Chapter 10

Analyzing genomic variation

Sections to study

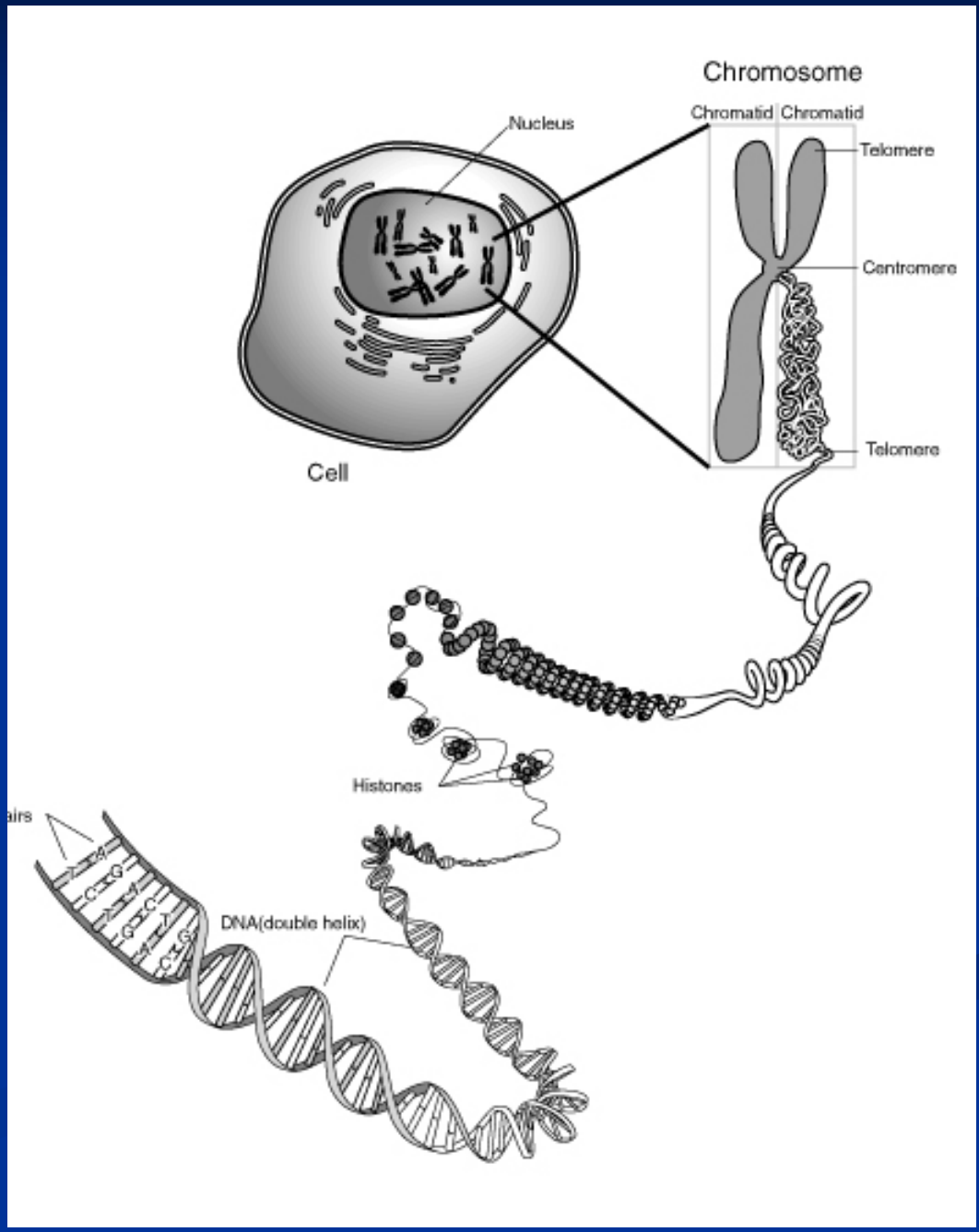
10.1 Variation among individual genomes

10.2 Four classes of DNA polymorphisms

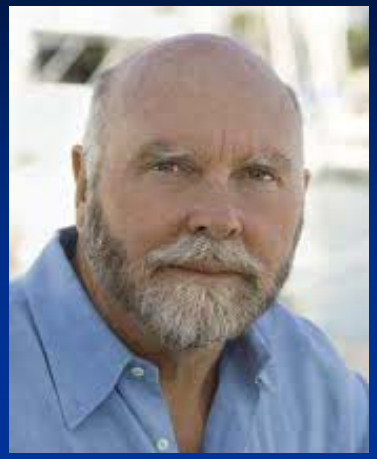
10.4 Positional cloning – Use of polymorphic DNA markers to clone disease-causing genes

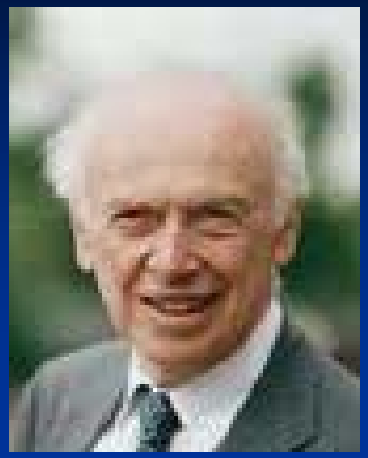
10.1 Variation among individual genomes

© Reprinted Full cover of Science 2/16/01.
© 2001 American Association for the Advancement of Science

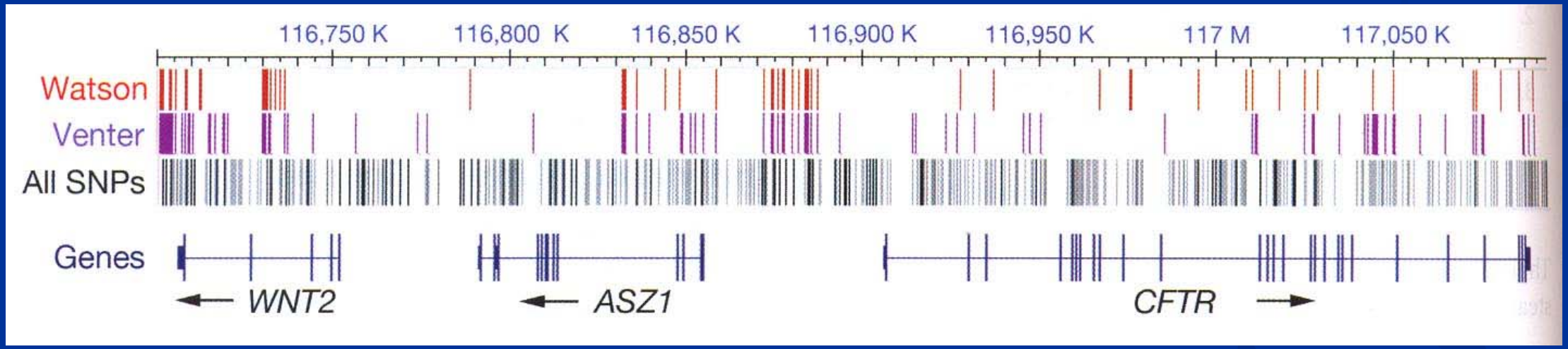


J. Craig Venter
(Genome published 2007-09)





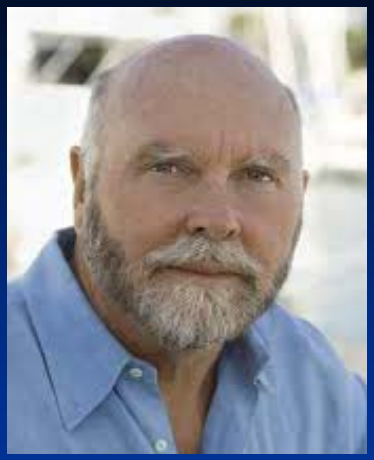
James Watson
(Genome published 2007-02)

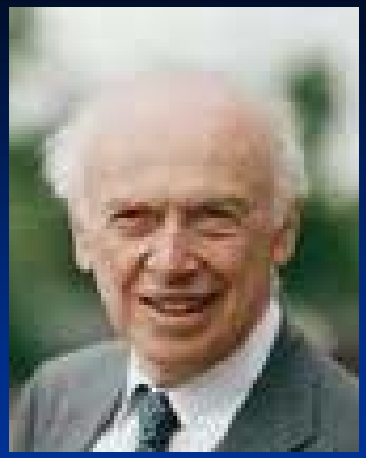


SNP variation in a region on chromosome 7

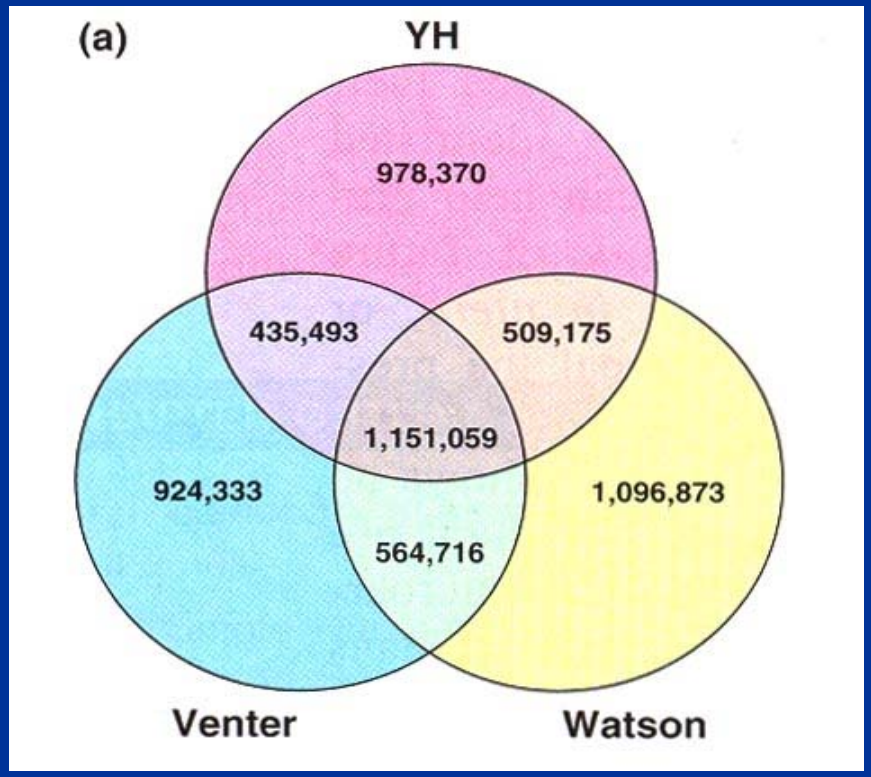
Fig. 10.5

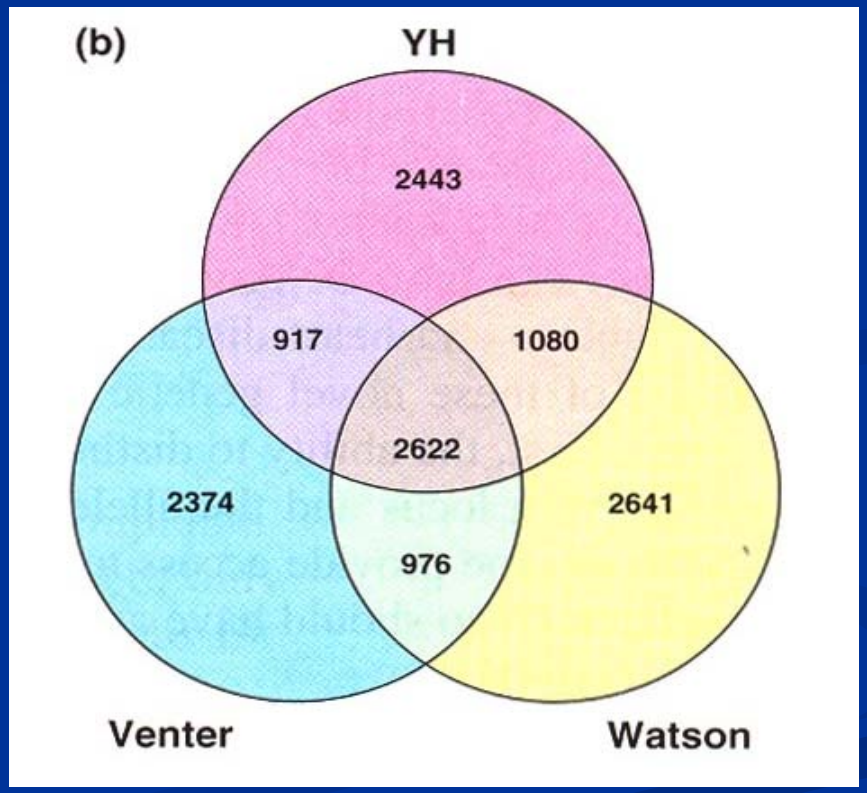
J. Craig Venter





James Watson





Single nucleotide substitutions

Amino acid-changing substitutions

Figure 11.2

Why the amino acid-changing differences are not abundant?

- Less than 2% of the human genome codes for genes.
- Even when mutations occur, many mutations of codons are silent (don't change the amino acid).
- If a particular mutation is not silent and has deleterious effects, natural selection could often lead to its disappearance from the human population.

- 1 in 1000 bp difference in any two randomly chosen humans.
- 3 million differences between any two randomly chosen individuals. (human haploid genome is 3×10^9 bp)
- Most differences are in non-coding, nonregulatory regions.

Extension to Mendelian definition of locus and allele in the post-genome era

	Mendelian genetics	Modern genetics
Locus	A gene	A designated location anywhere on a chromosome
Allele	Alternative forms of a gene	One of two or more alternative DNA sequences found at a locus – either coding or noncoding

- **(DNA) polymorphism** – a gene or noncoding region that has two or more alleles.
- **Polymorphic** – a locus with two or more alleles in a population.
- **DNA marker** – a polymorphic locus useful for mapping studies, disease diagnosis and others.
- **Anonymous locus** – a locus on the chromosome without known function.

10.2 Four classes of DNA polymorphisms

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

I. Single nucleotide polymorphism (SNP)

(a)

Allele	Genome	Single-Strand Notation
1	AGCTCTGATAC TCGAGACTATG	AGCTCTGATAC
2	AGCTCCGATAC TCGAGGCTATG	AGCTCCGATAC

(b)

Genotype	Notation
Homozygous T	TT
Heterozygous T	TC
Homozygous C	CC

- Single base-pair substitution.
- 1 per 1000 bp in the human genome.
- 2001 – over 5 million human SNPs identified.
- Mutation rate of 1×10^{-9} per locus per generation.
- Arise by mutagenic chemicals or mistakes in DNA replication.

- The origin of human SNP is determined by comparison to other species.
- The vast majority of SNPs occurs at anonymous loci.

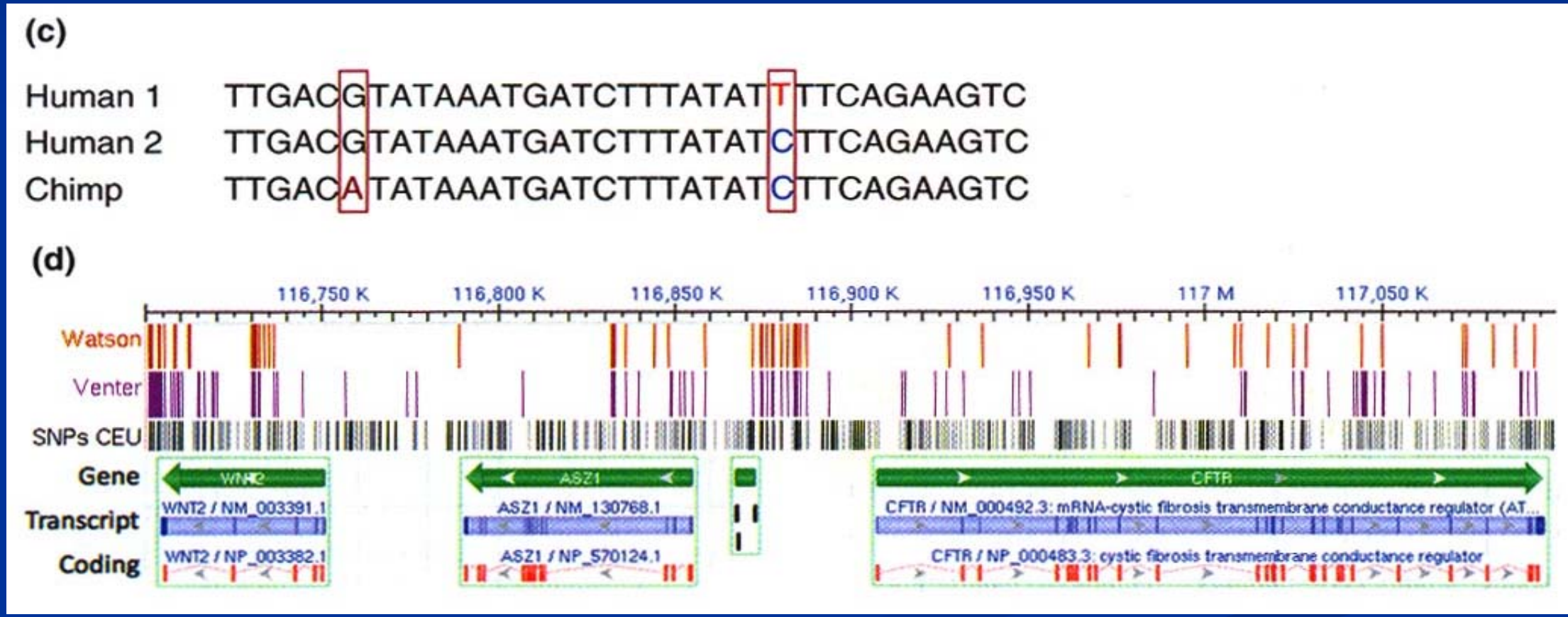


Fig 10.6

SNPs can be genotyped with several methods

1. Southern blot or PCR analysis of restriction site-altering SNPs
2. Allele-specific oligonucleotide (ASO) hybridization
3. DNA microarrays

1. Detection of SNPs that alter restriction sites by Southern blot analysis

Restriction fragment length polymorphism (RFLP)
is one type of SNP that changes DNA fragment size cut by specific restriction enzymes.

