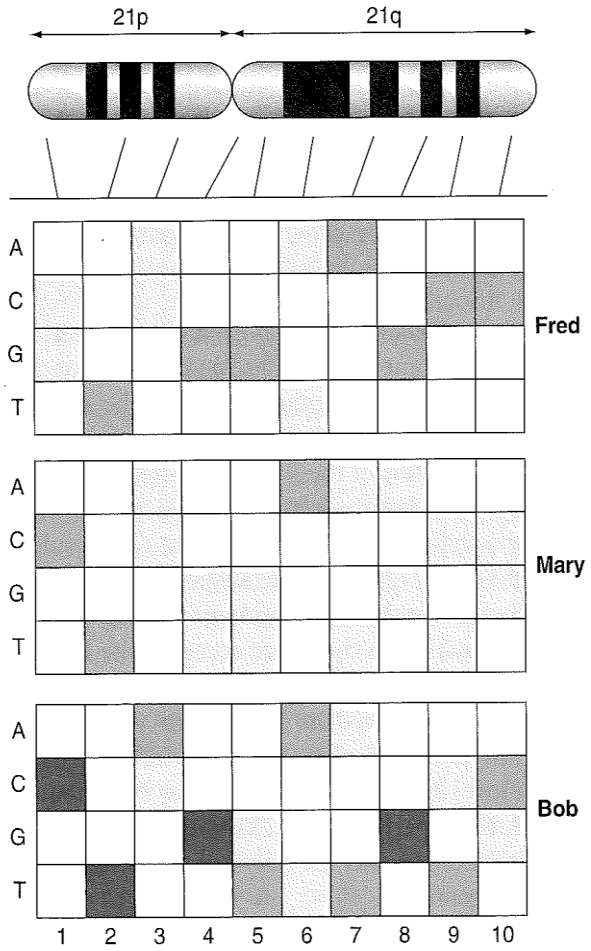
Large-scale multiplex ASO analysis with microarrays can detect *BRCA1* mutations



- Each column contains an ASO differing only at the nucleotide position under analysis
- BRCA1 DNA from any one allele can only be one of four ASOs in a column
- Heterozygotes are easily detected

Section 12.4

29. Fred and Mary have a child named Bob. The genomic DNAs of these three individuals are used as probes to ASO microarrays such as that shown in Fig. 10.17 on p. 356. The results with 10 SNP loci (1–10) that are located along chromosome 21 are shown in the figure that follows. *Yellow, orange,* and *red* indicate increasing intensities of hybridization.

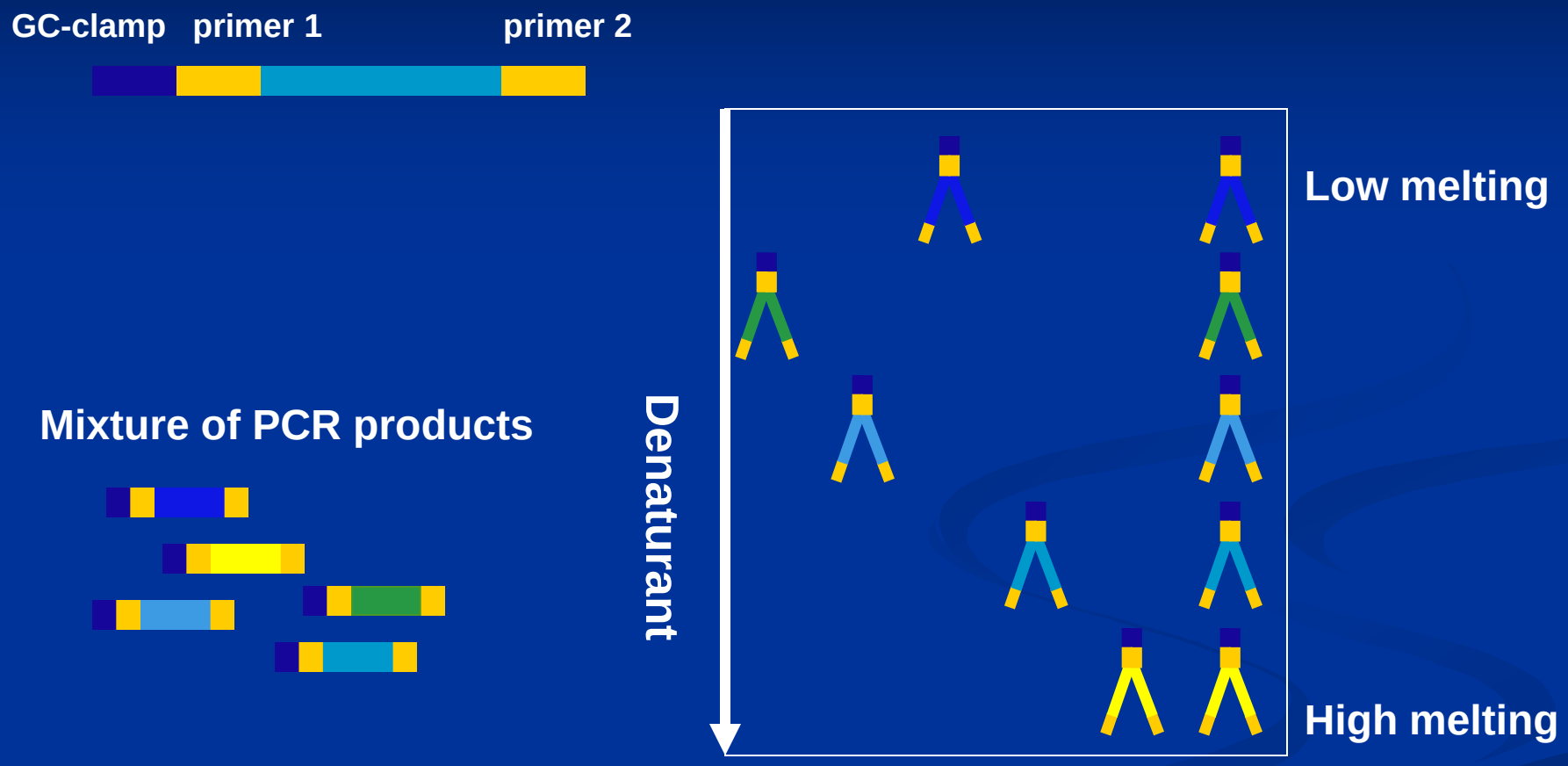


a. Is Bob likely to have Down syndrome? Explain.
b. Do the data provide evidence that a nondisjunction event occurred in one of the parents? If so, in which parent and during which meiotic division?
c. Do the data provide evidence that a recombination event occurred during meiosis in either parent? If so, in which parent, and describe the approximate location of the recombination event.
d. Describe how this example illustrates why it is important when tracking nondisjunction events to look particularly at loci near the centromere.

Other SNP detection methods

- Single strand conformation polymorphisms (SSCP)
  - Single strands with different sequences have different 3D shapes
- Denaturing gradient gel electrophoresis (DGGE) or temperature gradient gel electrophoresis (TGGE)
  - Double strands with different sequences dissociate at different concentrations of denaturant or different temperatures
- Direct sequencing

Denaturing gradient gel electrophoresis (DGGE)



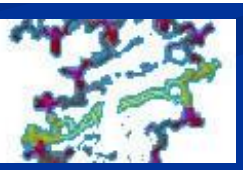
SNP loci and alleles in CFTR gene

Copyright © The McGraw-Hill Companies, Inc. Permission required for reproduction or display.

	Position			Allelle		Variant		Allele
	mRNA	Protein	Codon	Base	Amino Acid	Function	SNP ID	Frequency
Exon_1	133	1				start codon		
	156	8	3	A	Lys [K]	synonymous	rs1800071	N.D.
		8	3	G	Lys [K]			
	163	11	1	A	Ile [I]	missense	rs1800072	N.D.
		11	1	G	Val [V]			
Exon_2	204	24	3	A	Leu [L]	synonymous	rs55773134	N.D.
		24	3	G	Leu [L]			
	223	31	1	T	Cys [C]	missense	rs1800073	0.014
		31	1	C	Arg [R]			
	263	44	2	T	Val [V]	missense	rs1800074	N.D.
		44	2	A	Asp [D]			
Exon_3	345	71	3	C	Asn [N]	synonymous	rs1800075	N.D.
		71	3	T	Asn [N]			
	356	75	2	A	Gln [Q]	missense	rs1800076	0.012
		75	2	G	Arg [R]			
Exon_4	485	118	2	C	Ser [S]	frameshift	rs35871908	N.D.
		118	2	-/C	Ser [S]			
	492	120	3	A	Ala [A]	synonymous	rs1800077	N.D.
		120	3	G	Ala [A]			
	545	138	2	C	Pro [P]	missense	rs1800078	N.D.
		138	2	T	Leu [L]			
	575	148	2	C	Thr [T]	missense	rs35516286	0.026
		148	2	T	Ile [I]			

NCBI

dbSNP
Short Genetic Variations

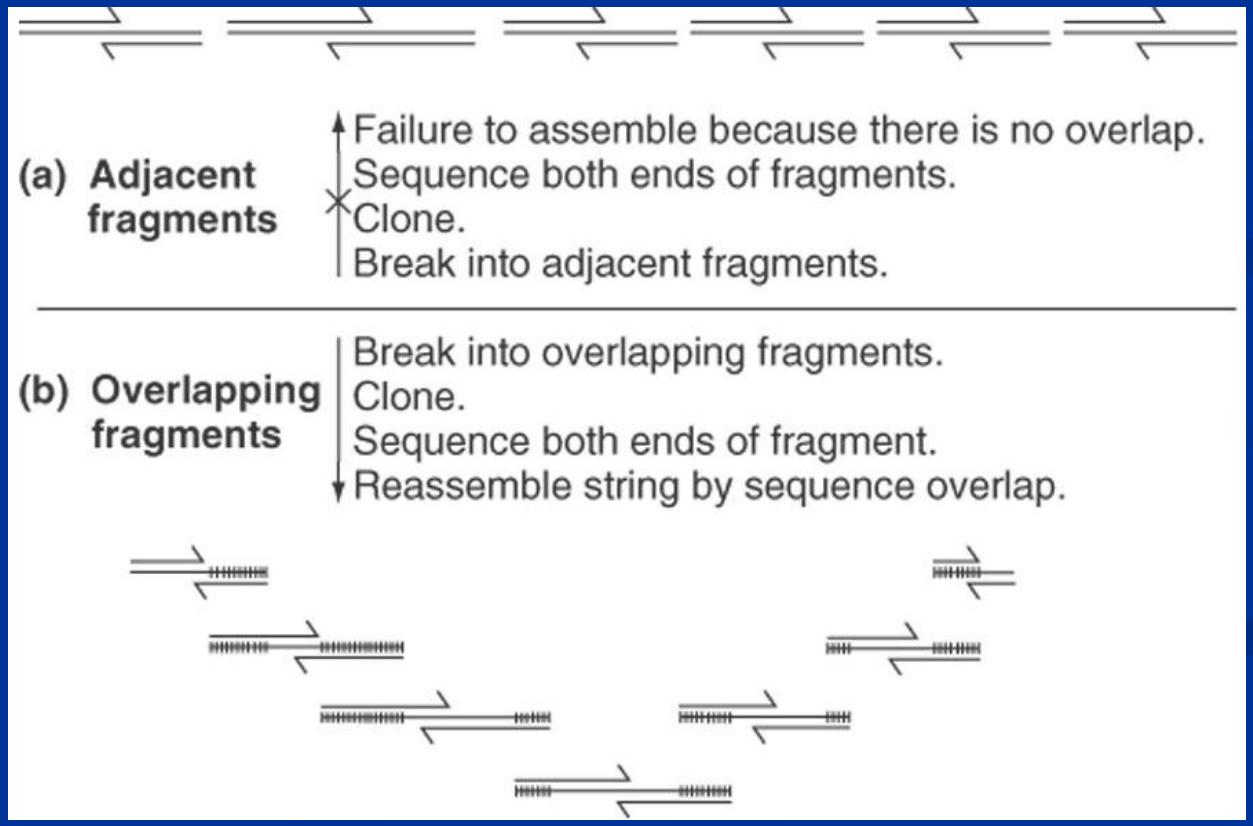


Whole genome sequencing

- Two approaches
  - Genetic (linkage) maps used to derive a physical maps (old)
    - High density linkage maps establish sets of DNA markers less than 1 cM apart
    - Use DNA markers to retrieve fragments, construct contigs and identify genes
  - Physical mapping techniques (new)
    - Direct analysis of DNA
    - Overlapping reads aligned

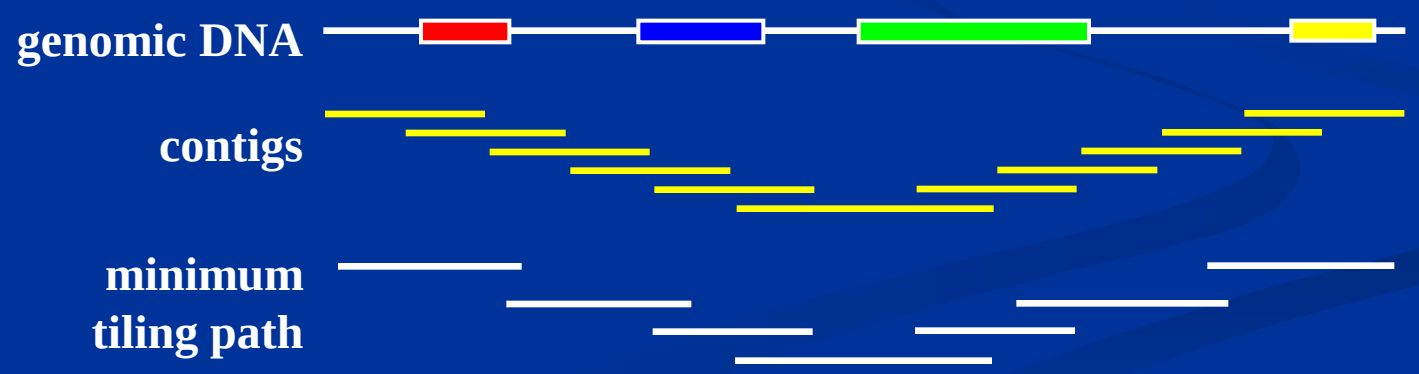
Divide and conquer strategy meets most challenges.

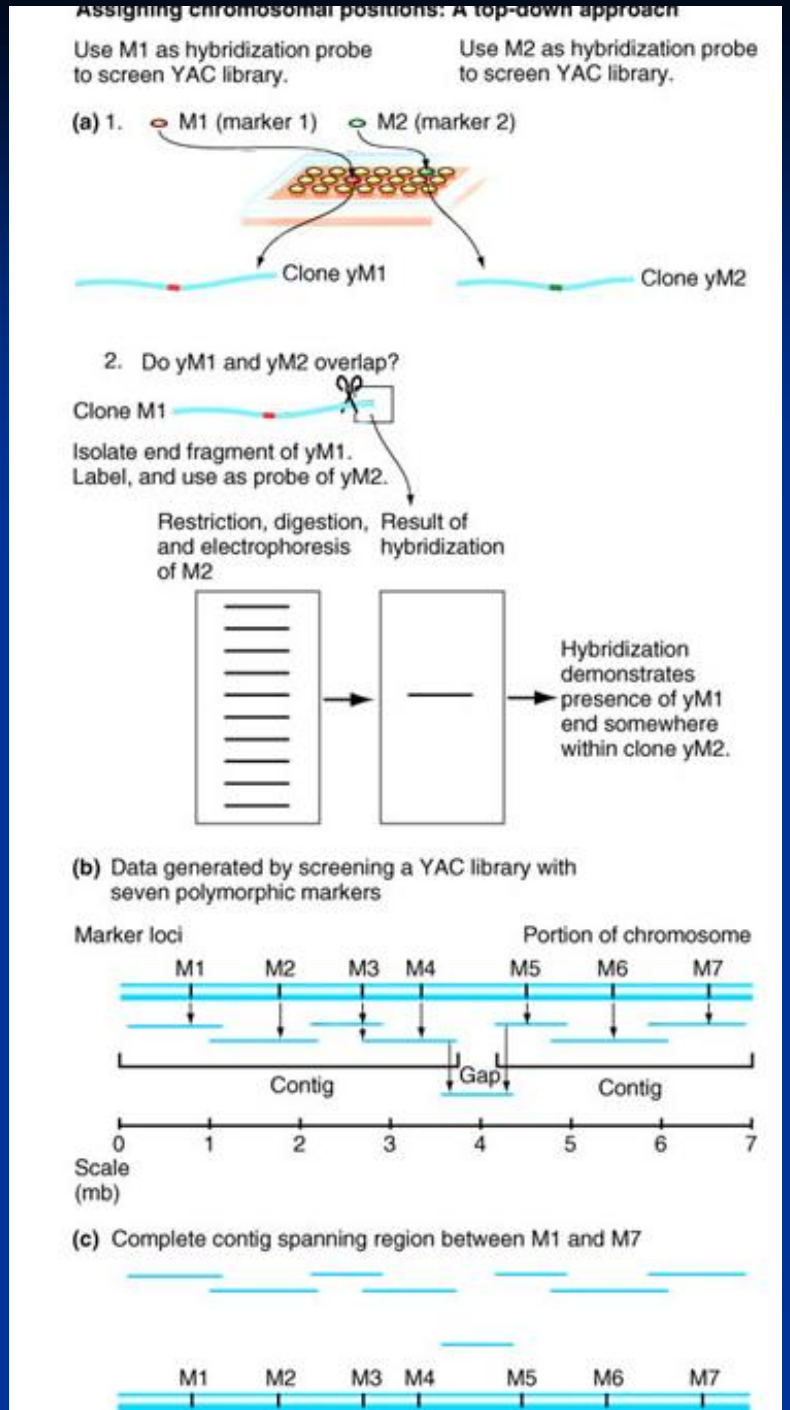
- Chromosomes broken or cut into small overlapping pieces and cloned
- Ends of clones sequenced and reassembled into original chromosome strings



Contigs

- Contigs are groups of overlapping pieces of chromosomal DNA
  - Make <u>contiguous</u> clones
- For sequencing, create a "minimum tiling path"
  - Contig of smallest number of inserts that covers a region of the chromosome





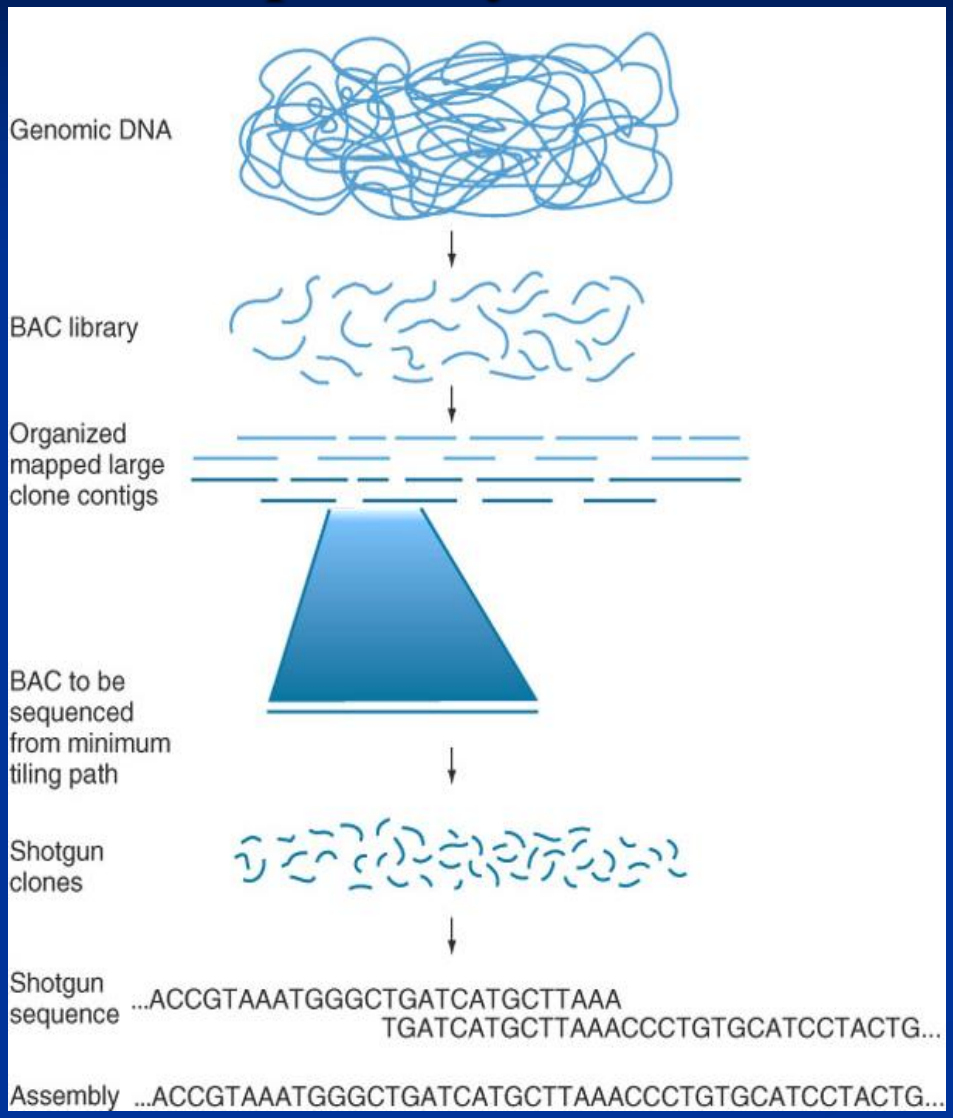
High density linkage mapping to build overlapping sets of genomic clones

Whole genome sequencing

- Two strategies used for sequencing human genome
  - Hierarchical shotgun sequencing (HSS)
  - Whole-genome shotgun sequencing (WGSS)
- Shotgun – randomly generated overlapping insert fragments
  - Fragments from mapped BACs (HSS)
  - Fragments from shearing whole genomes (WGSS)

Hierarchical shotgun strategy (map based)

Used in publicly funded effort to sequence human genome

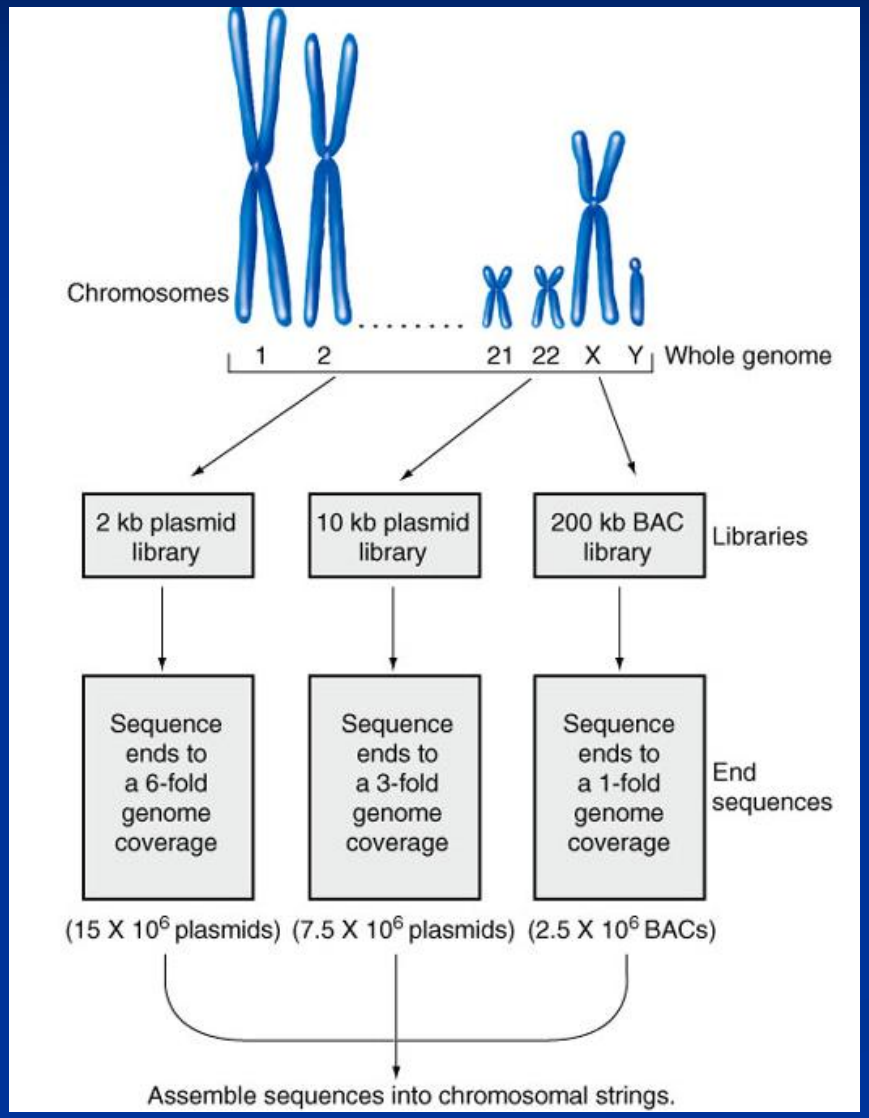


- Shear 200 kb BAC clone into ~2 kb fragments
- Sequence ends 10 times
- Linkage and physical maps used to assemble sequences into chromosomes
- Significant work to create libraries of each BAC and physically map BAC clones

Fig. 10.5 p. 340

Whole-genome shotgun sequencing

Private company (Celera) used to sequence human genome



- **Whole genome randomly sheared three times:**
  ~2 kb, ~10 kb, ~200 kb
- **Computer program assembles sequences into chromosomes**
- **No physical map construction**
- **Only one BAC library**

Fig. 10.6 p. 341

Sequencing of the human genome

- Most of draft took place during last year of project
  - ~ 345,600 bp/day
  - Automated factory-like production line generated sufficient DNA to supply sequencers on a daily basis
  - Large sequencing centers with 100-300 instruments – 103,680,000 bp/day (10-fold coverage in 30 days)

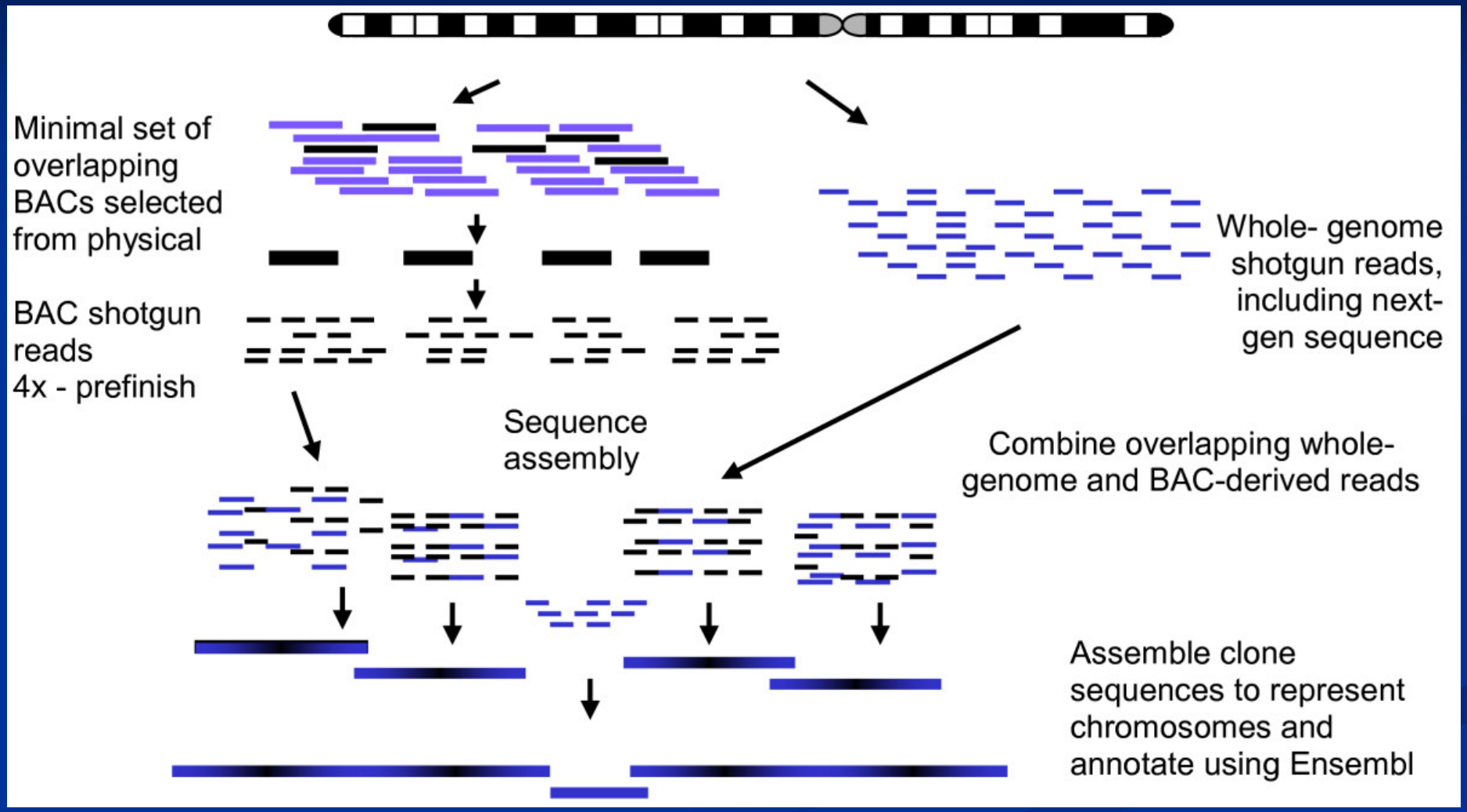
Integration of linkage, physical, and sequence maps

- Provides check on the correct order of each map against other two
- SSR and SNP DNA linkage markers integrated into physical map
- Linkage maps and physical maps have unique sequences allowing placement on sequence map

Hybrid approach

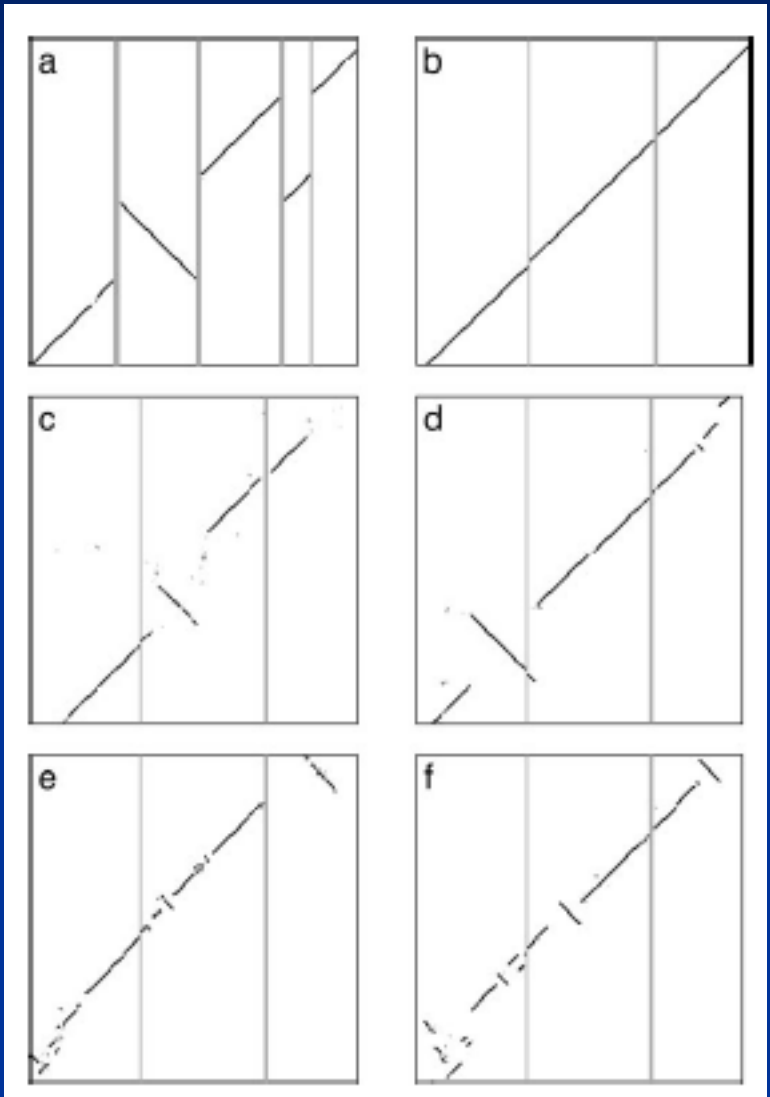
- Combines aspects of both map-based and whole-genome shotgun approaches
  - Map clones
  - Sequence some of the mapped clones
  - Whole-genome sequencing
  - Combine information from both methods

Hybrid approach



Archibald *et al. BMC Genomics 2010* ***11:438*** ***doi:10.1186/1471-2164-11-438***

Dot plots of sample assembly comparisons



Dot plot comparison of portions of NCBI-34 to:

a) WGSS - Sequence assembly using ONLY Celera generated data

b) WGA – Sequence assembly originally published by Celera

Istrail (2004) PNAS

High-throughput DNA sequencing

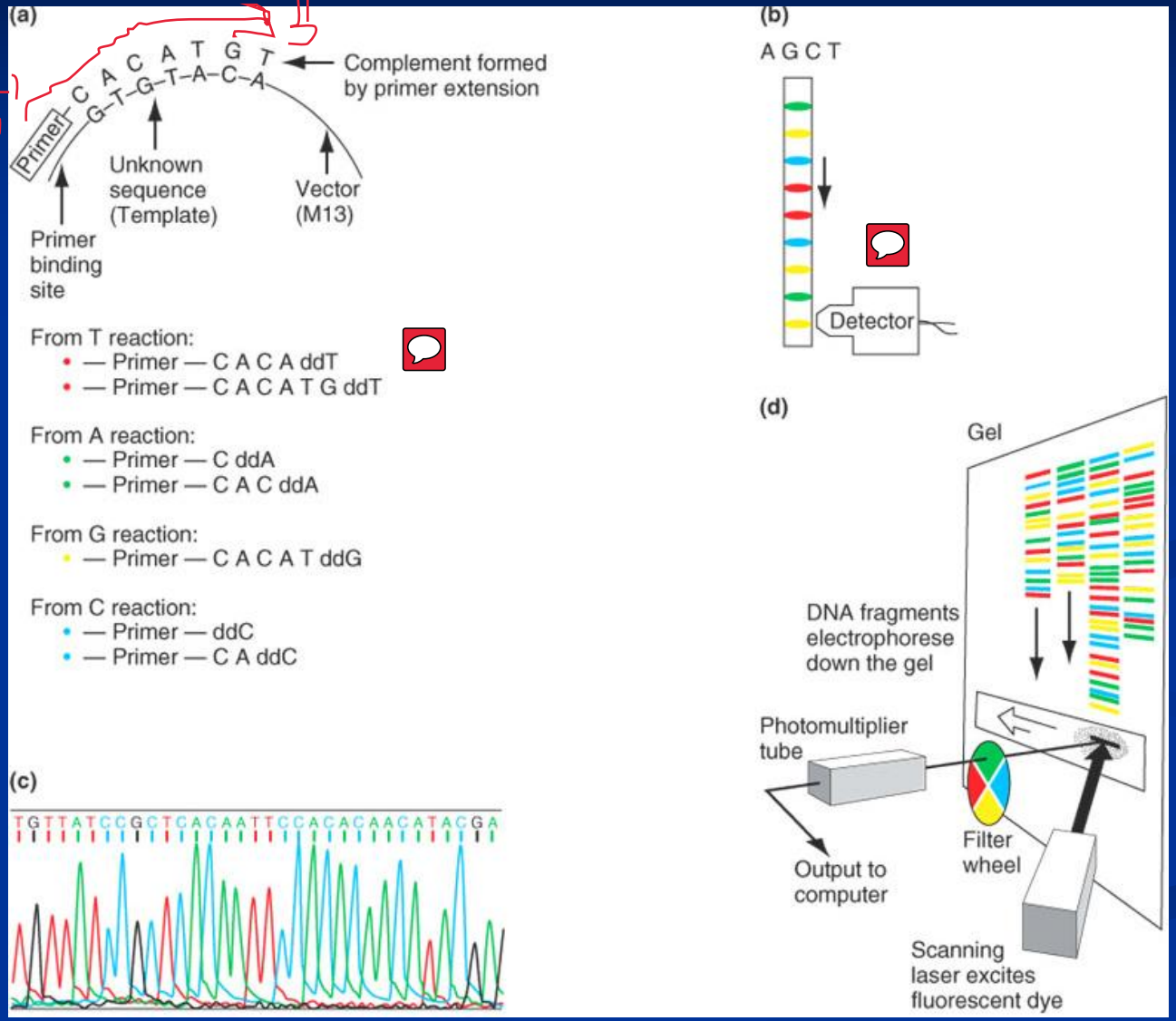
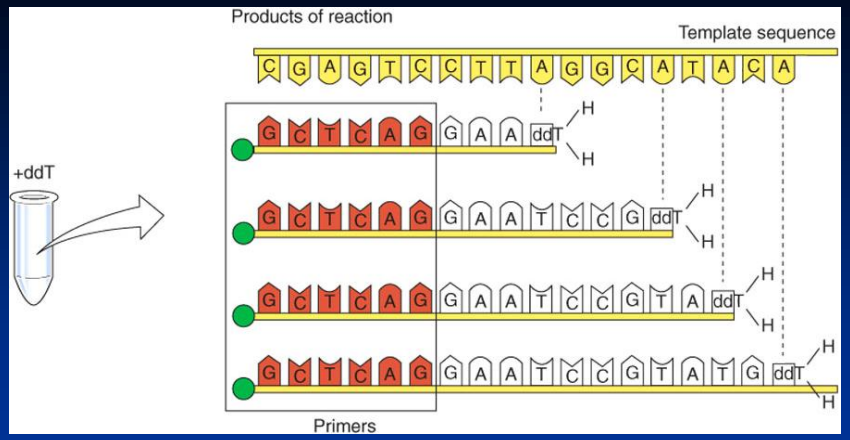


Fig. 10.20 p. 376



Sanger sequencing - chain termination

Ruled the DNA sequencing world for the past 30 years!

