

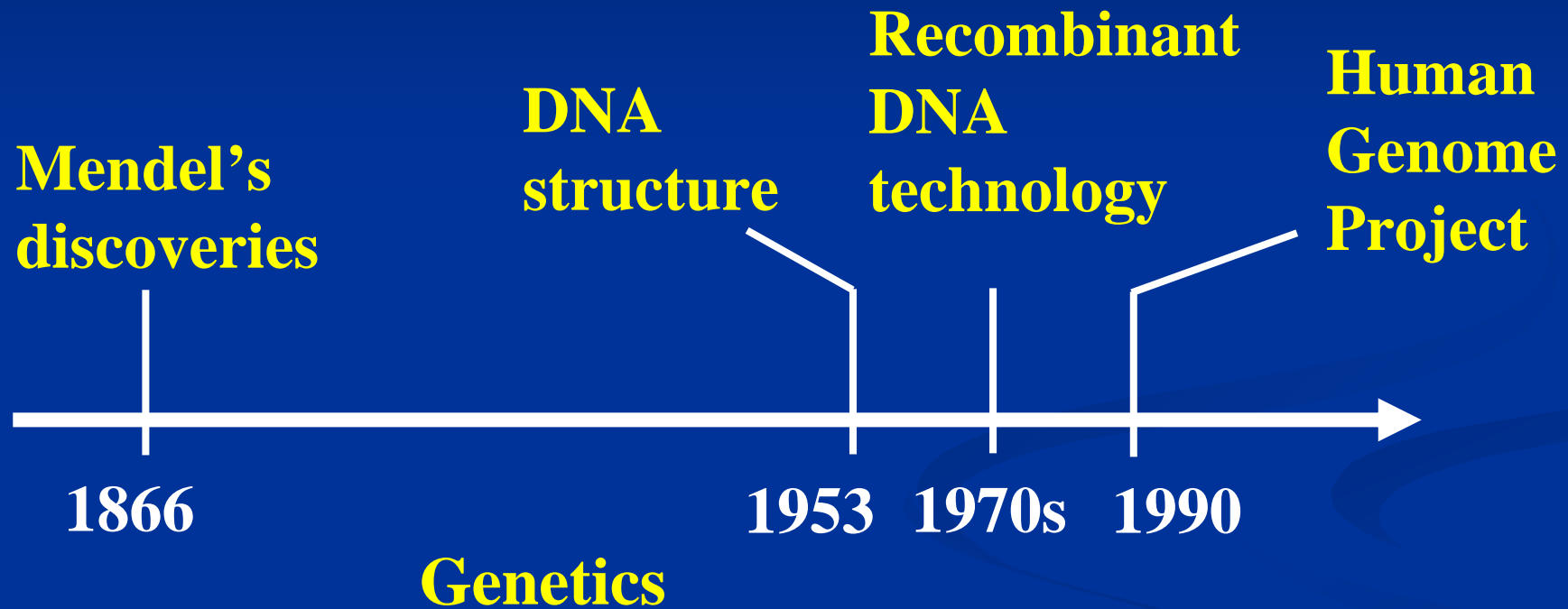
# Chapter 9B

## Genomes and Proteomes

## **Sections to study**

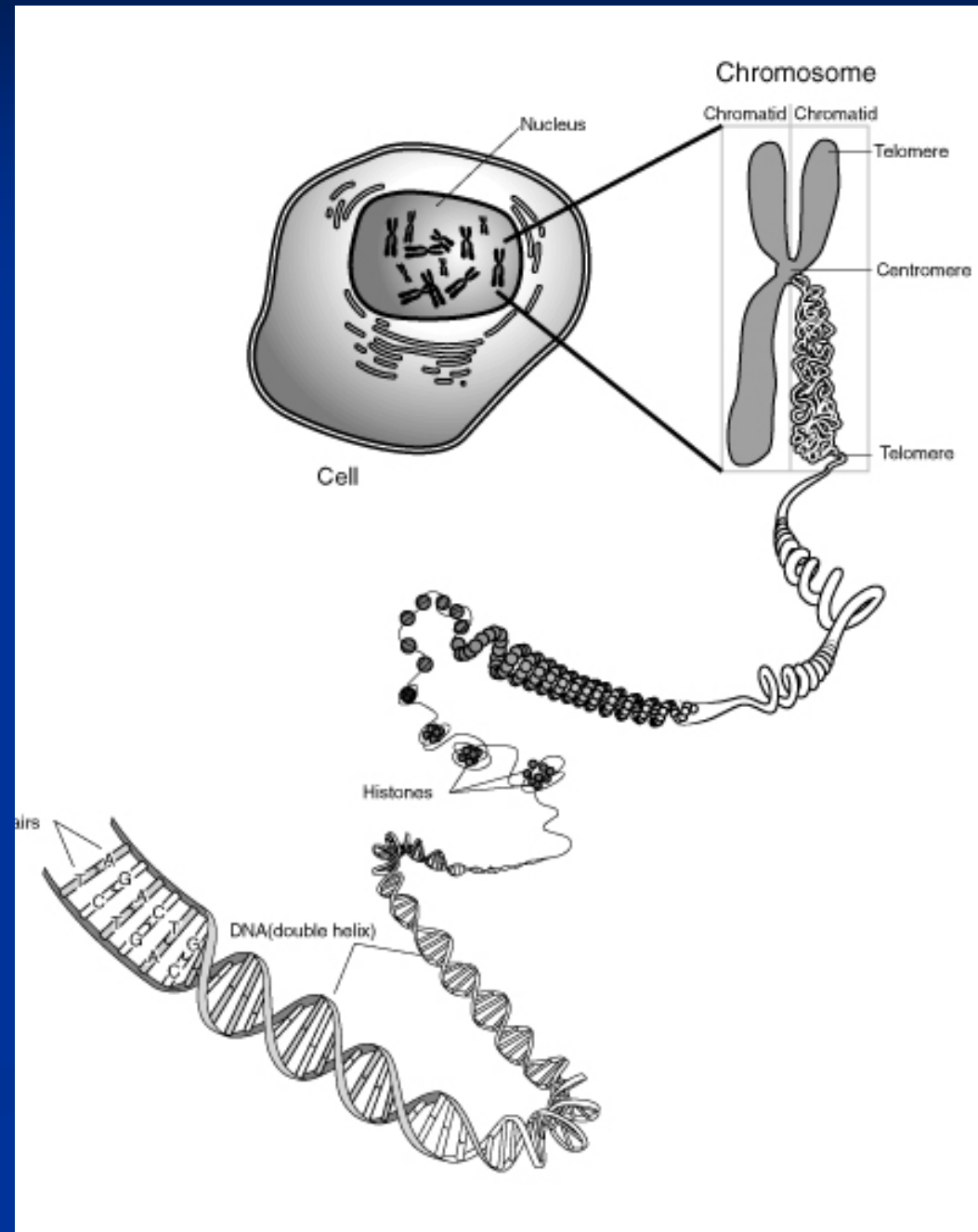
- 1 Large-scale genome mapping and analysis**
- 2 Major insights from the human and model organism genome sequences**
- 3 Global analysis of genes and their mRNAs**
- 4 Global analysis of proteomes**
- 5 Repercussions of the human genome project and high-throughput technology**

# The major advances that have transformed genetics



# 1 Large-scale genome mapping and analysis

- **Genome:** The entire genetic information in a particular cell or organism.
- **Genomics:** The study of whole genomes. It dedicated to the development and application of more effective mapping, sequencing, and computational tools.