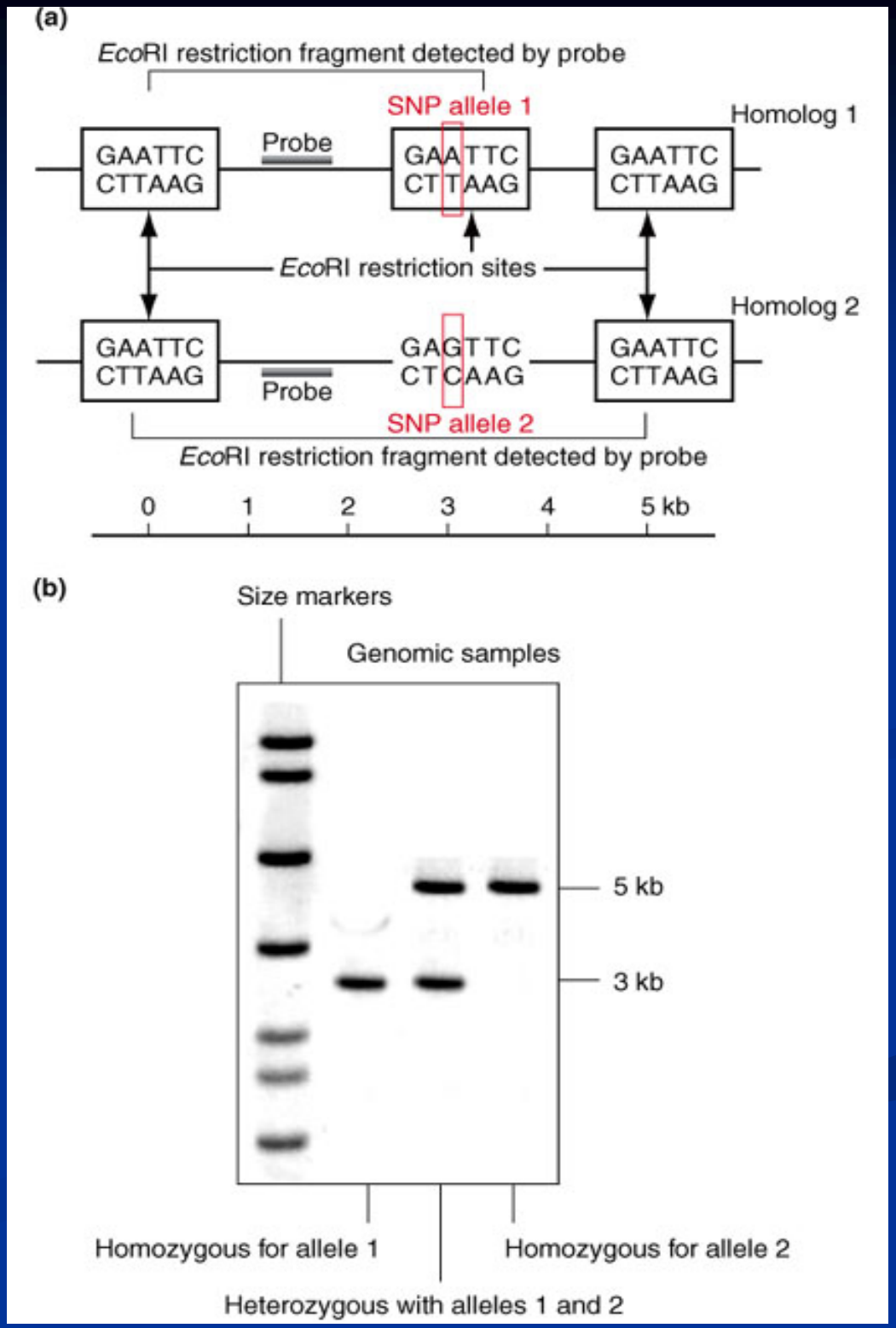


Fig. 11.4

Detection of SNPs that alter restriction sites by PCR analysis

- Must have sequence on either side of polymorphism
  - Amplification of fragment
  - Restriction enzyme digestion
  - Gel electrophoresis
- e.g., sickle-cell anemia genotyping with a PCR-based protocol

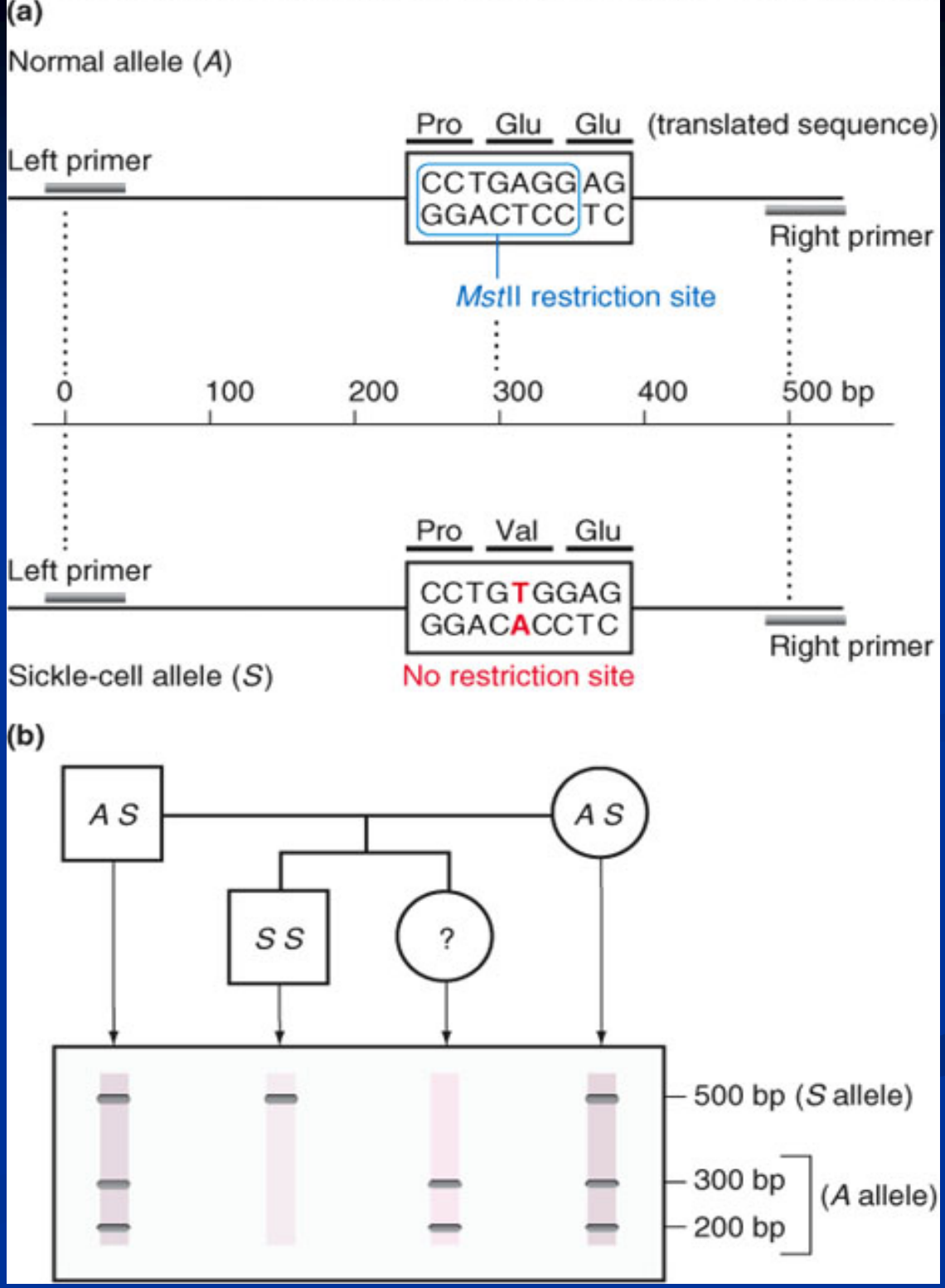


Fig. 11.5

2. SNP detection by ASO (allele-specific oligonucleotide) hybridization

ASOs are short oligonucleotides of 30-40 bases that hybridize to only one of the two alleles at a SNP locus under appropriate conditions.

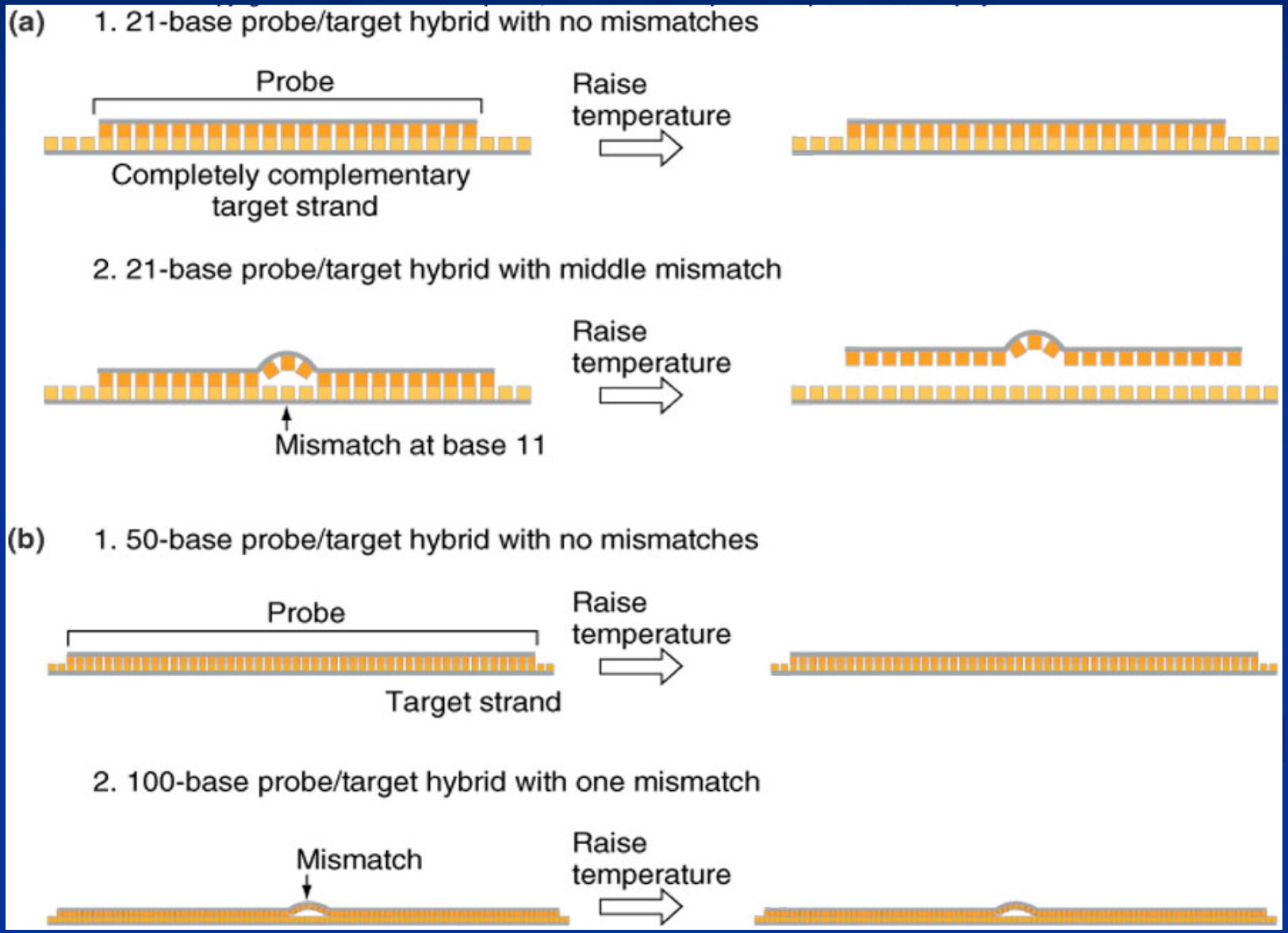
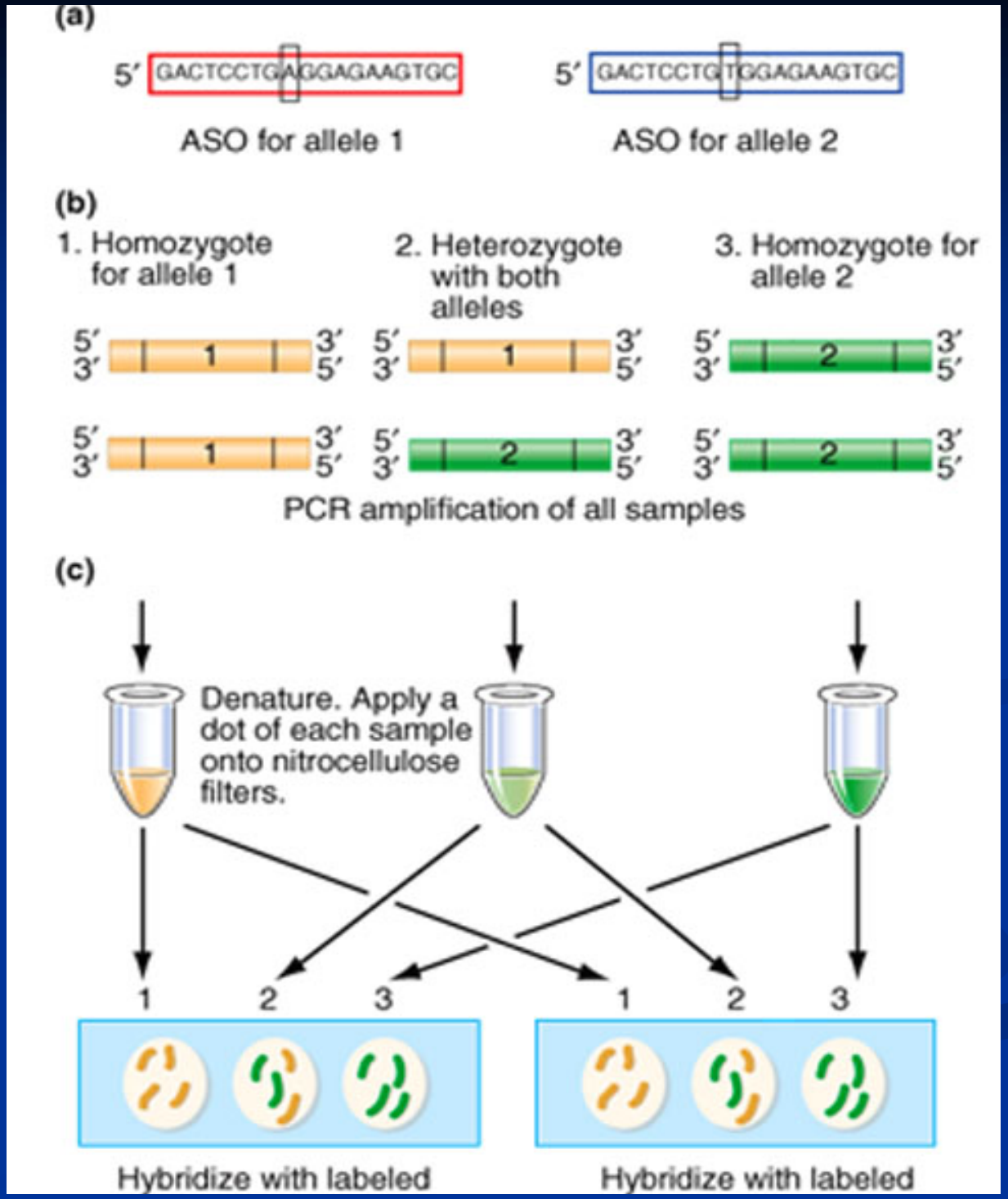