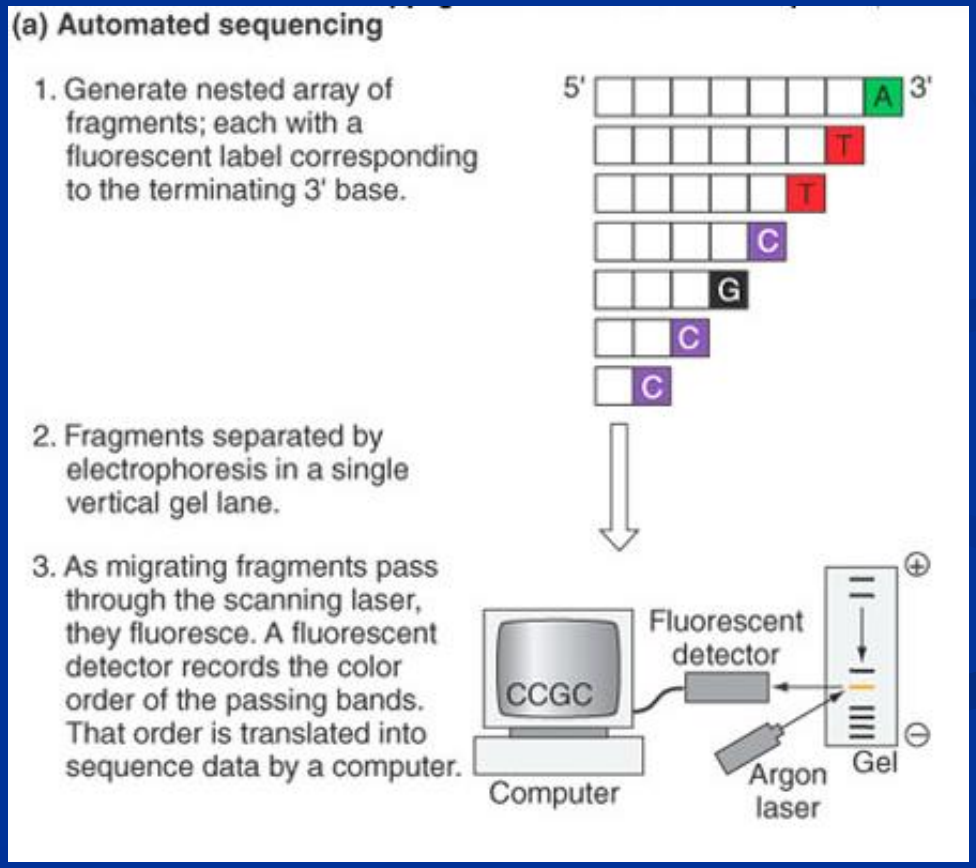
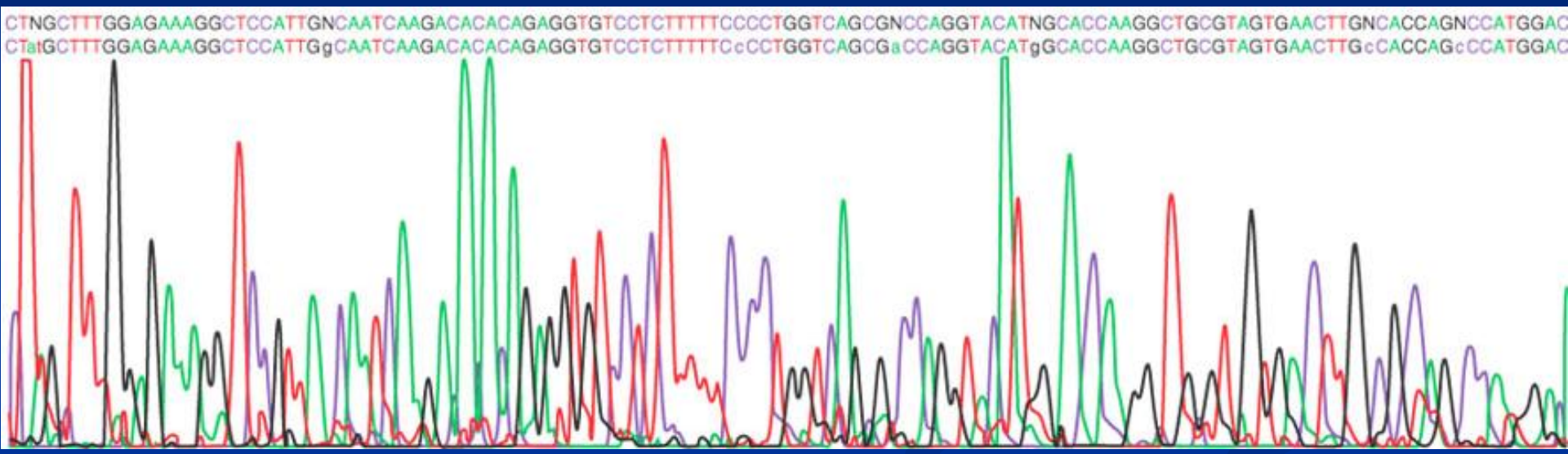


Fig. 9.14 p. 317

Automated Sanger DNA Sequencing

Electropherogram of DNA sequence



- Costs remained high
- Speed to generate remained low
- Involves labor intensive cloning procedures

Fig. 9.17c p. 317

Cost of sequencing whole genomes

http://archive.hpcwire.com/hpcwire/2013-06-05/dell_boxes_up_hpc_for_life_sciences.html

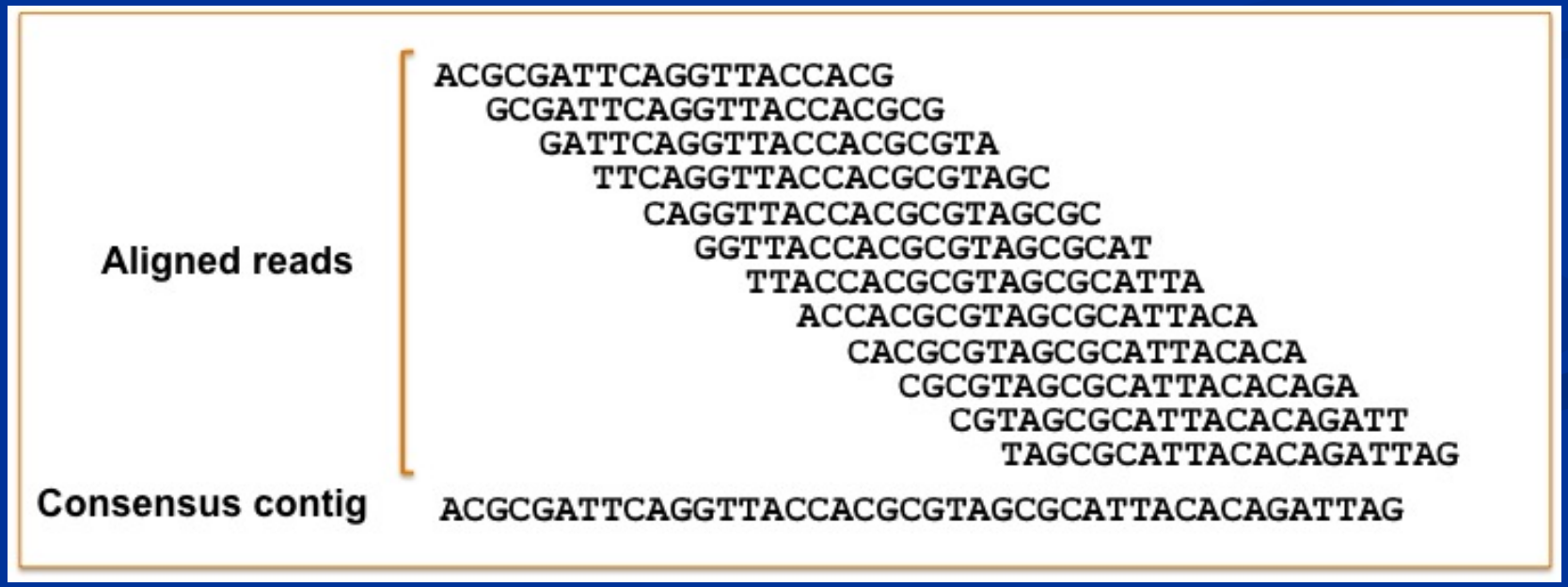
Next Generation Sequencing

- Made possible due to advances in nanotechnologies, synthetic chemistry, and bioinformatics
- Millions of parallel sequence reads performed simultaneously
- Single NGS platform generates data equal to ~200 hundred capillary sequencers (Sanger) in 24 hours

http://www.youtube.com/watch?v=g0vGrNjpyA8
http://www.youtube.com/watch?v=_ApDinCBt8g&feature=related

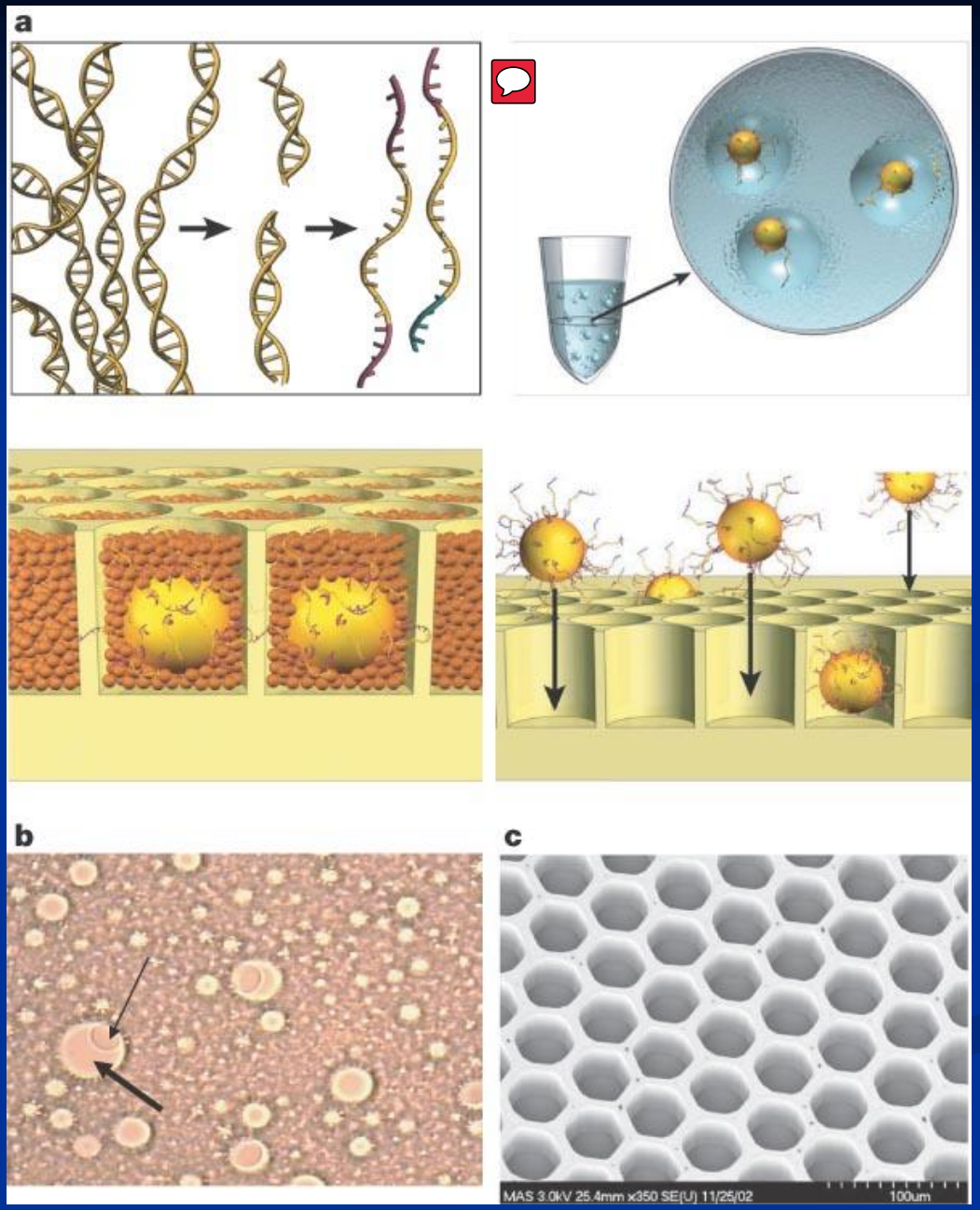
Next Generation Sequencing

- Main NGS platforms (now there are many others)
  - 454/FLX (Roche)
  - Solexa (Illumina)
  - SOLiD (Life Technologies)



454/FLX – Pyrosequencing

- Genome fragmented and adaptors added
- Single stranded DNA fragments immobilized on beads
- Emulsion PCR and one bead per well

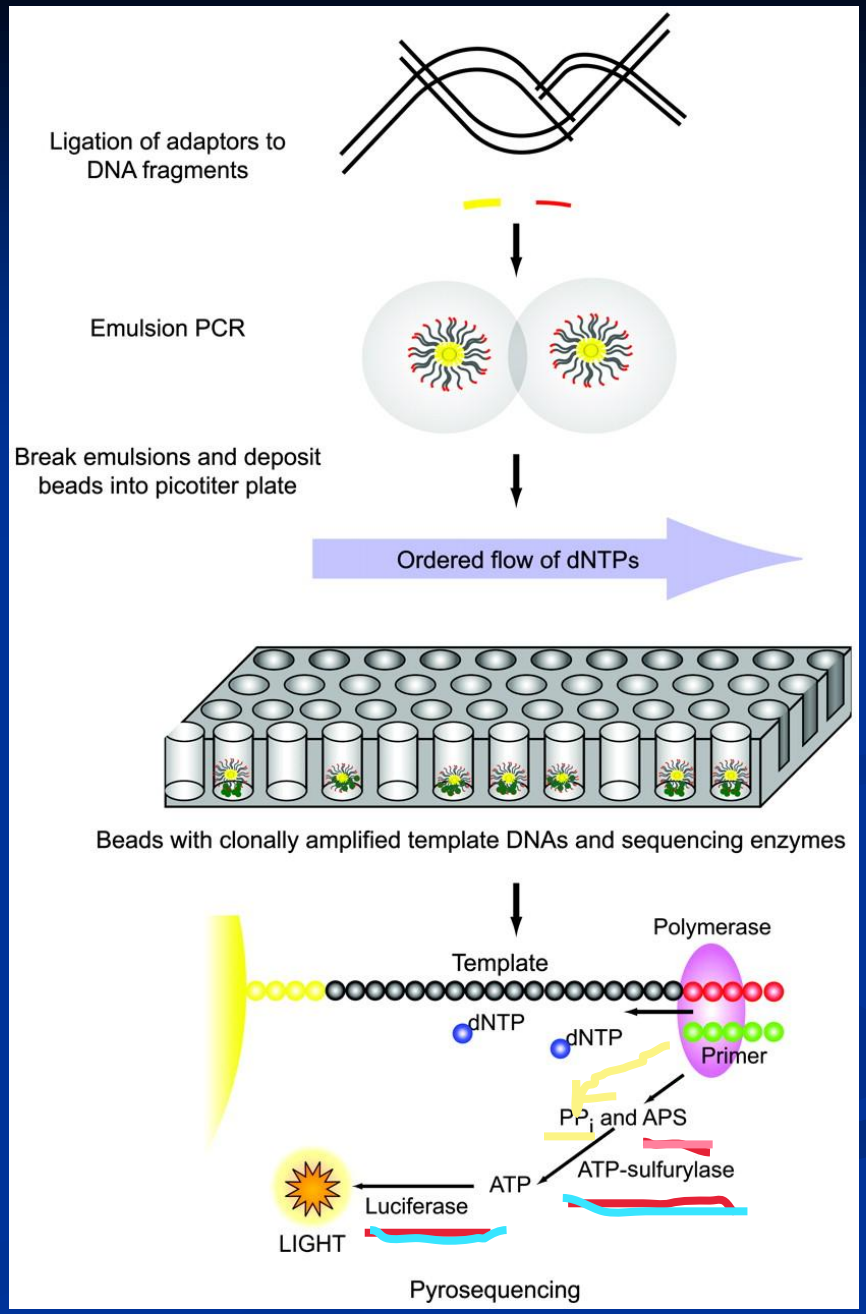


Margulies (2005) Nature

454/FLX – Pyrosequencing

- Single stranded DNA fragments immobilized on beads
- Emulsion PCR and one bead per well
- PPi is released converted to ATP by ATP sulfurylase
- ATP used by luciferase to produce light from the substrate luciferin
- PPi released proportional to number of nucleotides added

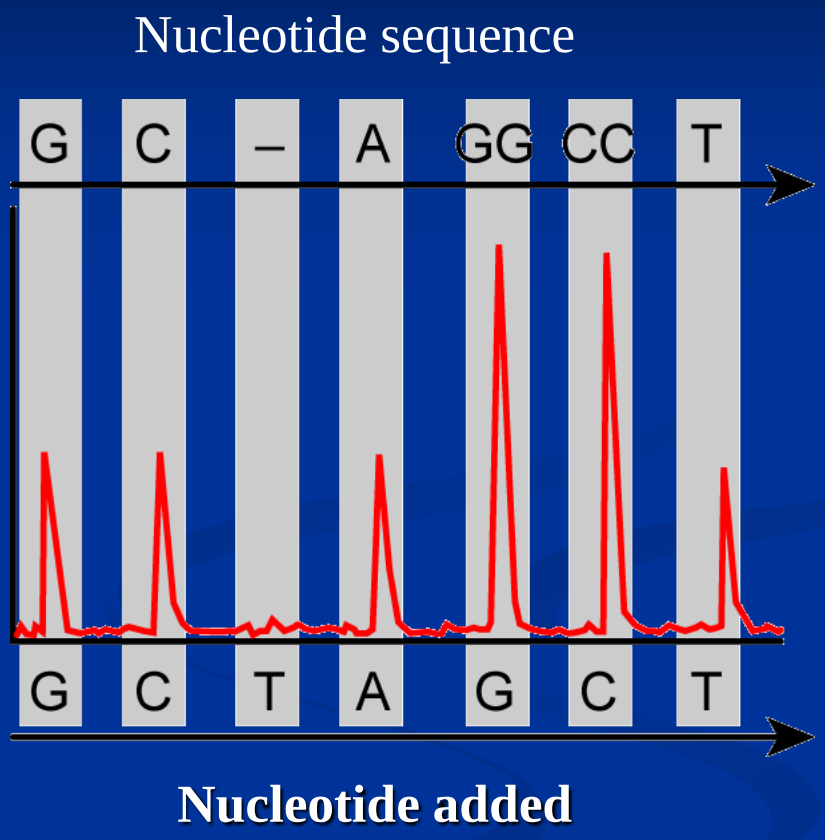
http://www.youtube.com/watch?v=bFNjxKHP8Jc&NR=1&feature=fvwp



Voelkerding (2009) Clin Chem

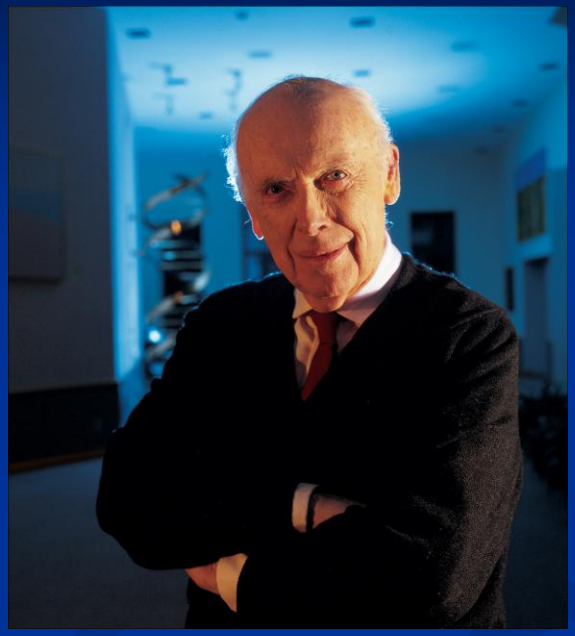
454/FLX - Pyrosequencing

- Sequential addition of each dNTP gives sequence
- Apyrase enzyme used to degrade dNTPs after reaction completed
- Sequence determined by amount of light emitted as each dNTP is added



http://www.youtube.com/watch?v=nFfgWGFe0aA

454 sequencing platform

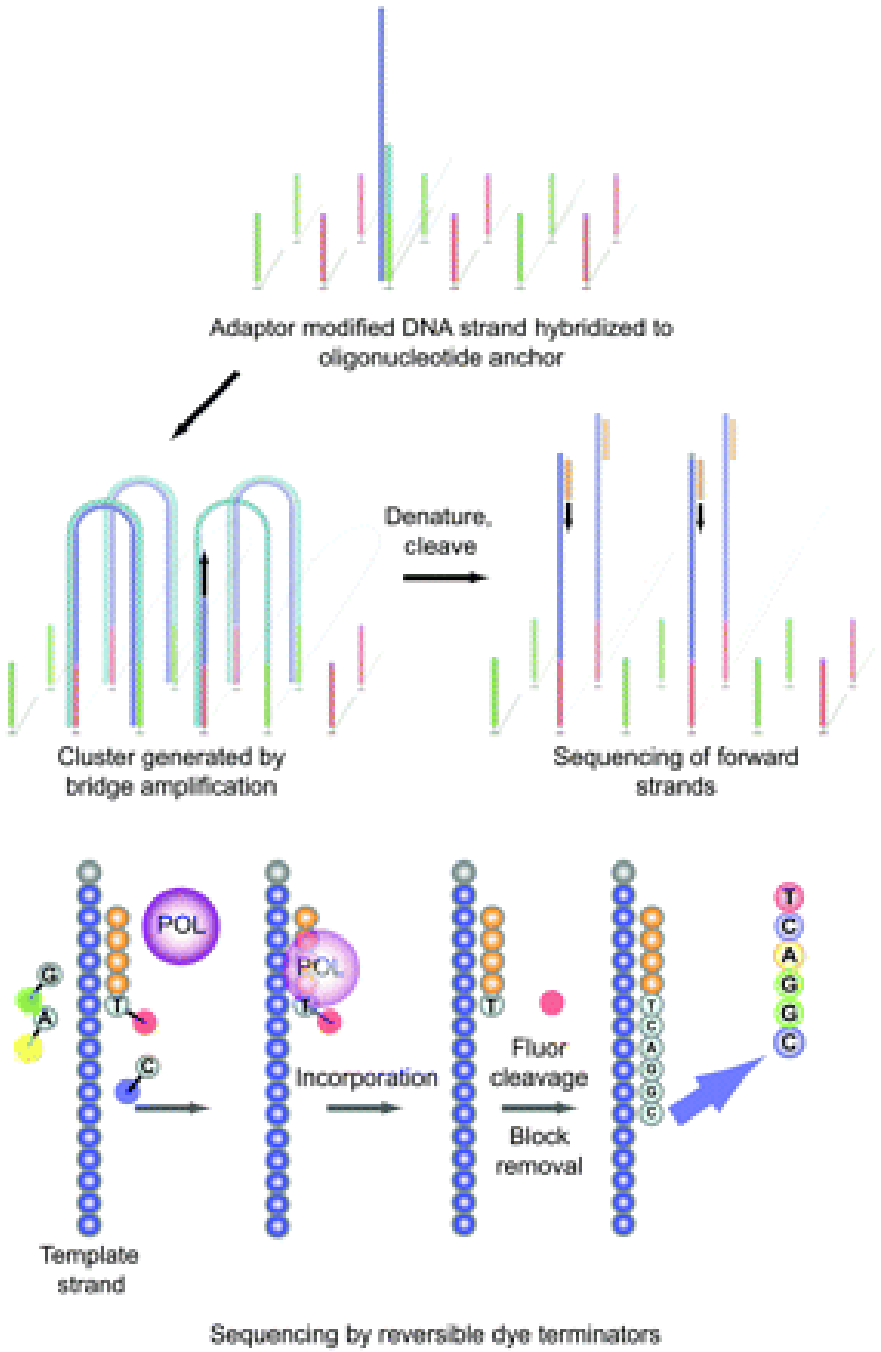


James Watson genome sequence

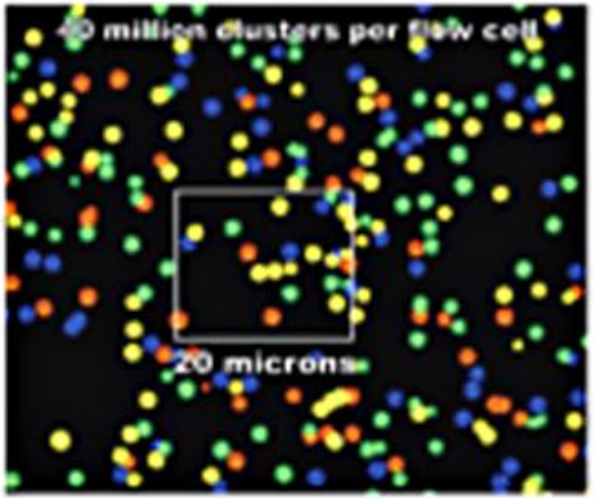
Neandertal genome sequence

- Genome fragmented and adaptors added
- Single stranded DNA fragments immobilized by hybridization on flow cells
- PCR by "bridge" amplification generates clusters of clonal molecules
- Nucleotides with reversible fluorophore and terminator properties are used
- Nucleotide incorporation, fluorescence imaging and cleavage

http://www.youtube.com/watch?v=77r5p8IBwJk&feature=related

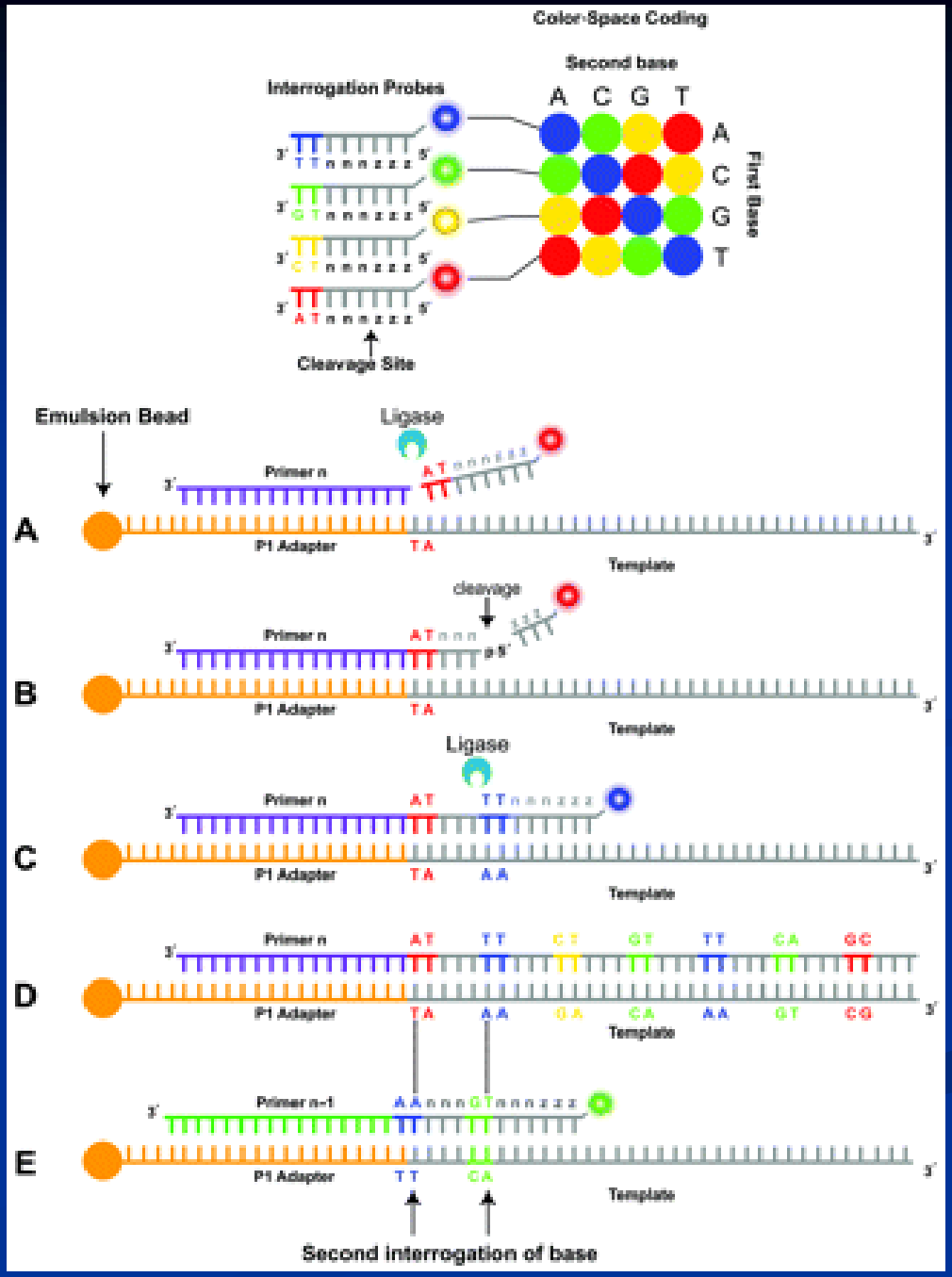


Voelkerding (2009) Clin Chem


40 million clusters per flow cell
20 microns

TGCTACGAT...
1
2
3
4
5
6
7
8
9
TTTTTTTGT...

- Emulsion PCR to attach genomic fragments to beads
- 16 different interrogation probes, each with 1 of 16 possible 2-base combinations
- Template with annealed primer (n) interrogated
- After annealing and ligation, fluorescence recorded
- 3 bases cleaved from probe, annealing and ligation of next probe



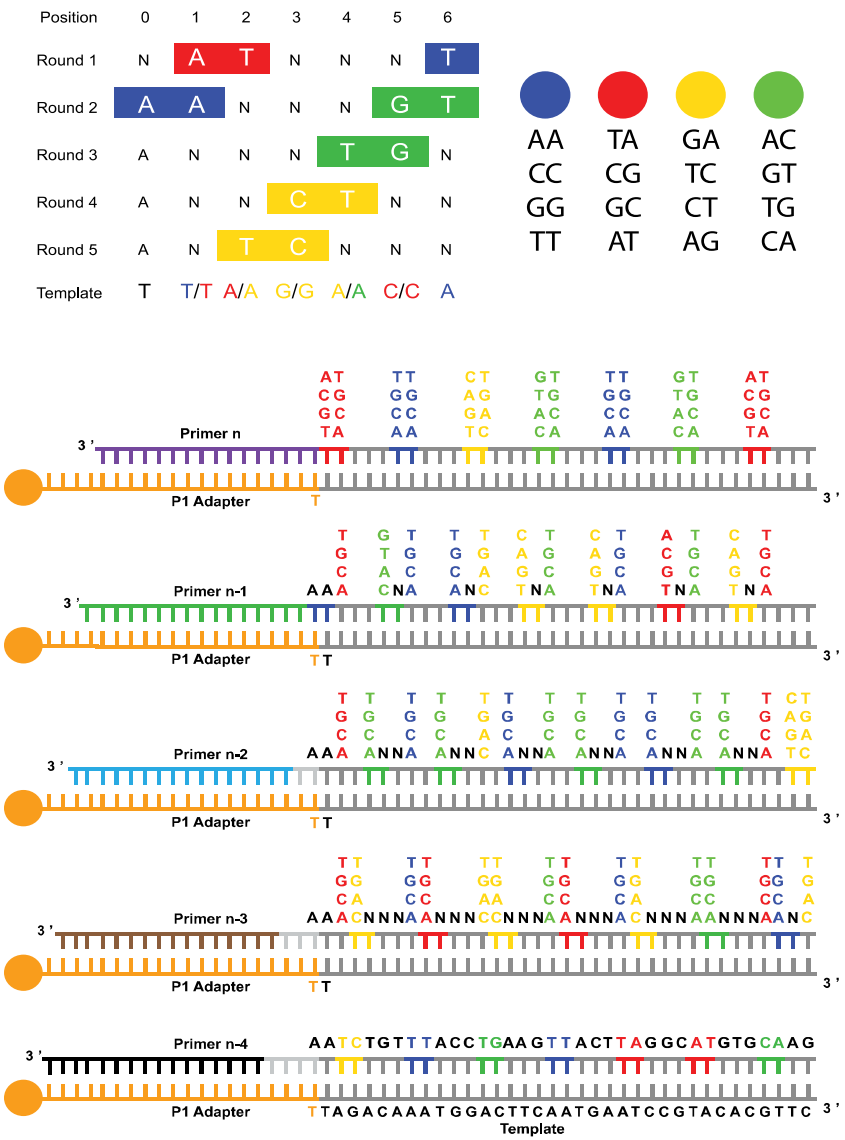
Voelkerding (2009) Clin Chem

- After complete extension, product with primer (n) denatured
- Second round performed with primer (n-1), third round primer (n-2), etc.
- Template sequence is then deconvoluted

http://www.youtube.com/watch?v=nlvyF8bFDwM&NR=1

Supplemental Text Voelkerding, Dames, and Durtschi

Supplemental Data Figure 1.



Voelkerding (2009) Clin Chem

	454	Solexa (Illumina)	SOLiD	Sanger
Sequencing method	Pyrosequencing	Reversible dye terminators	Sequencing by ligation	Dye terminators
Read lengths	400 bases	36 bases	50 bases	800 bases
Sequencing run time	8 hours	8 days	7 days	3 hours
Total bases per run	500 Mb	100-125 Gb	100 Gb	800 bases
Cost	$60/Mb	$2/Mb	$2/Mb	$0.50/kb

February 16, 2010
Elliot Margulies
NIH course - Current topics in Genome Analysis 2010
http://www.youtube.com/watch?v=g0vGrNjpyA8&feature=related

Voelkerding (2009) Clin Chem

Changes in biology, genetics and genomics

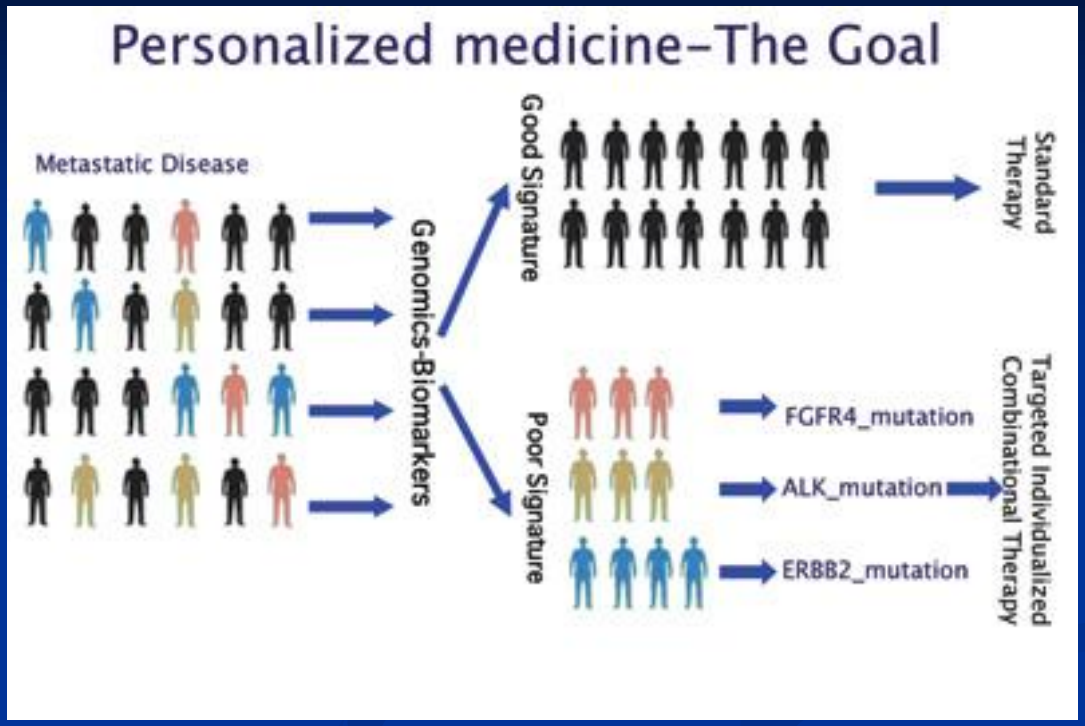
- Genetics parts list
- Speeds genome analysis
  - Sequence comparisons with other organisms through homology
  - Ready access to identification of known human polymorphisms
  - Sequence comparisons of normal versus tumor cell DNA
  - Sequence parts of genomes that are particularly interesting

Cost of sequencing whole genomes

http://archive.hpcwire.com/hpcwire/2013-06-05/dell_boxes_up_hpc_for_life_sciences.html

deCODE genetics

The race to a $100 genome



An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium* | NATURE | VOL 491 | 1 NOVEMBER 2012

23andMe

Four classes of DNA polymorphisms

p. 345

Copyright © McGraw-Hill Education. Permission required for reproduction or display.

TABLE 10.1 Categories of Genetic Variants

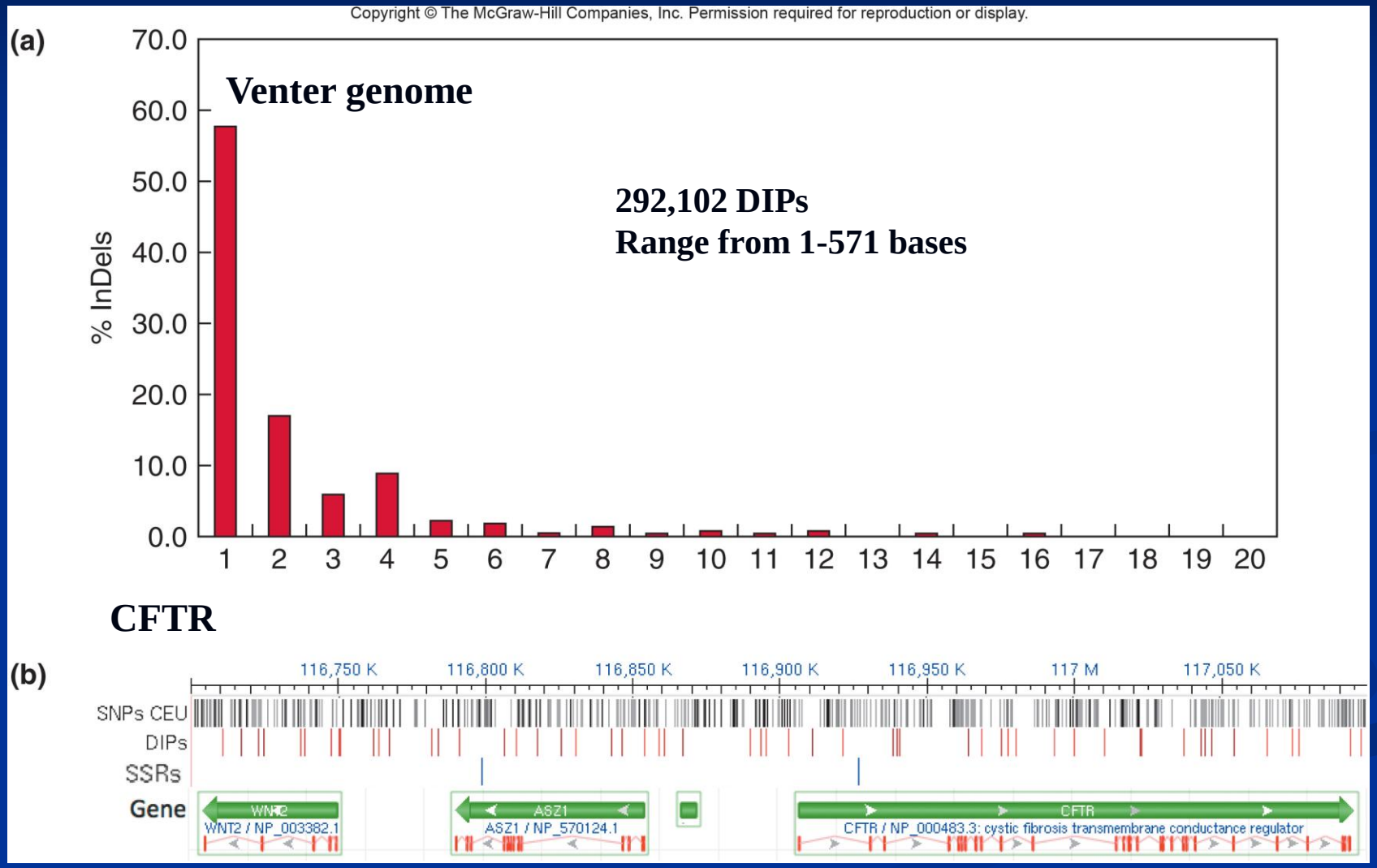
The right column shows how frequently on average you would find a polymorphism of the indicated class when comparing any two haploid human genomes.

	Size	Frequency (1 per. . . .)
SNP Single nucleotide polymorphism	1 bp	1 kb
DIP or **Indel** Insertion/deletion	1–100 bp	10 kb
SSR or **Microsatellite** Simple sequence repeat	1–10 bp repeat unit	30 kb
CNV Copy number variant	10 bp–1 Mb	3 Mb

Short deletions-insertions (Indels, DIPs)

- Expansion or contraction in the length of <u>nonrepetitive</u> DNA
- Range in length from 1 bp to >500 bps
- 75% are one or two bases
- Occur about 1 in every 10 kb

Genetic variants defined by changes in DNA length



Four classes of DNA polymorphisms

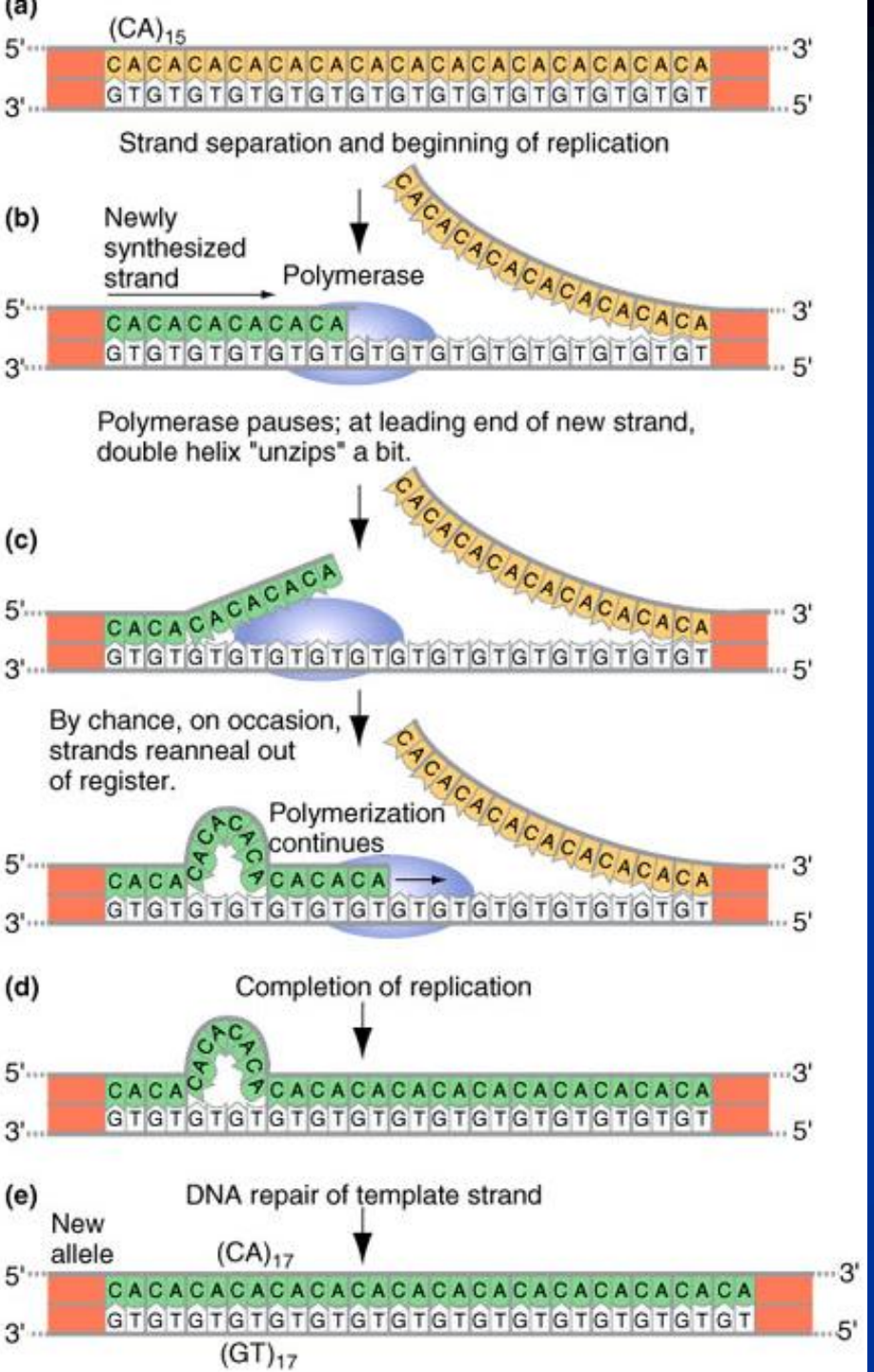
p. 345

Copyright © McGraw-Hill Education. Permission required for reproduction or display.

TABLE 10.1 Categories of Genetic Variants

The right column shows how frequently on average you would find a polymorphism of the indicated class when comparing any two haploid human genomes.

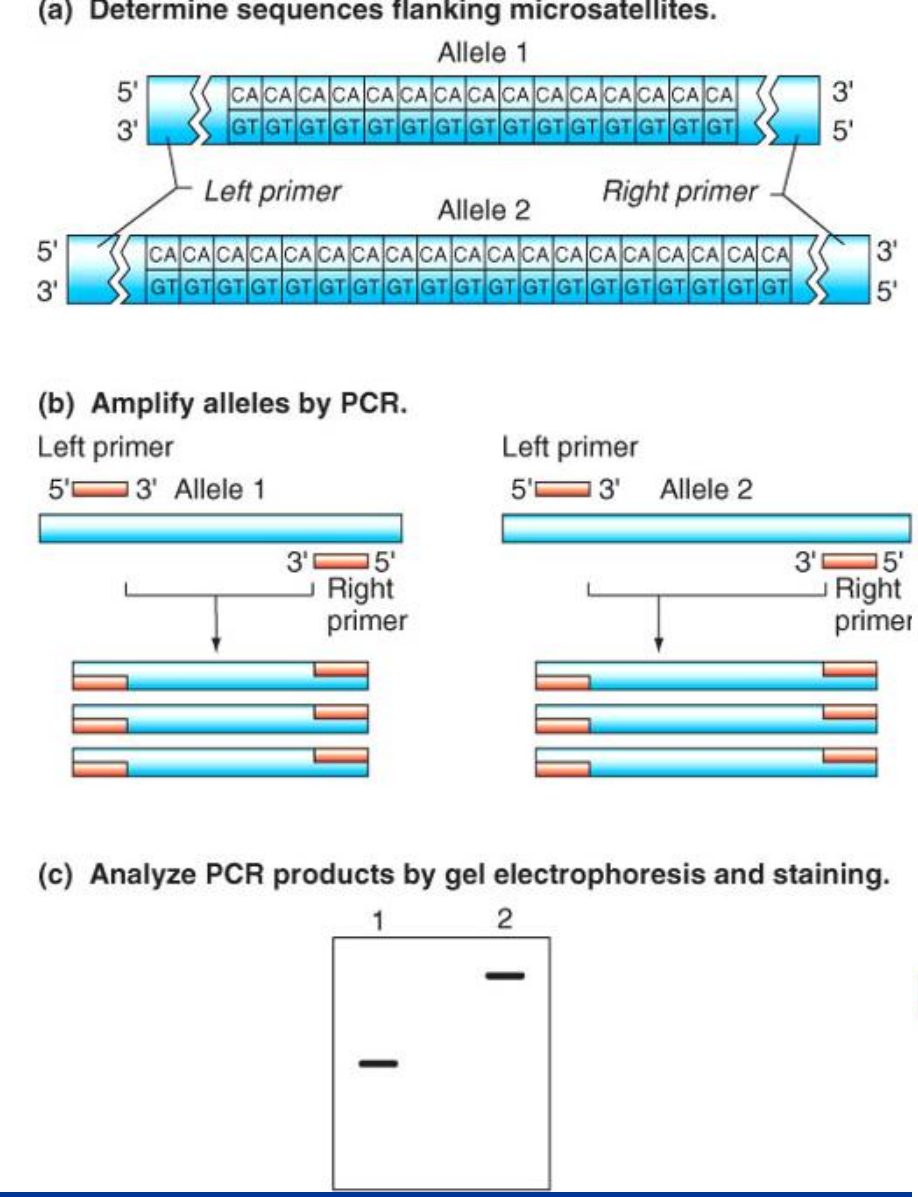
	Size	Frequency (1 per. . . .)
SNP Single nucleotide polymorphism	1 bp	1 kb
DIP or **Indel** Insertion/deletion	1–100 bp	10 kb
SSR or **Microsatellite** Simple sequence repeat	1–10 bp repeat unit	30 kb
CNV Copy number variant	10 bp–1 Mb	3 Mb



Simple sequence repeats (SSRs)

- Occur about 1 every 30,000 bp
- Repeated units 1-3 bp in length
- Tandem repeats 15-100 times
- Mutate by replication error into multiple alleles
- New alleles arise in 1 out of 1000 gametes

SSR allele detection analysis of size differences



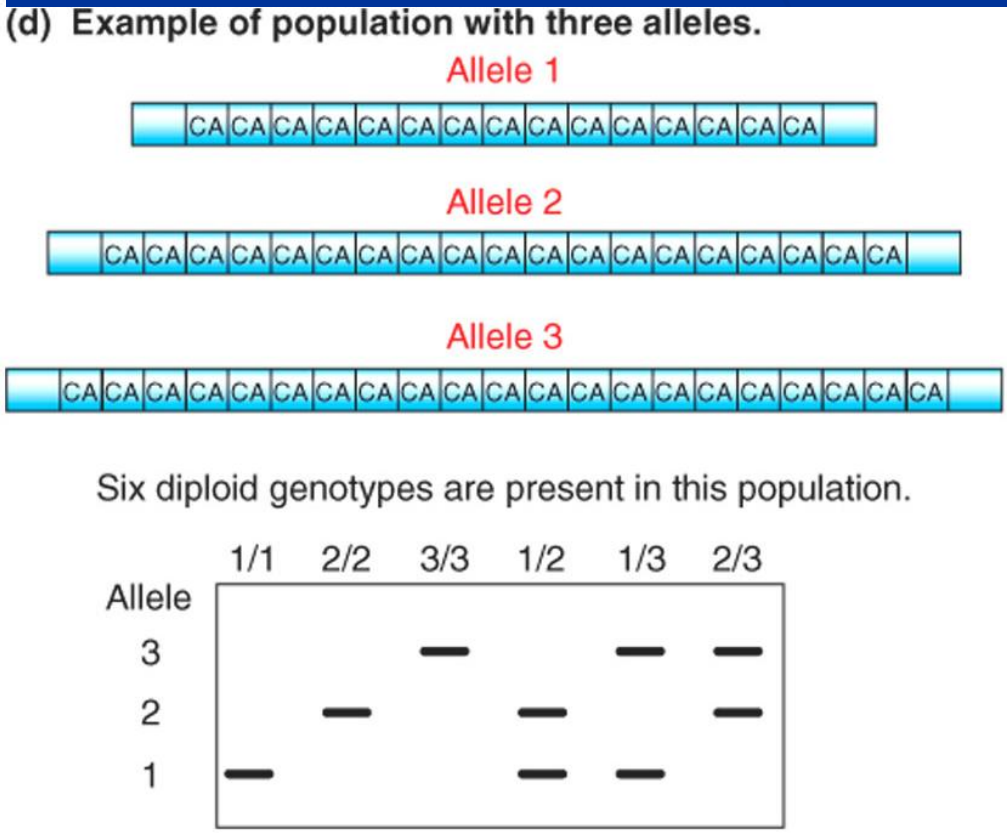
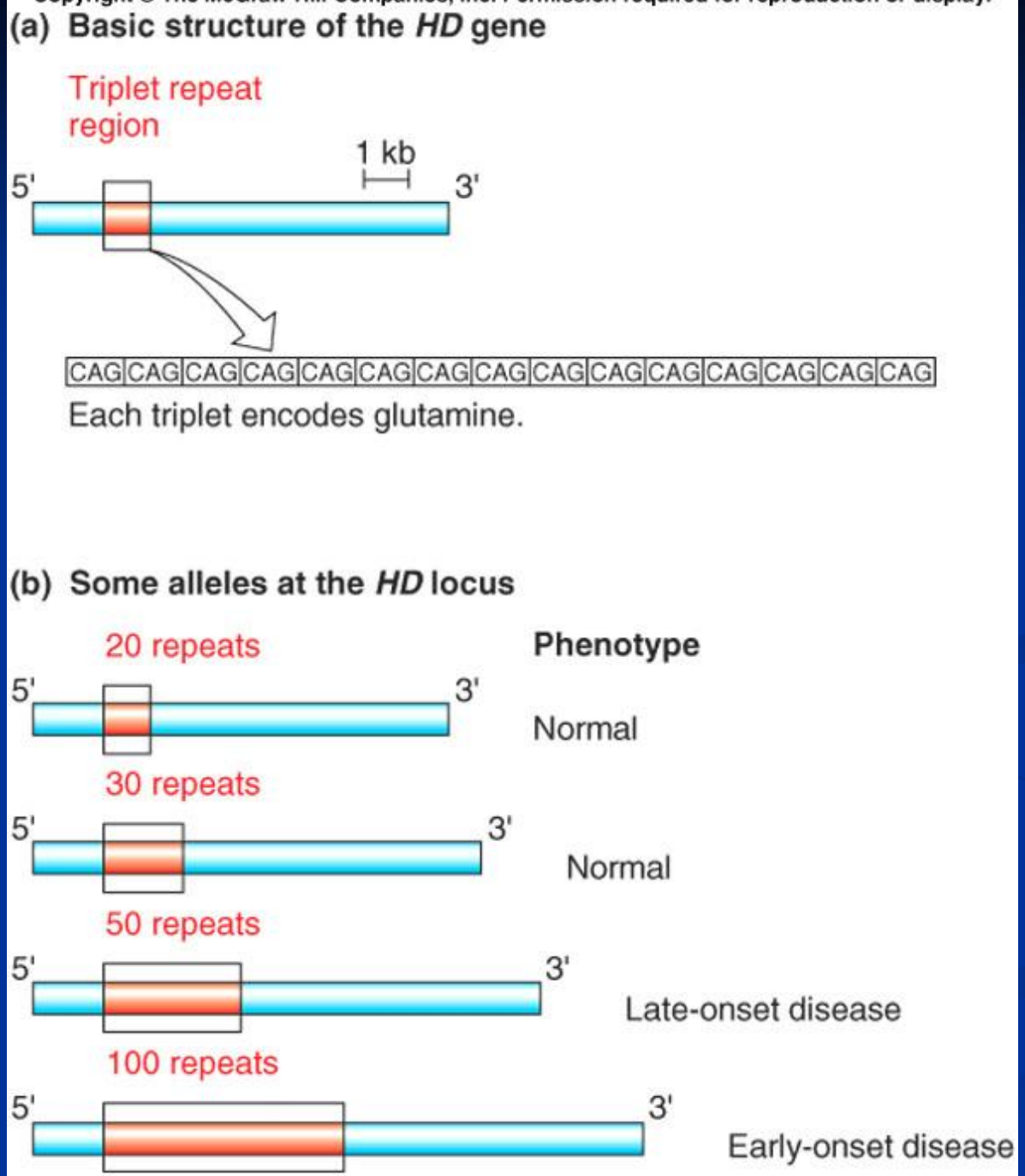


Fig. 10.12 p. 351

Copyright © The McGraw-Hill Companies, Inc. Permission required for reproduction or display.



Huntington's disease is an example of an SSR triplet repeat in a coding region

Displays variable expressivity but complete penetrance

Fig. 10.13 p. 352

Four classes of DNA polymorphisms

p. 345

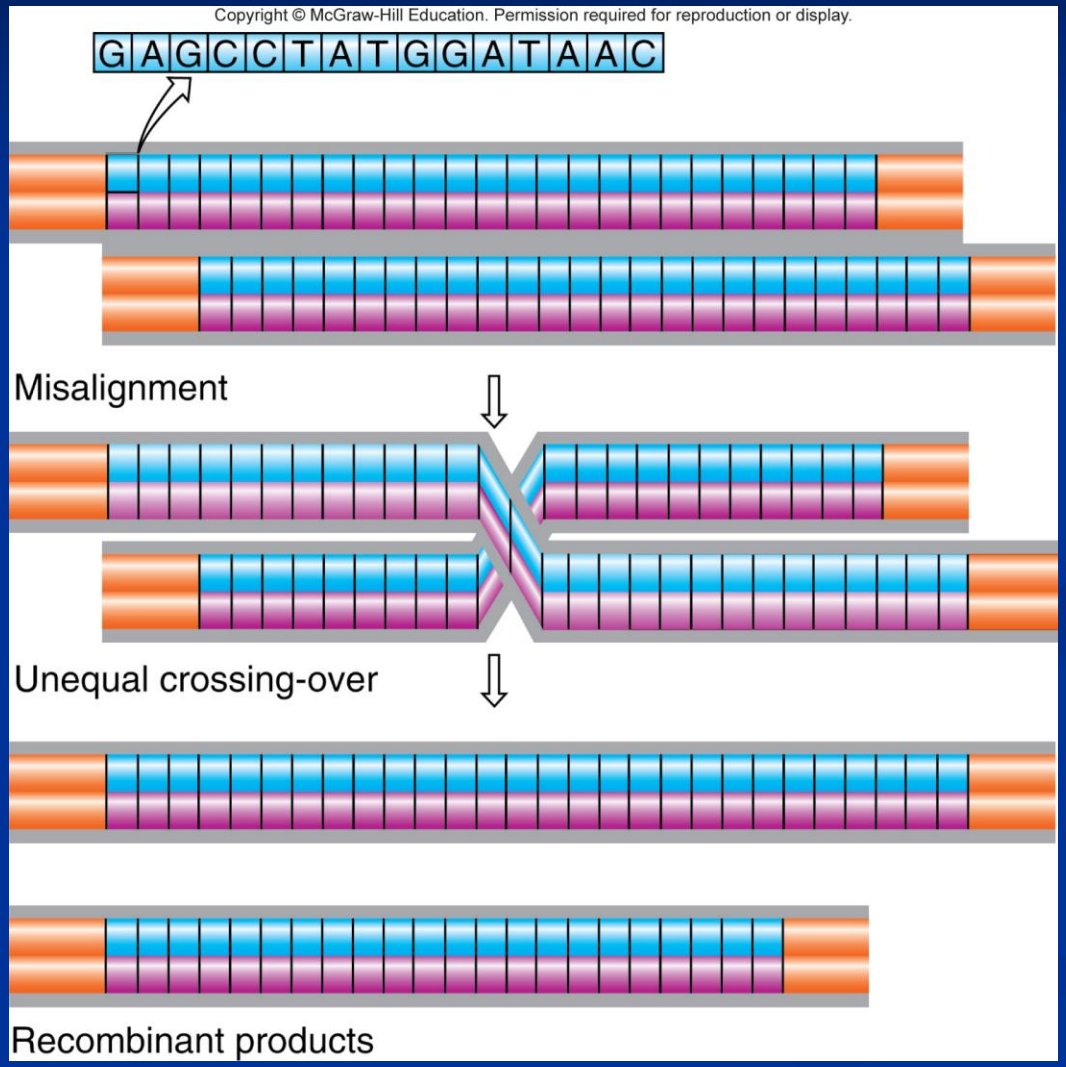
Copyright © McGraw-Hill Education. Permission required for reproduction or display.

TABLE 10.1 Categories of Genetic Variants

The right column shows how frequently on average you would find a polymorphism of the indicated class when comparing any two haploid human genomes.

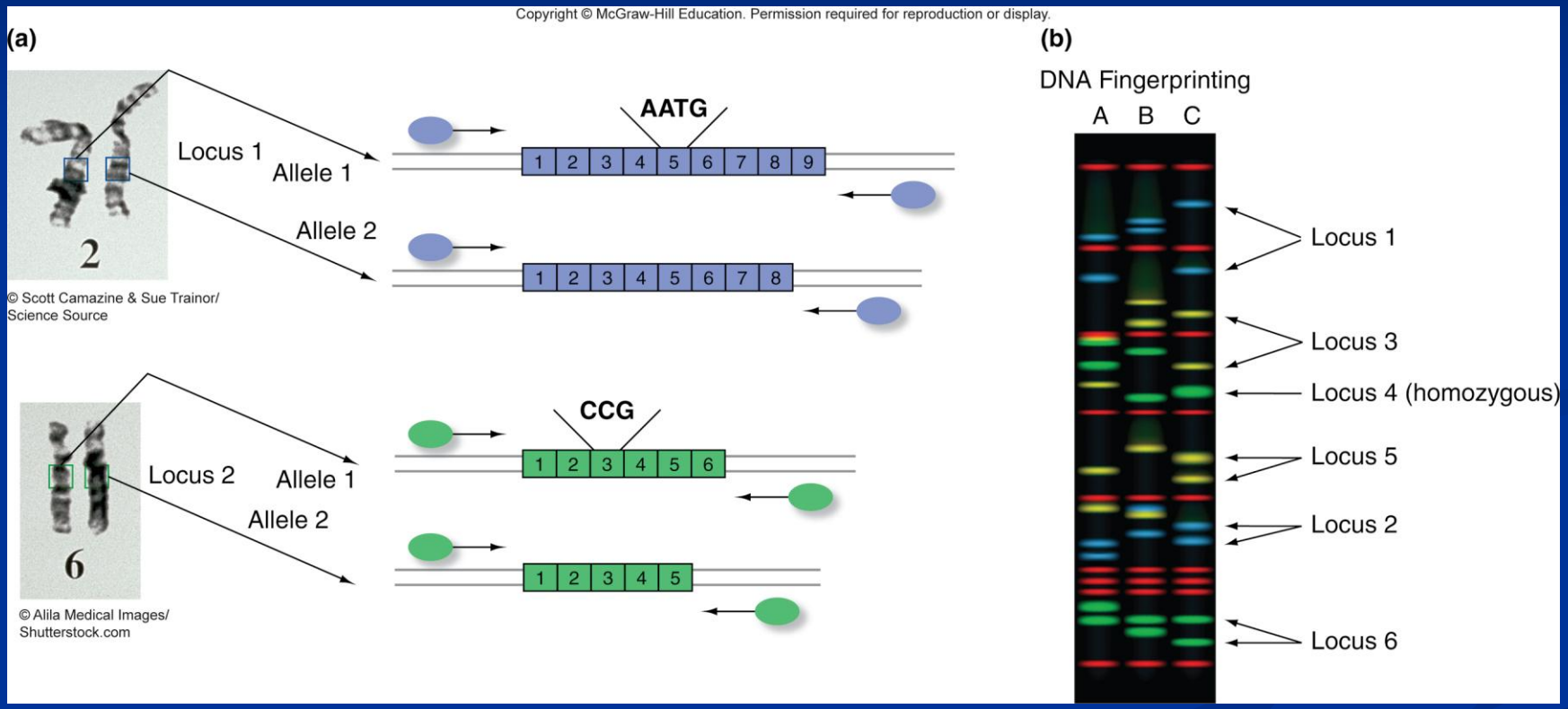
	Size	Frequency (1 per. . . .)
SNP Single nucleotide polymorphism	1 bp	1 kb
DIP or **Indel** Insertion/deletion	1–100 bp	10 kb
SSR or **Microsatellite** Simple sequence repeat	1–10 bp repeat unit	30 kb
CNV Copy number variant	10 bp–1 Mb	3 Mb

Minisatellites (Low-end CNVs)



- Repeating units 20-100 bp long
- Repeated 10-1000 times
- 1 per 100,000 bp, or about 30,000 in whole genome
- Highly polymorphic in length

Multiplex PCR is used for DNA fingerprinting

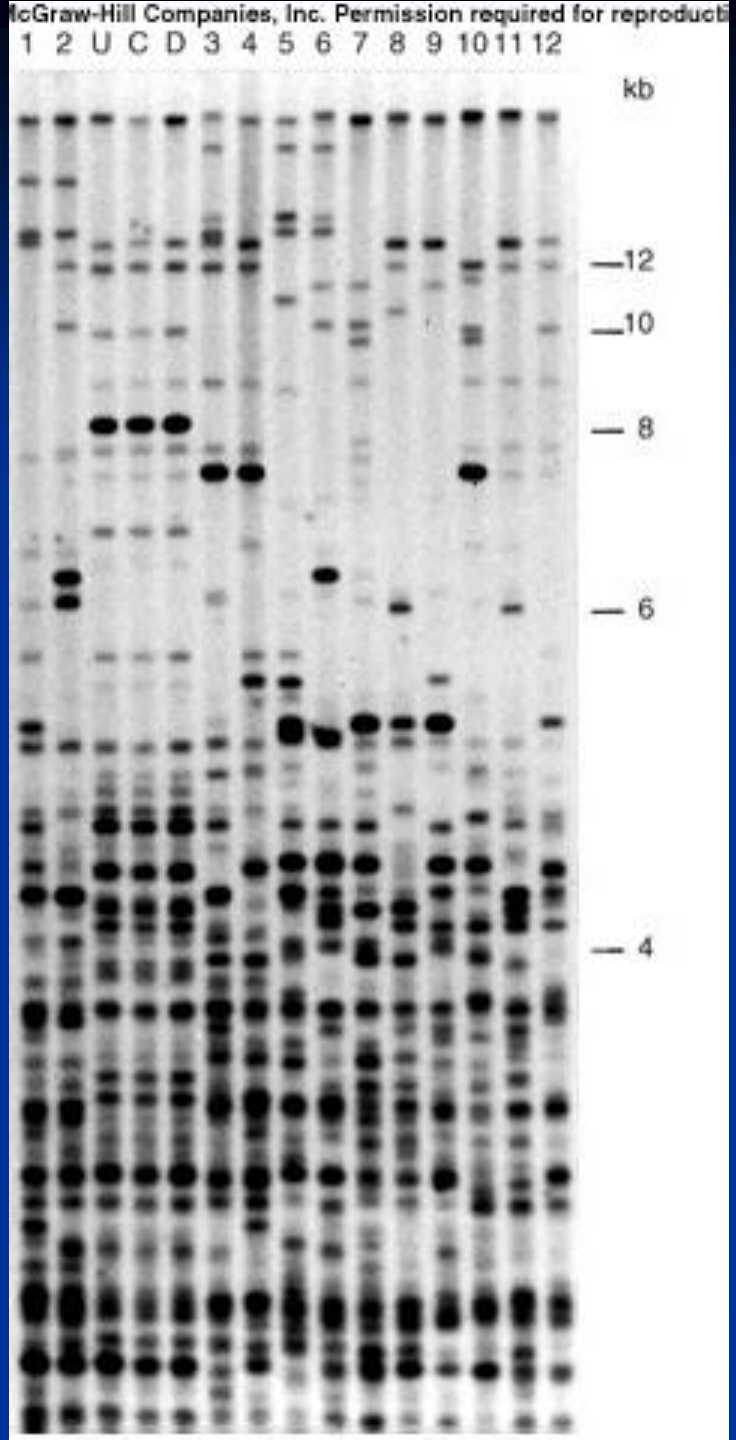


- Repeats of 500 bp to 20 kb
- Occur at a small number of different loci

Fig. 10.15 p. 354

Minisatellite detection and DNA fingerprinting

- 1985 – Alec Jeffreys made two key findings
  - Each minisatellite locus is highly polymorphic
  - Most minisatellites occur at multiple sites around the genome
  - DNA fingerprint – pattern of simultaneous genotypes at a group of unlinked loci
  - Use restriction enzymes and Southern blots to detect length differences at minisatellite loci



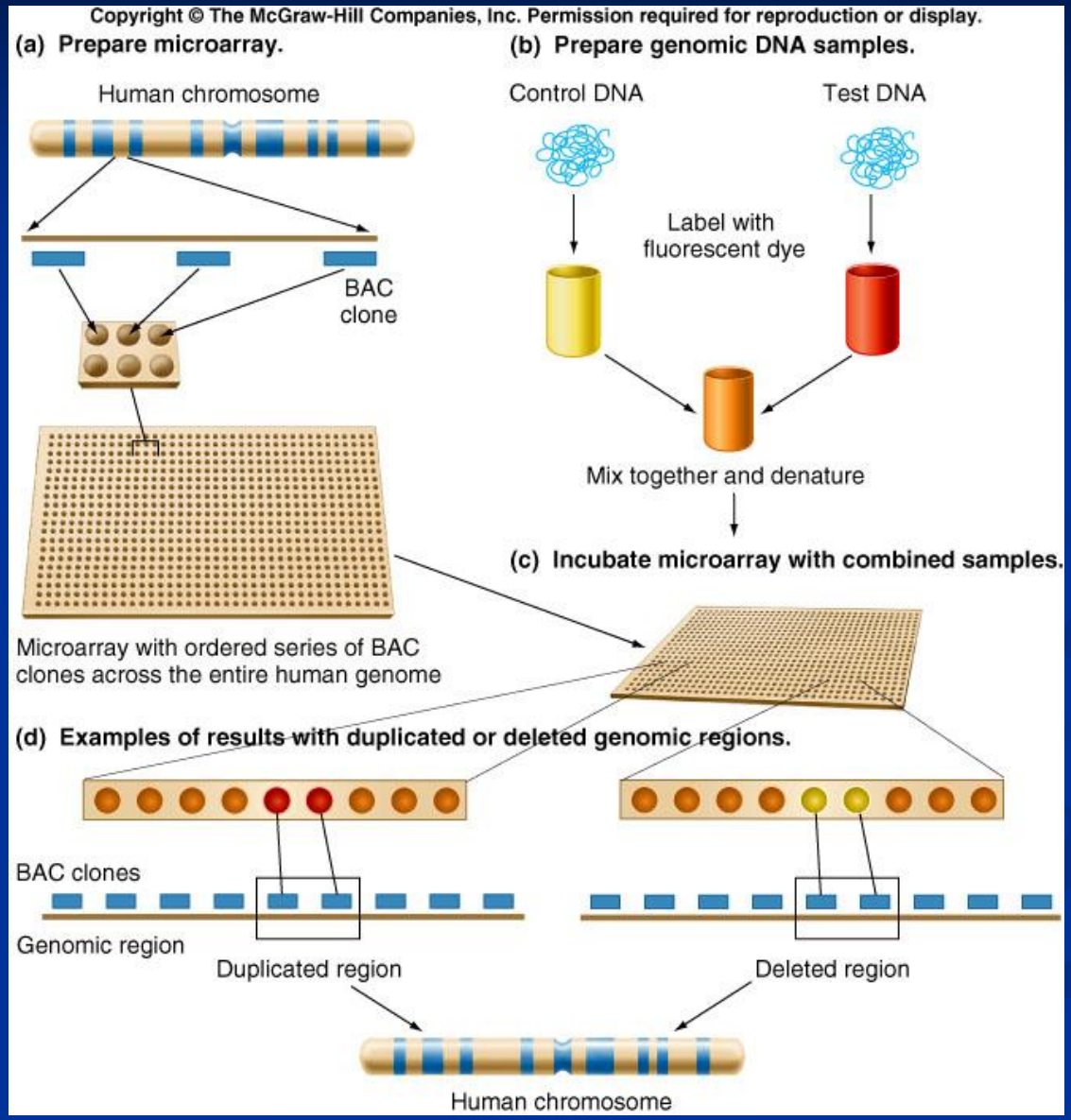
cGraw-Hill Companies, Inc. Permission required for reproducti

- DNA fingerprints can identify individuals and determine parentage
- E.g., DNA fingerprints confirmed Dolly the sheep was cloned from an adult udder cell
- Donor udder (U), cell culture from udder (C), Dolly's blood cell DNA (D), and control sheep 1-12

Larger scale deletions and duplications

- Deletions or duplications that cover one or more genes
  - Up to 1 Mb in length
  - Over 6000 loci identified in humans
  - Over 99% derived from inheritance rather than new mutations

Identifying large scale CNVs/CNPs



DNA microarrays and large scale CNVs

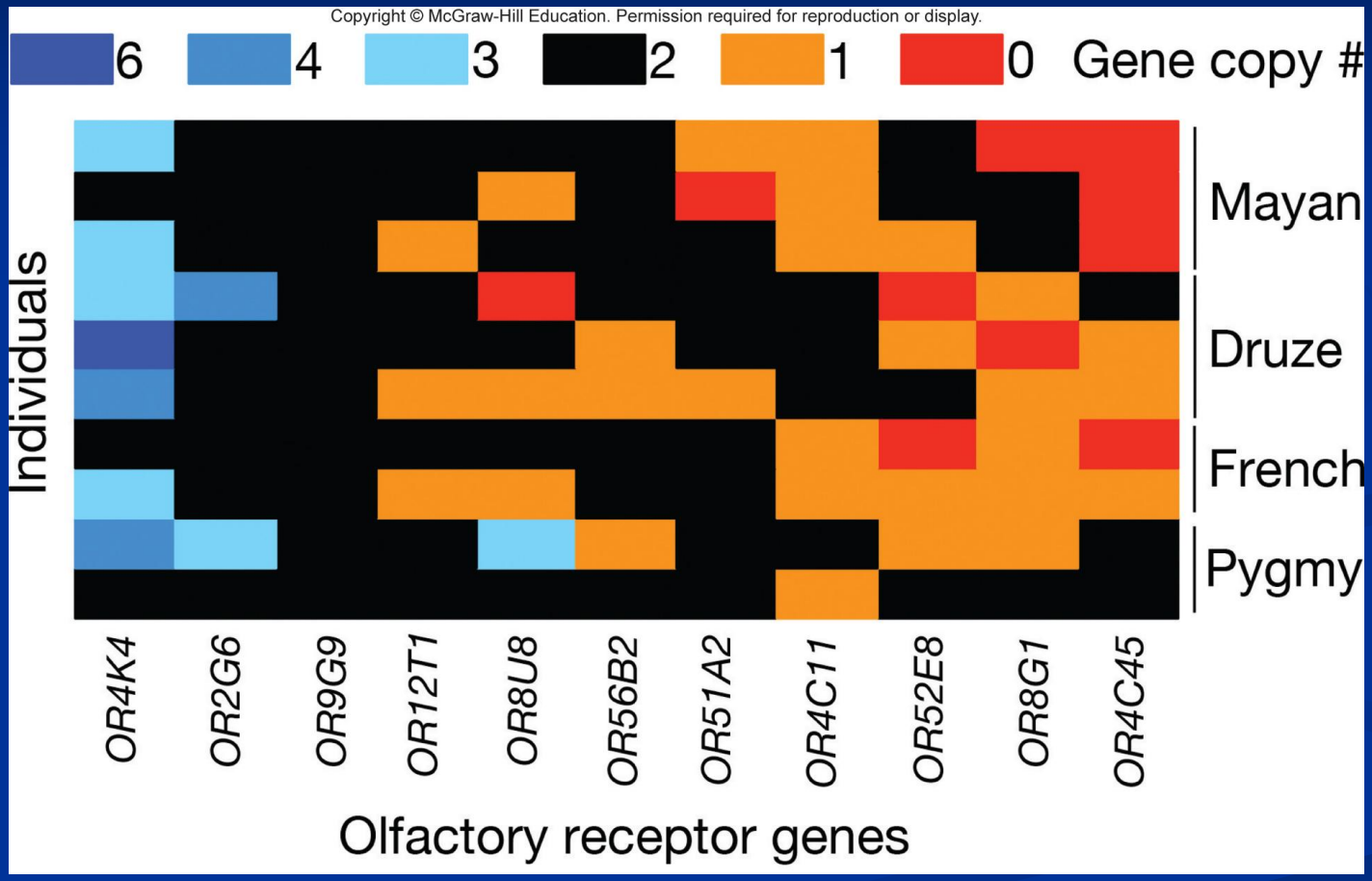
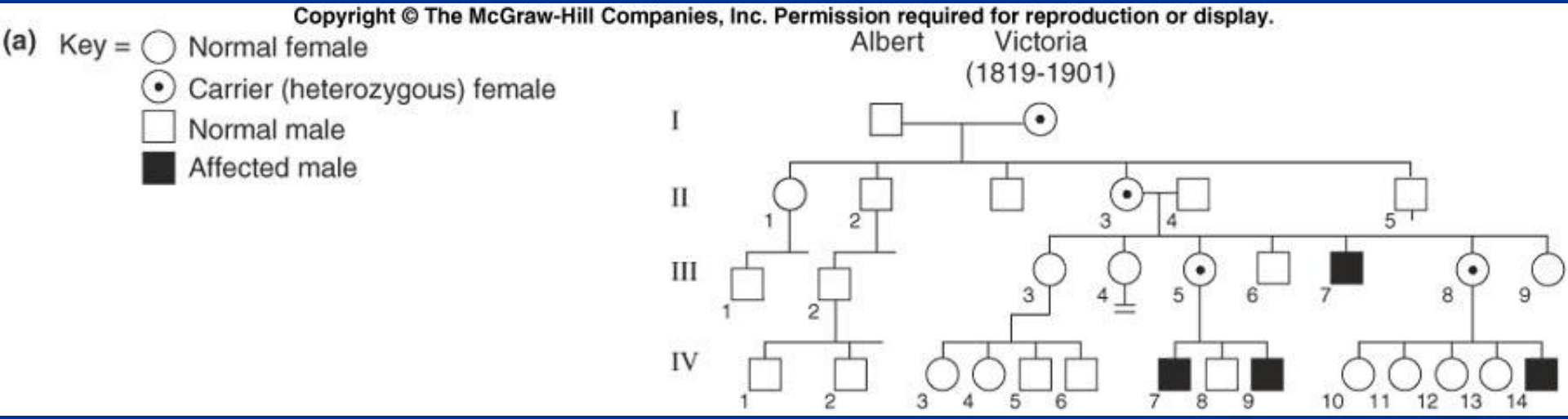