Fig. 10.16

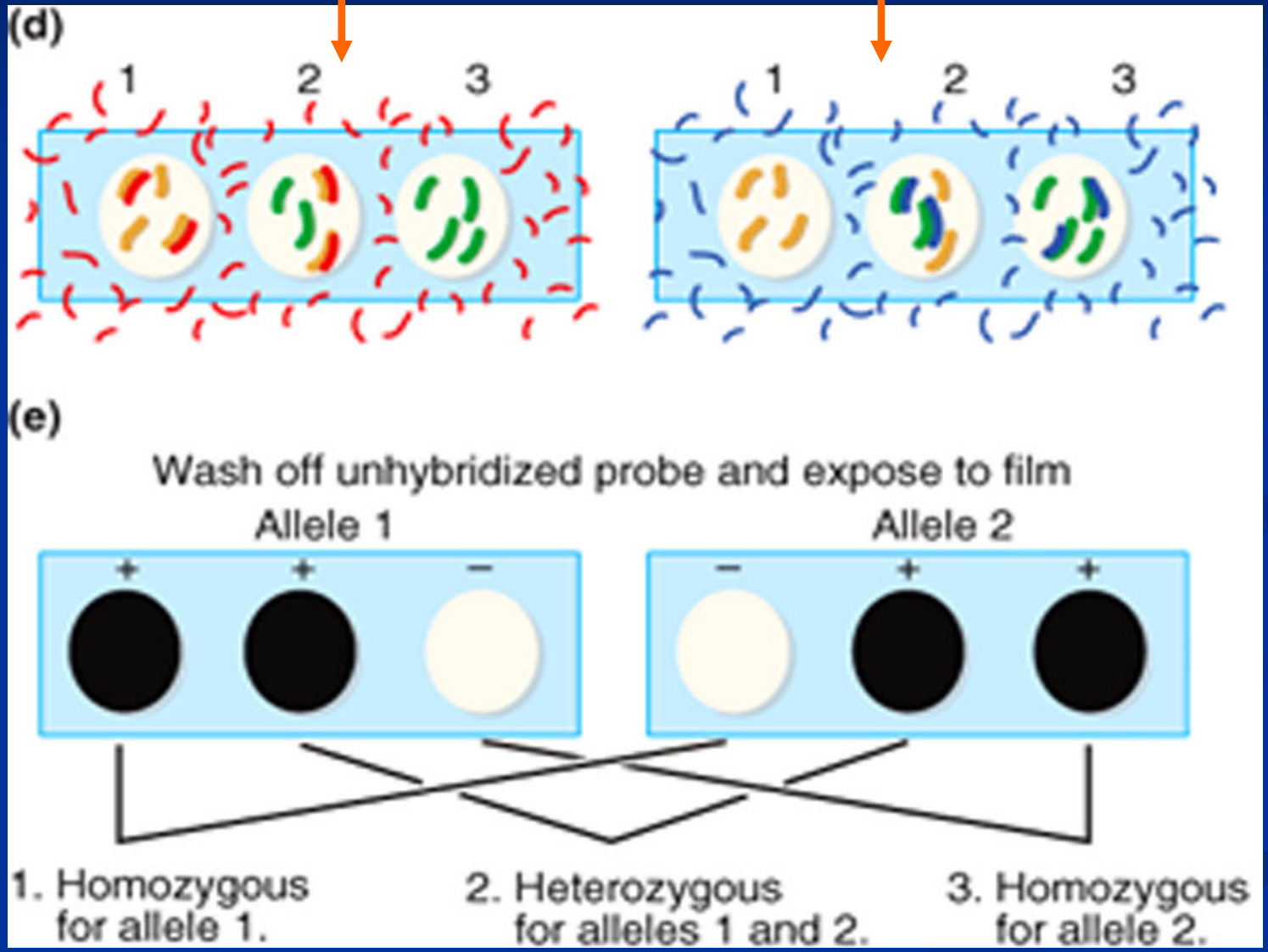
The PCR/ASO hybridization method

ASOs can determine genotype at any SNP locus



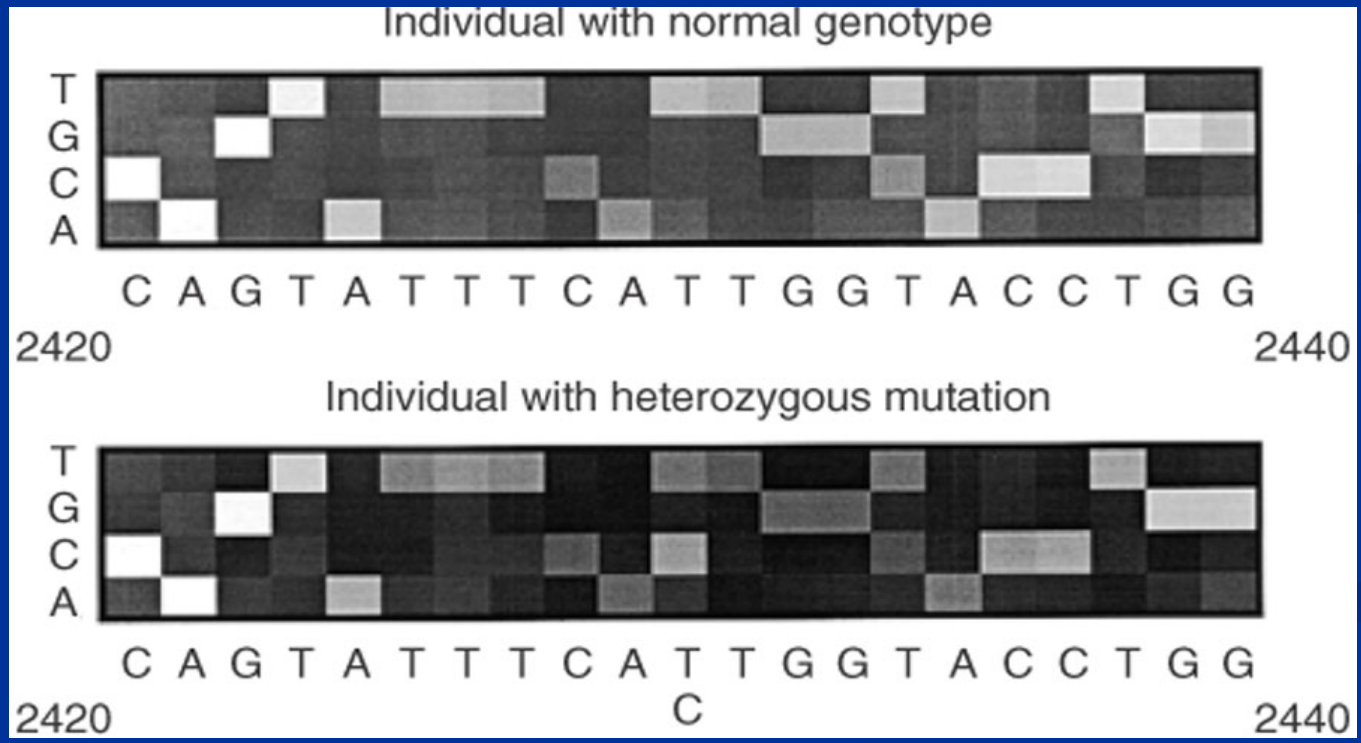
Hybridize with labeled ASO for allele 1

Hybridize with labeled ASO for allele 2



3. Rapid SNP detection with DNA microarrays

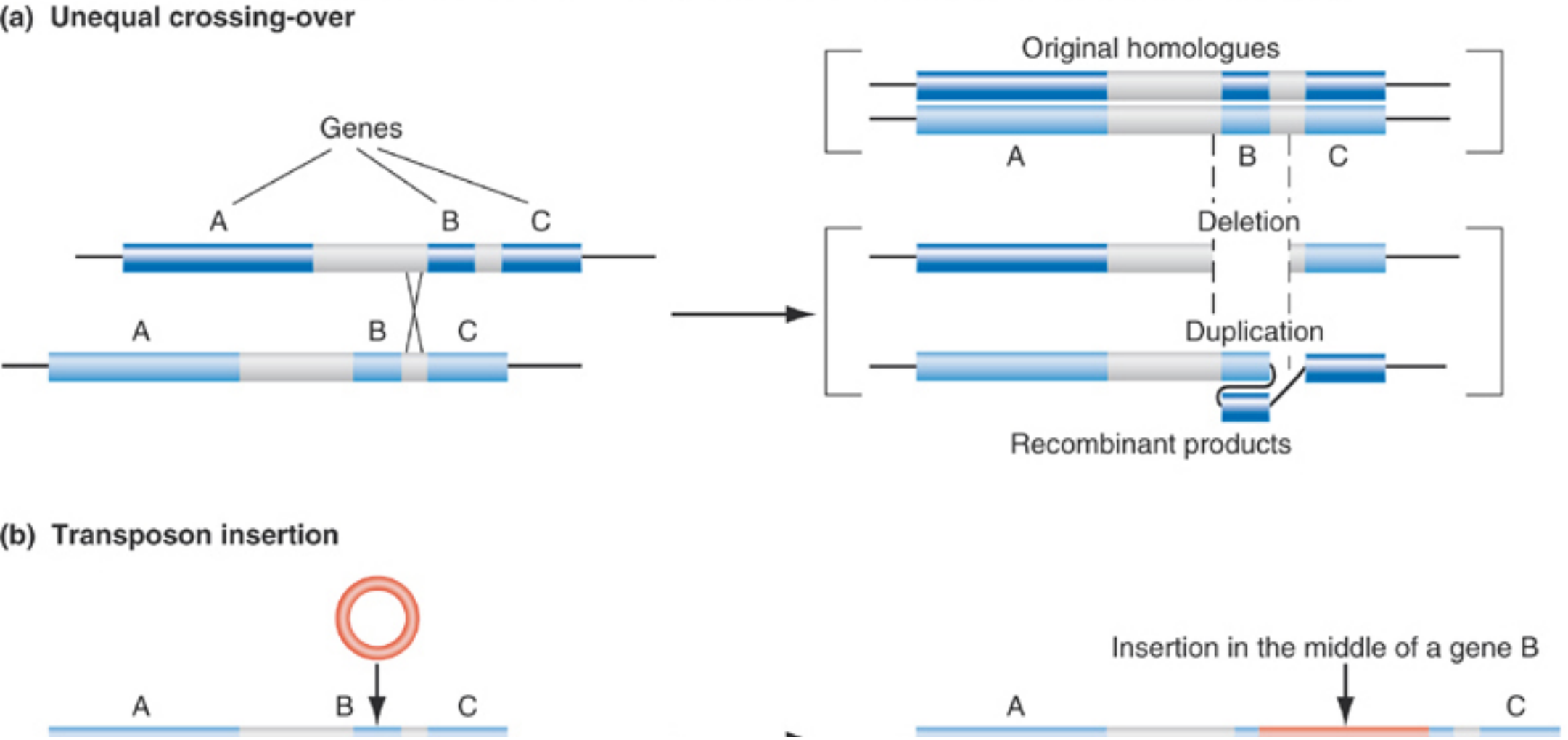
- 5500 bp *BRCA1* coding region requires 22000 ASOs.
- Each column contains an ASO differing only at the nucleotide position under analysis.
- BRCA1 DNA from any allele can hybridize with one of the four ASOs in a column.



II. Deletion-insertion polymorphisms (InDels or DIPs)

- Short insertions or deletions of DNA sequences that are 2-100 bp long.
- Distributed at about one in every 10 kb of DNA.
- Mutation rate of less than 10^{-9} per locus per gamete.
- Caused by mutagenic events that expand or contract the length of nonrepetitive DNA.
  - Small deletions and duplications arise by unequal crossing-over.
  - Small insertions can also be caused by transposable elements.

InDels arise from unequal crossing-over or transposon insertion



III. Simple sequence repeats (SSRs) a.k.a. microsatellites

- Repeating units 1–3 bp in length, repeated in tandem 15–100 times, *e.g.* AAAAAAAAAAAA or CACACACACACACA.
- CA-repeat occurs 1 per 30,000 bp in the mammalian genome.
- Originated from replication error.
- New alleles arise at a rate of 10^{-3} per locus per gamete.
- Useful as relatively stable, highly polymorphic DNA markers.

SSRs originated mainly from faulty DNA replication

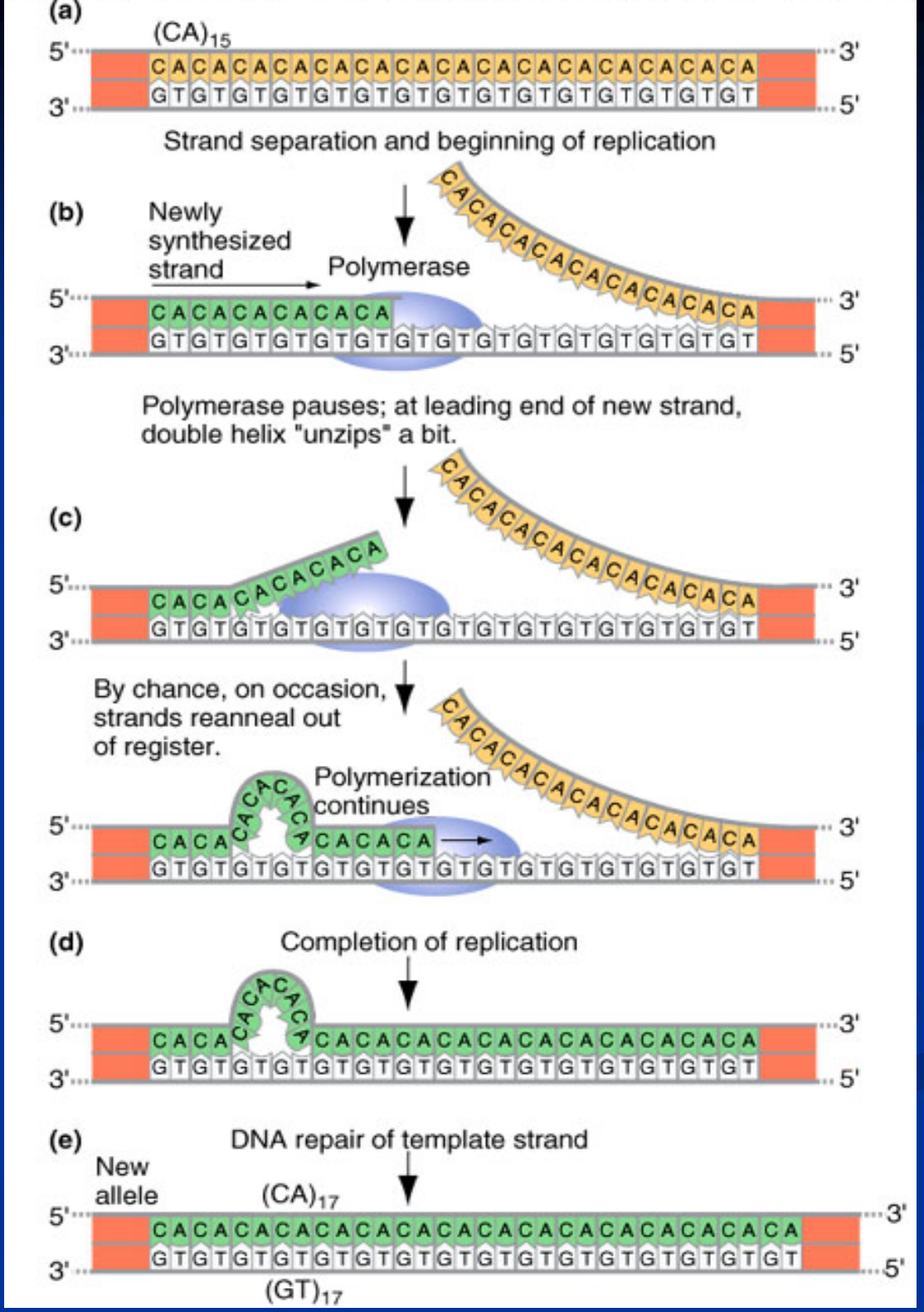


Fig. 11.9

SSR polymorphism detection

- PCR amplification
- Gel electrophoresis

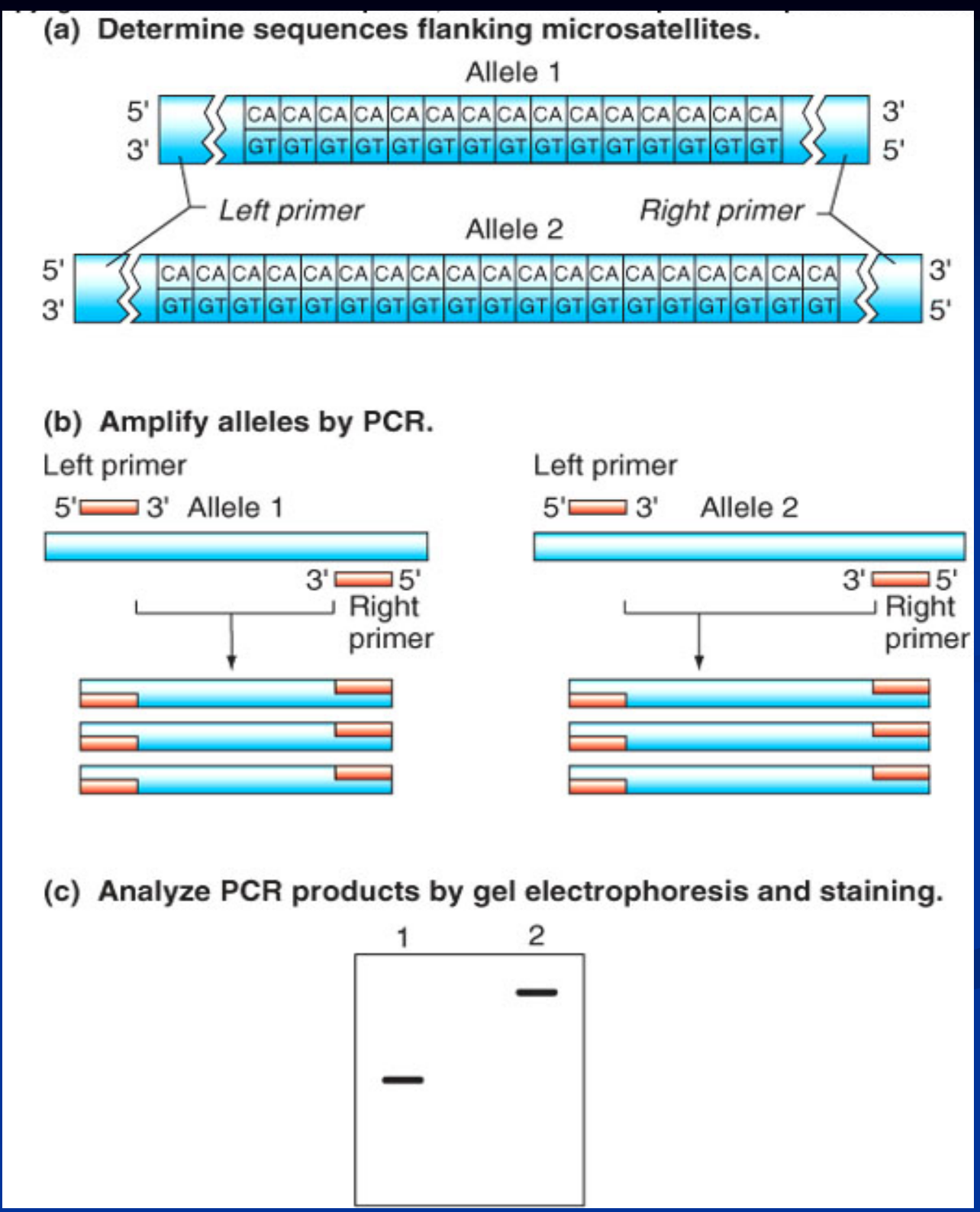


Fig. 10.12

SSRs are often highly polymorphic with many different alleles present in a population.