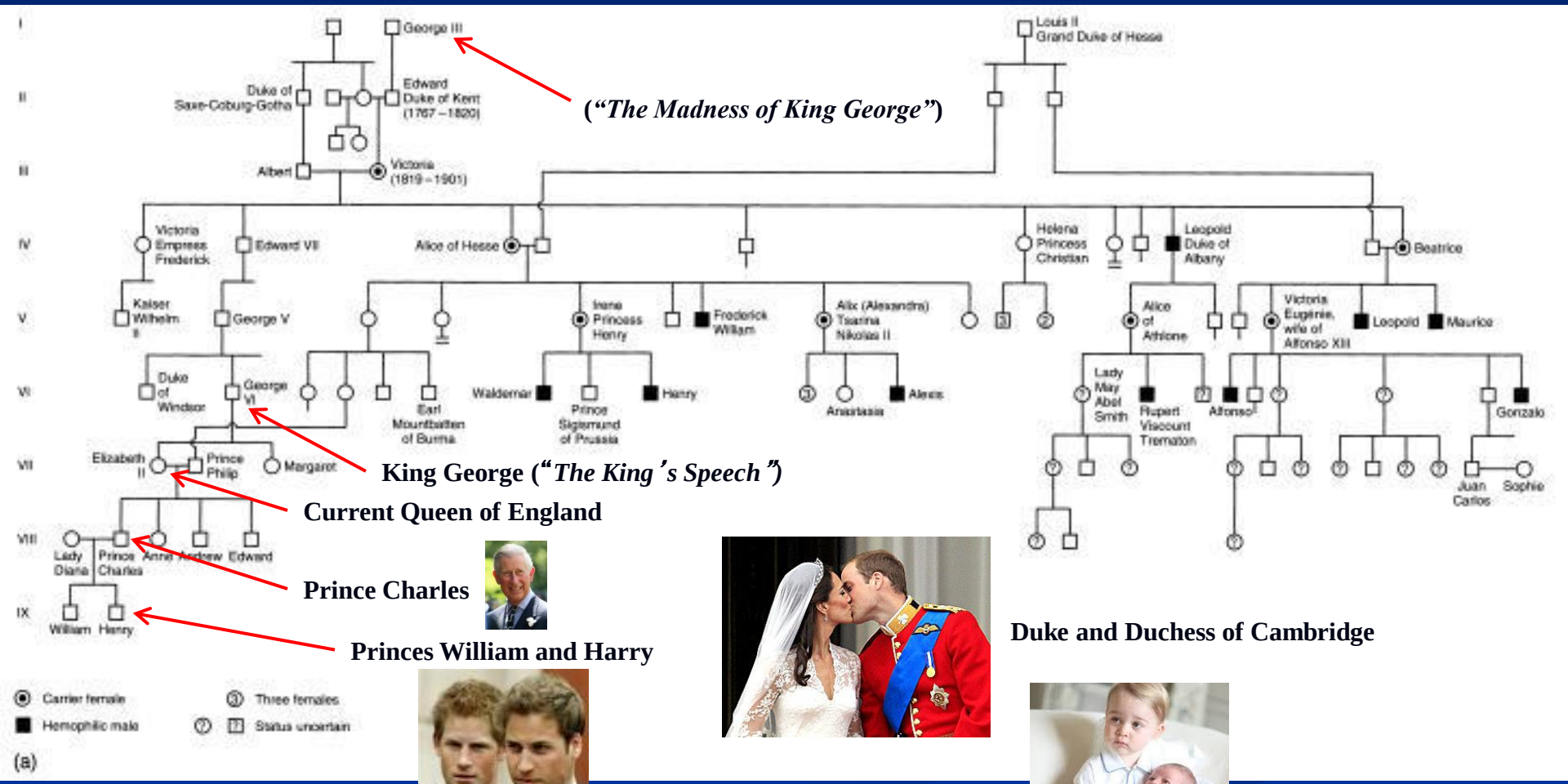
Fig. 10.8 p. 347

Cloning human genes

A pedigree of the royal family descended from Queen Victoria in which hemophilia A is segregating



British royal family and hemophilia



(*"The Madness of King George"*)

King George (*"The King's Speech"*)

Current Queen of England

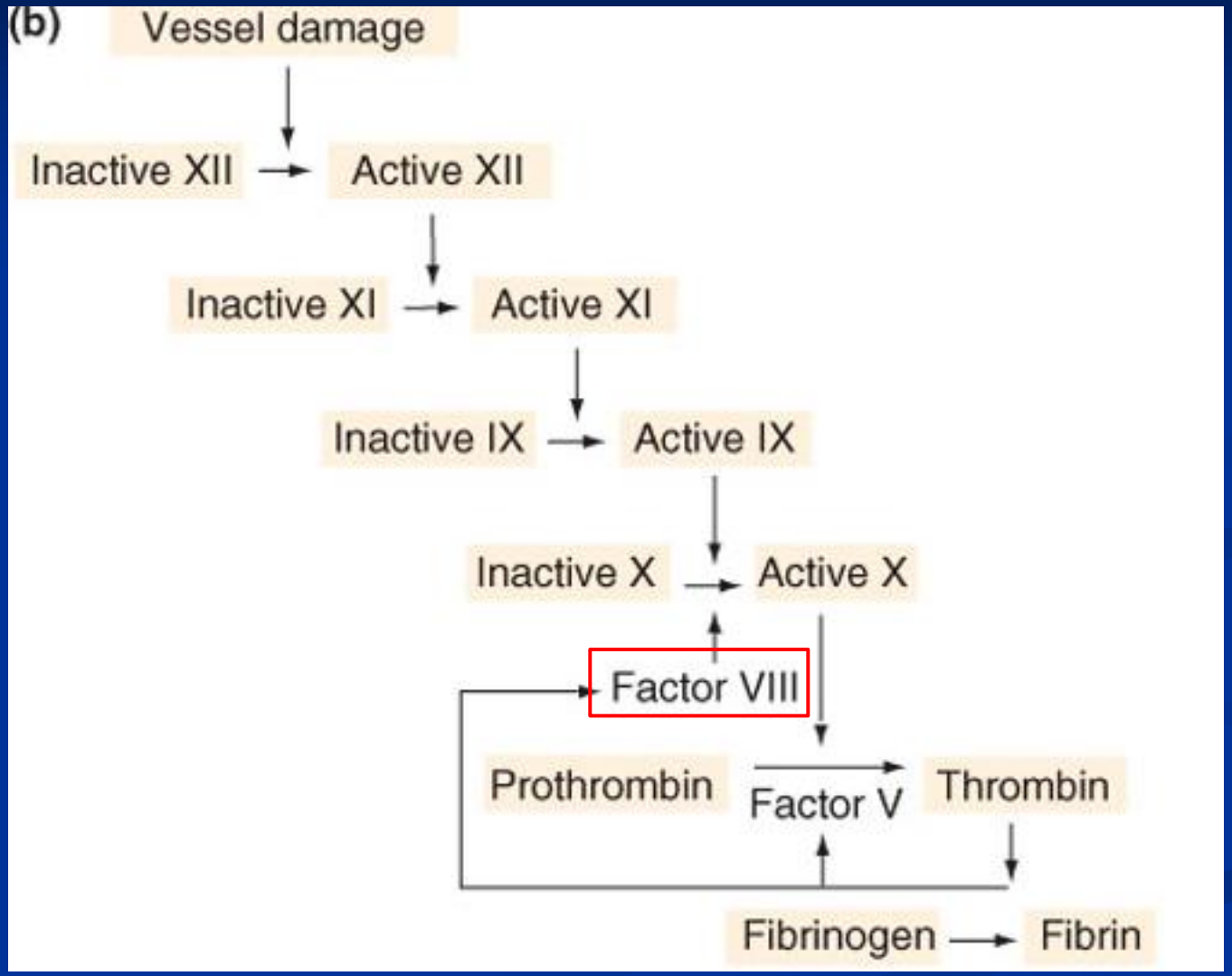
Prince Charles

Princes William and Harry

Duke and Duchess of Cambridge

Prince George and Princess Charlotte

Blood-Clotting Cascade



Blood tests determine if active form of each factor in the cascade is present

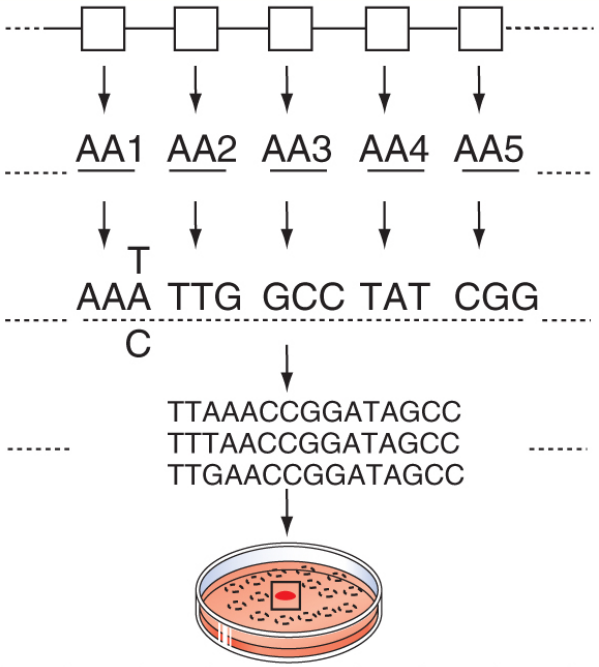
(c) Analysis of presence or absence of blood-clotting factors

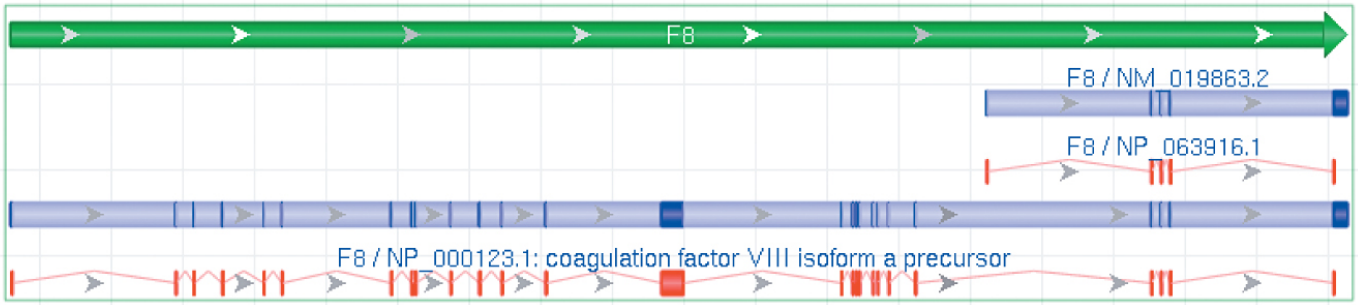
Factors	Wild-type person	Hemophiliac
XII	+	+
XI	+	+
IX	+	+
VIII	+	- ←
X	+	+
V	+	+
Prothrombin	+	+
Fibrinogen	+	+

Techniques used to purify Factor VIII and clone the gene

Copyright © The McGraw-Hill Companies, Inc. Permission required for reproduction or display.

(d) Purified factor VIII
↓
Obtain amino acid sequence
↓
Reverse translation into coding sequence
↓
Synthesize degenerate oligonucleotides
↓
Probe library
↓
Find factor VIII clone
↓
Structure of gene as determined from clone

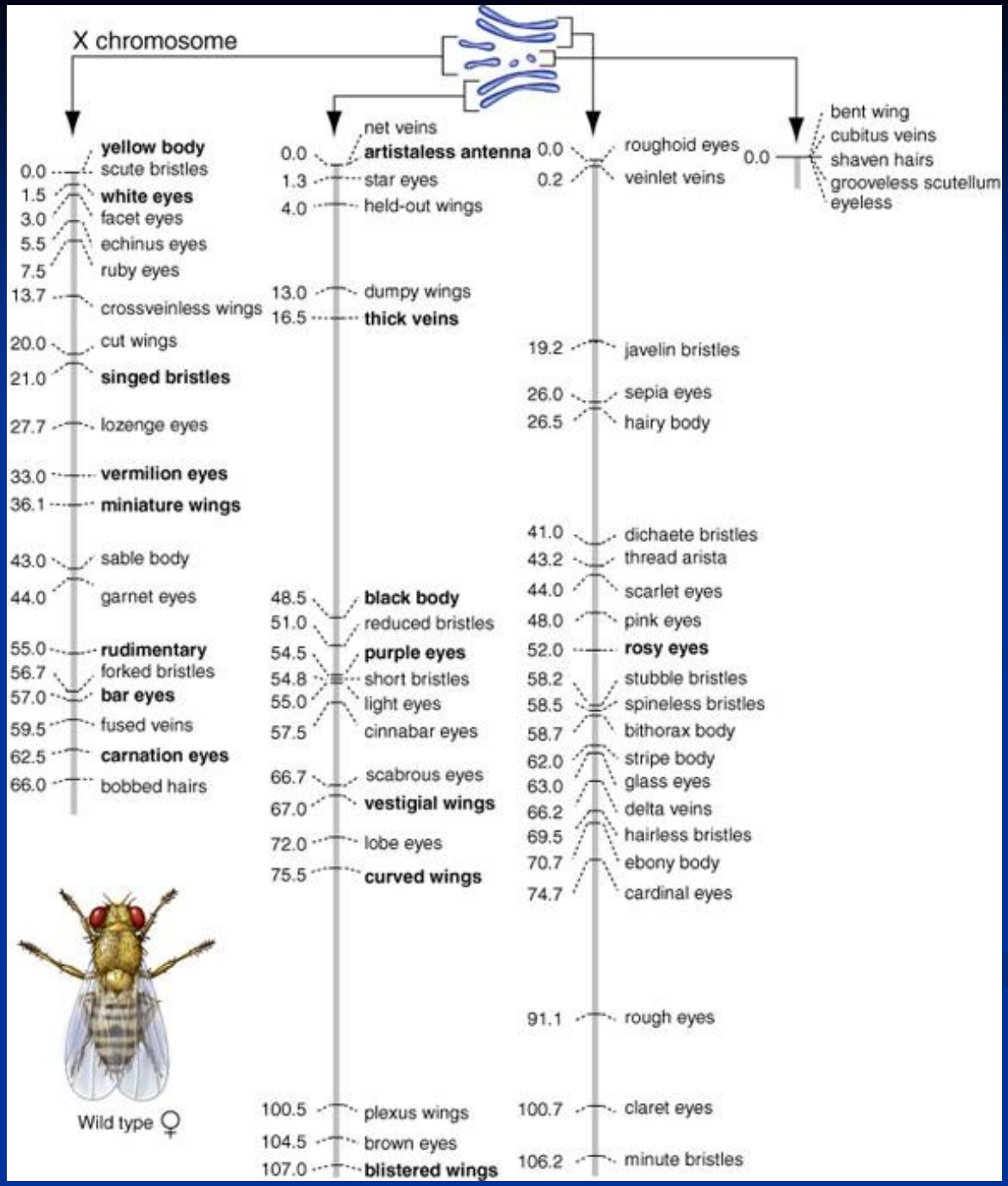




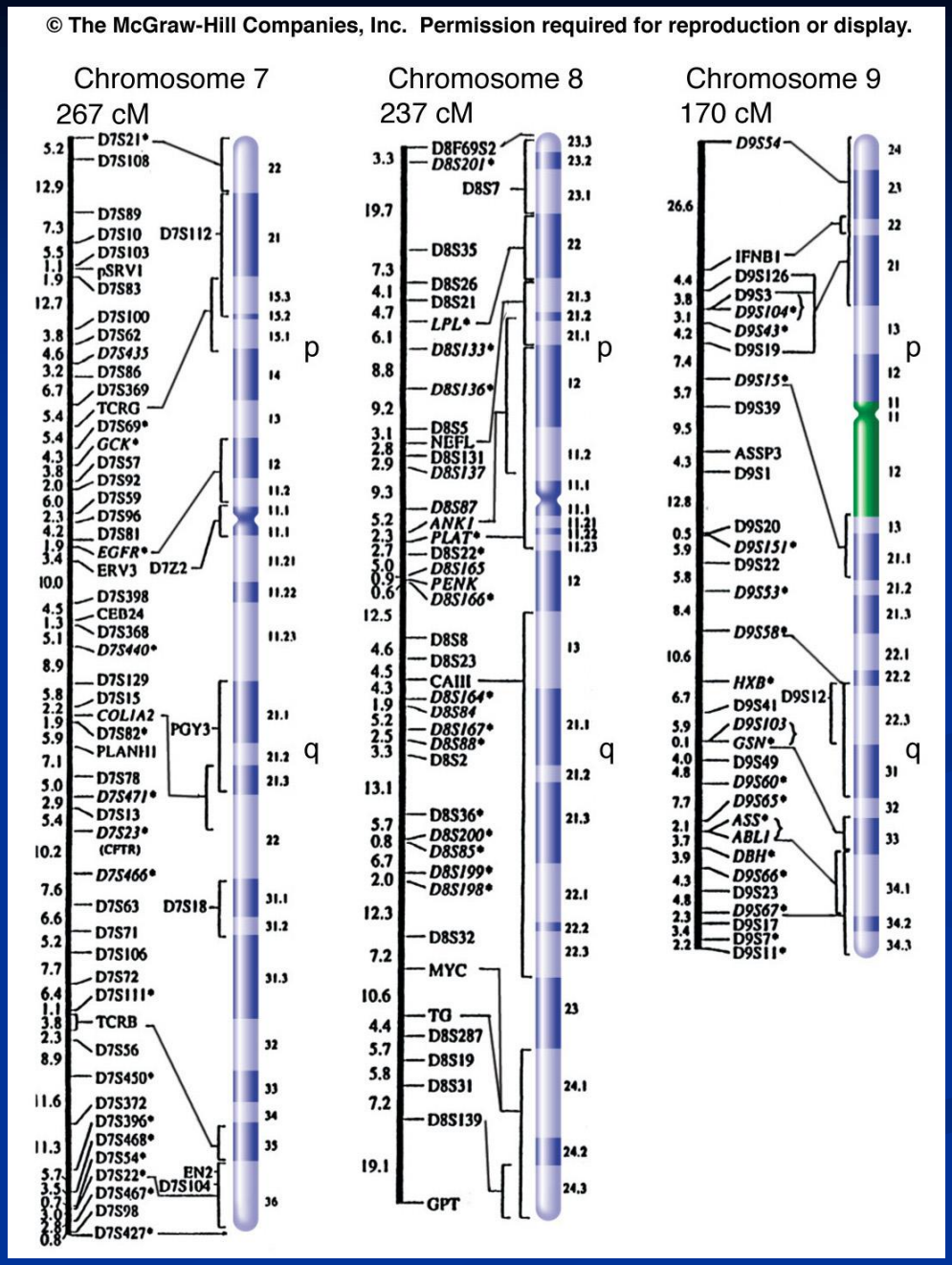
Summary of linkage and recombination

- Loci close together on the same chromosome are linked and do not segregate independently.
- Linked loci lead to a larger number of nonrecombinant offspring than expected in double heterozygotes.
- The farther away two loci are the greater the opportunity for crossing over to occur so….
- Recombination frequencies reflect relative distance between loci.
- Recombination frequencies between two loci vary from 0% to 50%.

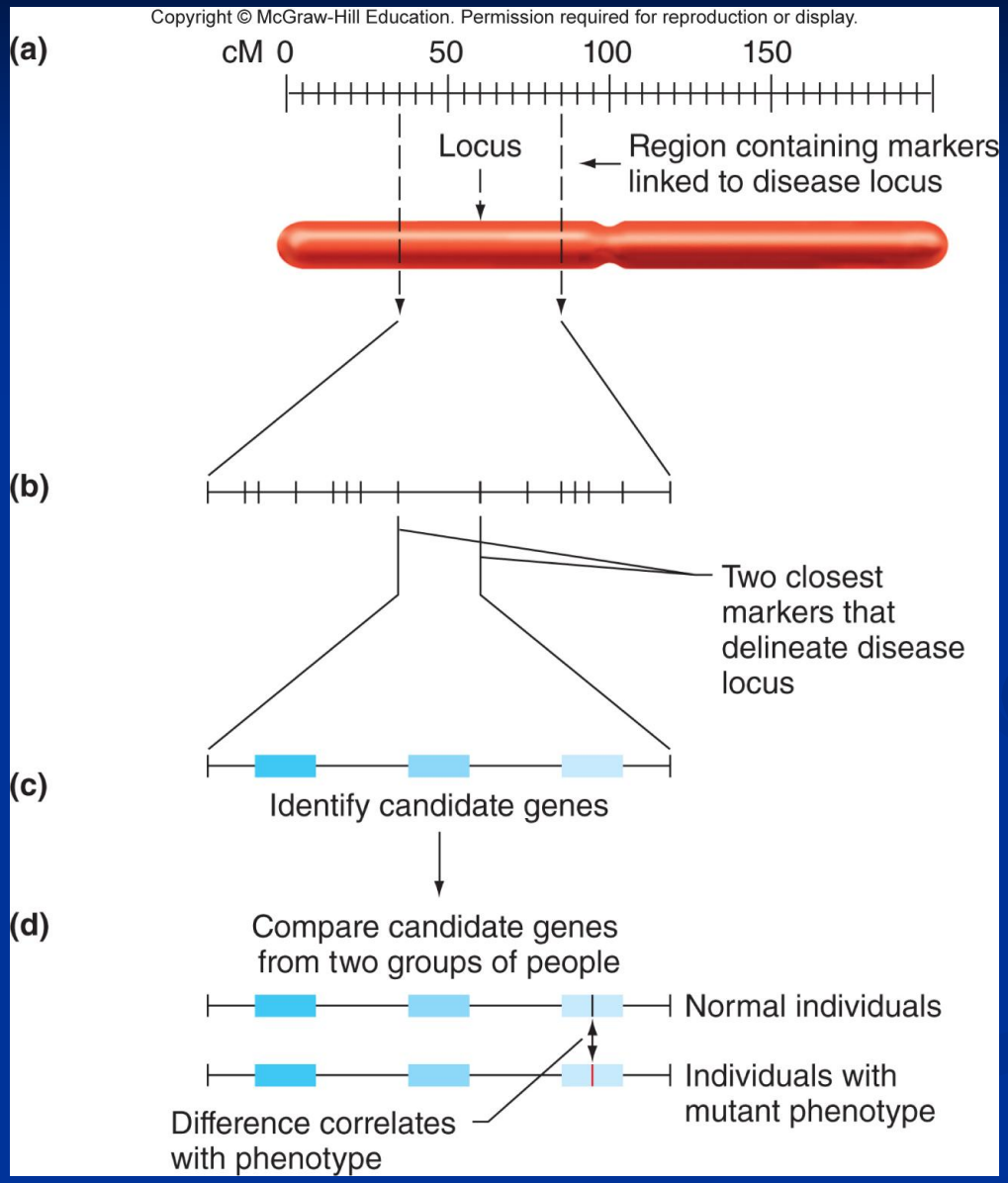
Genes or other DNA markers that are associated by linkage relationships are known as linkage groups.



Location of various DNA markers on human chromosomes



Positional Cloning



- **From phenotype to chromosomal location to defective gene**

Fig. 10.18 p. 357

Linkage to a DNA marker – Huntington disease

G8 marker – A, B, C or D alleles

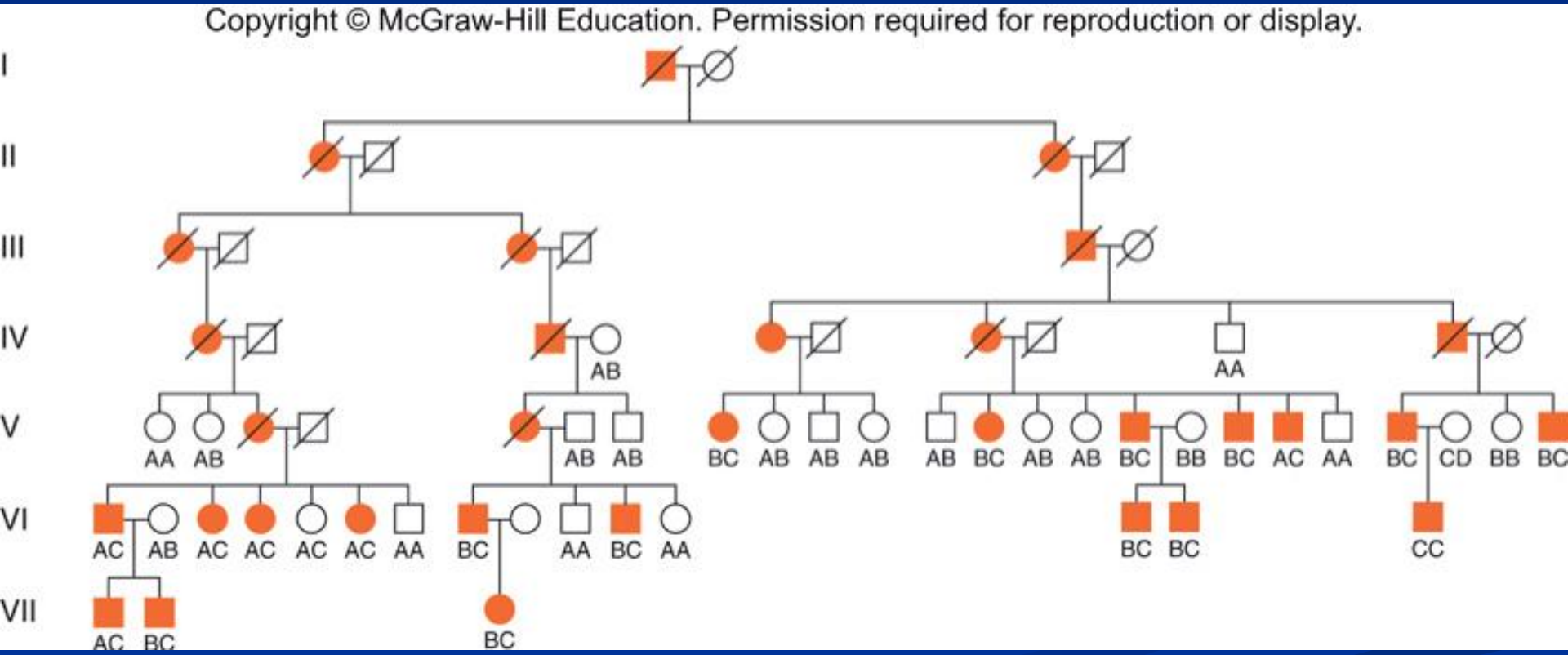
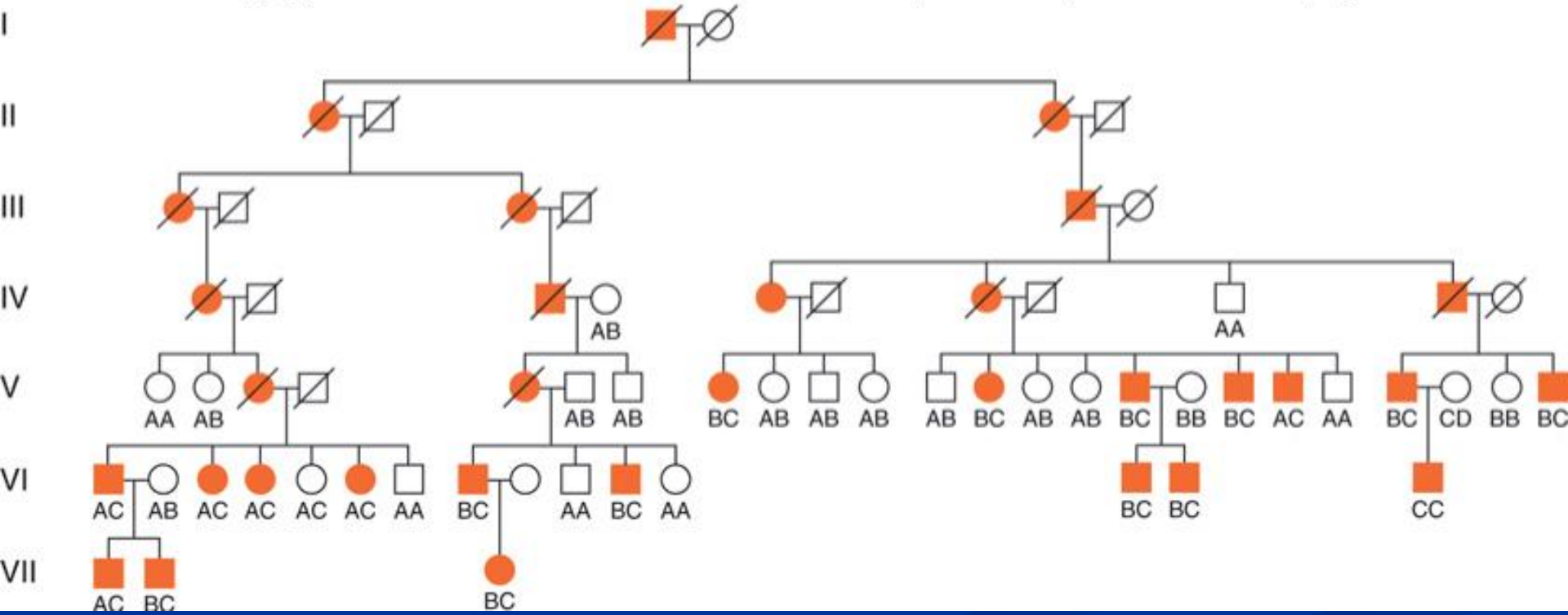


Fig. 10.22 p. 360

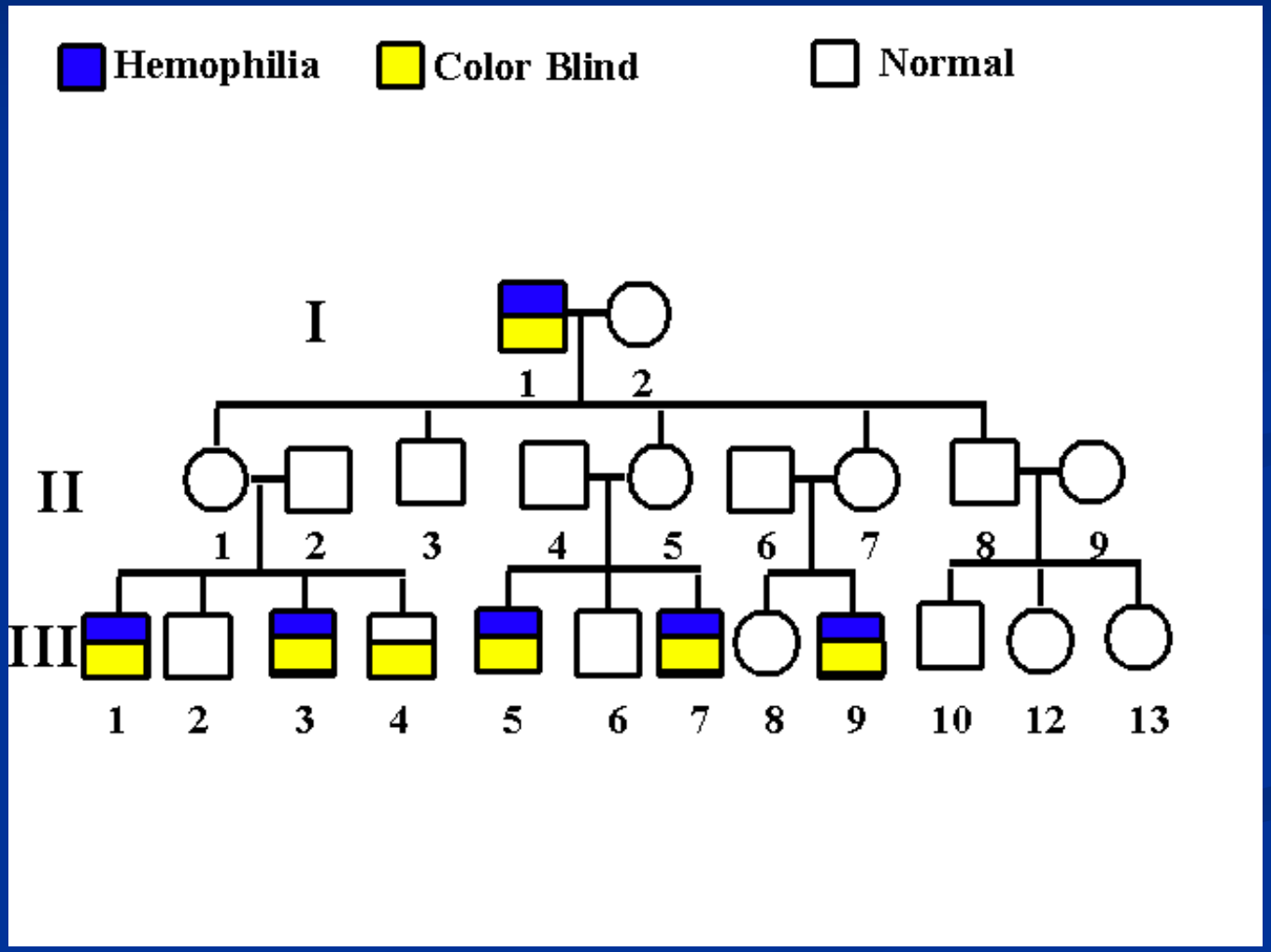
Copyright © McGraw-Hill Education. Permission required for reproduction or display.



1. For the large majority of people in the pedigree with Huntington disease, which allele of G8 (A, B C or D) did they inherit along with the dominant *HD* allele from their affected parent?

1. If you assume that G8 and *HD* are linked, how many people in this pedigree must be the product of a recombinant gamete from their affected parent?

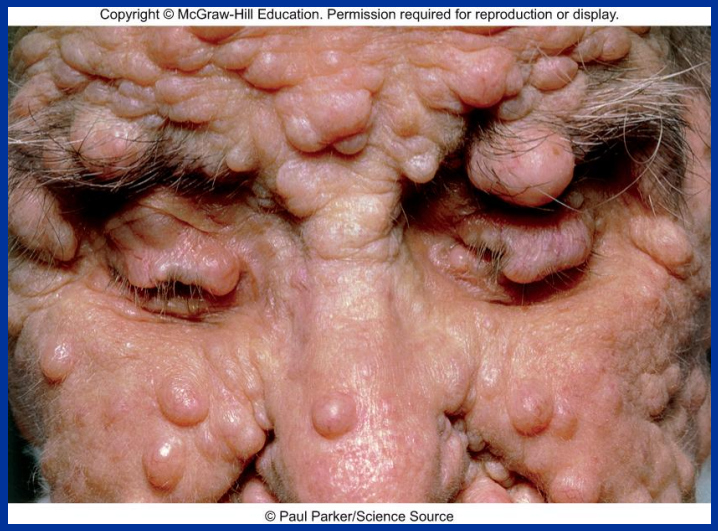
Linkage mapping in humans
"Grandfather method"



LOD score analysis

- Distinguishes between independent assortment (no linkage) and linkage
- Not dependent on knowledge of phase of doubly heterozygous parent
- Can combine information from a number of different families
- Gives estimate of recombination frequency (or genetic distance) in cases of linkage

Positional Cloning



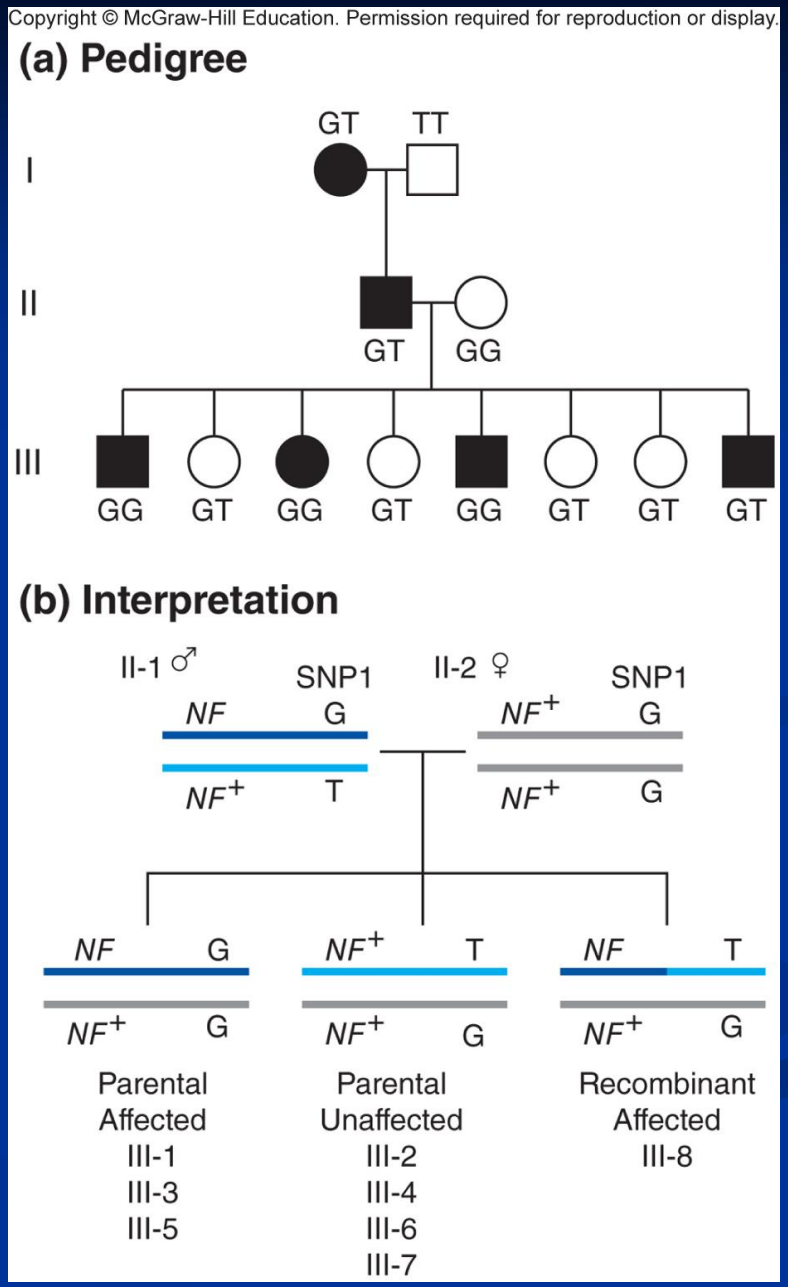
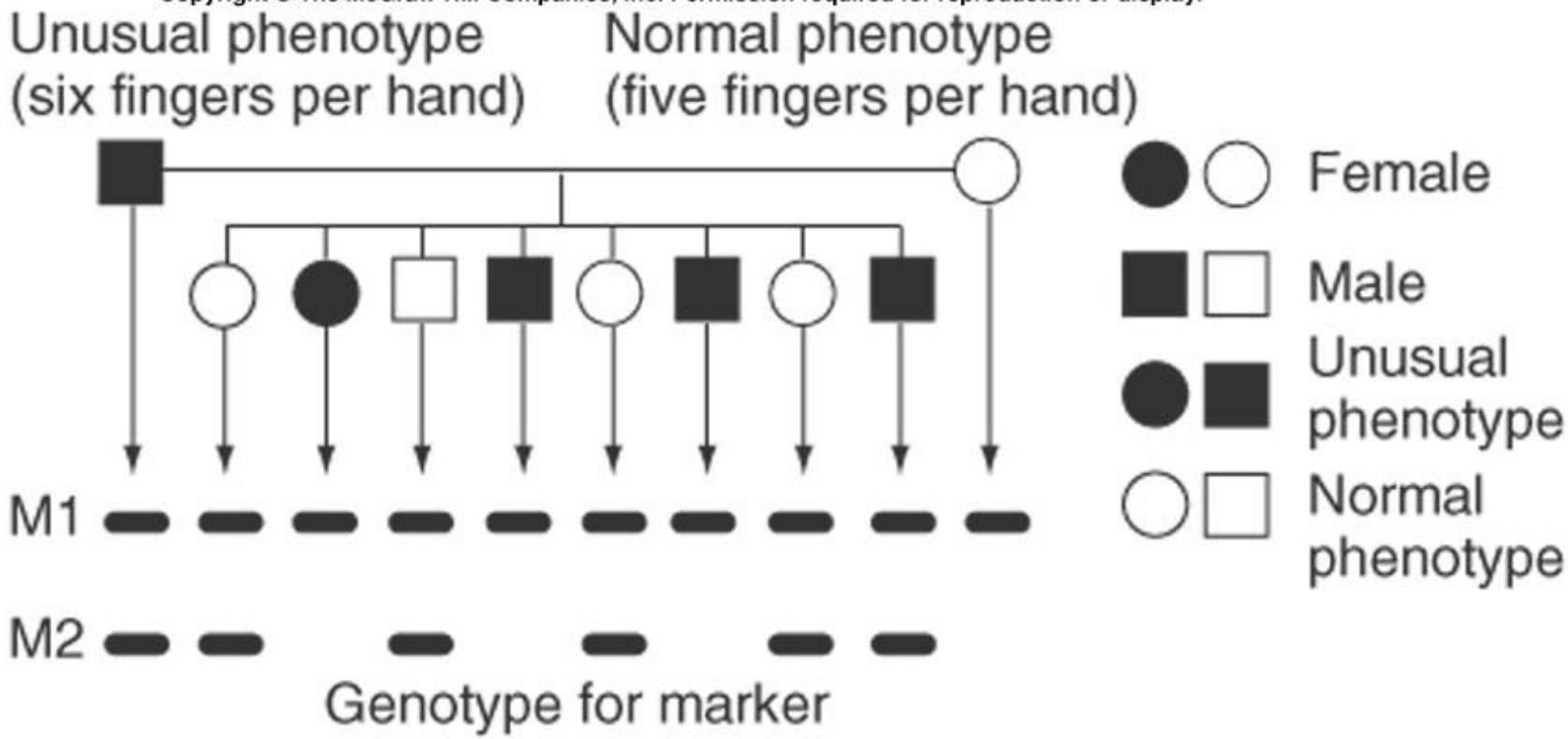


Fig. 10.19 p. 358

Copyright © The McGraw-Hill Companies, Inc. Permission required for reproduction or display.

(a)



(b)

Odds value (ratios)	LOD score (*Z*)
0.01 (1:100)	-2
0.1 (1:10)	-1
1 (1:1)	0
10 (10:1)	1
100 (100:1)	2
1000 (1000:1)	3

Phase 1:

1 nonrecombinant

7 recombinant

Phase 2:

7 nonrecombinant

1 recombinant

LOD score analysis

LOD score (Logarithm of the ODds) = log of the ratio of the likelihood of linkage versus the likelihood of no linkage

How much more likely is it that the data support linkage versus no linkage?

LOD score (Z) = Log [L (θ)/ L (0.50)]

theta θ = recombination fraction

LOD score analysis

Linkage:

When two loci are linked $0 \leq \theta \leq 0.50$

No linkage:

If $L(\theta) = L(0.50)$ then the two loci are considered NOT linked

LOD score analysis

How do we determine L (θ) in families with disease?

It depends on the probability of recombinant and nonrecombinant chromosomes among siblings in the study

LOD score analysis

Probability that offspring will receive either nonrecombinant chromosome

$$1/2\ (1-\theta) + 1/2\ (1-\theta) \text{ or } (1-\theta)$$

Probability that offspring will receive either recombinant chromosome

$$1/2\ (\theta) + 1/2\ (\theta) \text{ or } \theta$$

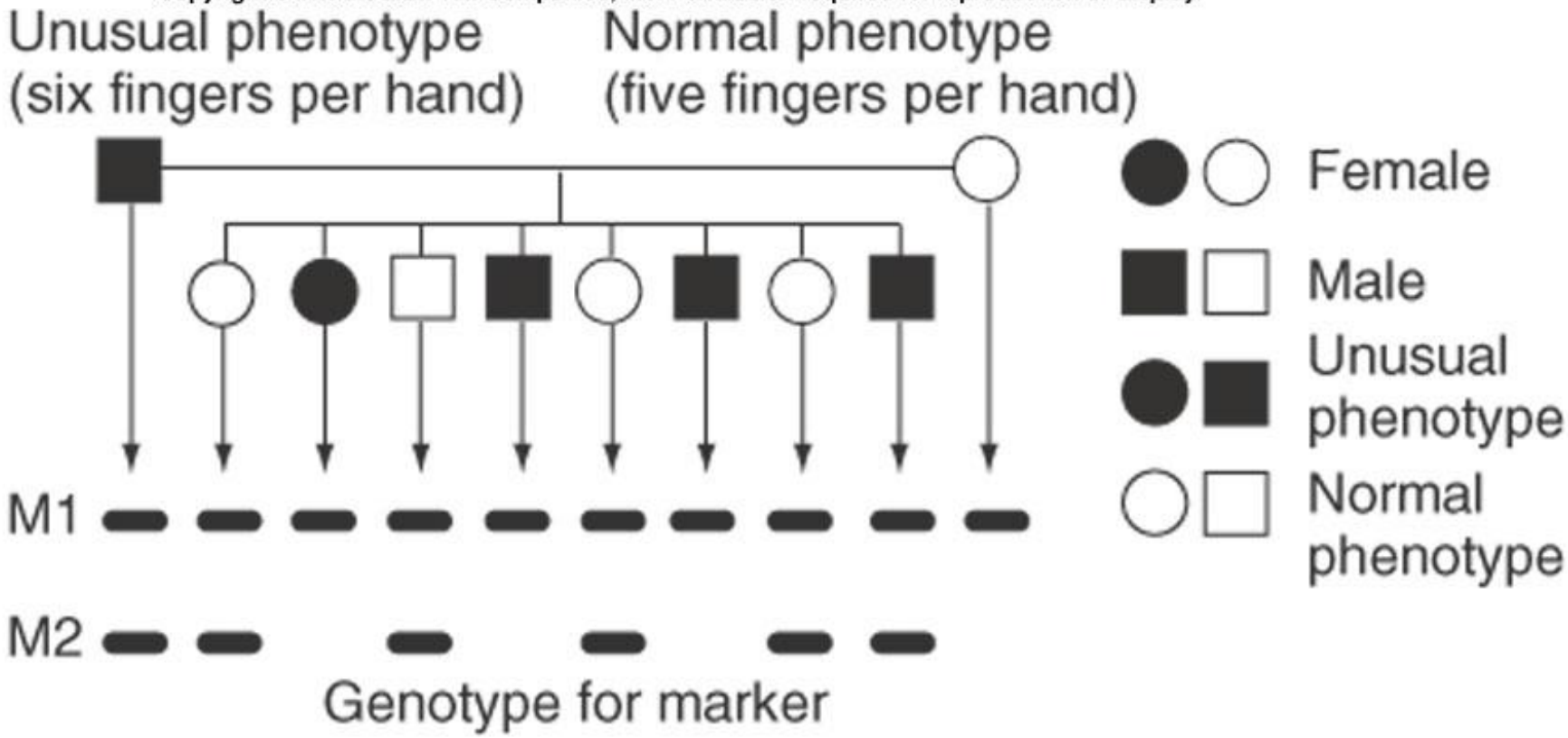
LOD score analysis

5 siblings: What is the probability of each individual receiving a <u>nonrecombinant</u> chromosome from the doubly heterozygous parent? $(1-\theta)^5$

4 siblings: What is the probablility of each individual receiving a <u>recombinant</u> chromosome from the doubly heterozygous parent? $(\theta)^4$

Copyright © The McGraw-Hill Companies, Inc. Permission required for reproduction or display.

(a)



(b)

Odds value (ratios)	LOD score (*Z*)
0.01 (1:100)	-2
0.1 (1:10)	-1
1 (1:1)	0
10 (10:1)	1
100 (100:1)	2
1000 (1000:1)	3

Phase 1:

1 nonrecombinant

7 recombinant

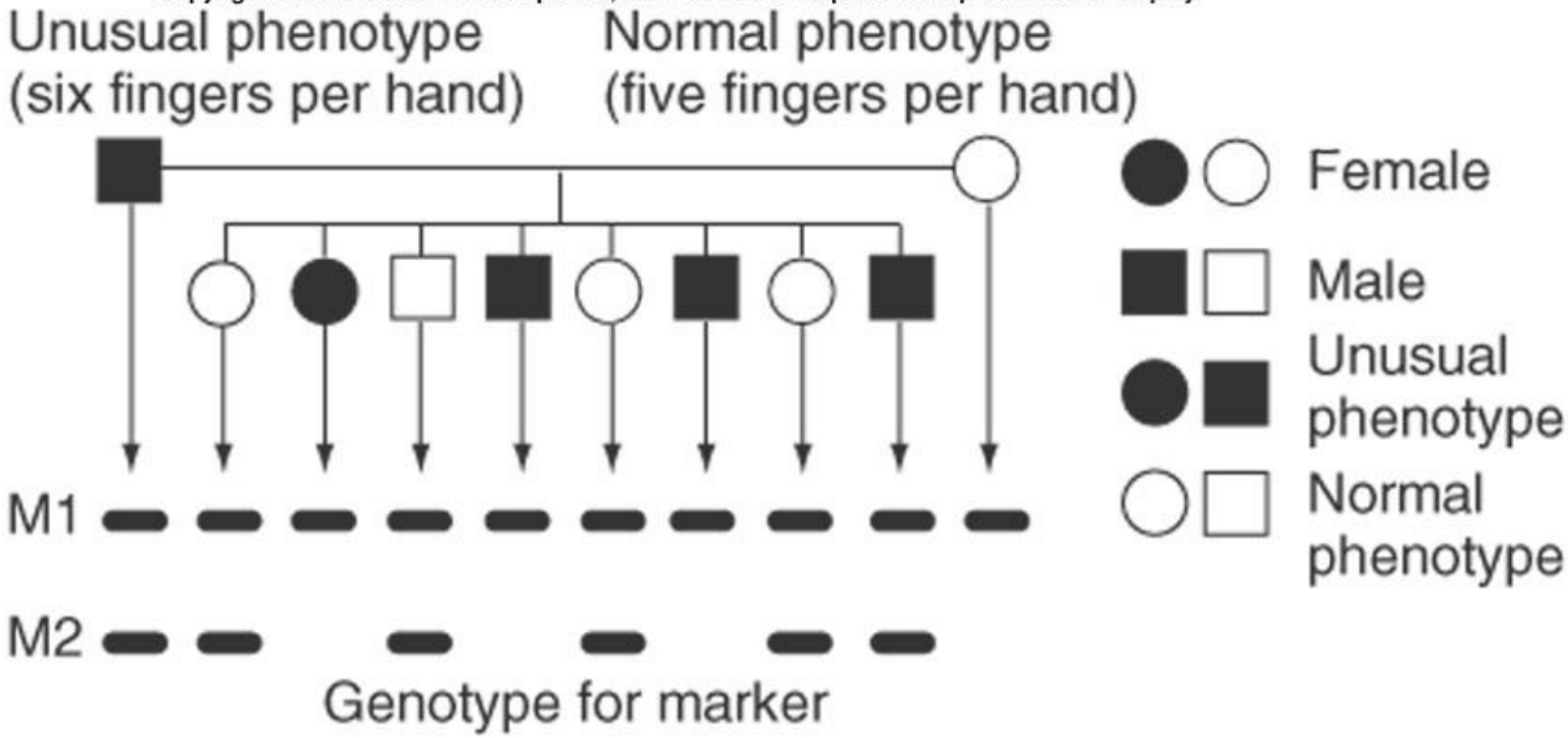
Phase 2:

7 nonrecombinant

1 recombinant

Copyright © The McGraw-Hill Companies, Inc. Permission required for reproduction or display.

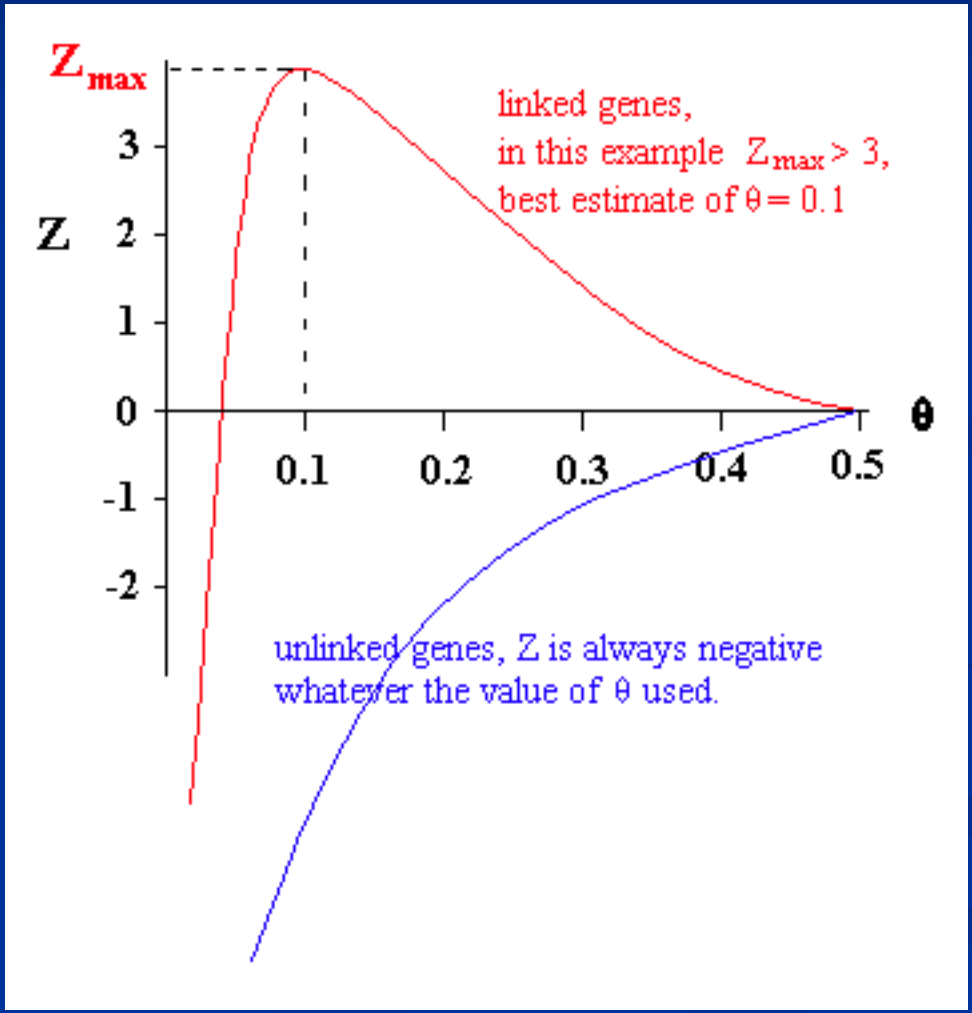
(a)



(b)

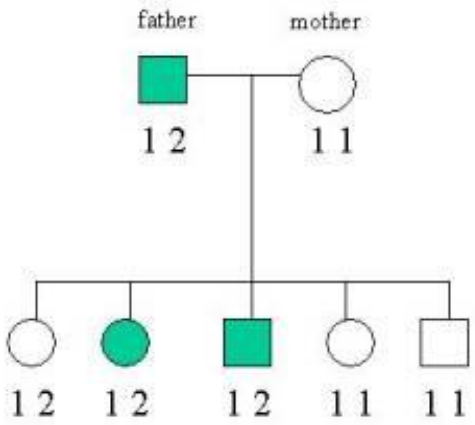
Odds value (ratios)	LOD score (*Z*)
0.01 (1:100)	-2
0.1 (1:10)	-1
1 (1:1)	0
10 (10:1)	1
100 (100:1)	2
1000 (1000:1)	3

LOD score curve



Example pedigree

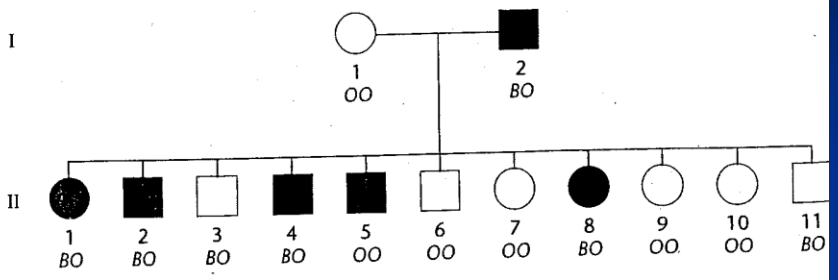
Dominant trait



- **Phase unknown**
- **How do we calculate LOD score (Z)?**

Example problem

Figure 3.19 A sample pedigree for demonstrating the LOD score method. Solid symbols represent individuals with nail-patella syndrome. The letters below each symbol designate the ABO blood group genotype. The ABO blood group alleles are abbreviated with *O* and *B* for *ABO*O* and *ABO*B*, respectively.



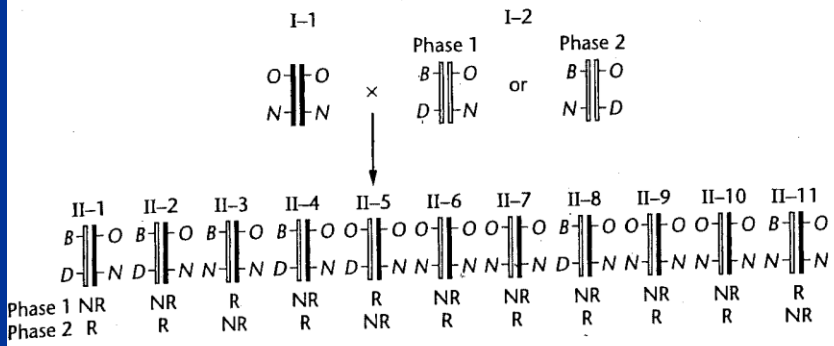
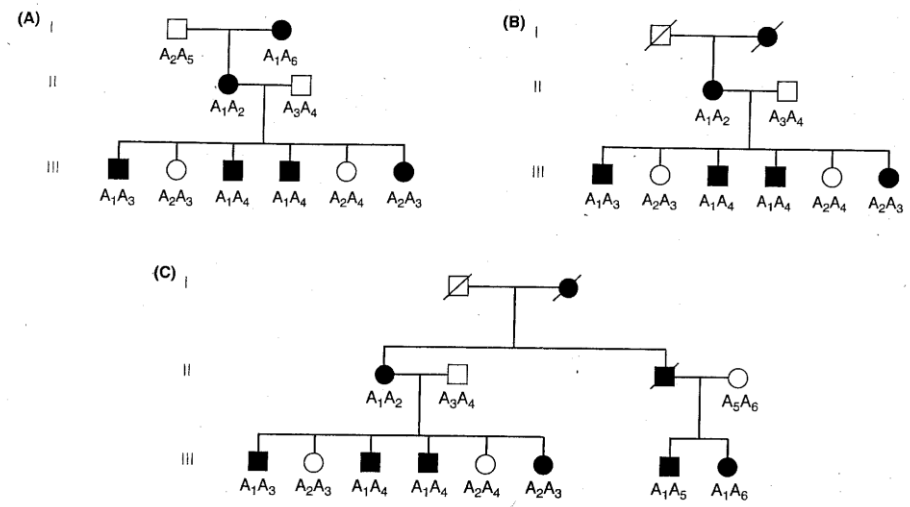


Figure 3.20 The genetic organization of nail-patella and ABO blood group alleles of each member of the pedigree shown in Figure 3.19, under the assumption that these two loci are linked. The ABO blood group allele designations are abbreviated with *O* and *B* for *ABO*O* and *ABO*B*, respectively. The normal and dominant alleles for the nail-patella syndrome locus are presented as *N* and *D*, respectively. Individual I-2 could be in either of two genetic phases (phase 1, phase 2). The bars with no fill represent the chromosomes of individual I-2 and the chromosomes inherited by the offspring from I-2. The filled bars represent chromosomes of individual I-1 and the chromosomes inherited by offspring from I-1. For phase 1 and phase 2, the chromosome inherited from individual I-2 by each offspring is identified as nonrecombinant (NR) or recombinant (R).

Example problem



Box 11.3

Calculation of lod scores for the families in *Figure 11.4*

- Given that the loci are truly linked, with recombination fraction θ, the likelihood of a meiosis being nonrecombinant is $1 - \theta$ and the likelihood of it being recombinant is θ.
- If the loci are in fact unlinked, the likelihood of a meiosis being either recombinant or nonrecombinant is 1/2.

Family A

There are five recombinants and one nonrecombinant.
The overall likelihood, given linkage, is $(1 - \theta)^5.\theta$
The likelihood given no linkage is $(1/2)^6$
The likelihood ratio is $(1 - \theta)^5.\theta / (1/2)^6$
The lod score, Z, is the logarithm of the likelihood ratio.

θ	0	0.1	0.2	0.3	0.4	0.5
Z	– infinity	0.577	0.623	0.509	0.299	0

Family B

II_1 is phase-unknown.
If she inherited A_1 with the disease, there are five nonrecombinants and one recombinant.
If she inherited A_2 with the disease, there are five recombinants and one nonrecombinant.
The overall likelihood is $\frac{1}{2}[(1 - \theta)^5.\theta / (1/2)^6] + \frac{1}{2}[(1 - \theta).\theta^5 / (1/2)^6]$. This allows for either possible phase, with equal prior probability.
The lod score, Z, is the logarithm of the likelihood ratio.

θ	0	0.1	0.2	0.3	0.4	0.5
Z	– infinity	0.276	0.323	0.222	0.076	0

Family C

At this point nonmasochists turn to the computer.

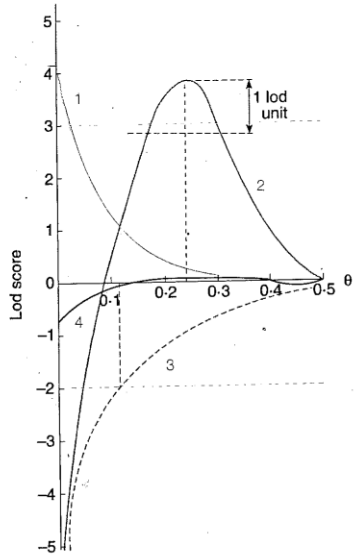


Figure 11.5: Lod score curves.

Graphs of lod score against recombination fraction from a hypothetical set of linkage experiments. Curve 1: evidence of linkage ($Z > 3$) with no recombinants. Curve 2: evidence of linkage ($Z > 3$) with the most likely recombination fraction being 0.23. Curve 3: linkage excluded ($Z < -2$) for recombination fractions below 0.12; inconclusive for larger recombination fractions. Curve 4: inconclusive at all recombination fractions.

3. Individuals from three moderately sized families with the same disorder were screened with polymorphic markers, and LOD scores were determined. Which marker(s) is linked to the disease-causing locus?

Family	D7S472	D7S641	D7S2508	D7S2458	D7S496
A	−1.79	−0.29	1.76	1.36	0.68
B	1.28	−0.69	0.76	2.54	−0.39
C	−0.43	1.24	−0.39	0.69	1.12

4. All the members of two very large families with the same phenotype were screened with a set of polymorphic markers. LOD scores for one of the markers are shown below. Explain these results.

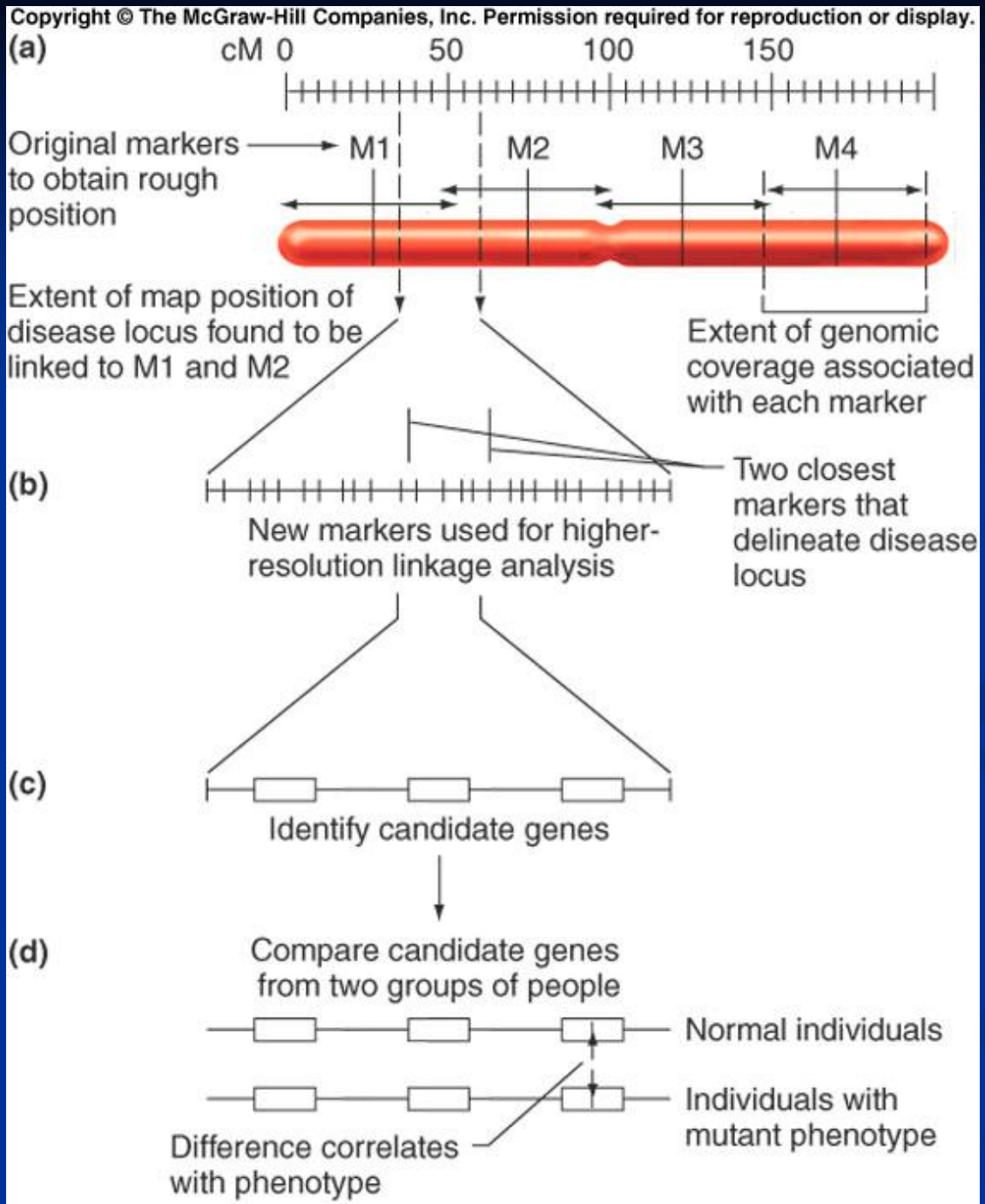
	$\theta =$ 0.05	0.10	0.15	0.20	0.25	0.30	0.35
Family A	3.96	4.35	5.65	4.45	3.25	2.98	1.54
Family B	−2.91	−1.69	0.79	−0.25	−1.34	0.78	1.12

Positional Cloning – Step 1

- Find extended families in which disease is segregating
- Use panel of polymorphic markers spaced at ~10 cM intervals across all chromosomes
- Determine genotype for all individuals in families for each DNA marker
- Look for linkage between a marker and disease phenotype by calculating LOD (Logarithm of the ODds) scores

Positional cloning – Step 2 identifying candidate genes

- Once region of chromosome has been narrowed down by linkage analysis to 1000 kb or less, all genes in that region are identified
- Candidate genes
  - Usually about 17 genes per 1000 kb fragment
  - Identify coding regions
    - Computational analysis to identify conserved sequences between species
    - Computational analysis to identify exon-like sequences by looking for codon usage, ORFs, and splice sites
    - Appearance in one or more EST (expressed sequence tag) databases



- **Once region of chromosome is determined, candidate genes are identified**

Fig. 11.17 p. 386

1503 genes
on chromosome 7

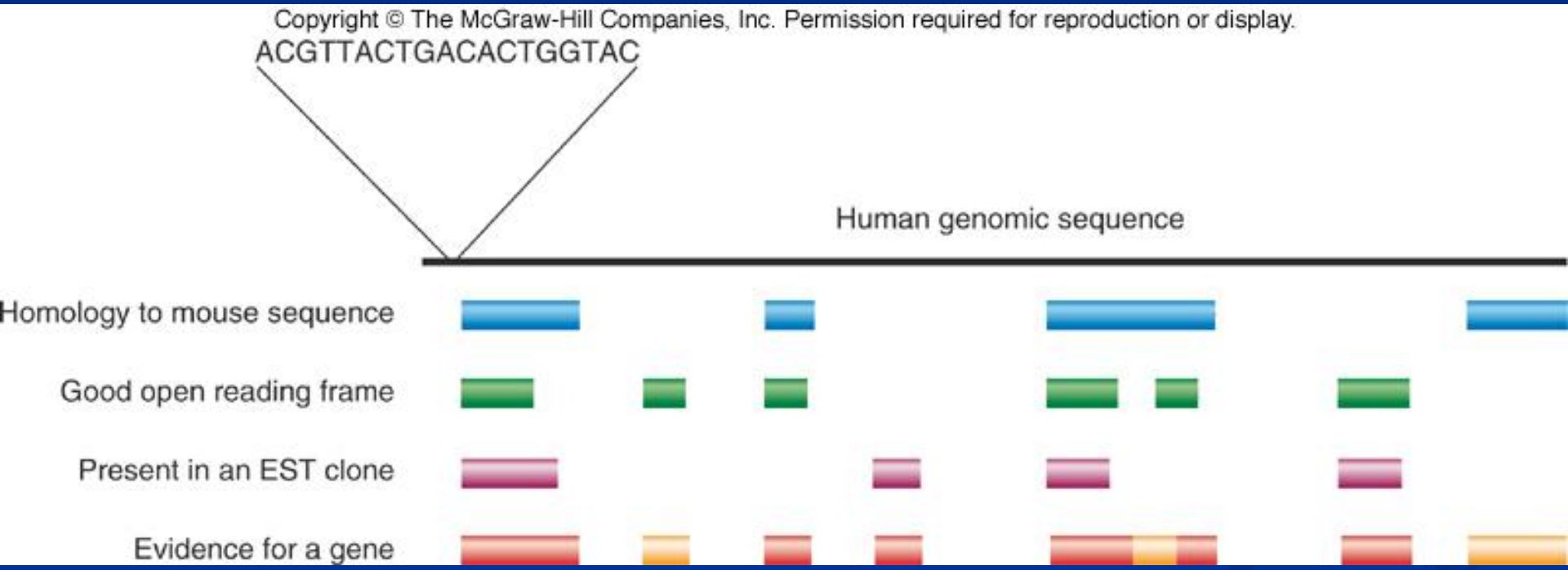
9 genes in 3 Mb
region

3 genes in 540 kb
region

Copyright © The McGraw-Hill Companies, Inc. Permission required for reproduction or display.

Fig. 9.16 p. 320

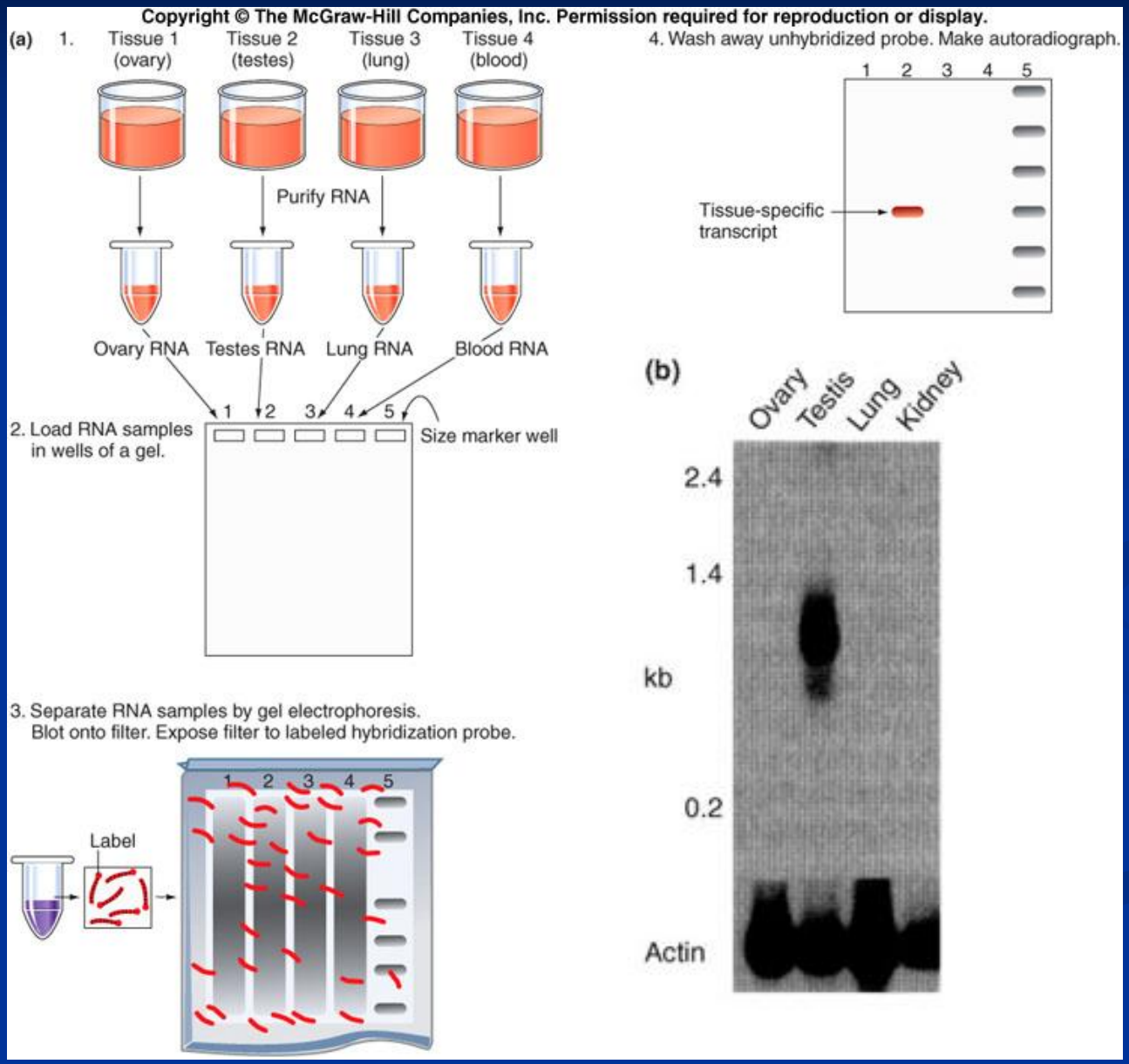
Computational analysis of genomic sequences to identify candidate genes



Gene expression patterns can pinpoint candidate genes

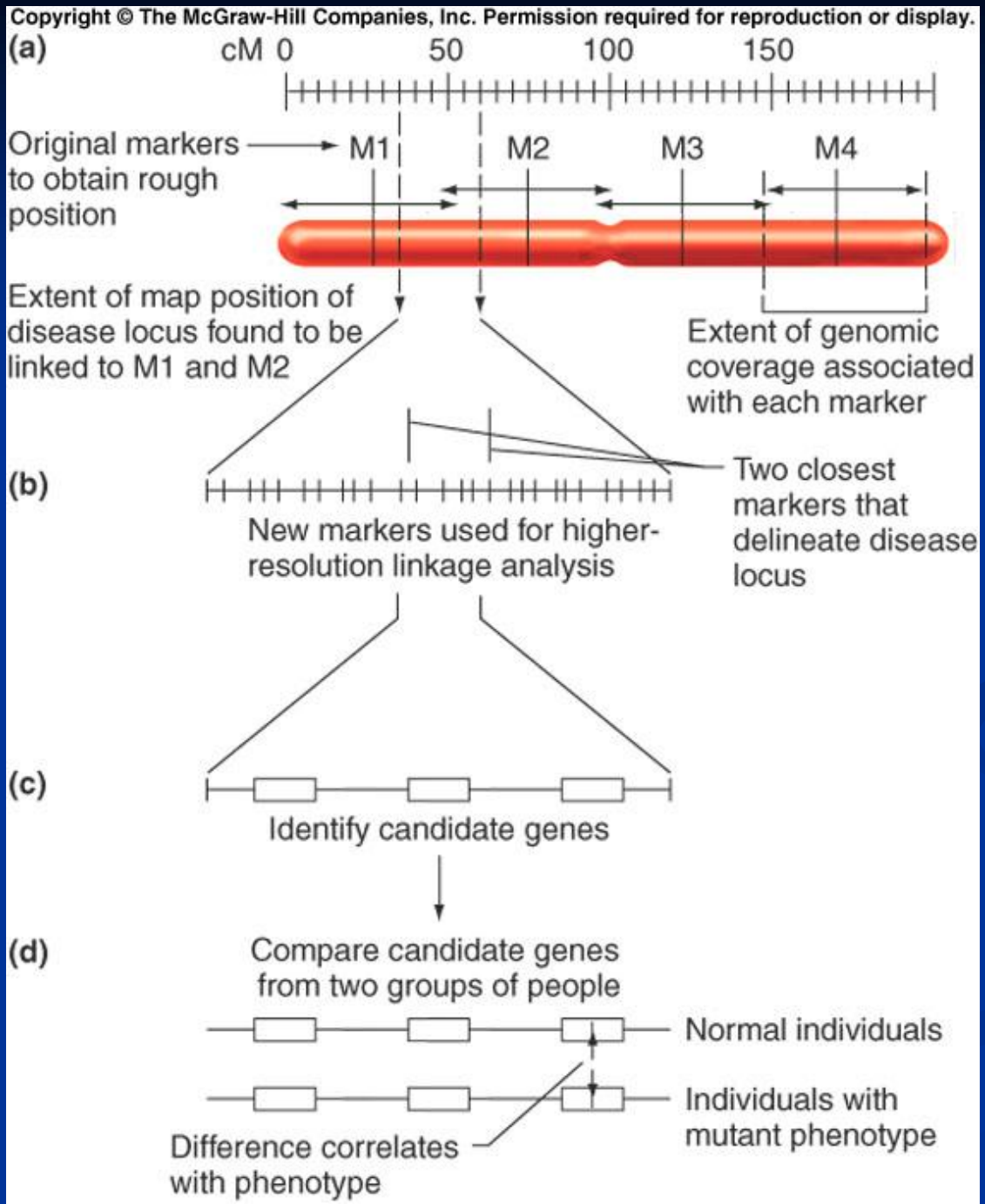
- Search EST (expressed sequence tag) databases for sequences representing specific tissues
- Northern blot
- RT-PCR

Northern blot example showing *SRY* candidate for testes determining factor is expressed in testes, but not lung, ovary, or kidney



Positional cloning – Step 3

- Find the gene responsible for the phenotype
  - Expression patterns in affected individuals
    - RNA expression assayed by Northern blot or RT-PCR with primers specific to candidate transcript
    - Look for misexpression (no expression, underexpression, overexpression)
  - Sequence differences
    - Missense mutations identified by sequencing coding region of candidate gene from normal and abnormal individuals
  - Transgenic modification of phenotype
    - Assess the mutant gene in a model organism