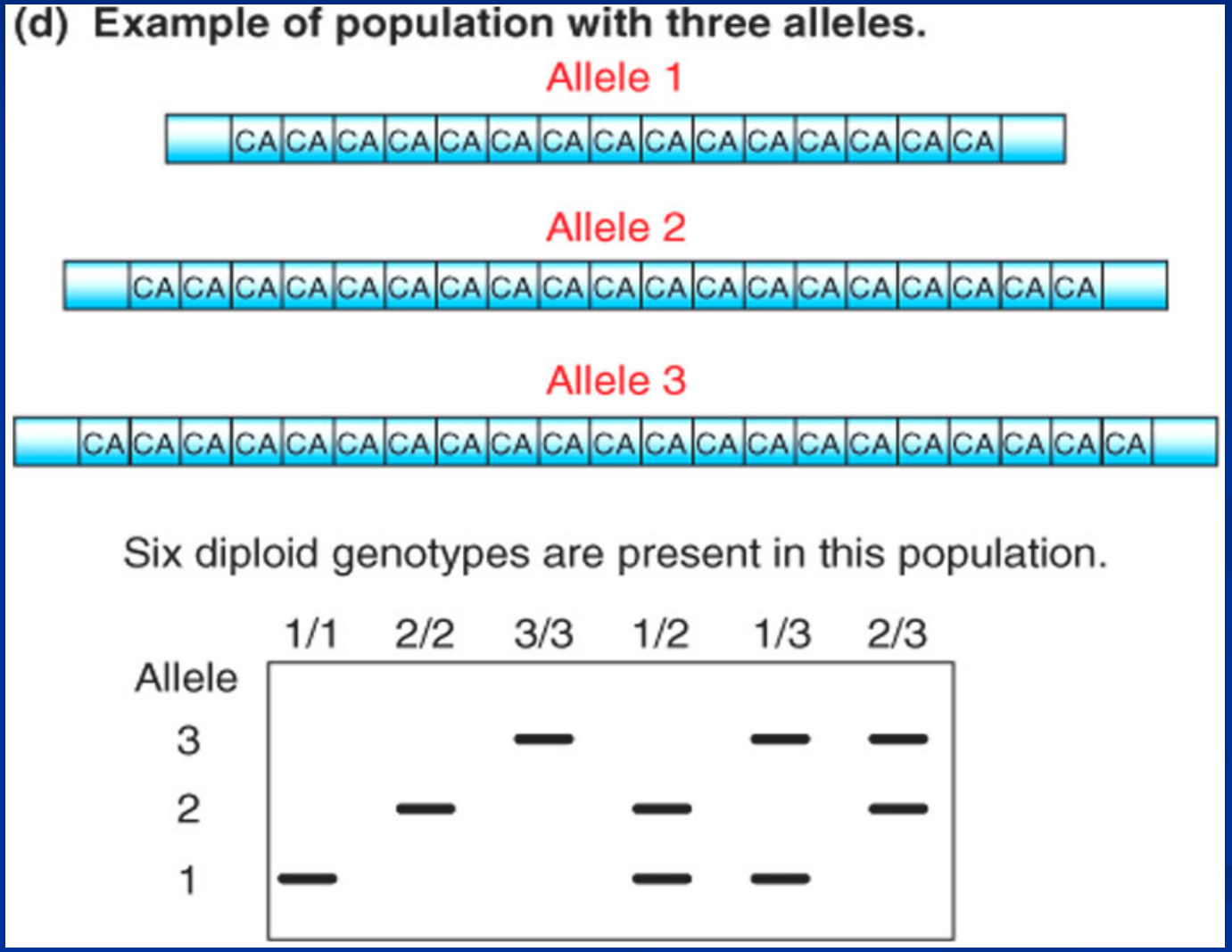


Fig. 10.12

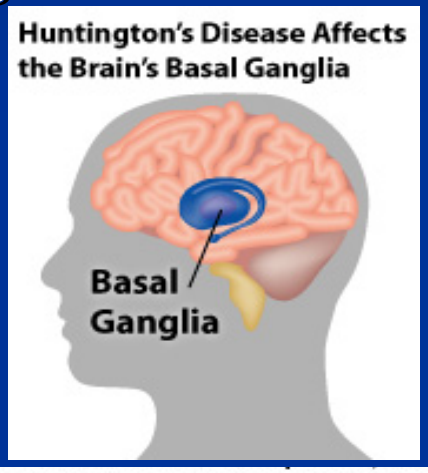
SSRs and disease

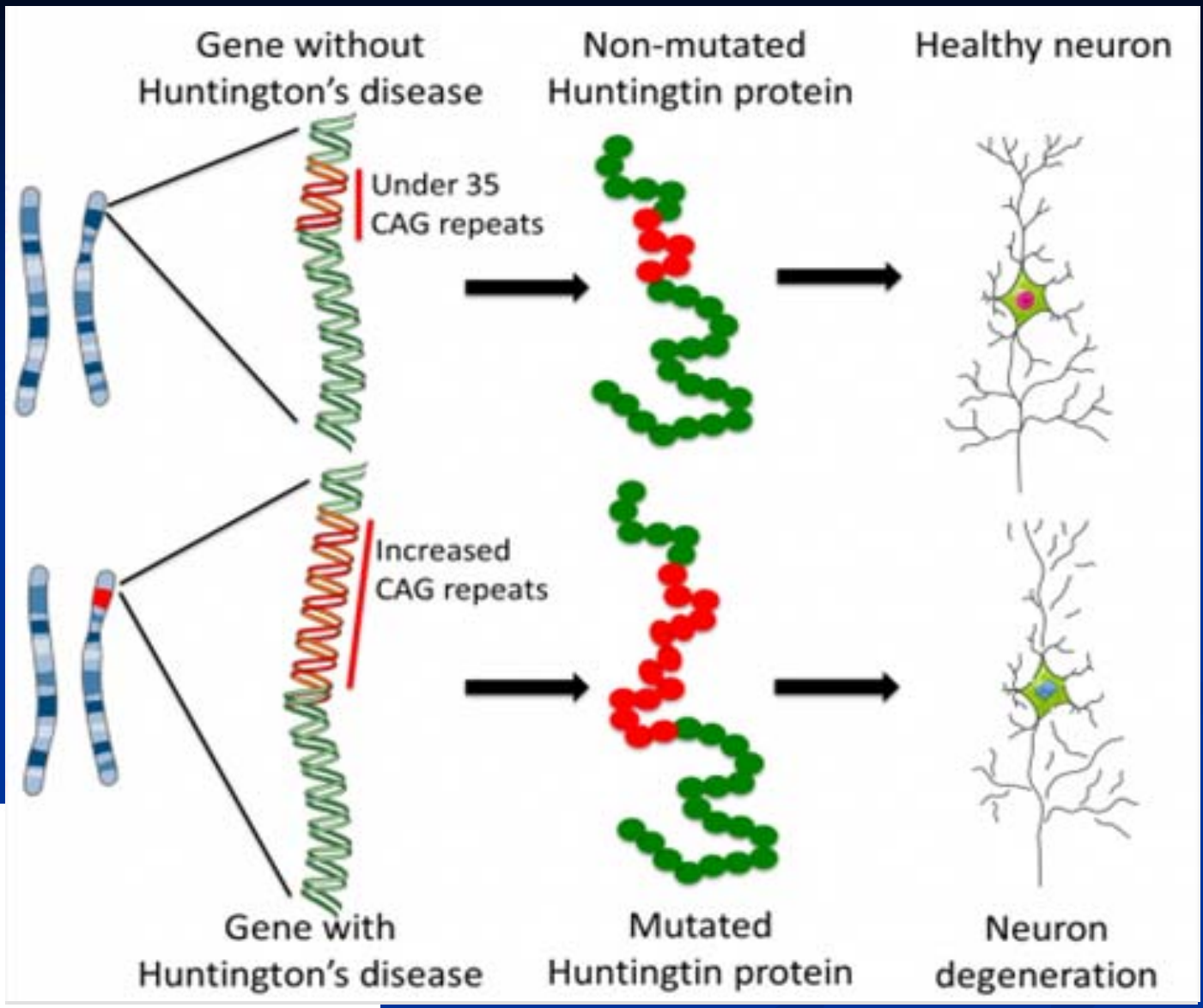
- **Huntington disease: expansion of CAG repeats in the *HD* gene.**
- **Fragile X syndrome: expansion of CGG repeats in *FMR-1* gene.**

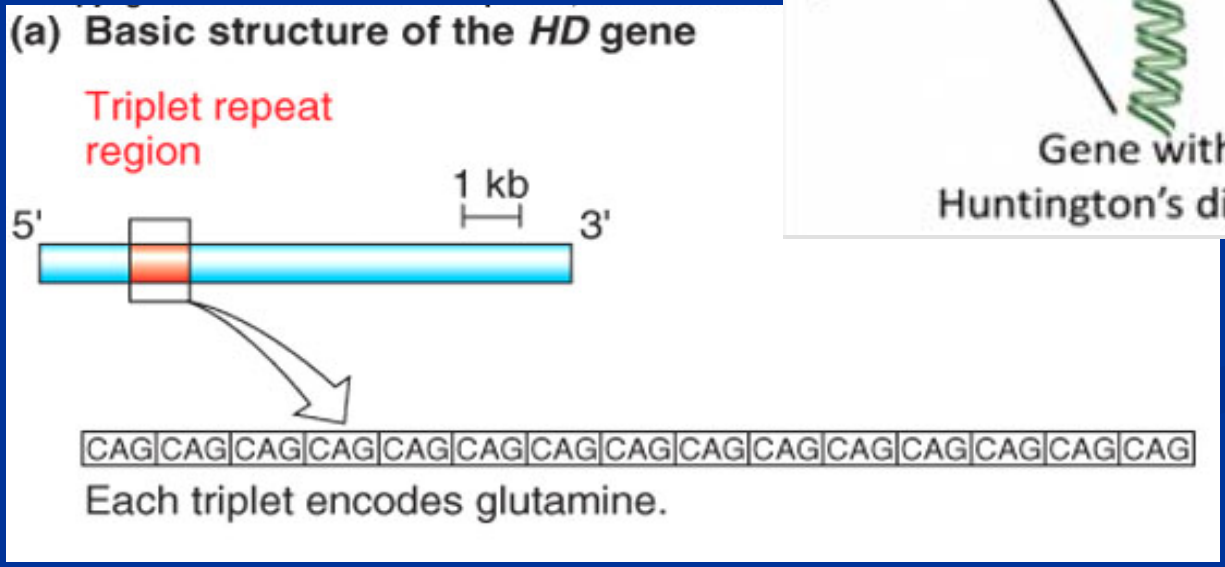
Huntington disease (HD, 亨廷顿舞蹈症):

- First described by George Huntington in 1872.
- Autosomal dominant disease.
- Symptoms include involuntary, jerky movements; unsteady gait; mood swings; personality changes; slurred speech; impaired judgment.
- Symptoms usually show up between the ages of 30 and 50.
- In 1993, after 10 years of intensive research, investigators identified and cloned the *HD* (or *HTT*) gene.

Huntington disease is caused by the expansion of a CAG triplet repeat in the *HD* gene







- < 34 repeats: normal
- > 42 repeats: disease
- More repeats, earlier disease onset.
- Highest reported CAG repeat copies is about 250.

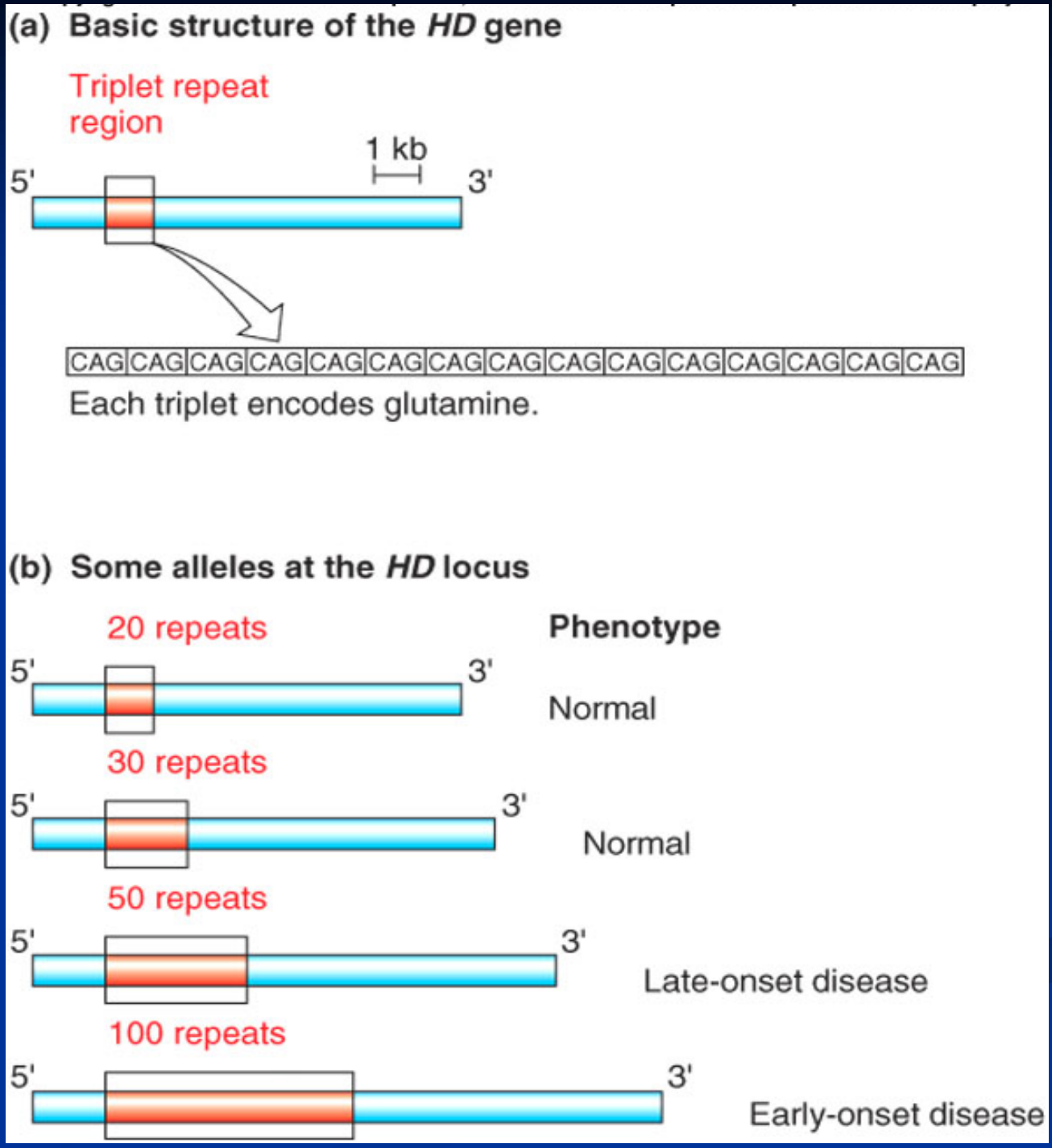
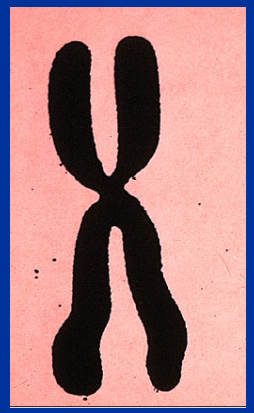


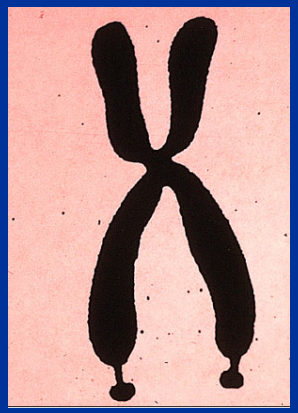
Fig. 10.13

Fragile X syndrome:

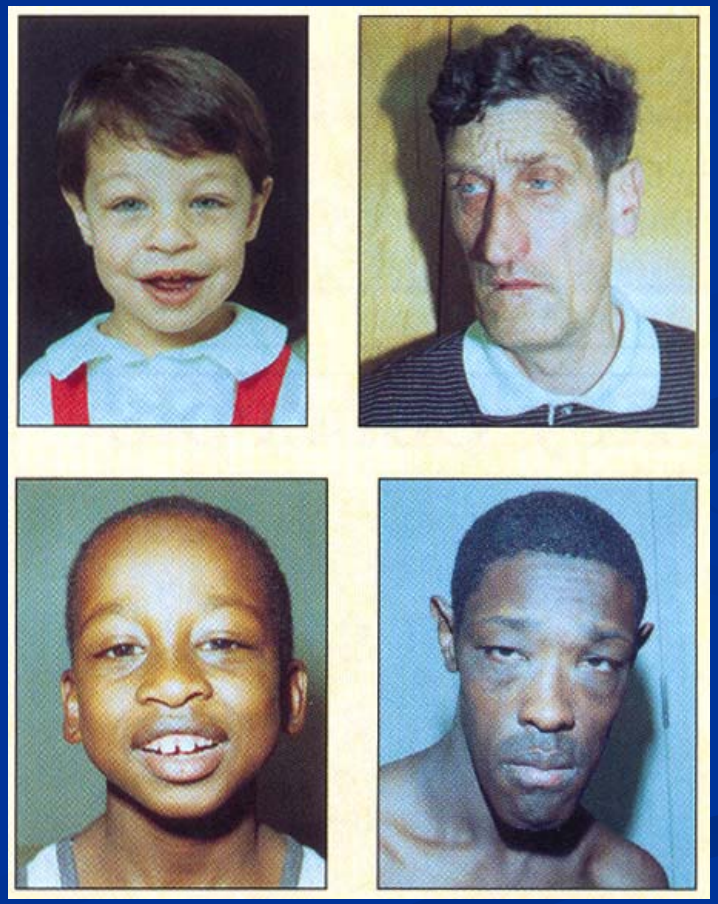
- Moderate to severe mental retardation, second only to Down syndrome.
- Affects about one in 4,000 males and one in 8,000 females.



Normal
X chromosome



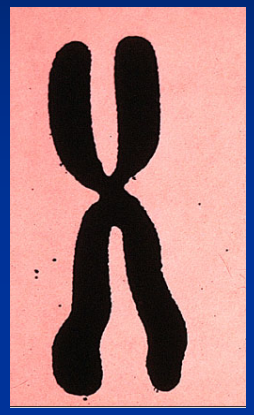
FMR patient
X chromosome



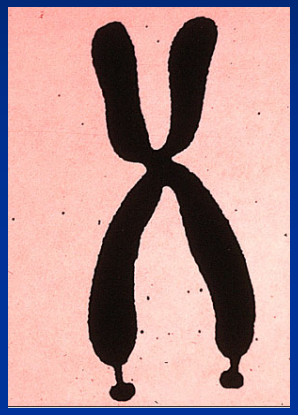
Young Older

Genetics and Society

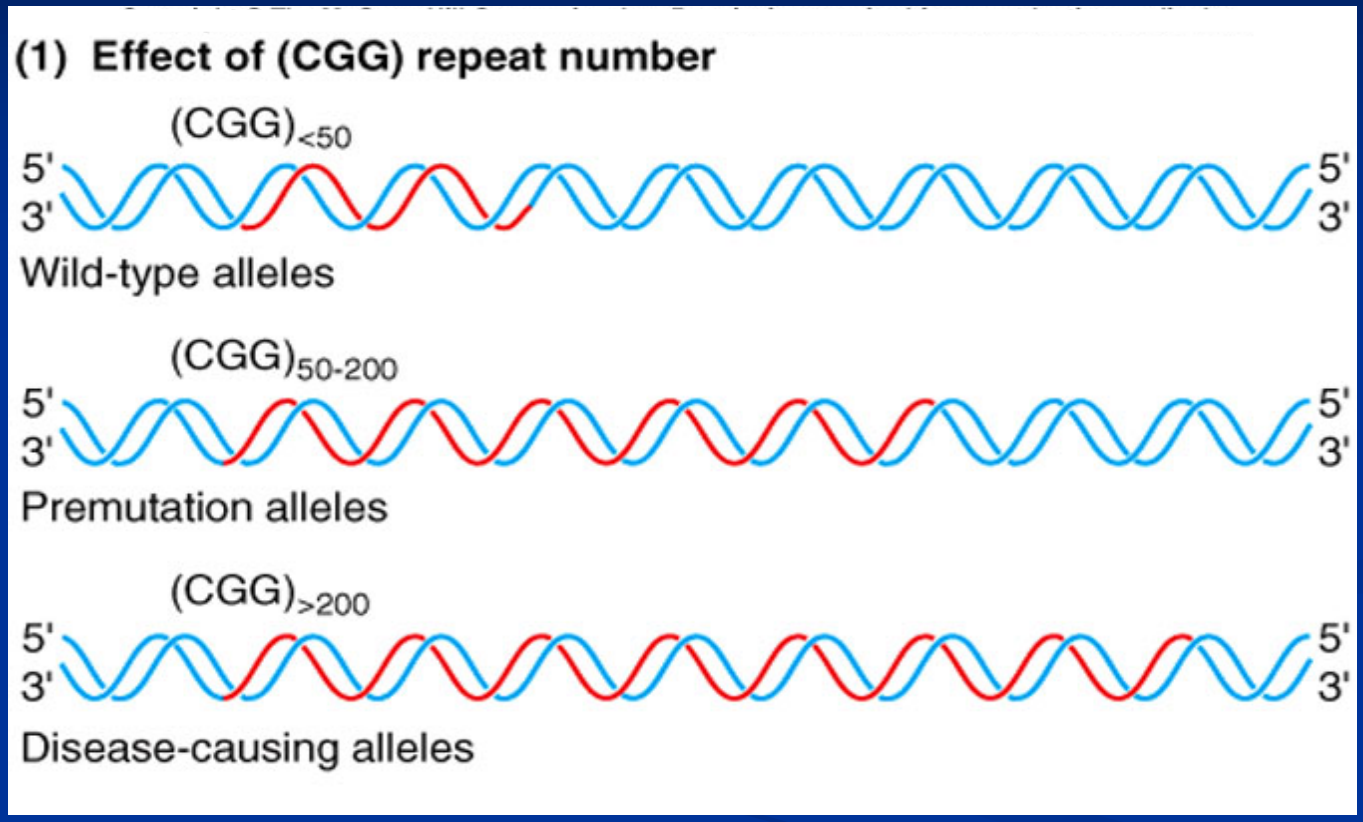
Fig. A, B(2)



Normal X chromosome



FMR patient X chromosome



Genetics and Society

Fig. A, B(2)

Triplet Repeat Disorders

Disorder	OMIM	mRNA Repeat	Normal Number of Copies	Disease Number of Copies	Signs and Symptoms (Phenotype)
Fragile X syndrome	309550	CGG or CCG	6–50	200–2,000	Mental retardation, large testicles, long face
Friedreich ataxia	229300	GAA	6–29	200–900	Loss of coordination and certain reflexes, spine curvature, knee and ankle jerks
Haw River syndrome	140340	CAG	7–25	49–75	Loss of coordination, uncontrollable movements, dementia
Huntington disease	143100	CAG	10–34	40–121	Personality changes, uncontrollable movements, dementia
Jacobsen syndrome	147791	CGG	11	100–1,000	Poor growth, abnormal face, slow movement
Myotonic dystrophy type I	160900	CTG	5–37	80–1,000	Progressive muscle weakness; heart, brain, and hormone abnormalities
Myotonic dystrophy type II	602668	CCTG	<10	>100	Progressive muscle weakness; heart, brain, and hormone abnormalities
Spinal and bulbar muscular atrophy	313200	CAG	14–32	40–55	Muscle weakness and wasting in adulthood
Spinocerebellar ataxia (5 types)	271245	CAG	4–44	40–130	Loss of coordination

IV. Copy number variants (CNVs)

- Short copy number repeats (minisatellites)
- Large-scale deletions and duplications – copy number variants (CNVs) or copy number polymorphisms (CNPs)

Short copy number repeats are ideal for DNA fingerprinting

Minisatellites:

- Size 0.5 – 20 kb. Repeating units 20-100 bp. Repeated up to thousands of times per locus.
- Particular minisatellite sequences often occur at a small number of different genomic loci.
- Mutation rate of 10^{-3} per locus per gamete.
- Can be analyzed by PCR analysis, but usually analyzed by restriction enzyme digestion and Southern blotting.

Highly polymorphic minisatellites are generated by unequal crossing-over

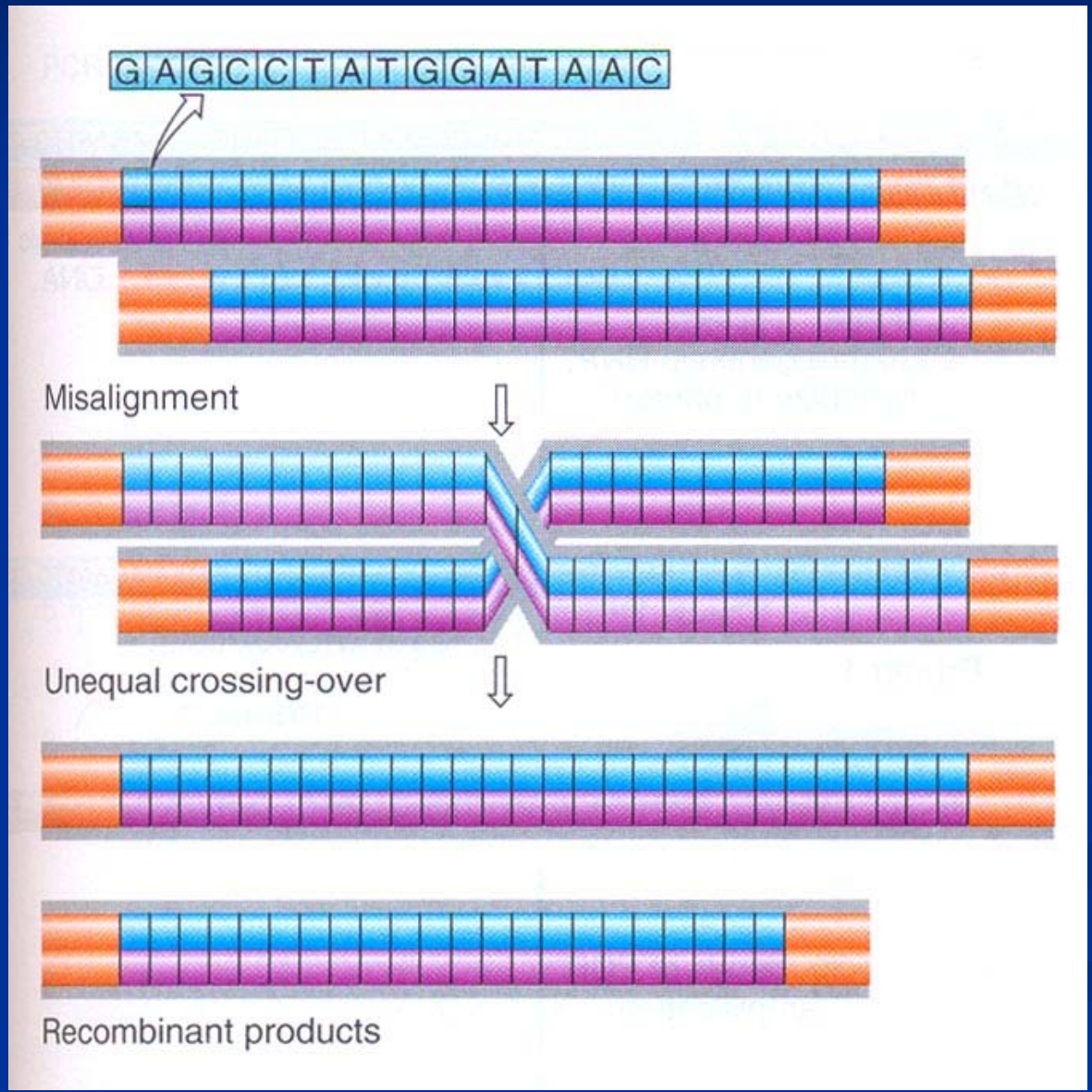
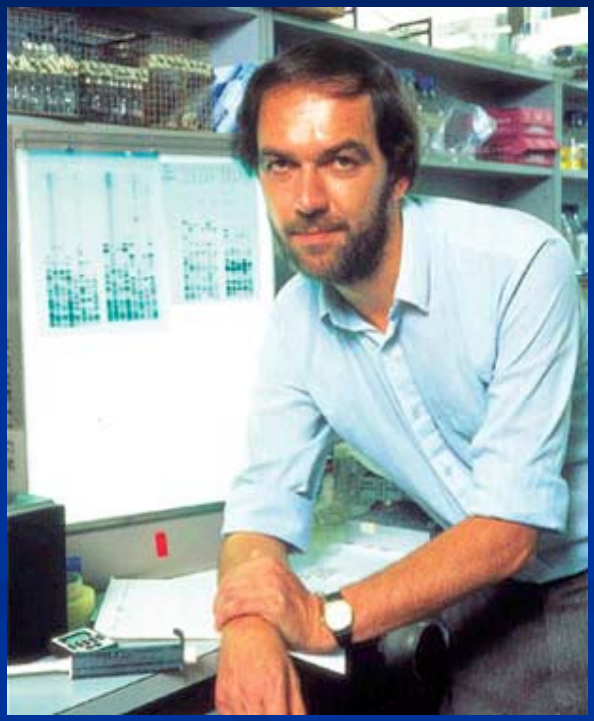
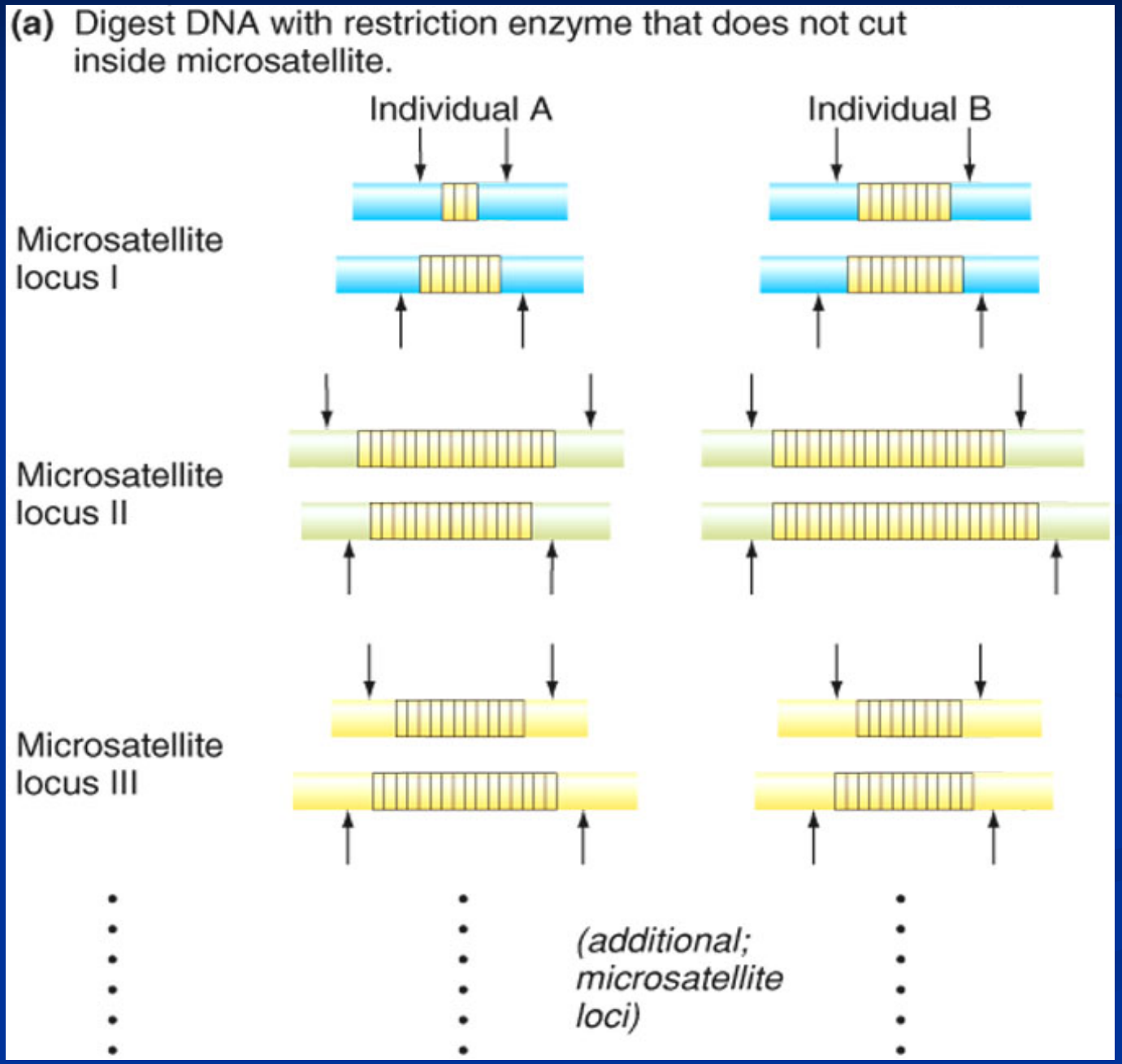


Fig. 10.7

- 1985 – Alec Jeffreys made two key findings.
  - Each minisatellite locus is highly polymorphic.
  - Most minisatellites occur at multiple sites (2-50) around the genome.
- Minisatellites that can cross-hybridize to 5 – 10 loci in the genome are useful for DNA fingerprinting.
  - Occurs 1 per 100,000 bp.
  - ~ 30,000 minisatellites in the human genome.