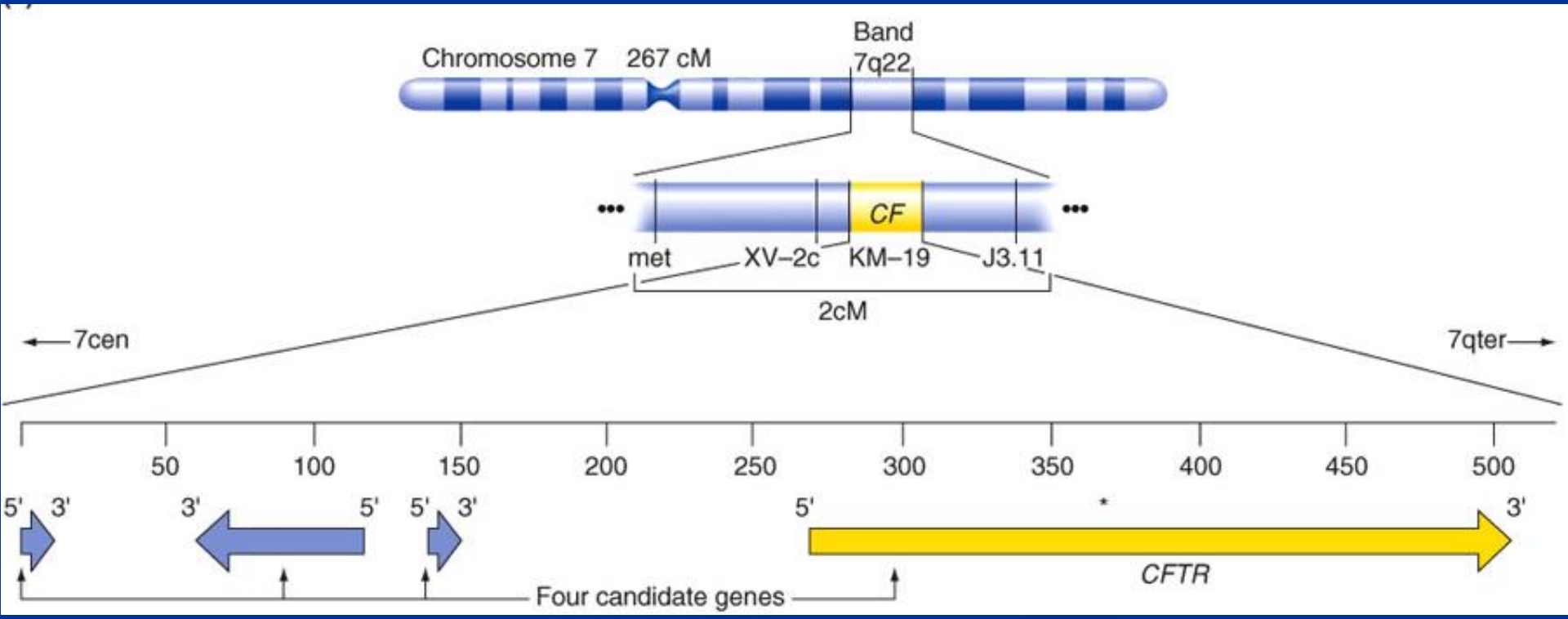


- **Identify the defective gene in individuals with the mutant phenotype**

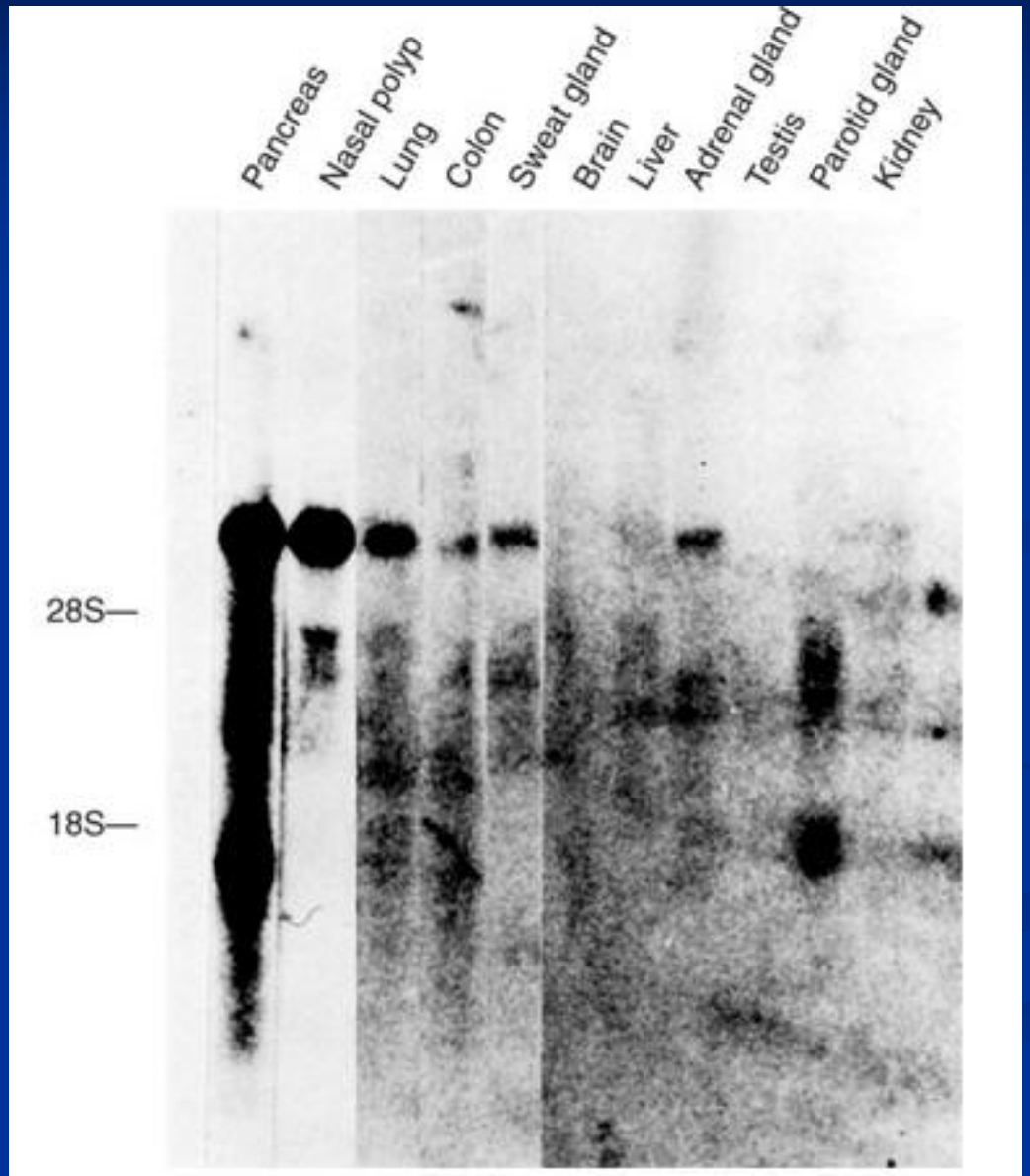
Fig. 10.18 p. 357

Example: Positional Cloning of Cystic Fibrosis Gene

- Linkage analysis places CF within a 2 cM region on chromosome 7



Northern blot analysis reveals only one of candidate genes is expressed in lungs and pancreas



Every CF patient is homozygous for a mutation in the CFTR gene

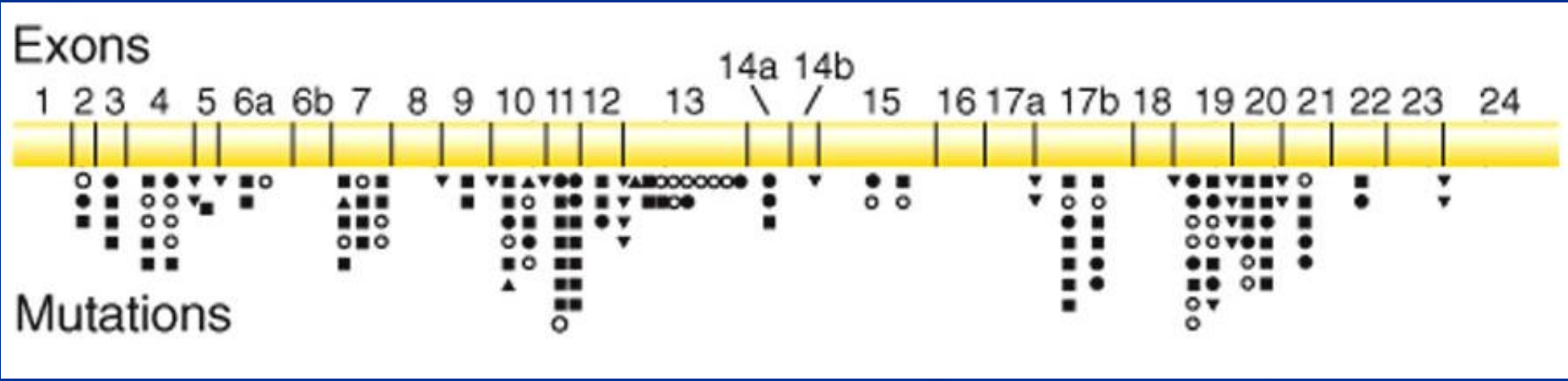
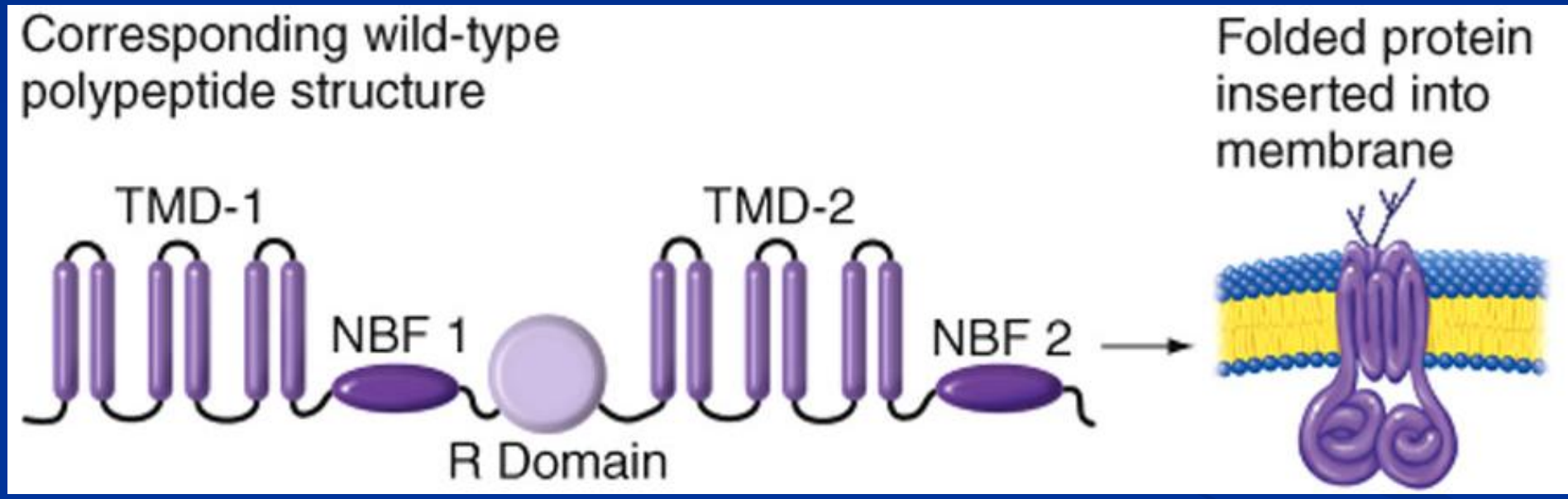


Fig. 10.23 p. 362

CFTR is a membrane protein



Proving CFTR is the right gene

- Phenotype results from loss-of-function
  - Cannot use transgenic “add-in” technology
- Perform CFTR gene “knockout” in mice to examine phenotype without CFTR gene
  - Targeted mutagenesis

Genetic dissection of complex traits

Copyright © The McGraw-Hill Companies, Inc. Permission required for reproduction or display.

TABLE 11.2 Complexities That Alter Traditional Mendelian Ratios

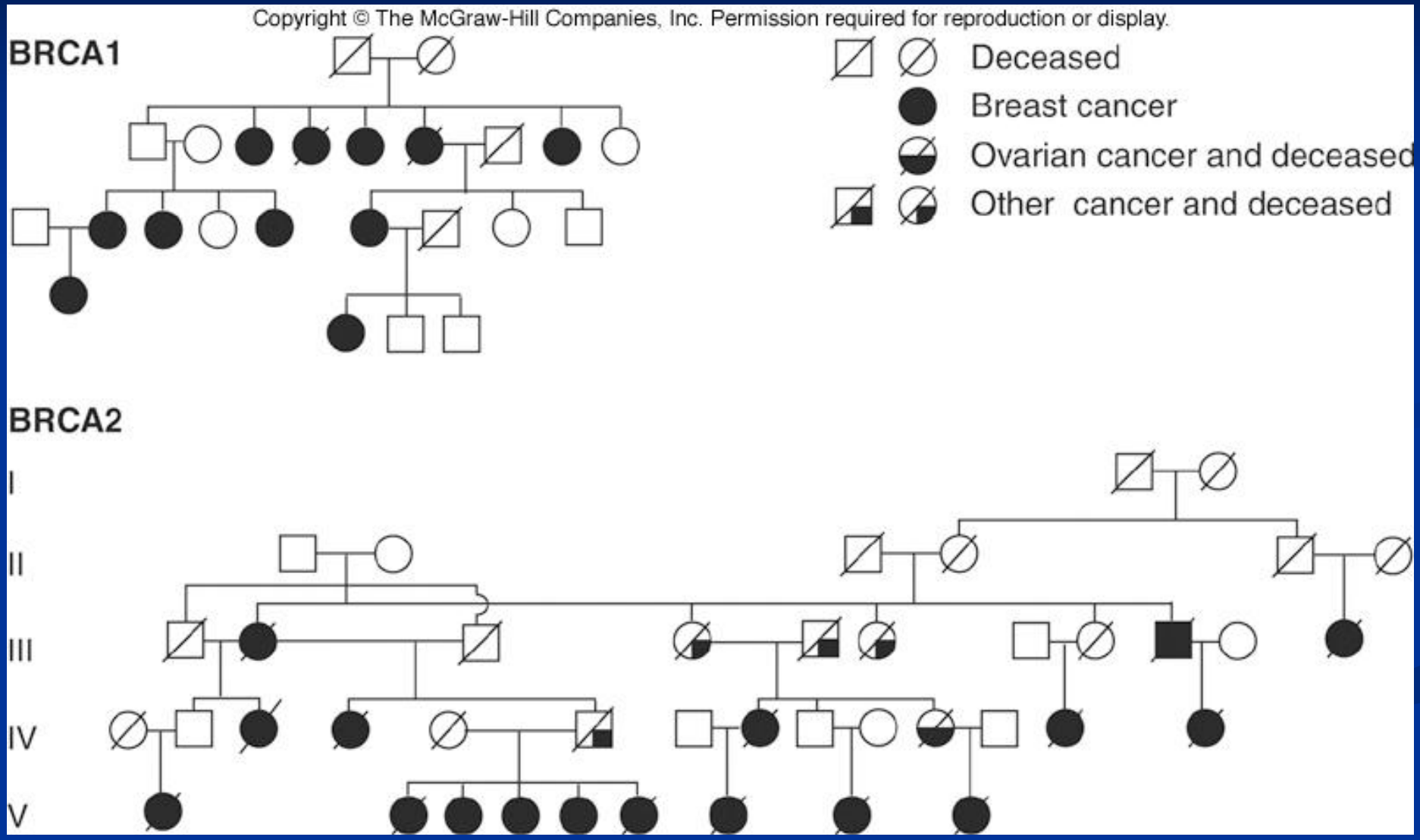
Category of Complexity	Problem in Observed Relationship Between Genotype and Phenotype	Changes Necessitated in Mapping Strategy
Incomplete penetrance	Disease genotype can occur in an individual who does not express the disease phenotype	Eliminate nondiseased individuals from analysis
Phenocopy	Disease phenotype can be expressed by an individual who does not have the disease genotype	Limit studies to families that show evidence for inheritance of the trait
Genetic heterogeneity	In different families, different disease genotypes are responsible for the same disease phenotype	Divide complete set of disease-transmitting families into subgroups (based on various parameters such as average age of onset) and perform linkage analysis separately on each subgroup
Polygenic determination	Mutant alleles at more than one locus influence expression of the disease phenotype in a single individual	To search for complex patterns of association between the disease trait and multiple DNA markers

Incomplete penetrance – when a mutant genotype does not always cause a mutant phenotype

- Positional cloning identified BRCA1 as one gene causing breast cancer.
  - Only 66% of women who carry BRCA1 mutation develop breast cancer by age 55
- Incomplete penetrance hampers linkage mapping and positional cloning
  - Solution – exclude all nondisease individuals from analysis
  - Requires many more families for study

- Genetic heterogeneity
  - Mutations at more that one locus cause the same disease phenotype
    - Multiple families used in most studies
    - If different families have different gene mutations, power of statistics to detect linkage drops significantly
      - BRCA1 and BRCA2 genes
- Phenocopy
  - The disease phenotype is not caused by any inherited predisposing mutation
    - Decreases ability to detect correlation between inheritance of disease locus and expression of the disease
      - Women without BRCA mutations develop breast cancer

Incomplete penetrance and genetic heterogeneity in breast cancer



- **Polygenic inheritance**
  - Two or more genes interact in the expression of phenotype
    - QTLs, or quantitative trait loci
      - Unlimited number of transmission patterns for QTLs
        - Discrete traits – penetrance may increase with number of mutant loci
        - Expressivity (or severity) may vary with number of loci
      - Many other factors complicate analysis
        - Some mutant genes may have large effect
        - Mutations at some loci may be recessive while others are dominant or codominant

Linkage disequilibrium (linkage in populations)

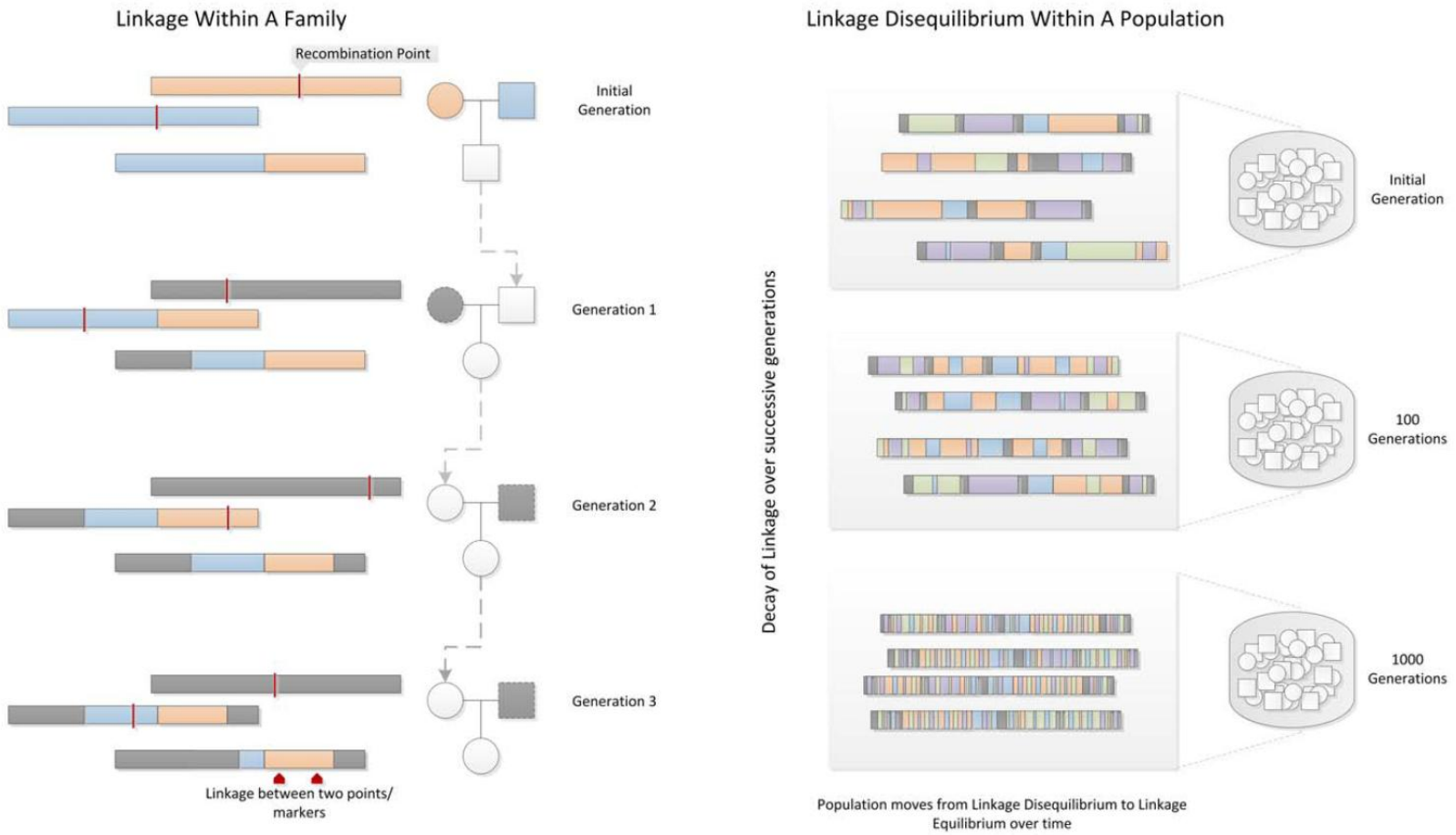


Figure 2. Linkage and Linkage Disequilibrium. Within a family, linkage occurs when two genetic markers (points on a chromosome) remain linked on a chromosome rather than being broken apart by recombination events during meiosis, shown as red lines. In a population, contiguous stretches of founder chromosomes from the initial generation are sequentially reduced in size by recombination events. Over time, a pair of markers or points on a chromosome in the population move from linkage disequilibrium to linkage equilibrium, as recombination events eventually occur between every possible point on the chromosome.
doi:10.1371/journal.pcbi.1002822.g002

Permission required for reproduction or display.

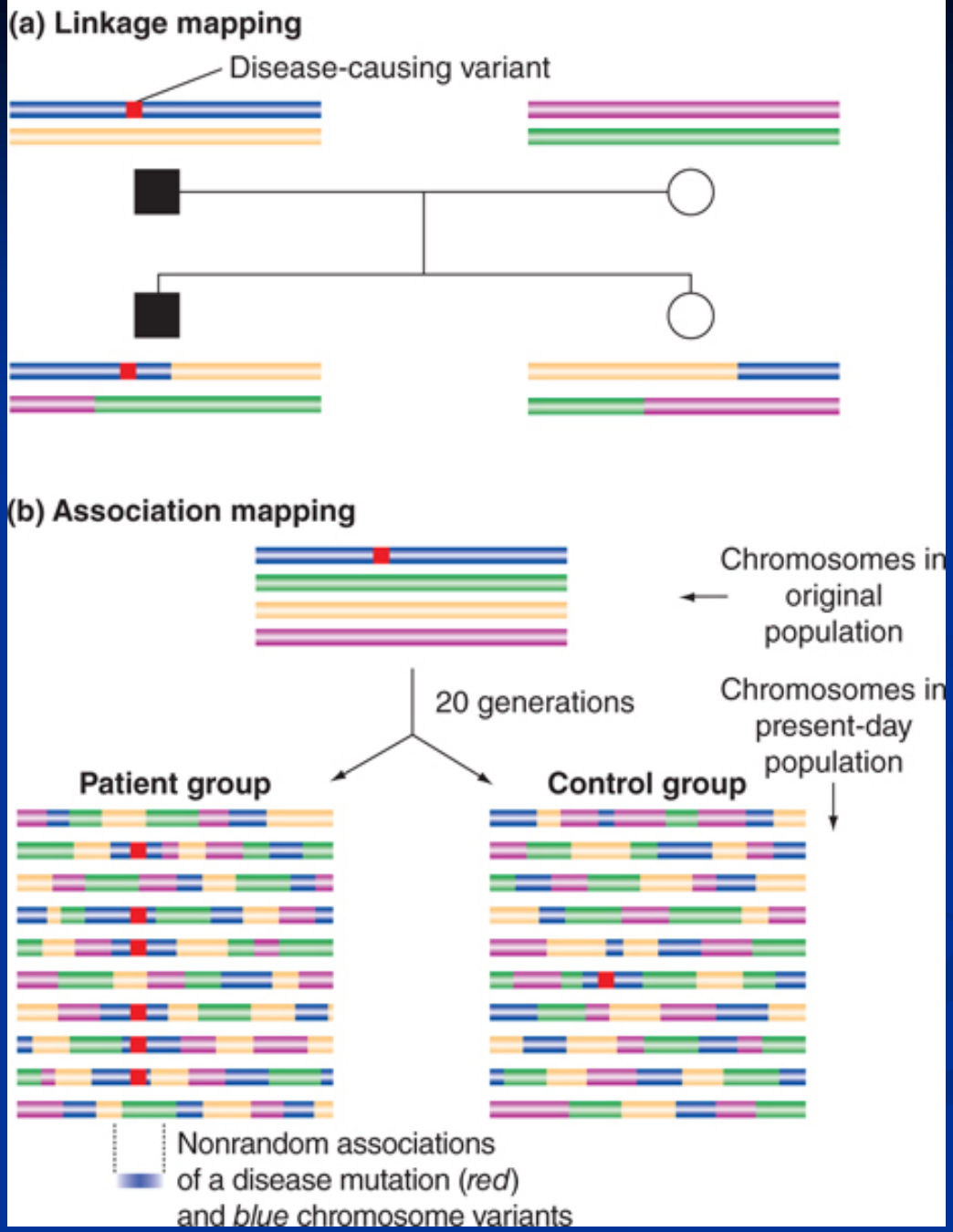
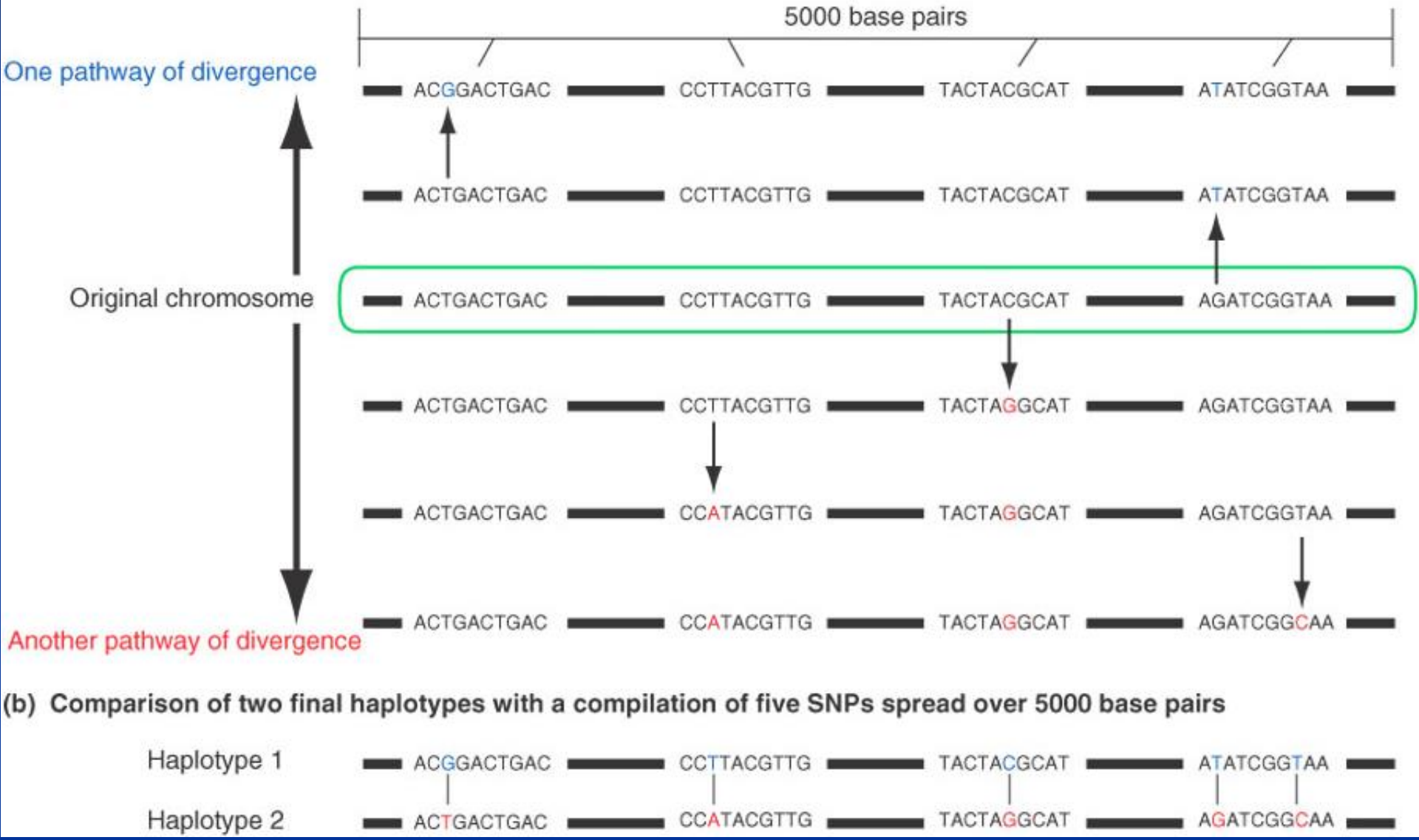


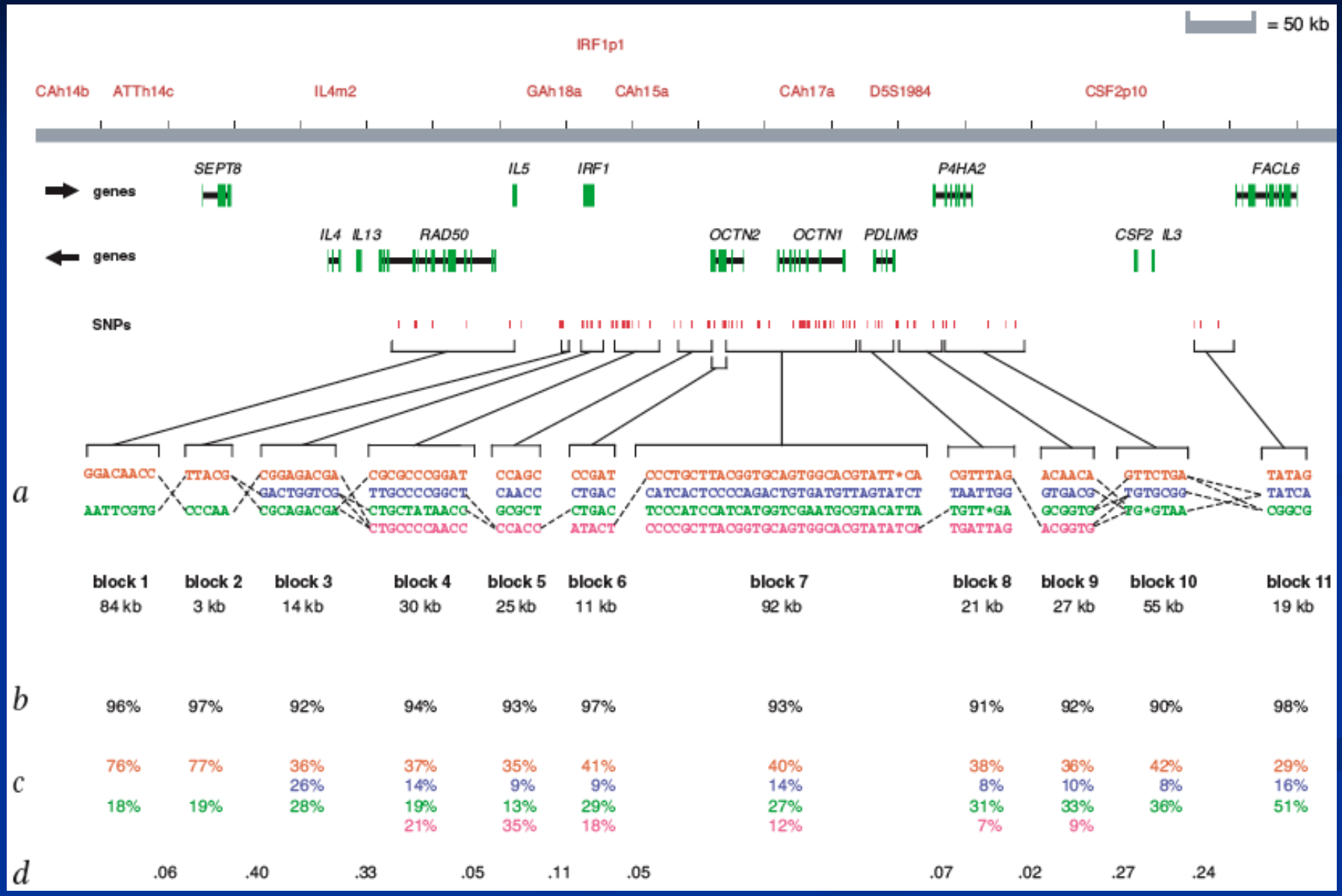
Fig. 21.13 p. 707

Haplotyping allows the world population to be seen as one giant pedigree

- Haplotypes
  - Unique combinations of common SNP alleles over extended regions of the genome
    - Inheritance of blocks of DNA that remain intact over many generations
    - Extended versions of alleles that cover regions containing multiple genes
- A small number of "tag SNPs" can be used to obtain a nearly complete whole-genome profile of individuals
  - DNA microarrays with 500,000 tag SNPS will cover the entire human genome

Formation of haplotypes over time



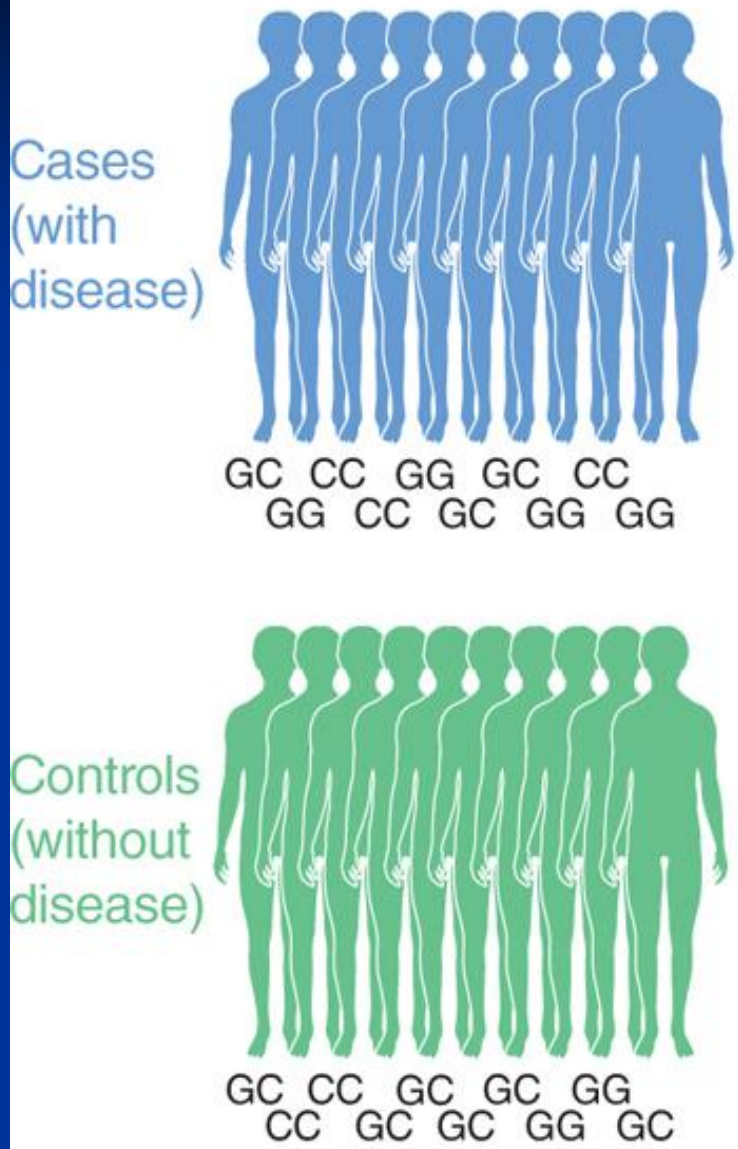

= 50 kb
IRF1p1
CAh14b
ATTh14c
IL4m2
GAh18a
CAh15a
CAh17a
D5S1984
CSF2p10
genes
SEPT8
IL5
IRF1
P4HA2
FACL6
genes
IL4
IL13
RAD50
OCTN2
OCTN1
PDLIM3
CSF2
IL3
SNPs
a
GGACAACC
AATTCGTG
TTACG
CCCAA
CGGAGACGA
GACTGGTCG
CGCAGACGA
CGCGCCCGGAT
TTGCCCCGGCT
CTGCTATAACG
CTGCCCCAACC
CCAGC
CAACC
GCGCT
CCACC
CCGAT
CTGAC
CTGAC
ATACT
CCCTGCTTACGGTGCAGTGGCACGTATT*CA
CATCACTCCCCAGACTGTGATGTTAGTATCT
TCCCATCCATCATGGTCGAATGCGTACATTA
CCCCGCTTACGGTGCAGTGGCACGTATATCA
CGTTTAG
TAATTGG
TGTT*GA
TGATTAG
ACAACA
GTGACG
GCGGTG
ACGGTG
GTTCTGA
TGTGCGG
TG*GTAA
TATAG
TATCA
CGGCG
block 1
84 kb
block 2
3 kb
block 3
14 kb
block 4
30 kb
block 5
25 kb
block 6
11 kb
block 7
92 kb
block 8
21 kb
block 9
27 kb
block 10
55 kb
block 11
19 kb
b
96%
97%
92%
94%
93%
97%
93%
91%
92%
90%
98%
c
76%
77%
36%
37%
35%
41%
40%
38%
36%
42%
29%
26%
14%
9%
9%
14%
8%
10%
8%
16%
18%
19%
28%
19%
13%
29%
27%
31%
33%
36%
51%
21%
35%
18%
12%
7%
9%
d
.06
.40
.33
.05
.11
.05
.07
.02
.27
.24

Genome-wide association studies (GWAS)

- Does not depend on traditional pedigree analysis but are computationally intensive
- Thousands of individuals make up a study group
- Each individual is observed or tested for expression of one or more traits of interest
- DNA microarrays are used to obtain whole-genome profiles for each member of the study population
- Genotypes at each tag SNP are tested for association with each trait
- Only a small number of tag SNPs will show a significant association with a trait

http://www.youtube.com/watch?v=9cZkIZ1YtMs&feature=relmfu
https://www.youtube.com/watch?v=b9GIBWqpinE

Copyright © McGraw-Hill Education. Permission required for reproduction or display.



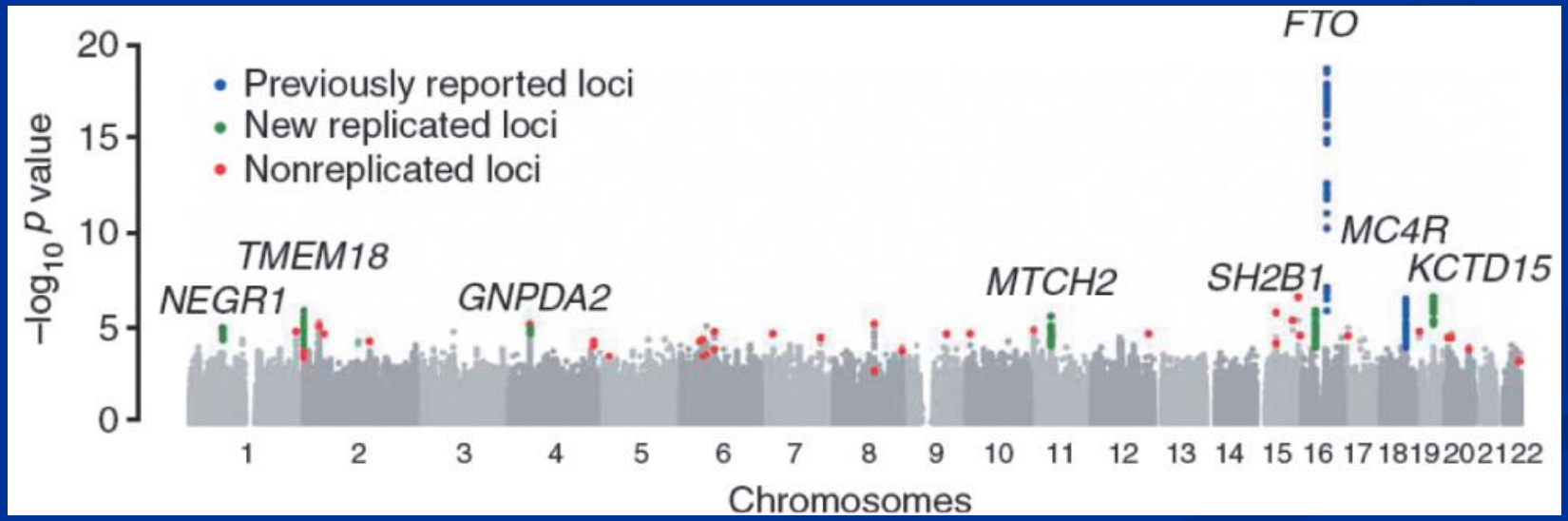
SNP1	*SNP2*	*SNP ...*
Cases	**Cases**	*Repeat for all SNPs*
Count of G: 2104 of 4000	Count of G: 1648 of 4000	
Frequency of G: 52.6%	Frequency of G: 41.2%	
Controls	**Controls**	
Count of G: 2676 of 6000	Count of G: 2532 of 6000	
Frequency of G: 44.6%	Frequency of G: 44.2%	
P-value: $5.0 \cdot 10^{-15}$	**P-value:** 0.33	

Fig. 21.16 p. 708

GWAS of body mass index (BMI)

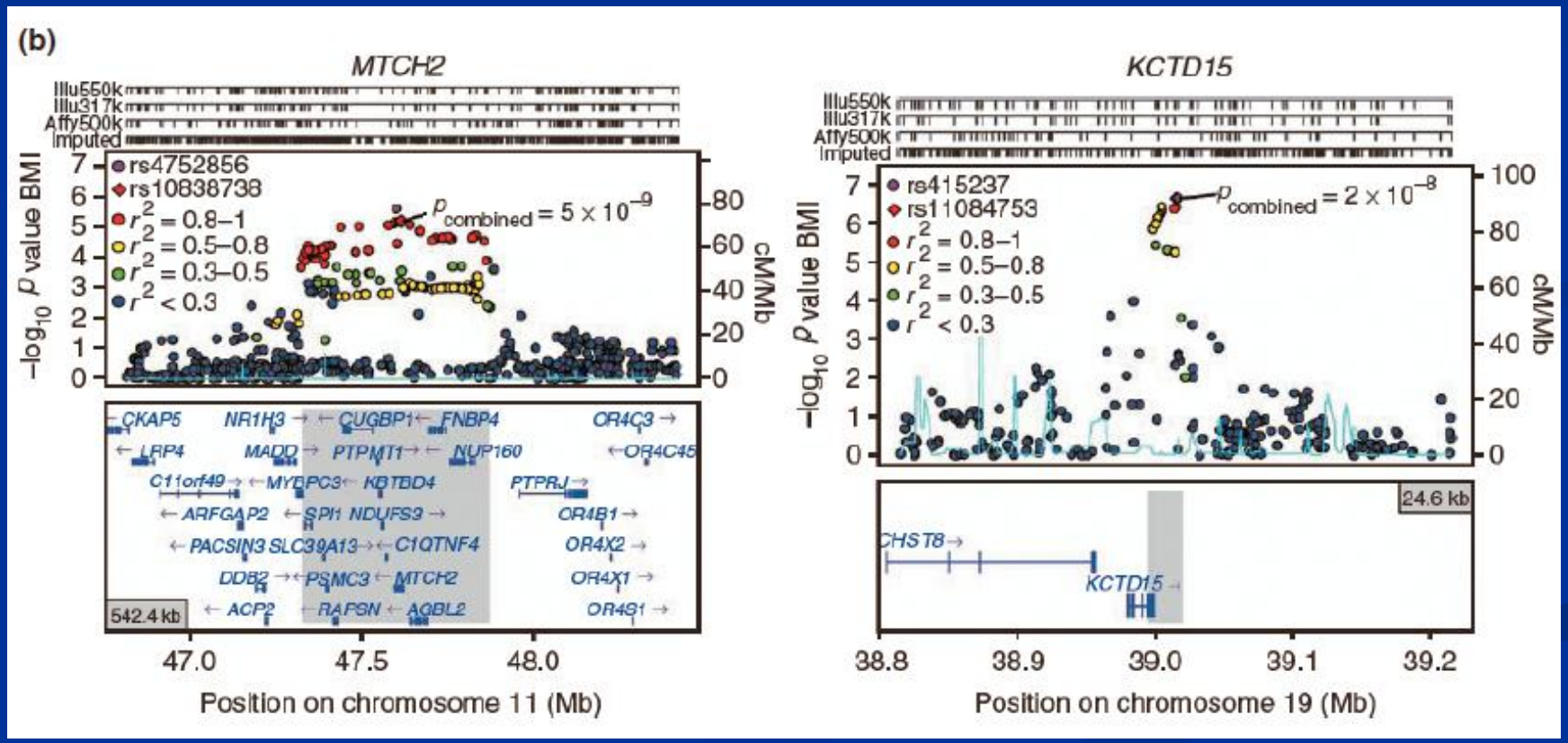
P values for all SNPs tested for association with high BMI across all chromosomes

- Each dot represents a single SNP test
- Lowest P values are shown as the highest dots



GWAS of BMI (cont)

Fine-scale mapping of two regions associate with high BMI



The open source human genome

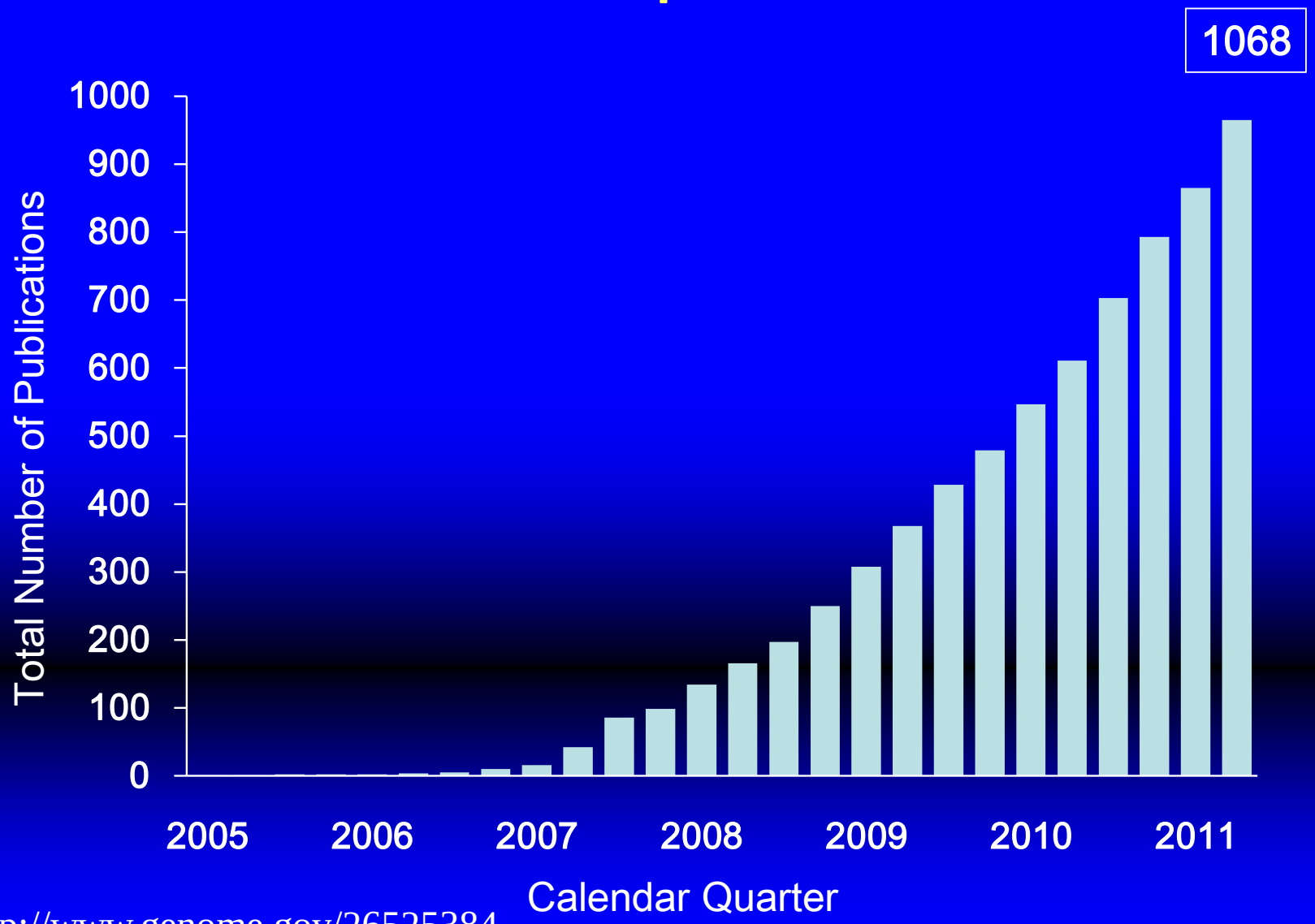
All newly-obtained SNP information is deposited public databases maintained by the NIH

29 interlinked databases at NCBI

Examples:

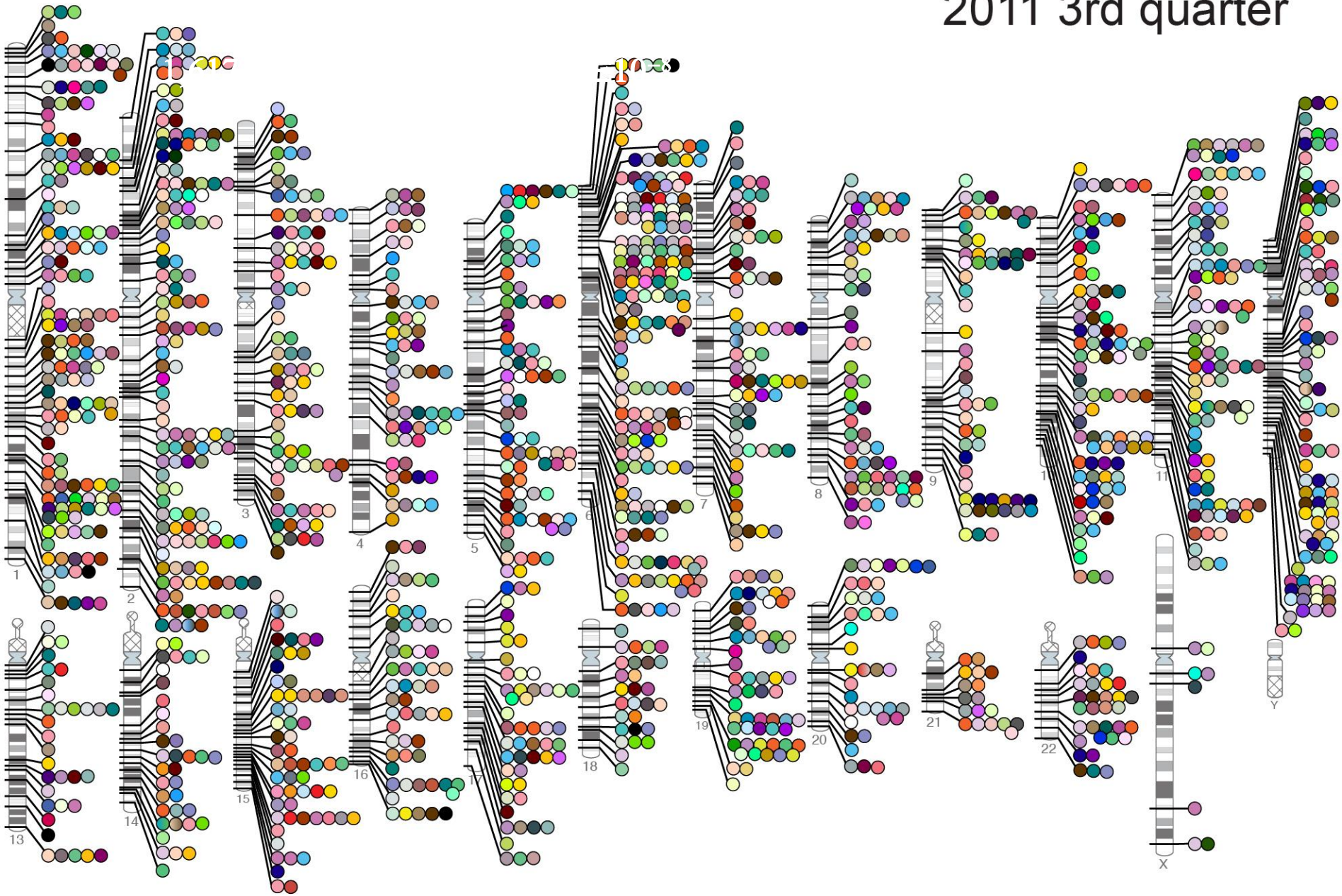
- dbSNP – comprehensive catalog of all human SNPs
- dbGaP – database of studies involving GWAS and other studies investigating the association of genotype and phenotype
- OMIM – online compendium of annotated records of each heritable trait and gene in humans
- Numerous software tools to query and retrieve genetic data

Published GWA Reports, 2005 – 9/2011



http://www.genome.gov/26525384

2011 3rd quarter



http://www.genome.gov/26525384

- Abdominal aortic aneurysm
- Acute lymphoblastic leukemia
- Adhesion molecules
- Adiponectin levels
- Age-related macular degeneration
- AIDS progression
- Alcohol dependence
- Alopecia areata
- Alzheimer disease
- Amyloid A levels
- Amyotrophic lateral sclerosis
- Angiotensin-converting enzyme activity
- Ankylosing spondylitis
- Arterial stiffness
- Asparagus anosmia
- Asthma
- Atherosclerosis in HIV
- Atrial fibrillation
- Attention deficit hyperactivity disorder
- Autism
- Basal cell cancer
- Behcet's disease
- Bipolar disorder
- Biliary atresia
- Bilirubin
- Bitter taste response
- Birth weight
- Bladder cancer
- Bleomycin sensitivity
- Blond or brown hair
- Blood pressure
- Blue or green eyes
- BMI, waist circumference
- Bone density
- Breast cancer
- Butyrylcholinesterase levels
- C-reactive protein
- Calcium levels
- Cardiac structure/function
- Cardiovascular risk factors
- Carnitine levels
- Carotenoid/tocopherol levels
- Carotid atherosclerosis
- Celiac disease
- Celiac disease and rheumatoid arthritis
- Cerebral atrophy measures
- Chronic lymphocytic leukemia
- Chronic myeloid leukemia
- Cleft lip/palate
- Coffee consumption
- Cognitive function
- Conduct disorder
- Colorectal cancer
- Corneal thickness
- Coronary disease
- Cortical thickness
- Creutzfeldt-Jakob disease
- Crohn's disease
- Crohn's disease and celiac disease
- Cutaneous nevi
- Cystic fibrosis severity
- Dermatitis
- DHEA-s levels
- Diabetic retinopathy
- Dilated cardiomyopathy
- Drug-induced liver injury
- Drug-induced liver injury (amoxicillin-clavulanate)
- Endometrial cancer
- Endometriosis
- Eosinophil count
- Eosinophilic esophagitis
- Epirubicin-induced leukopenia
- Erectile dysfunction and prostate cancer treatment
- Erythrocyte parameters
- Esophageal cancer
- Essential tremor
- Exfoliation glaucoma
- Eye color traits
- F cell distribution
- Fibrinogen levels
- Folate pathway vitamins
- Follicular lymphoma
- Fuch's corneal dystrophy
- Freckles and burning
- Gallstones
- Gastric cancer
- Glioma
- Glycemic traits
- Graves disease
- Hair color
- Hair morphology
- Handedness in dyslexia
- HDL cholesterol
- Heart failure
- Heart rate
- Height
- Hemostasis parameters
- Hepatic steatosis
- Hepatitis
- Hepatitis B vaccine response
- Hepatocellular carcinoma
- Hirschsprung's disease
- HIV-1 control
- Hodgkin's lymphoma
- Homocysteine levels
- HPV seropositivity
- Hypospadias
- Idiopathic pulmonary fibrosis
- IFN-related cytopeni
- IgA levels
- IgE levels
- Inflammatory bowel disease
- Insulin-like growth factors
- Intracranial aneurysm
- Iris color
- Iron status markers
- Ischemic stroke
- Juvenile idiopathic arthritis
- Keloid
- Kidney stones
- LDL cholesterol
- Leprosy
- Leptin receptor levels
- Liver enzymes
- Longevity
- LP (a) levels
- LpPLA(2) activity and mass
- Lung cancer
- Magnesium levels
- Major mood disorders
- Malaria
- Male pattern baldness
- Mammographic density
- Matrix metalloproteinase levels
- MCP-1
- Melanoma
- Menarche & menopause
- Meningioma
- Meningococcal disease
- Metabolic syndrome
- Migraine
- Moyamoya disease
- Multiple sclerosis
- Myeloproliferative neoplasms
- Myopia (pathological)
- N-glycan levels
- Narcolepsy
- Nasopharyngeal cancer
- Natriuretic peptide levels
- Neuroblastoma
- Nicotine dependence
- Obesity
- Open angle glaucoma
- Open personality
- Optic disc parameters
- Osteoarthritis
- Osteoporosis
- Otosclerosis
- Other metabolic traits
- Ovarian cancer
- Pancreatic cancer
- Pain
- Paget's disease
- Panic disorder
- Parkinson's disease
- Periodontitis
- Peripheral arterial disease
- Personality dimensions
- Phosphatidylcholine levels
- Phosphorus levels
- Photic sneeze
- Phytosterol levels
- Platelet count
- Polycystic ovary syndrome
- Primary biliary cirrhosis
- Primary sclerosing cholangitis
- PR interval
- Progranulin levels
- Progressive supranuclear palsy
- Prostate cancer
- Protein levels
- PSA levels
- Psoriasis
- Psoriatic arthritis
- Pulmonary funct. COPD
- QRS interval
- QT interval
- Quantitative traits
- Recombination rate
- Red vs.non-red hair
- Refractive error
- Renal cell carcinoma
- Renal function
- Response to antidepressants
- Response to antipsychotic therapy
- Response to carbamazepine
- Response to clopidogrel therapy
- Response to hepatitis C treat
- Response to interferon beta therapy
- Response to metaformin
- Response to statin therapy
- Restless legs syndrome
- Retinal vascular caliber
- Retinol levels
- Rheumatoid arthritis
- Ribavirin-induced anemia
- Schizophrenia
- Serum metabolites
- Skin pigmentation
- Smoking behavior
- Speech perception
- Sphingolipid levels
- Statin-induced myopathy
- Stevens-Johnson syndrome
- Stroke
- Sudden cardiac arrest
- Suicide attempts
- Systemic lupus erythematosus
- Systemic sclerosis
- T-tau levels
- Tau AB1-42 levels
- Telomere length
- Testicular germ cell tumor
- Thyroid cancer
- Thyroid volume
- Tooth development
- Total cholesterol
- Triglycerides
- Tuberculosis
- Type 1 diabetes
- Type 2 diabetes
- Ulcerative colitis
- Urate
- Urinary albumin excretion
- Urinary metabolites
- Uterine fibroids
- Venous thromboembolism
- Ventricular conduction
- VEGF levels
- Vertical cup-disc ratio
- Vitamin B12 levels
- Vitamin D insuffiency
- Vitamin E levels
- Vitiligo
- Warfarin dose
- Weight
- White cell count
- White matter hyperintensity
- YKL-40 levels

Two powerful transgenic techniques

- Untargeted transgenics
  - Foreign DNA injected into pronucleus of fertilized egg
  - Place injected one-cell embryo back into oviduct
  - 25-50% of time DNA integrates randomly into chromosome

- Targeted transgenics
  - Embryonic stem cells
  - Integrate engineered gene into germ line
  - Homologous recombination may occur
  - Knockout constructs- nonfunctional gene exchanged for normal gene by homologous recombination

Untargeted transgenics in mice

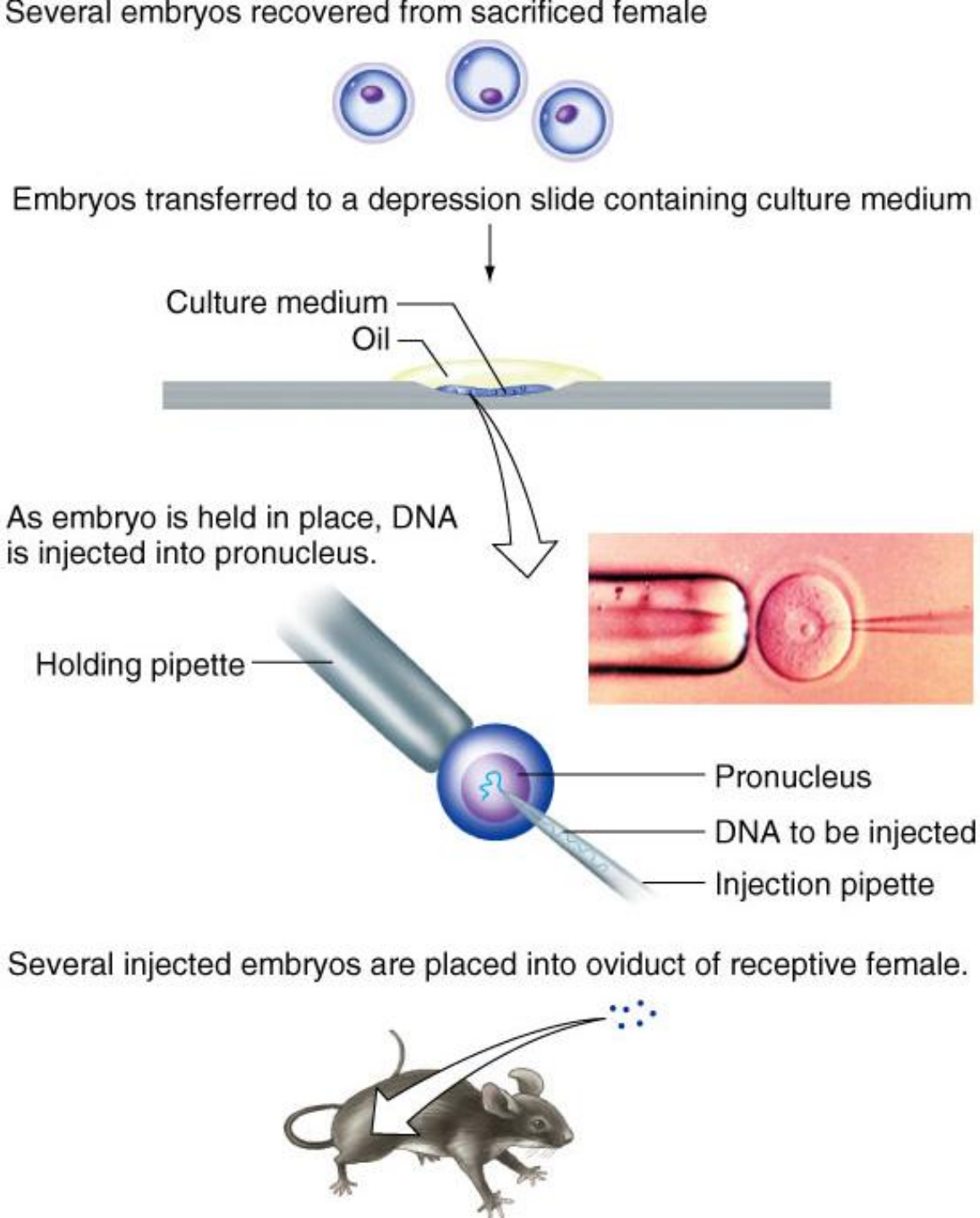
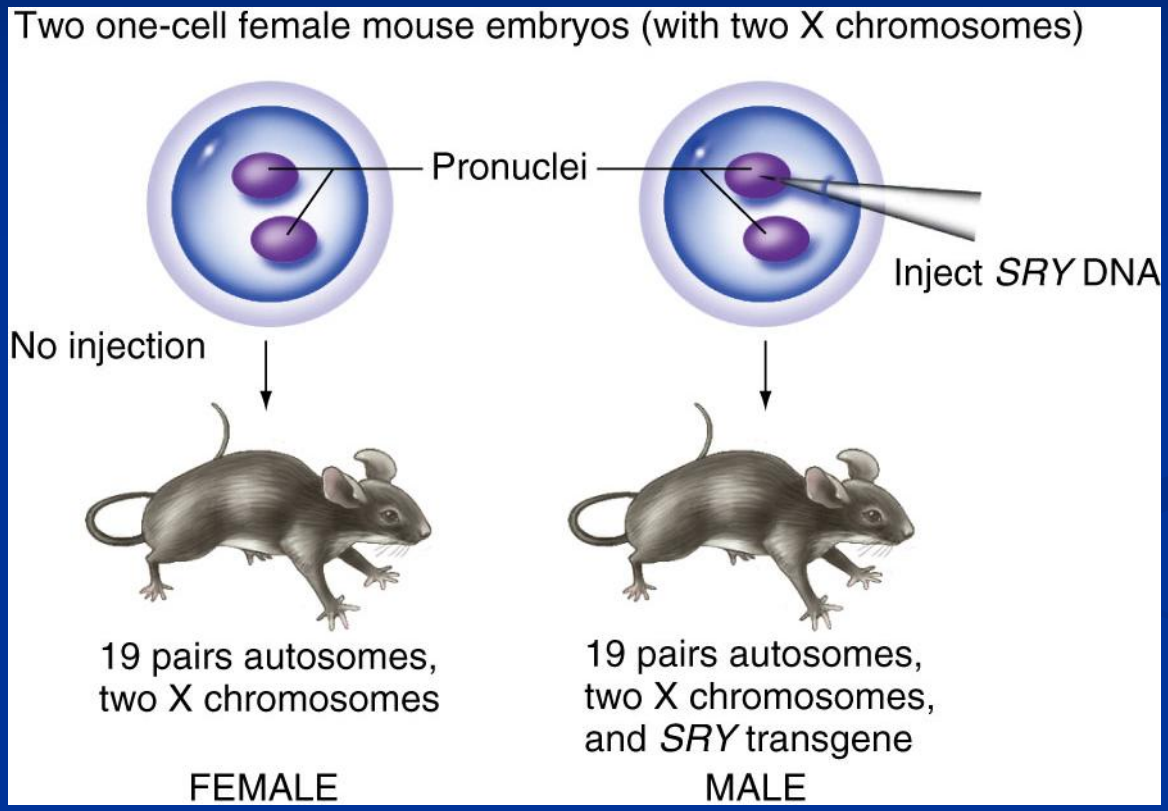


Fig. E.7 p. 116

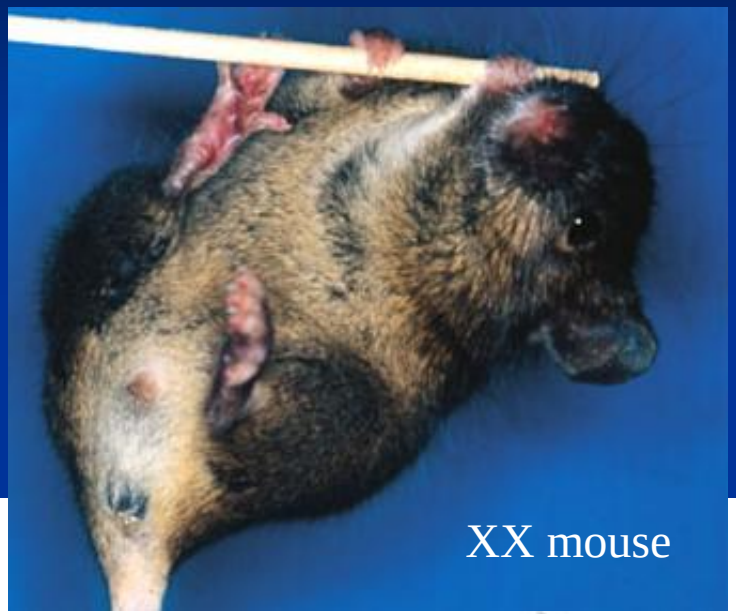
Using transgenic tools to determine gene function

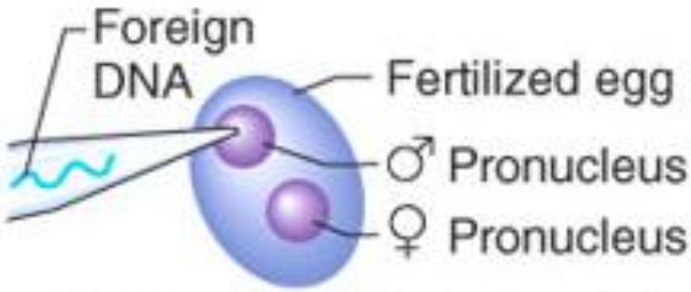


- **The *SRY* locus is responsible for the expression of maleness**

Fig. E.9 p. 117

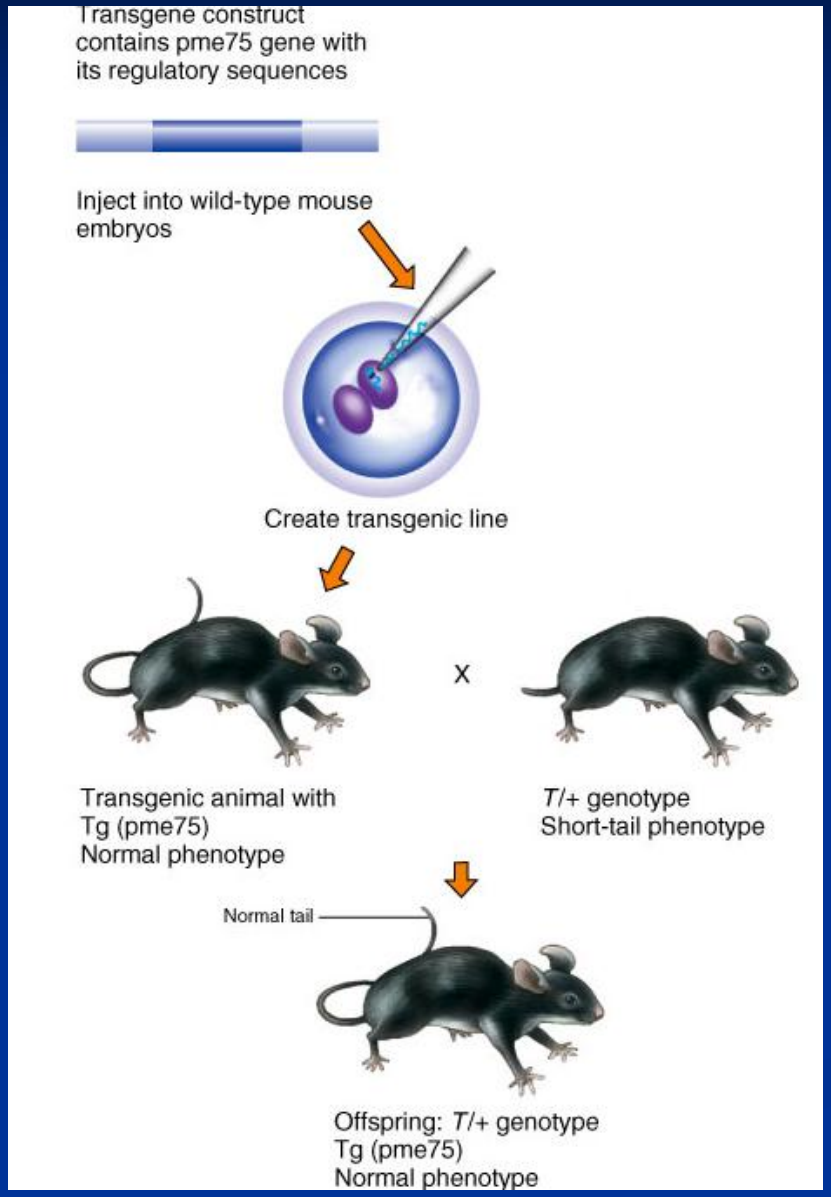
Transgenic analysis to test candidate genes





Foreign DNA injected into male pronucleus of newly fertilized egg. → Injected eggs surgically implanted into uterus of "foster" mother and allowed to develop. → Mice are born with foreign DNA in every cell nucleus.