Minisatellite analysis provides a broad comparison of whole genomes



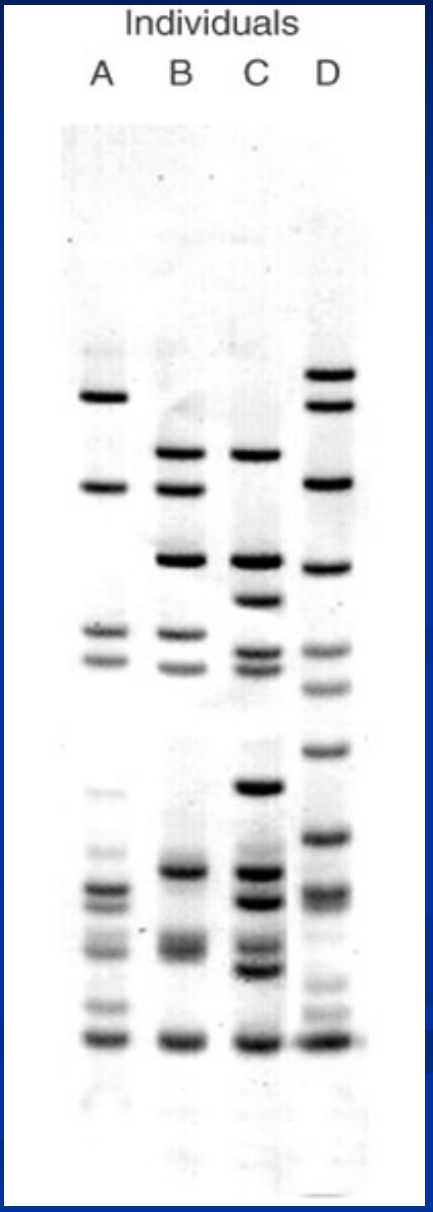


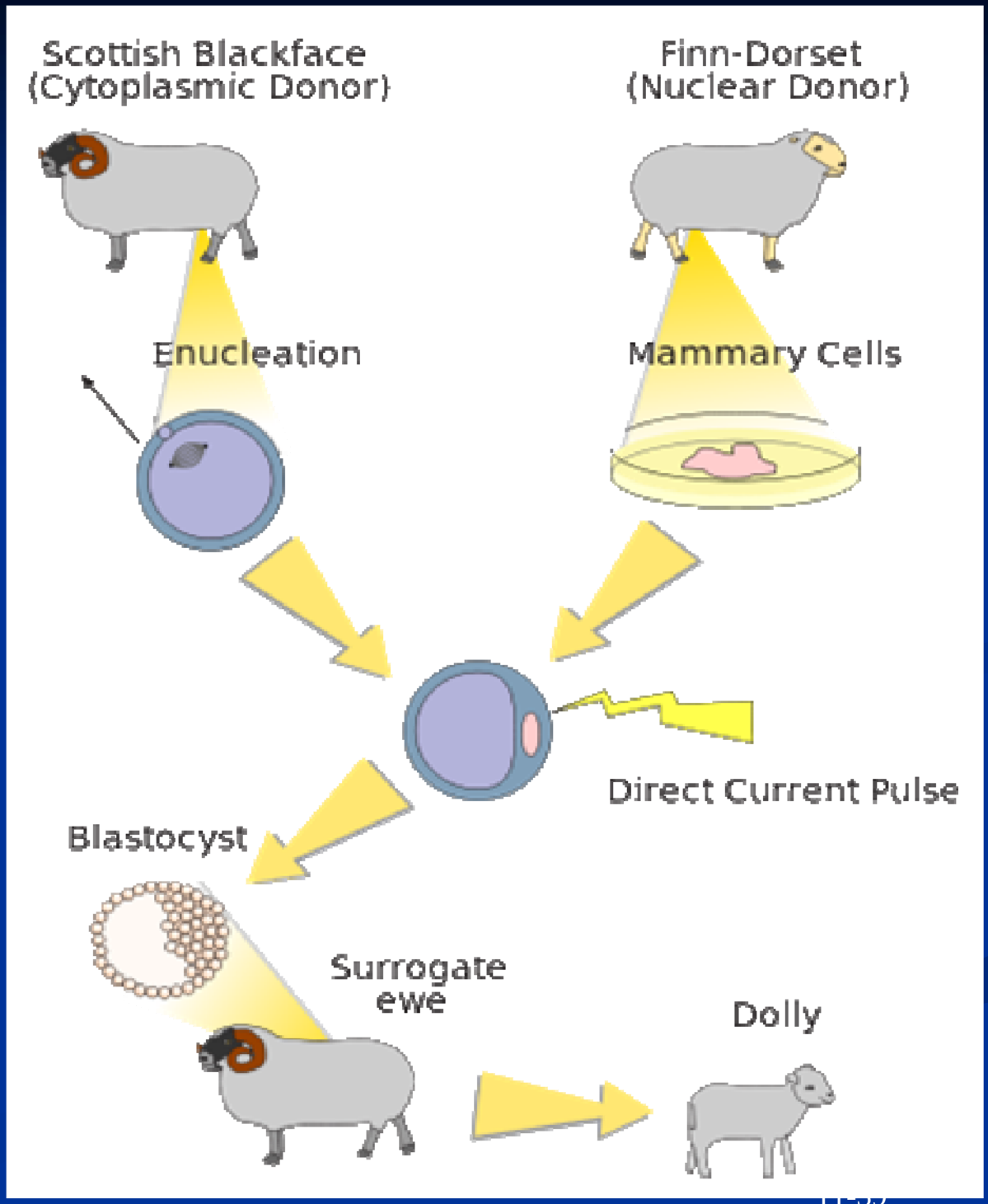
Fig. 11.12

DNA fingerprinting using minisatellites

- Probability of two unrelated individuals having identical minisatellite alleles at one locus (diploid) is 37.5%.
- 24 unrelated loci contribute chance: 0.375^{24} = 1 in 17 billion. (Total population$\leqslant$8 billion)
- DNA fingerprint – a pattern produced by the simultaneous detection of genotype at a group of unlinked, highly polymorphic loci.
  - Use restriction enzymes and Southern blots to detect length differences at minisatellite loci.
  - Most useful minisatellites have 10 – 20 sites around genome and can be analyzed on one gel.

DNA fingerprints can identify individuals and determine parentage

Dolly, the cloned sheep (1996-2003)



- 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

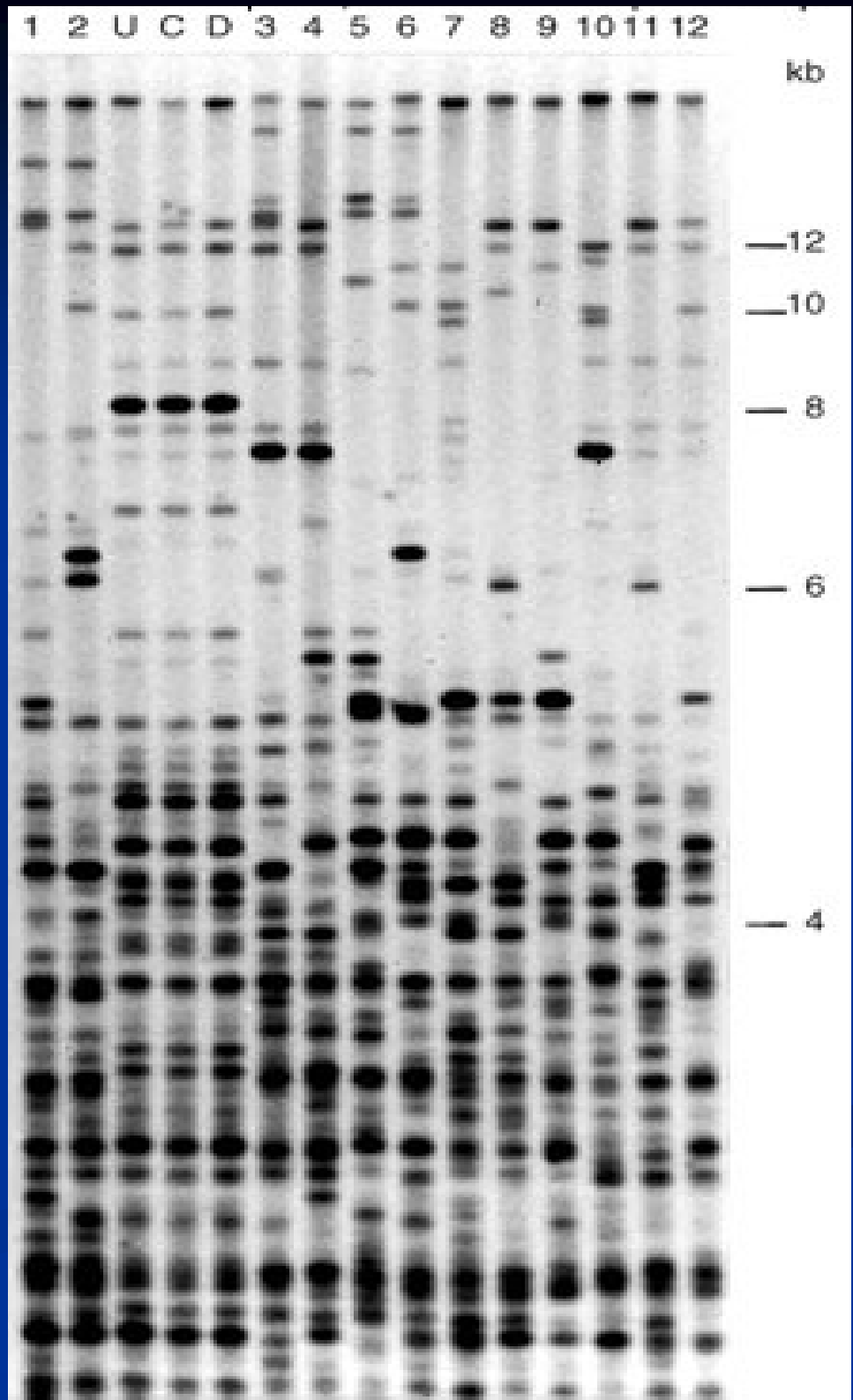


Fig. 11.13

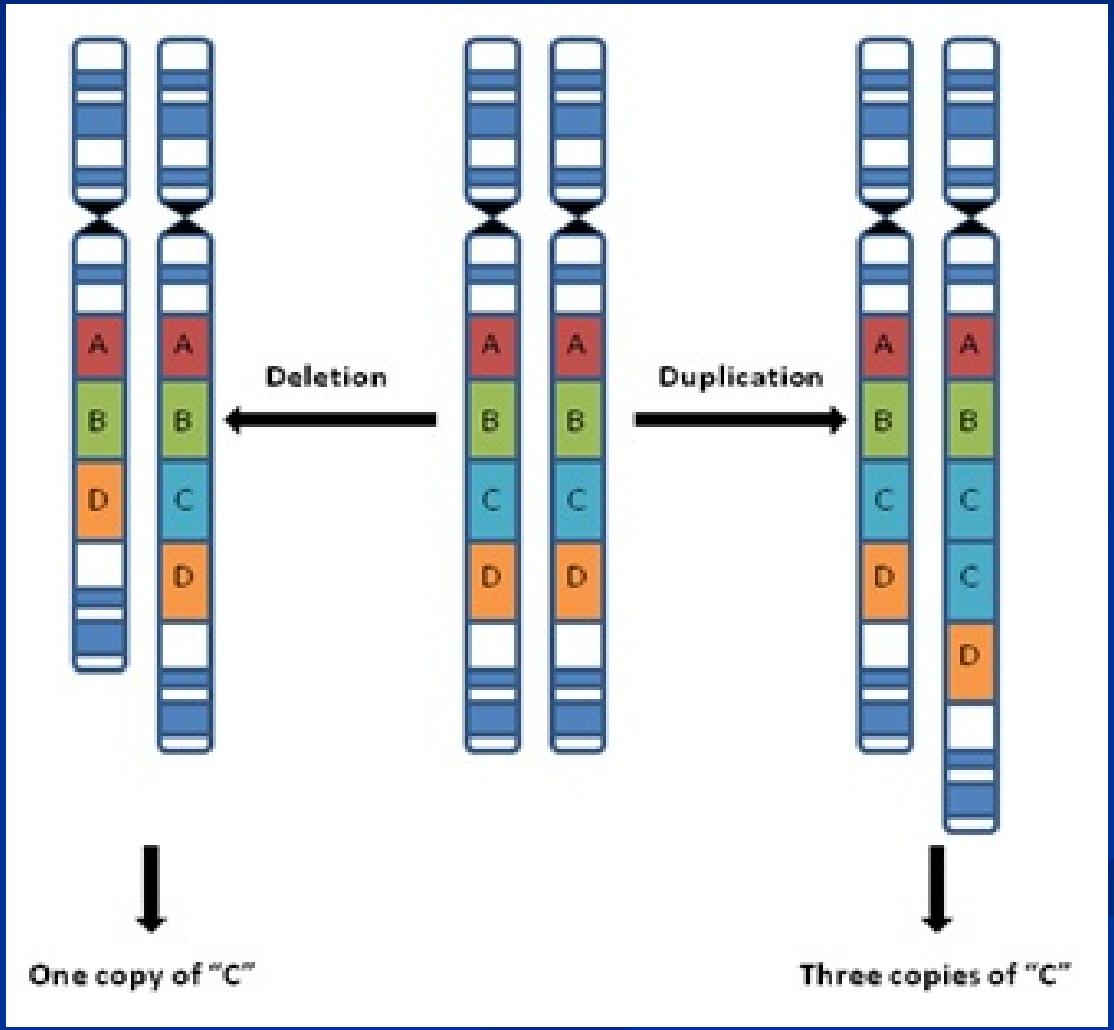
DNA fingerprints are a powerful tool for forensic analysis

- Identification of crime suspects.
- DNA isolated from blood, hair, skin cells, or other genetic evidence left at the scene of a crime can be compared with the DNA of a criminal suspect to determine guilty or innocence.
  - The first criminal conviction based on DNA evidence in the United States occurred in 1988.
  - To the present, more than 50,000 cases worldwide have been solved through the use of this DNA-based technology.

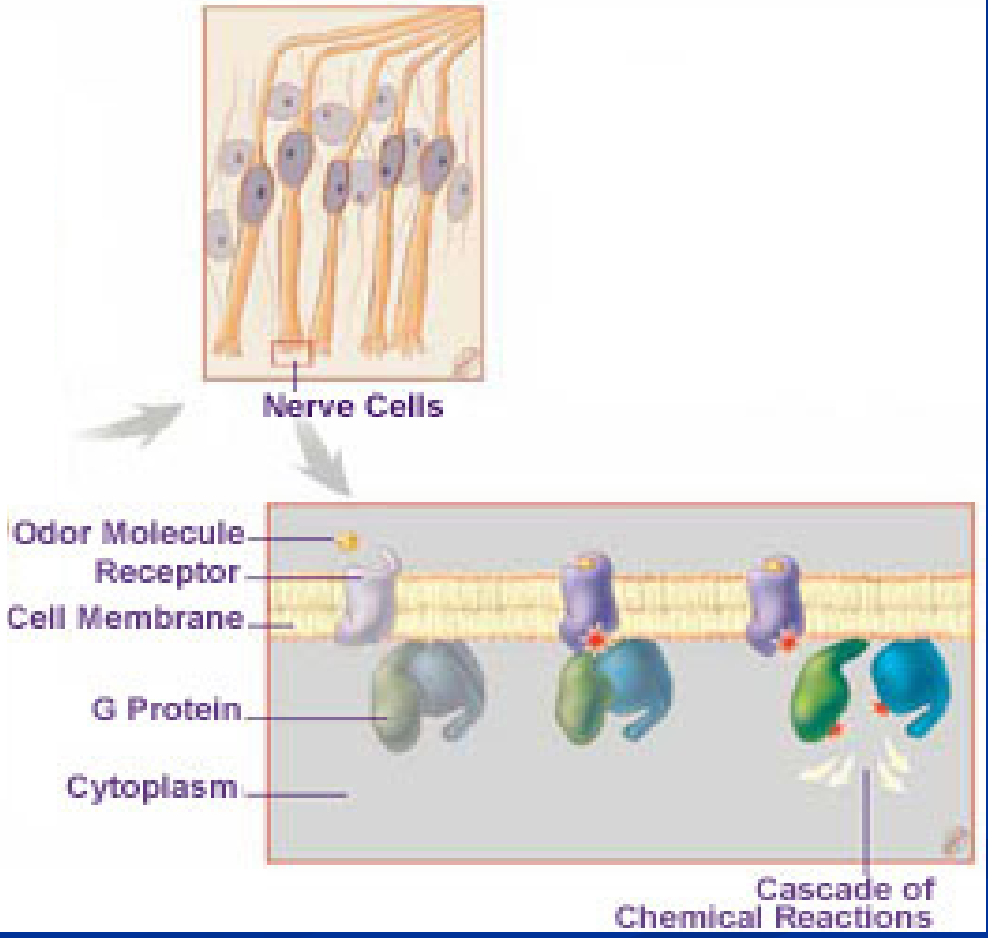
Large-scale deletions and duplications commonly differentiate human genomes

Copy number variants (CNVs):

- Deletions or duplications of 0.1-1000 kb genomic sequences without causing disease.
- Occurrence at 1 per 30 Mb.
- Can be detected by microarray analysis.



- Quite common, over 6000 CNV loci have been identified.
- One example is the olfactory receptor (OR) gene family.
  - Copy number of OR genes (less than 1000) varies in humans.



The variation in copy number among 10 people at 11 representative *OR* loci.

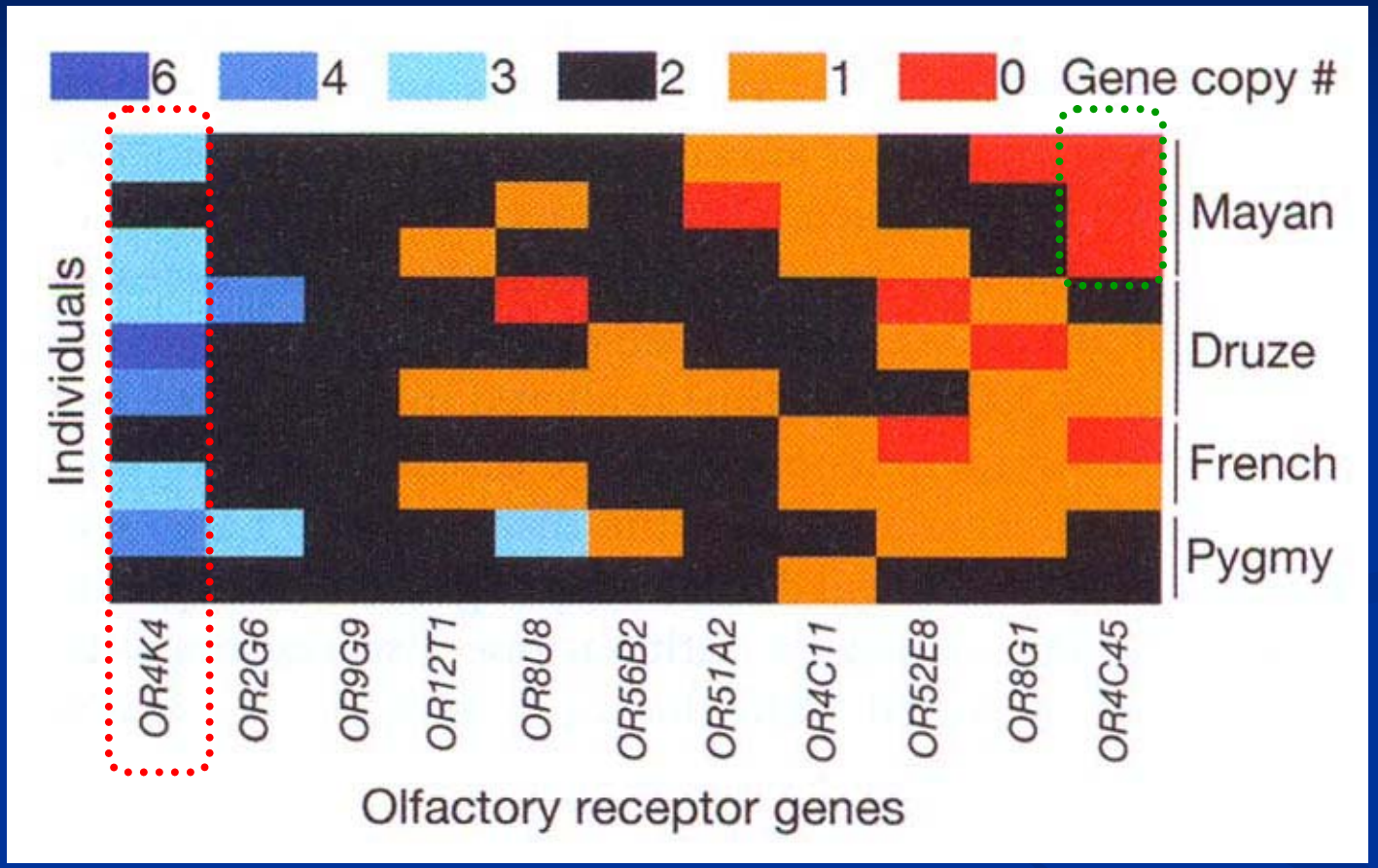
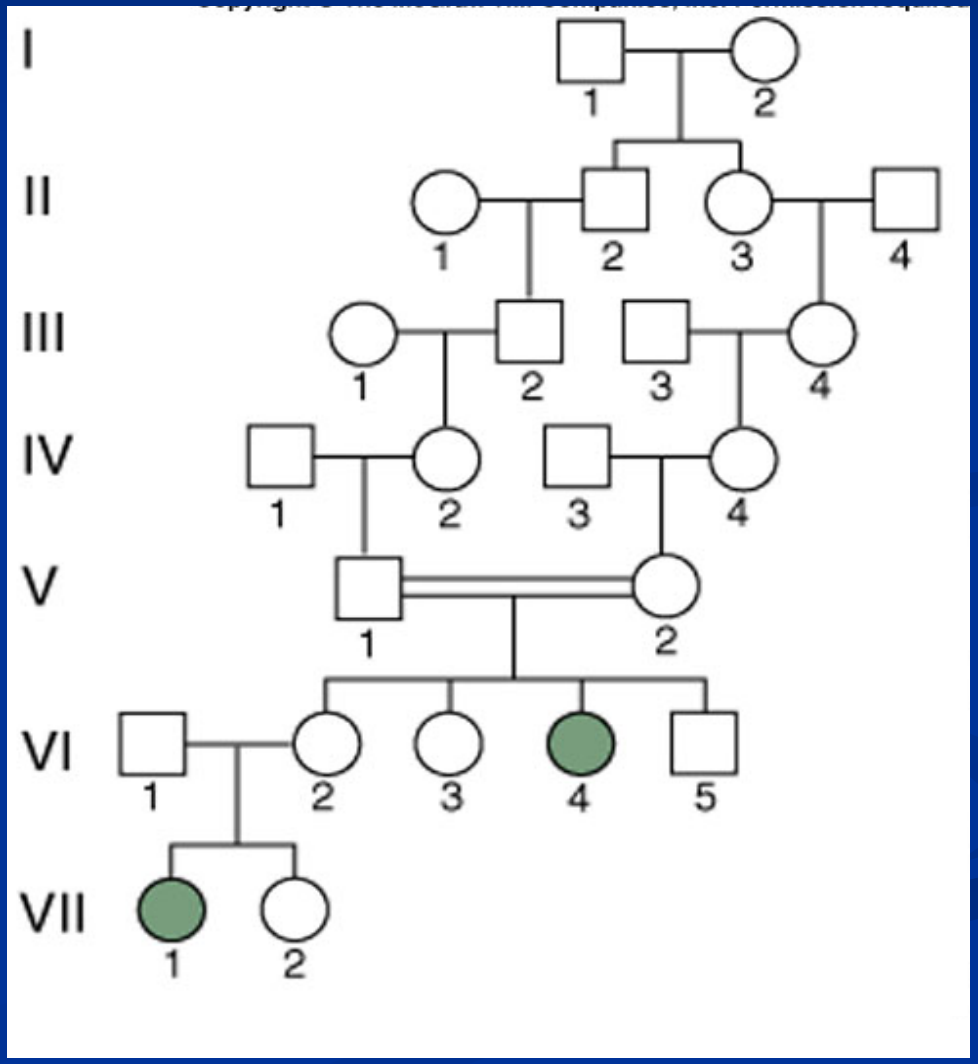


Fig. 10.8

10.4 Positional cloning – Use of polymorphic DNA markers to clone disease-causing genes

Positional cloning – Step 1
Correlating phenotypic transmission with one area of the genome

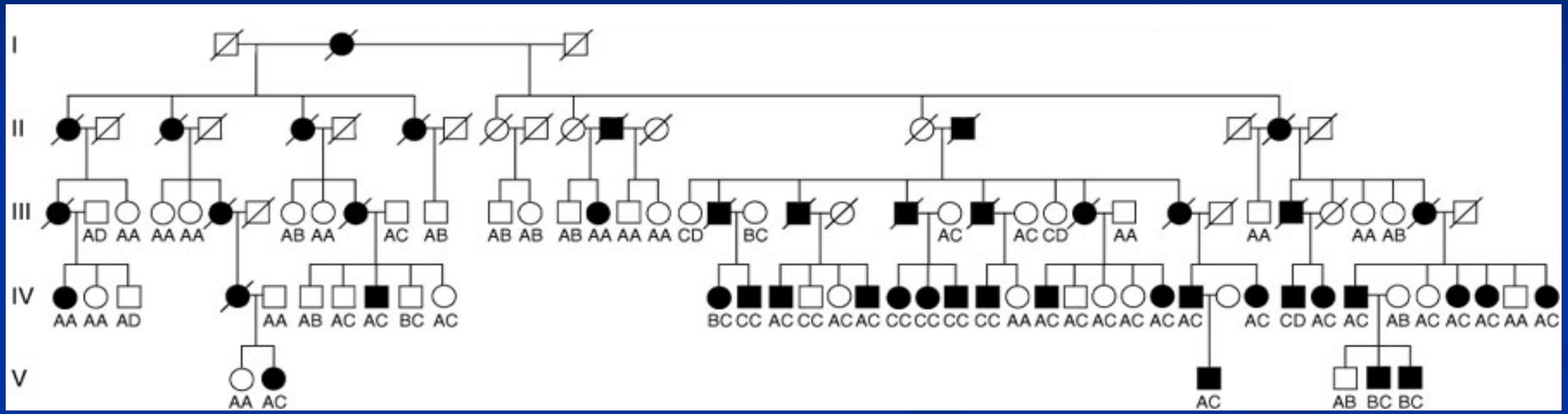
- **Find extended families in which disease is segregating.**



A pedigree of cystic fibrosis disease

Fig.2.22a

Portion of a Venezuelan pedigree affected by Huntington disease



HD **locus is within 5 cM of the G8 DNA marker (alleles indicated by A, B, C, D)**

- Use panel of polymorphic markers spaced at 10 cM intervals across all chromosomes.
  - 300 markers total (human genome 3000 cM).
- Determine genotype for all individuals in families for each DNA marker.
- Look for linkage between a marker and disease phenotype.

In humans, 1 cM = 1000 kb

- **Once region of chromosome is identified, a high resolution mapping is performed with additional markers to narrow down region where gene may lie.**

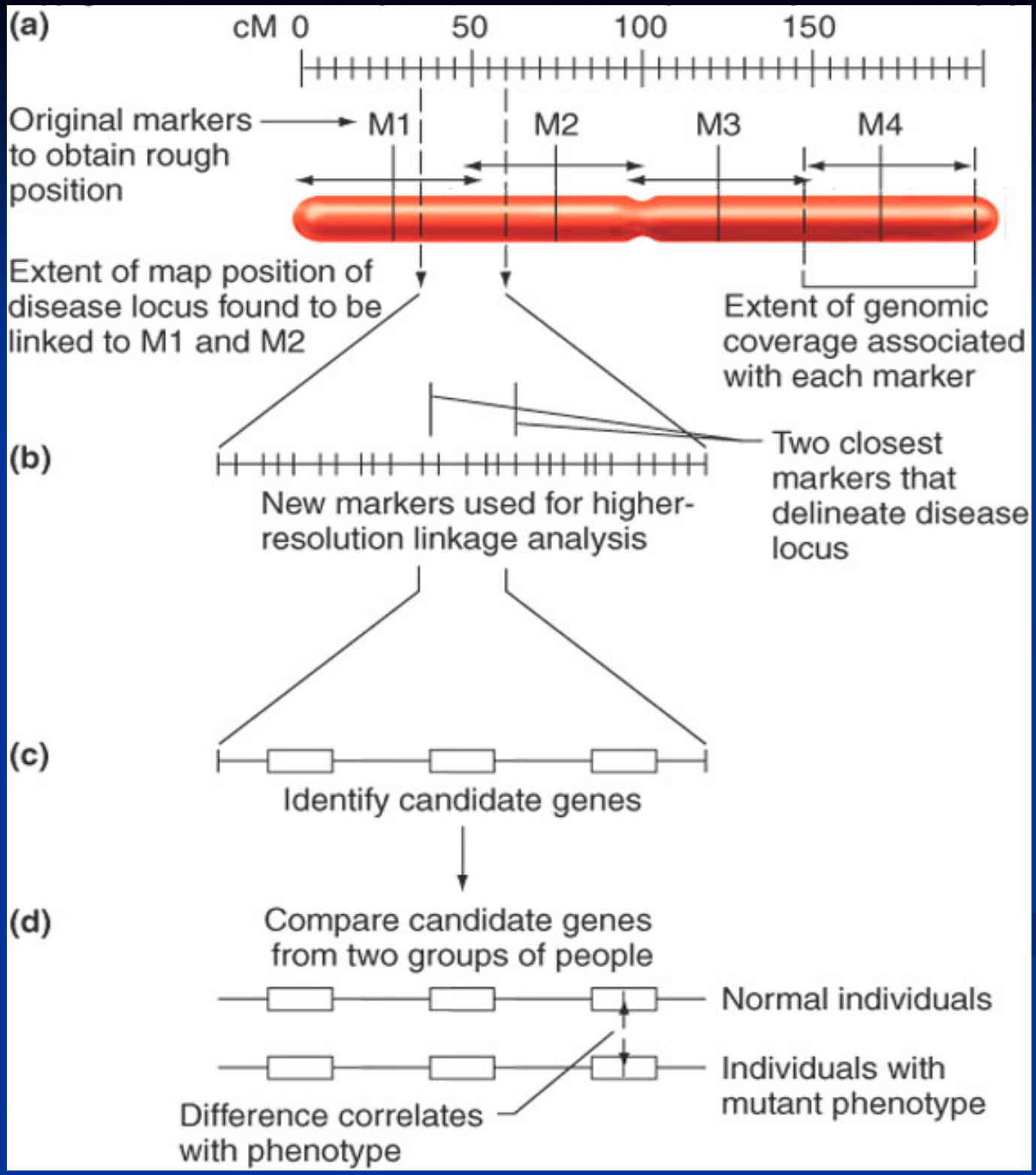
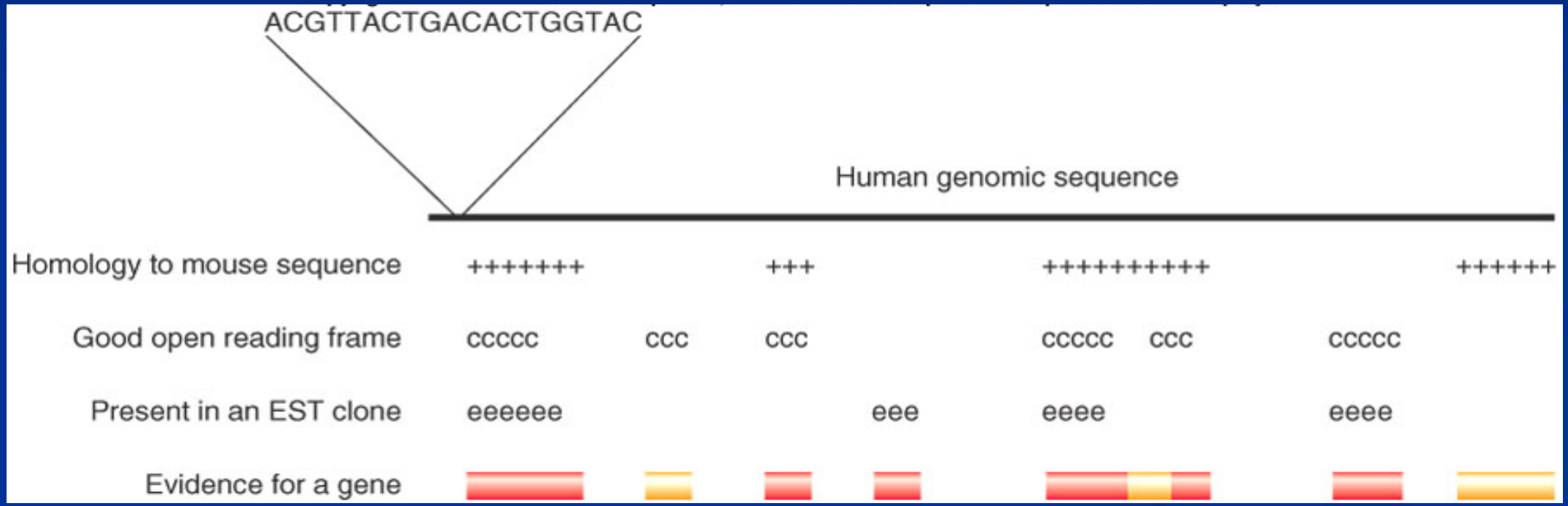


Fig. 11.17

Positional cloning – Step 2
Identifying candidate genes

- Region of chromosome is narrowed down by linkage analysis to 1000 kb or less. All genes within are identified.
- Identify 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 on one or more EST clones derived from cDNA