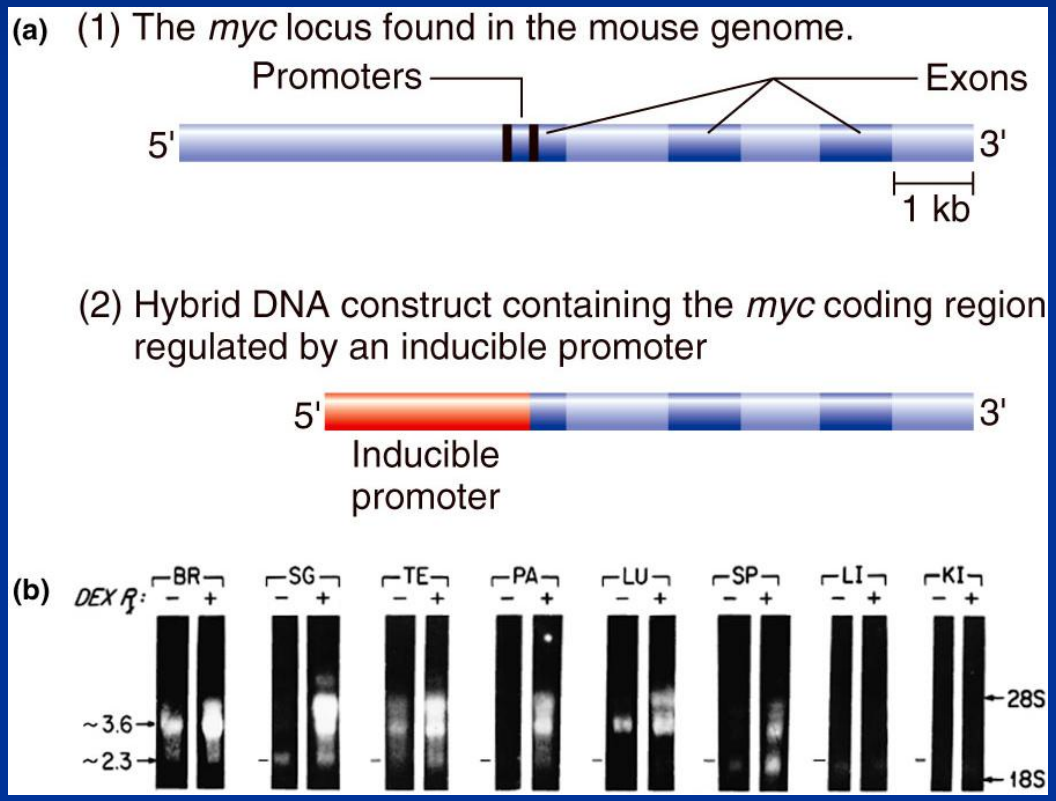
Transgenic analysis to test candidate genes



- **Dominant deletion mutation at T locus causes short tail**
- **Transgenic mouse with wild-type *pme75* transgene mated to mutant**
- **Normal phenotype demonstrates transgene corrects the short tail phenotype**

Fig. E.13 p. 121

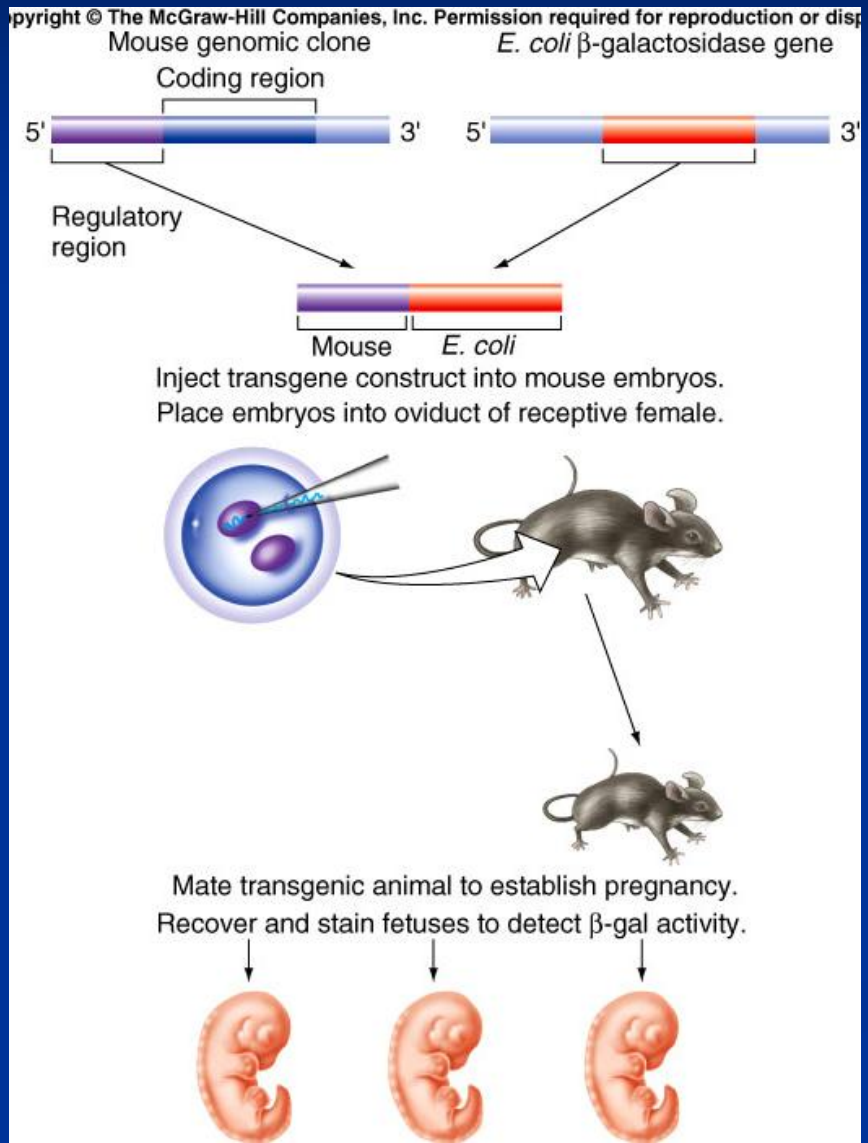
Using transgenic tools to determine gene function



- Transgenic overexpression of myc provides information on role of myc in tumor formation
- Structure of gene (a)
- Northern blot analysis (b)
- Rate of tumor formation increased in most tissues

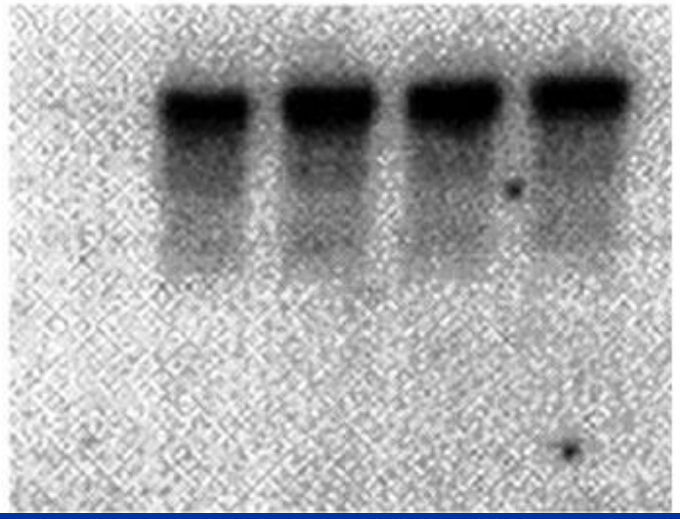
Fig. E.10 p. 118

Transgenics to characterize gene expression patterns – reporter genes



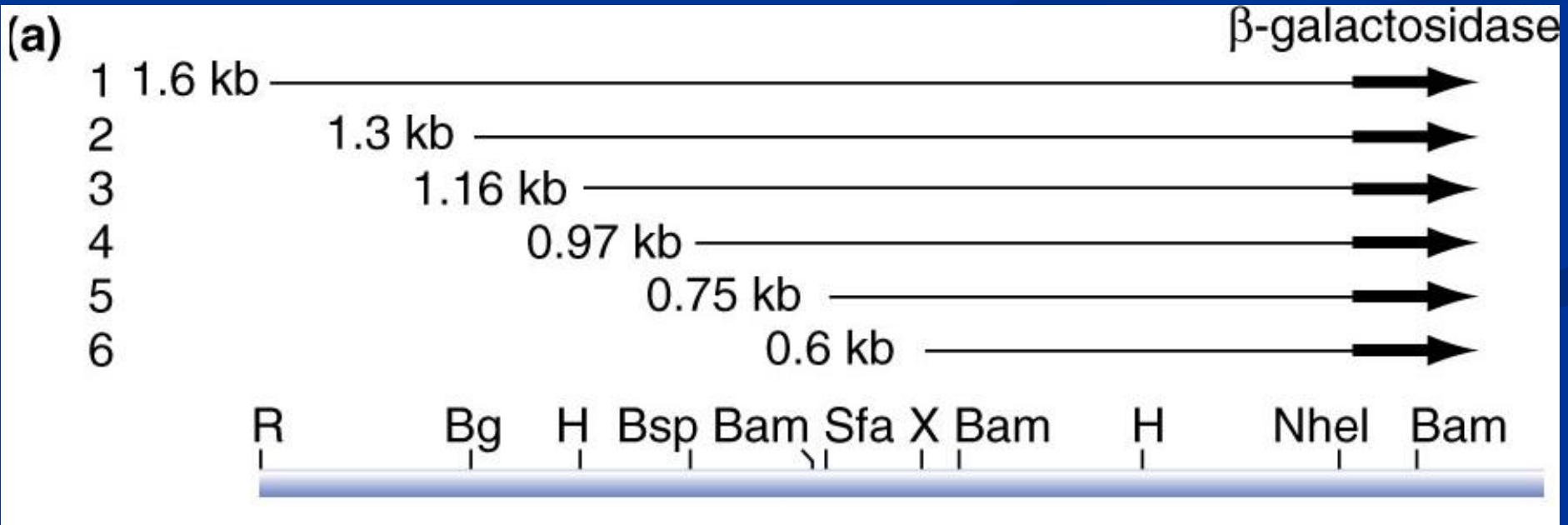
- **DNA construct containing mouse regulatory region of interest is attached to β-galactosidase gene from *E. coli***
- **Function assessed by β-gal expression in transgene fetus**

Fig. E..11 p. 119

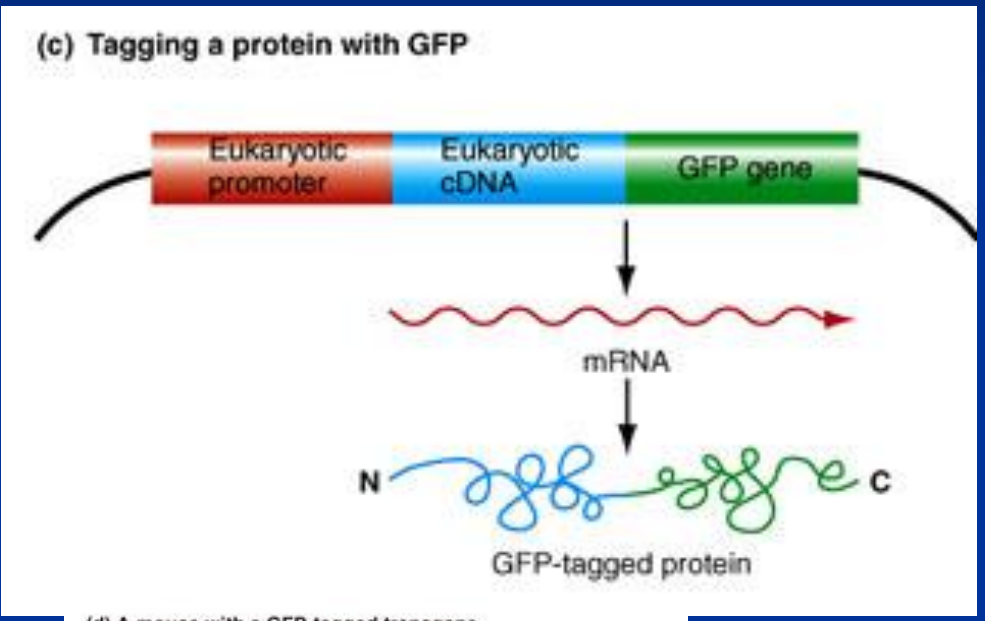


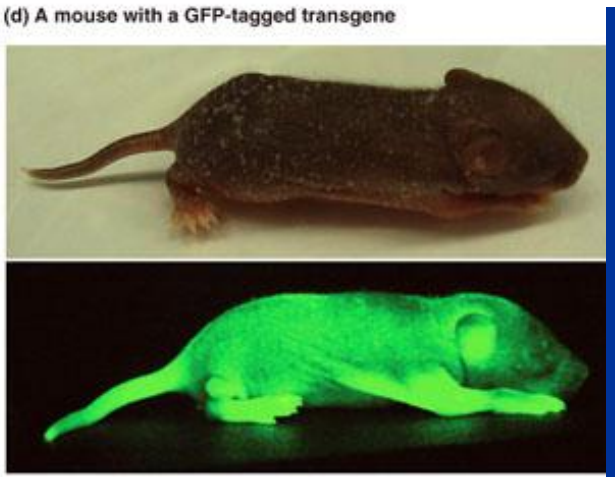
- **Use of transgene technology to map cis-acting regulatory region of *Tcp 10b*^{t} gene**
- **Important for spermatogenesis**

Fig. E.12 p. 120



Transgenics to characterize gene expression patterns – reporter genes

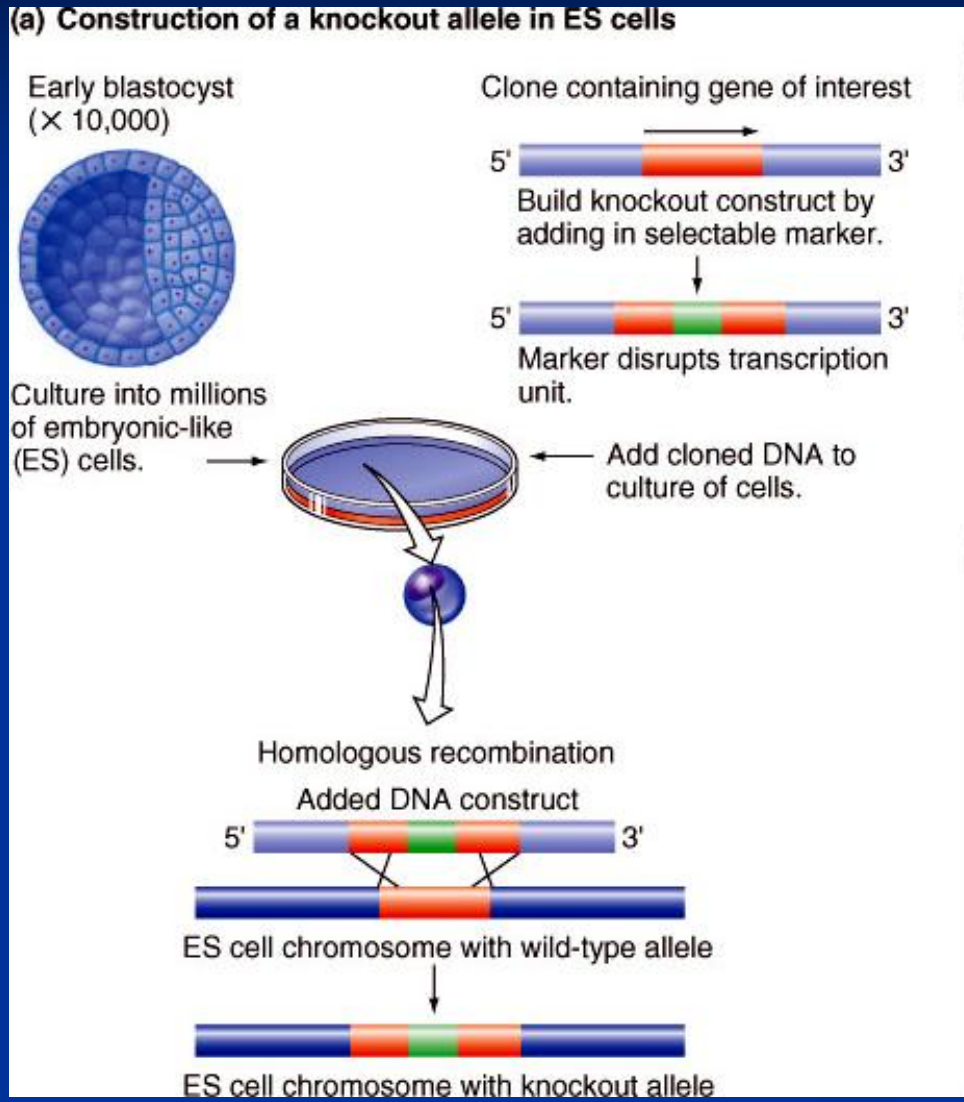


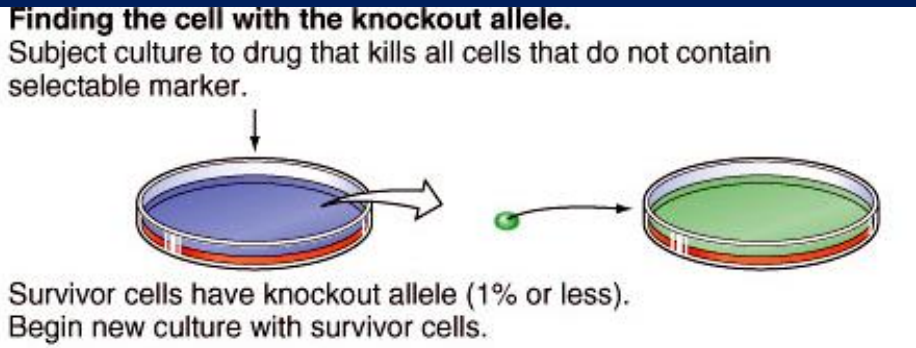


- Recombinant gene encoding a GFP fusion protein at C terminus
- Mouse with GFP-labeled transgene expressed throughout body

Fig. 18.13c,d p. 630

Targeted transgenics - Mouse knock-outs





(b)

Targeted mutagenesis to create a mouse model for human disease

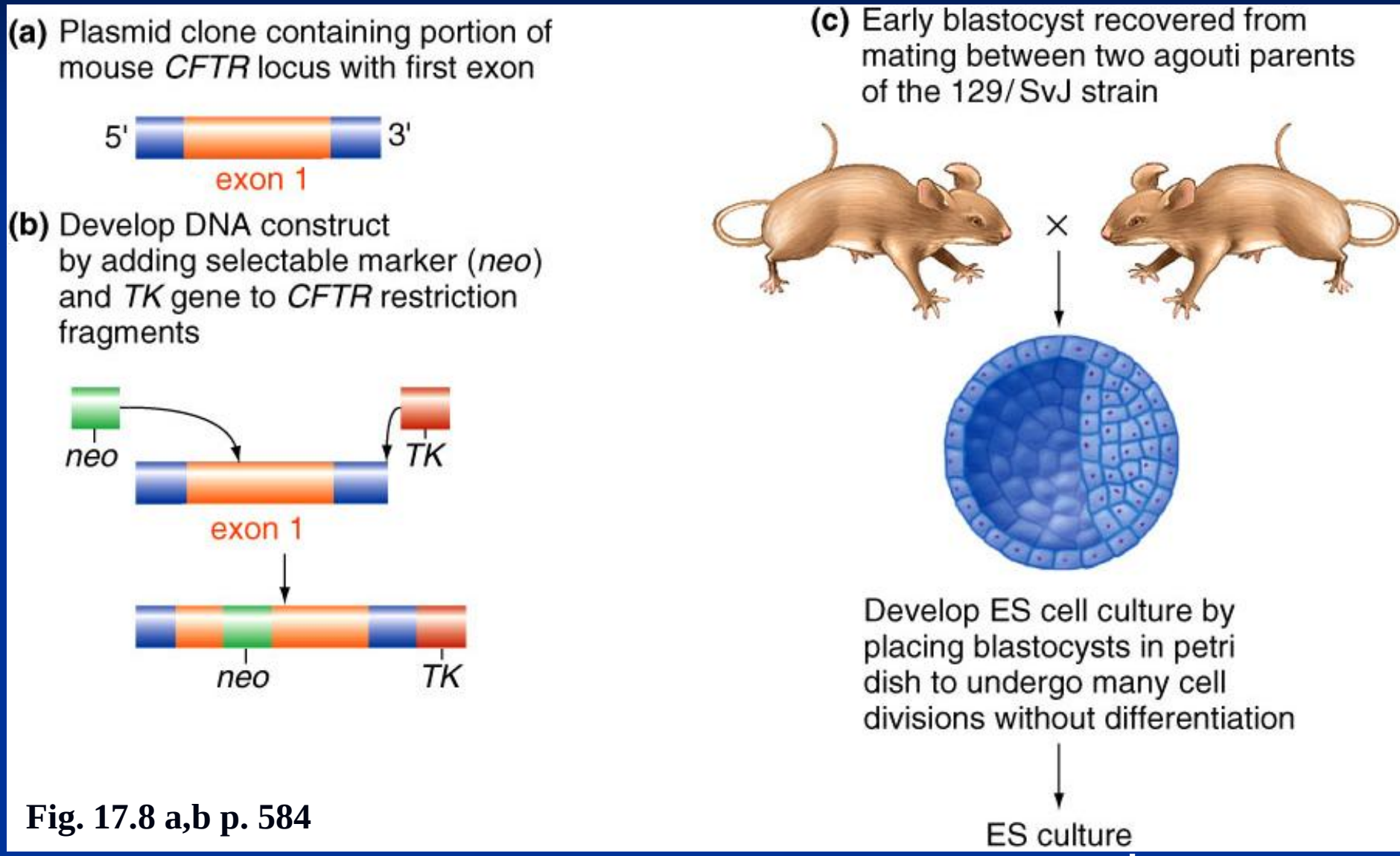
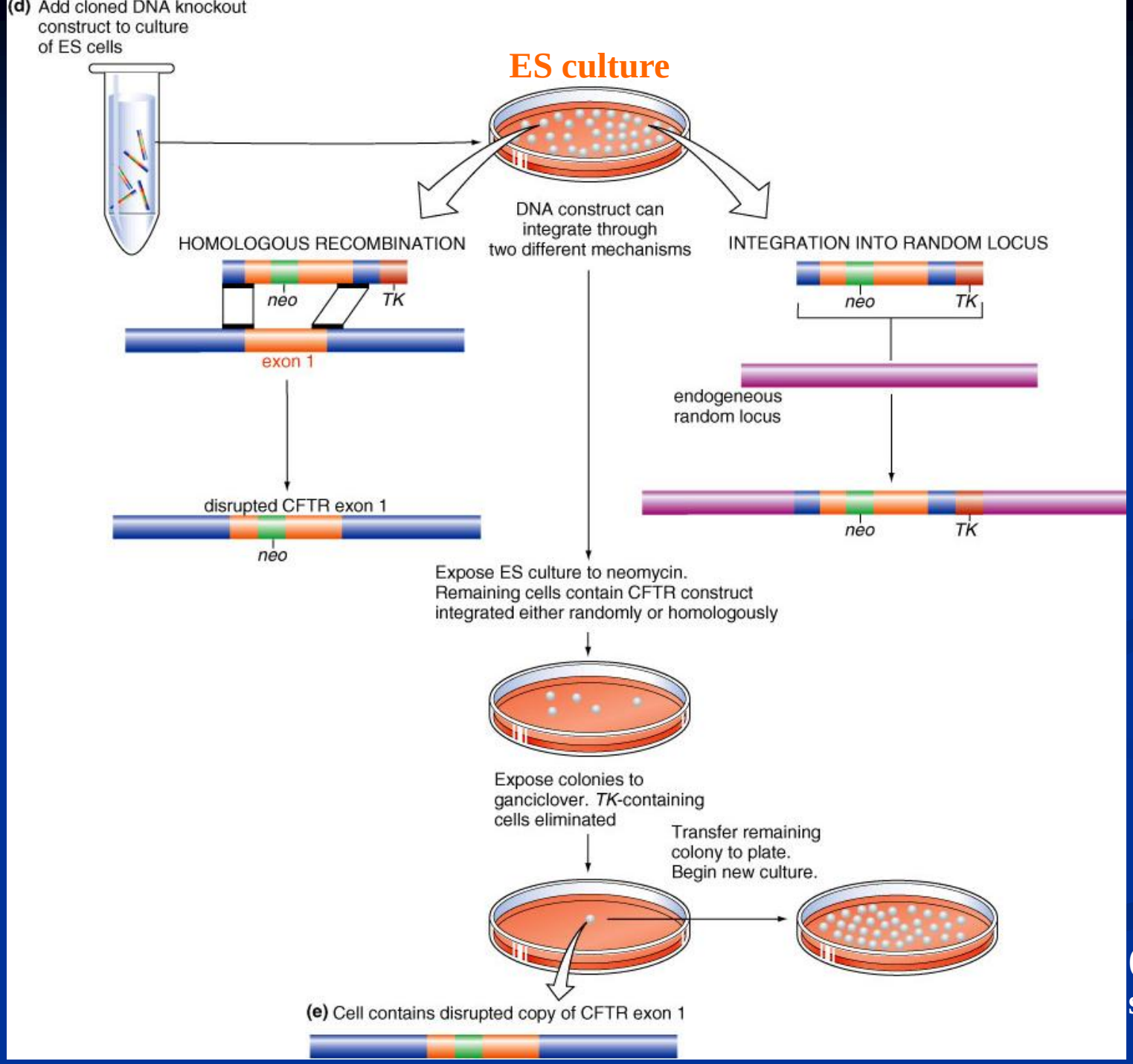


Fig. 17.8 a,b p. 584

(Cont'd next slide)



(cont'd next slide)

Copyright © The McGraw-Hill Companies, Inc. Permission required for reproduction or display.

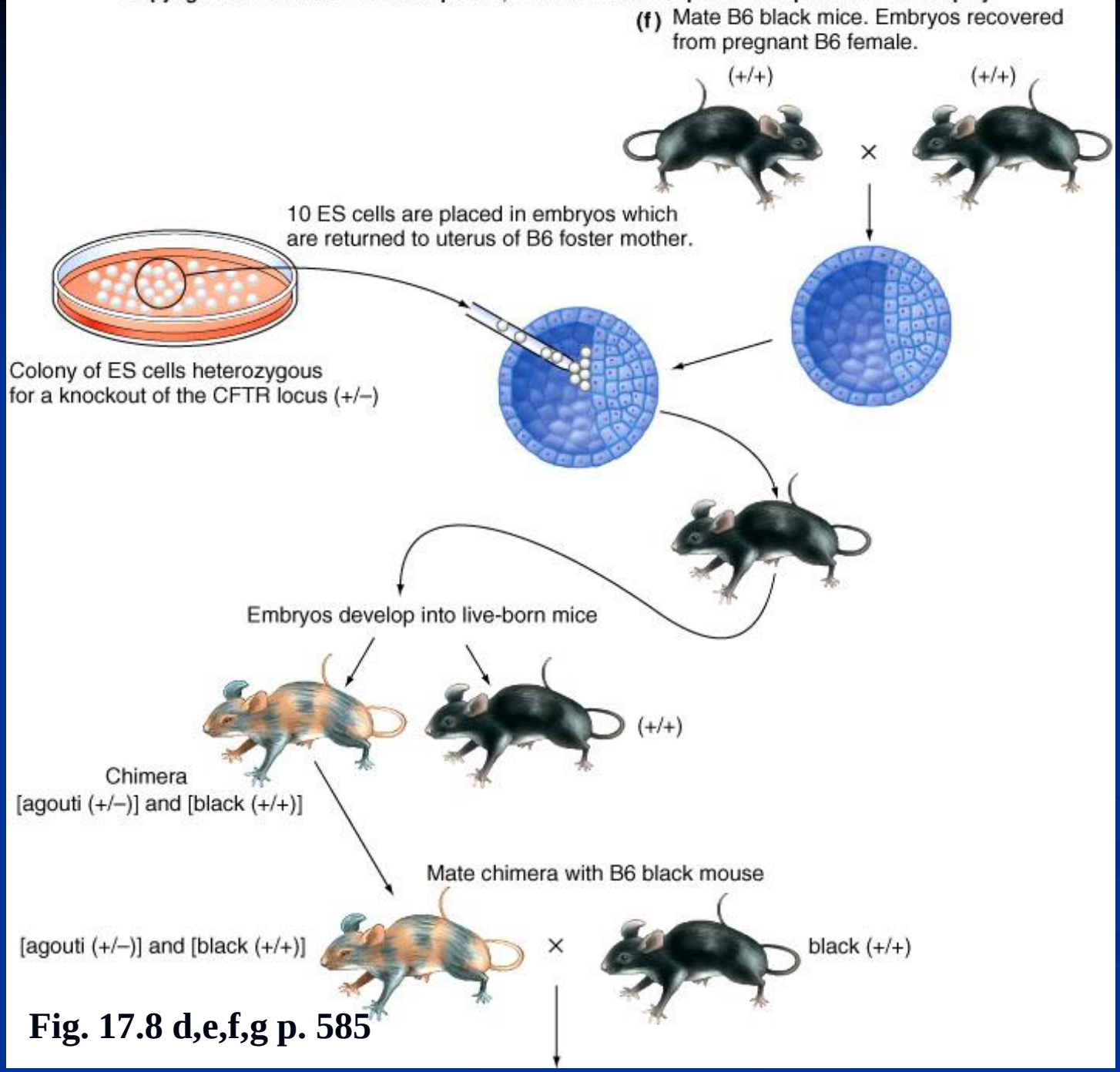


Fig. 17.8 d,e,f,g p. 585

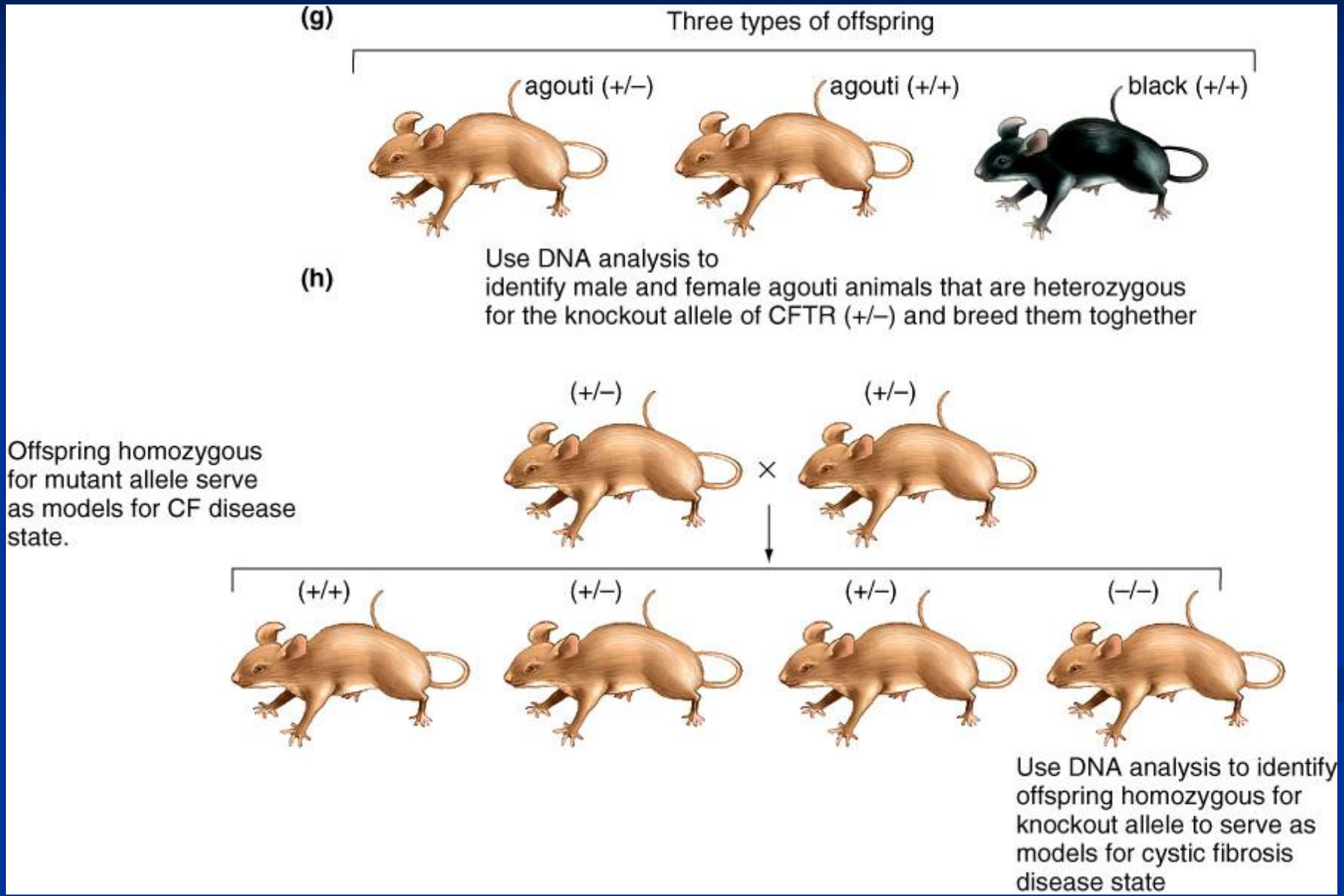
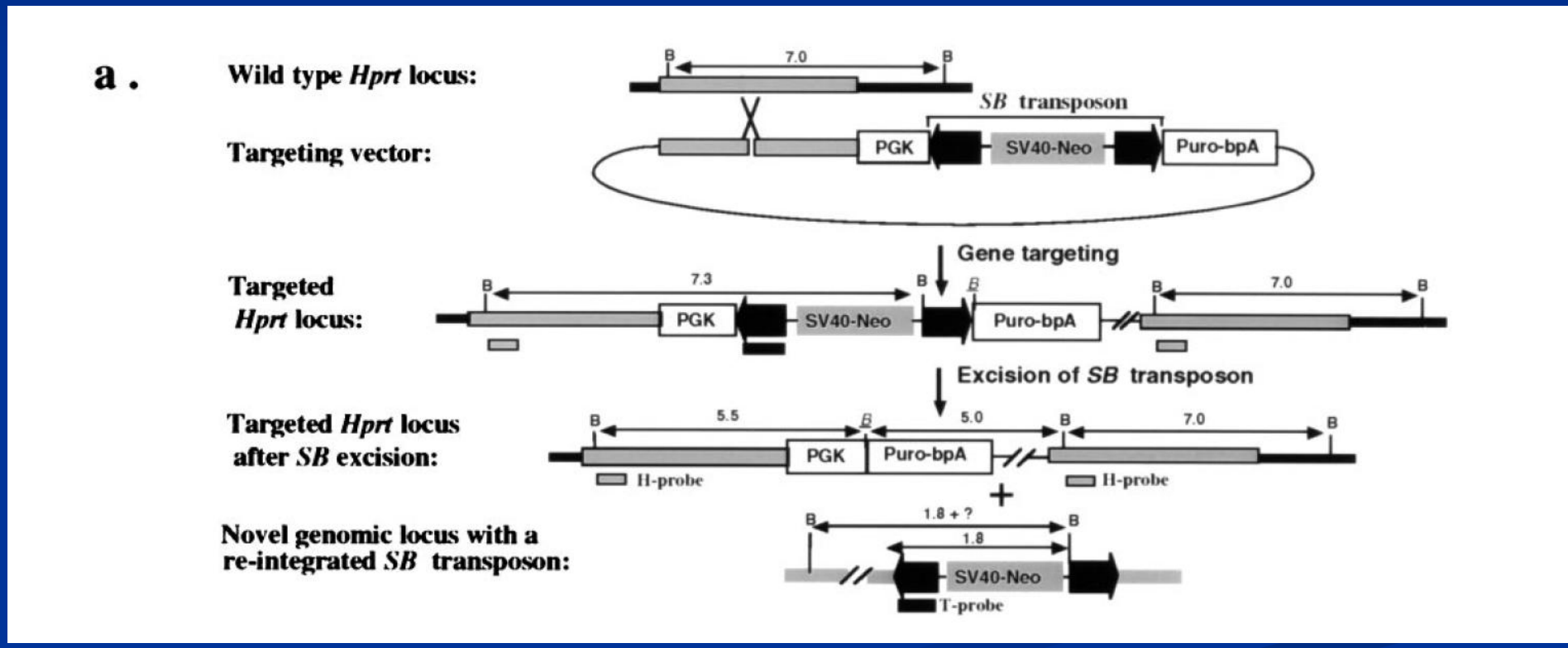

(g)
Three types of offspring
agouti (+/–)
agouti (+/+)
black (+/+)
(h)
Use DNA analysis to
identify male and female agouti animals that are heterozygous
for the knockout allele of CFTR (+/–) and breed them toghether
(+/–)
(+/–)
×
Offspring homozygous
for mutant allele serve
as models for CF disease
state.
(+/+)
(+/–)
(+/–)
(–/–)
Use DNA analysis to identify
offspring homozygous for
knockout allele to serve as
models for cystic fibrosis
disease state

Figure 1a from Sleeping Beauty paper



Cre-loxP system

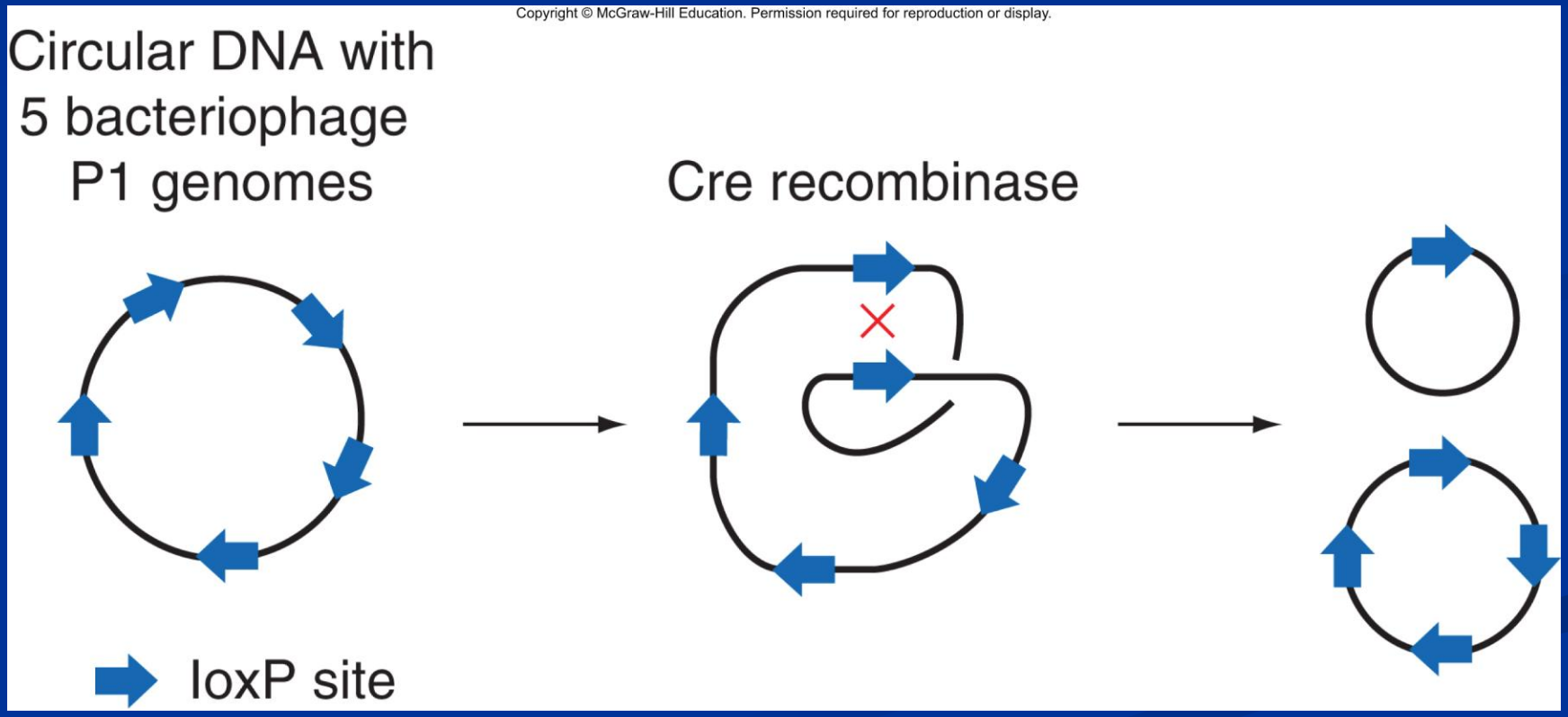


Fig. 17.9 p. 587

Cre-loxP system

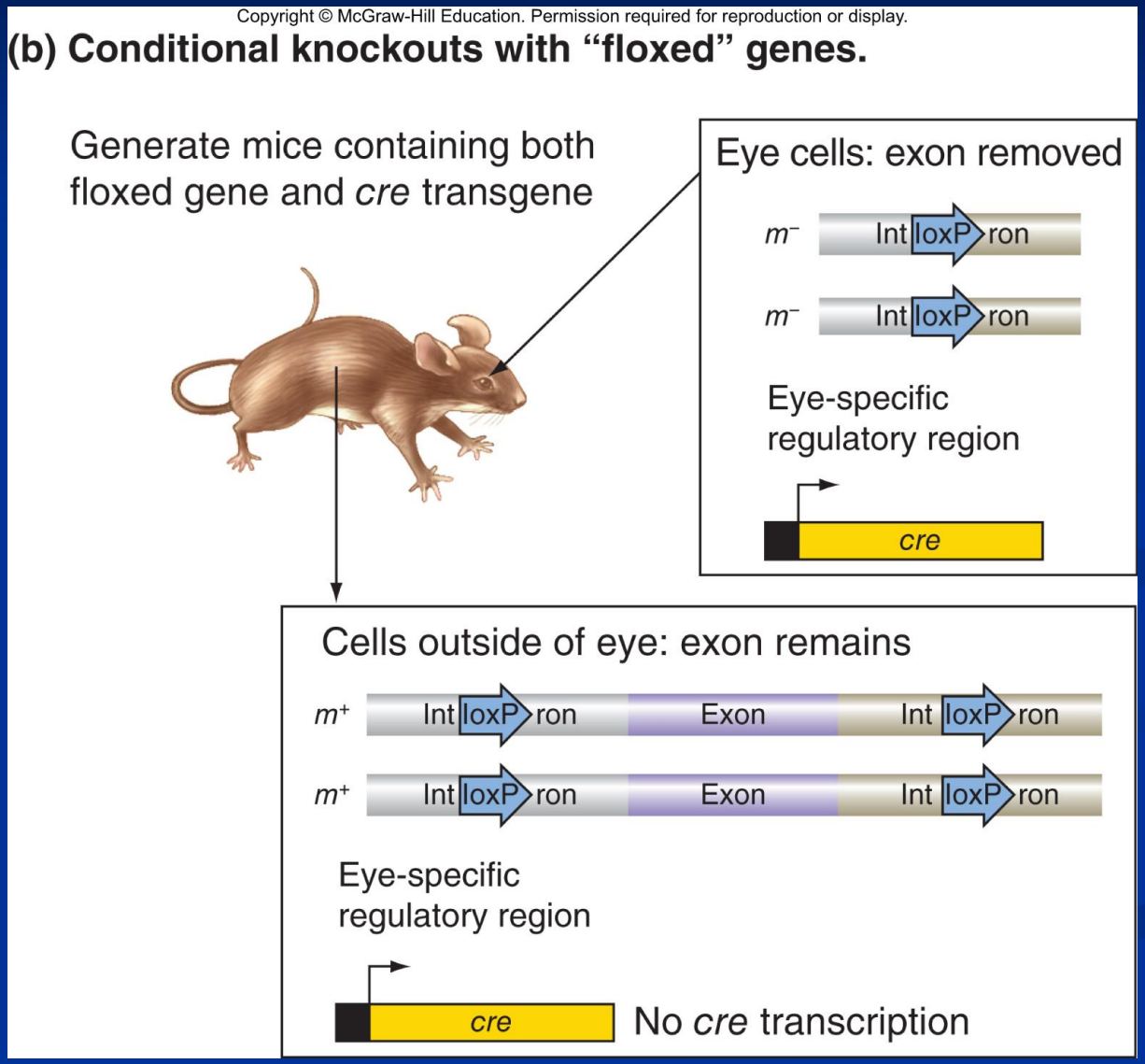


Fig. 17.10 p. 587