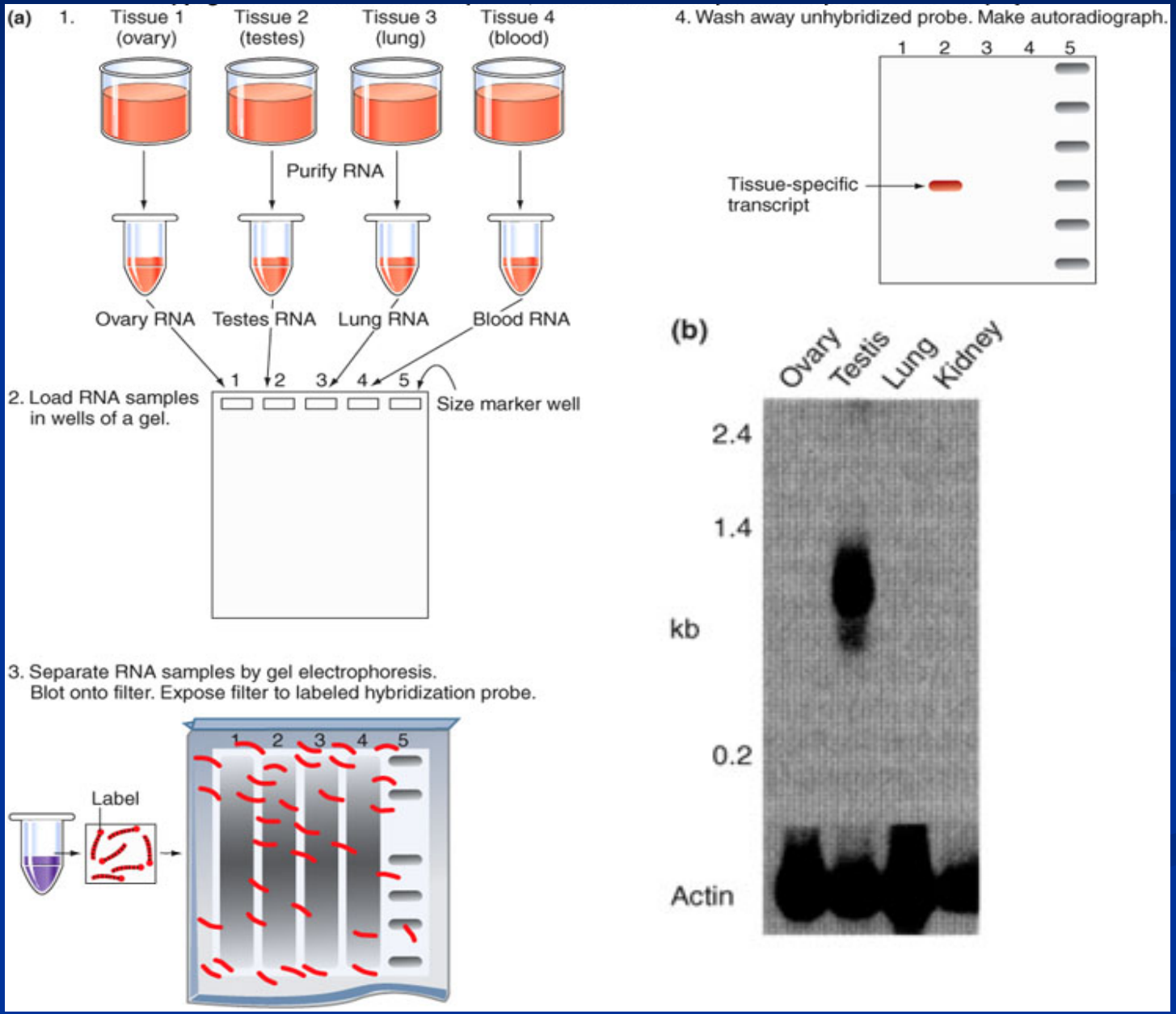
Computational analysis of genomic sequences to identify candidate genes



Positional cloning – Step 3
Finding the gene responsible for the phenotype

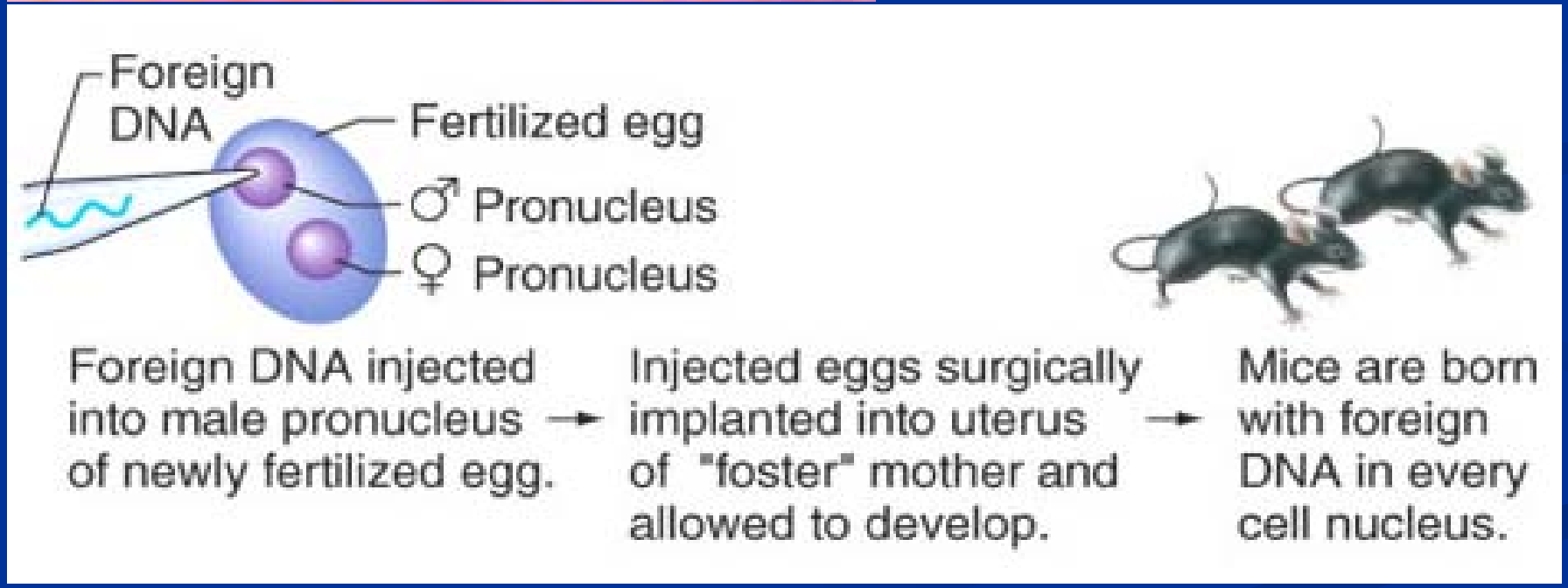
- Gene expression patterns can help to pinpoint candidate genes
  - Look in public database of EST sequences representing certain tissues.
  - RNA expression assayed by Northern blot or PCR amplification of cDNA with primers specific to candidate transcript
  - Look for misexpression (no expression, underexpression, overexpression).

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



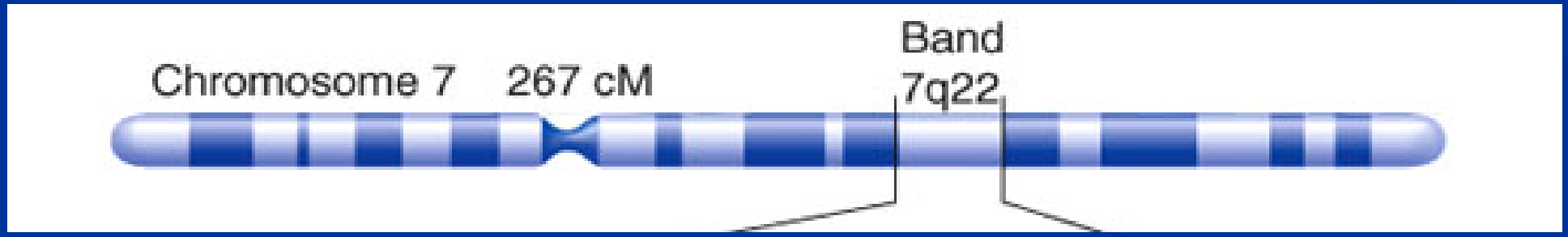
- **Sequence differences**
  - **Missense mutations identified by sequencing coding region of candidate gene from normal and abnormal individuals**
- **Transgenic modification of phenotype**
  - **Insert the mutant gene into a model organism.**

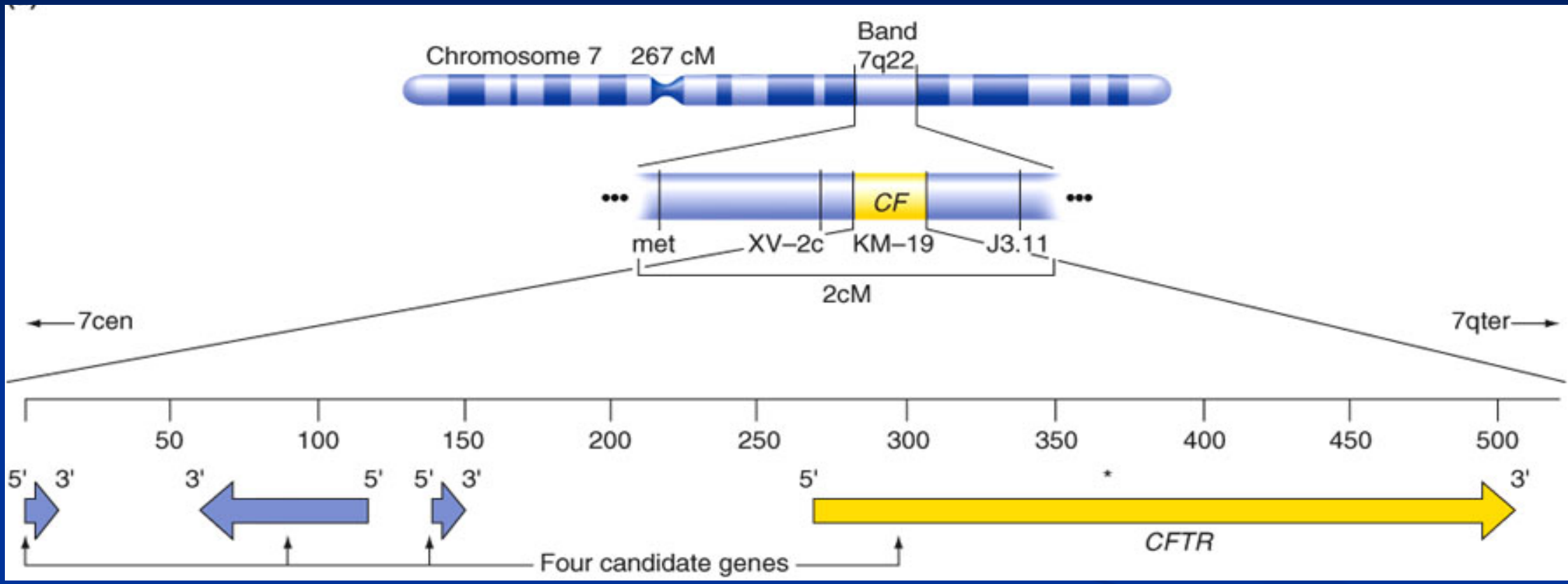
Transgenic analysis can prove the function of a candidate gene *SRY* as the testes-determining factor (*TDF*)



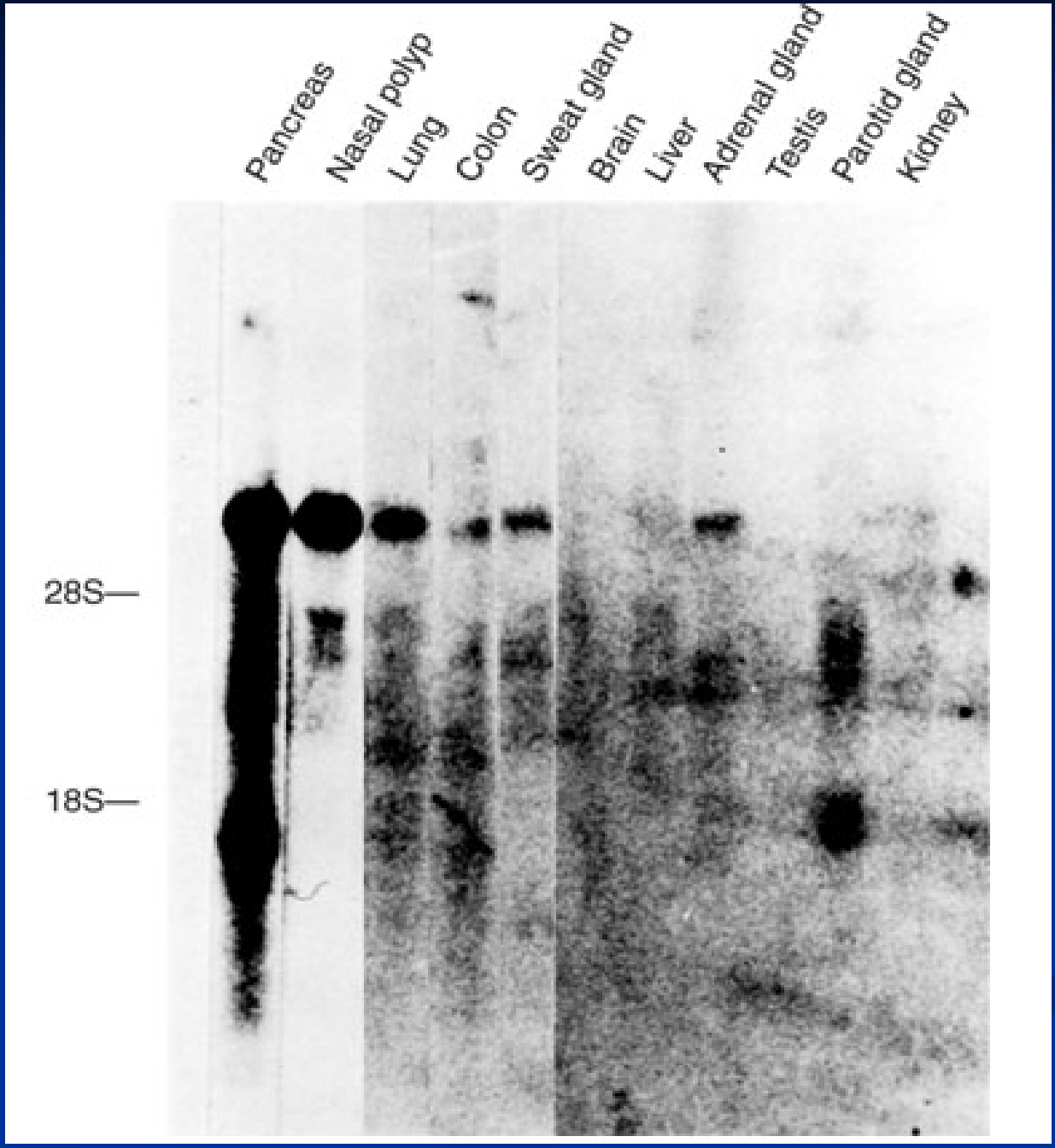
Example: Positional cloning of cystic fibrosis gene

- Afflicted children have a variety of symptoms arising from abnormally viscous secretions in the lungs, pancreas, and sweat glands.
- Linkage analysis places CF on chromosome 7

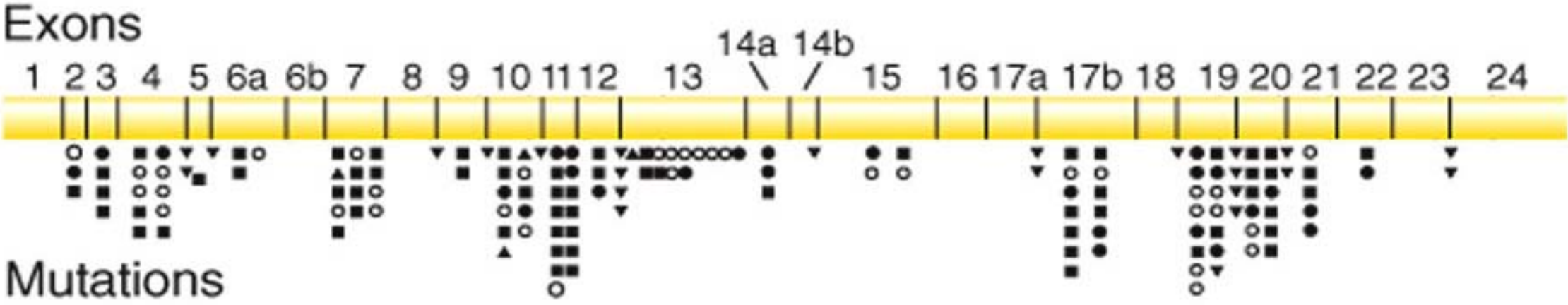



Chromosome 7
267 cM
Band
7q22
CF
met
XV–2c
KM–19
J3.11
2cM
7cen
7qter
50
100
150
200
250
300
350
400
450
500
5'
3'
3'
5'
5'
3'
5'
3'
CFTR
Four candidate genes

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



Every CF patient has a mutated allele of the CFTR gene on both chromosome 7



Location and number of mutations indicated under diagram of chromosome

Fig. 10.23

Proving CFTR is the right gene

- Look for phenotype when gene function is eliminated.
  - Cannot use transgenic technology
- Instead perform CFTR gene "knockout" in mouse to examine phenotype without CFTR.
  - Targeted mutagenesis
    - Introduce mutant CFTR into mouse embryonic cells in culture.
    - Rare double recombinant events with homologous wild-type CFTR gene are selected for.
    - Mutant cell is introduced into normal mouse embryos where they incorporate into germ line.
    - Knockout mouse created
    - Offspring of knockout mouse exhibits the disease symptoms.

CFTR is a membrane protein that regulates passage of chloride ions into and out of cells