# Techniques for genome mapping and sequencing

## ■ Cloning

- Library of DNA fragments 500 – 1,000,000 bp
- Insert into one vector

## ■ Hybridization

- Location of a particular DNA sequence within a library of fragments

## ■ PCR amplification

- Direct amplification of a particular region of DNA ranging from 100 bp to > 20kb

## ■ DNA sequencing

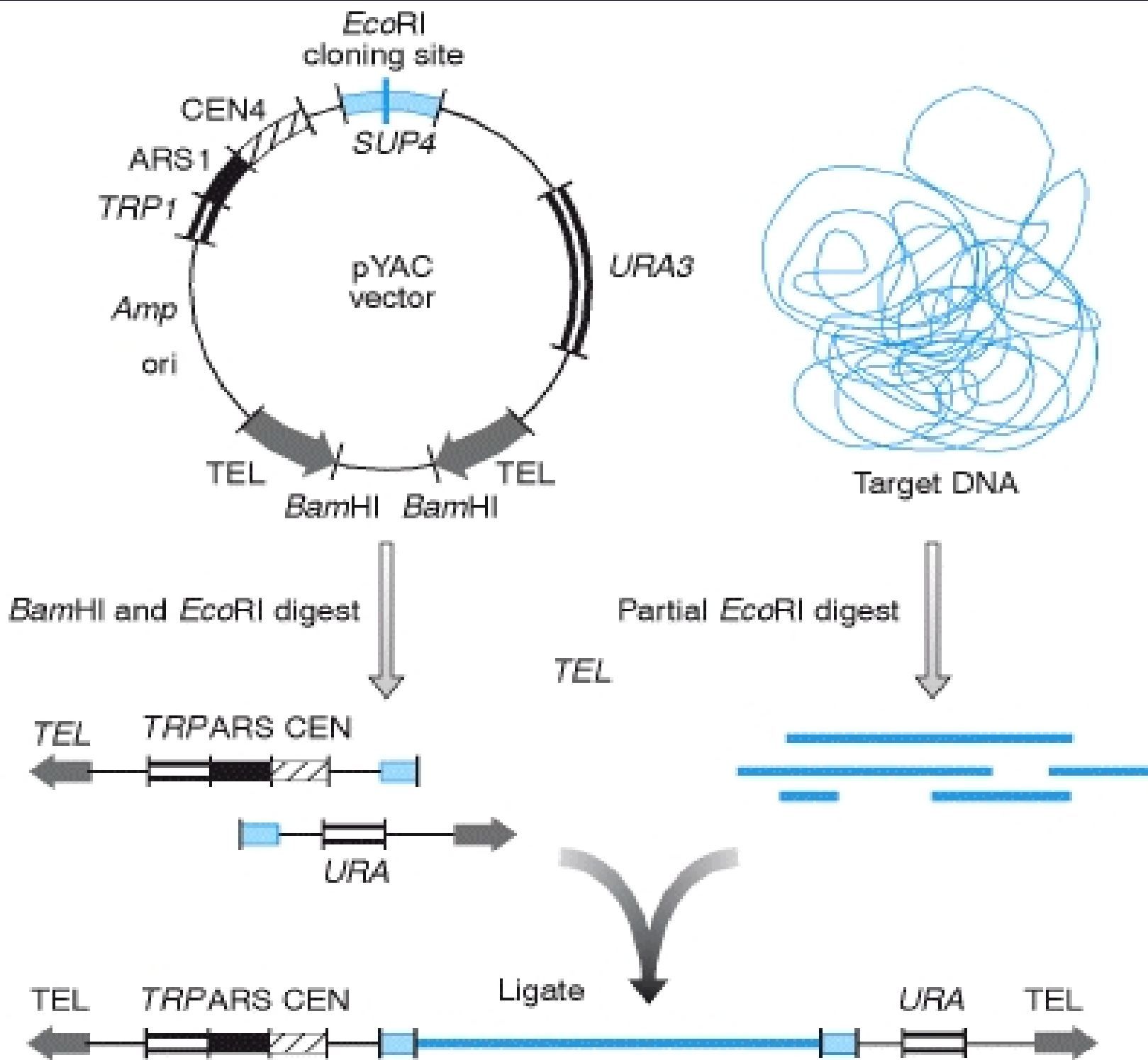
- Automated DNA sequencer using Sanger method determines sequences 1000 bp at a time.

## ■ Computational tools

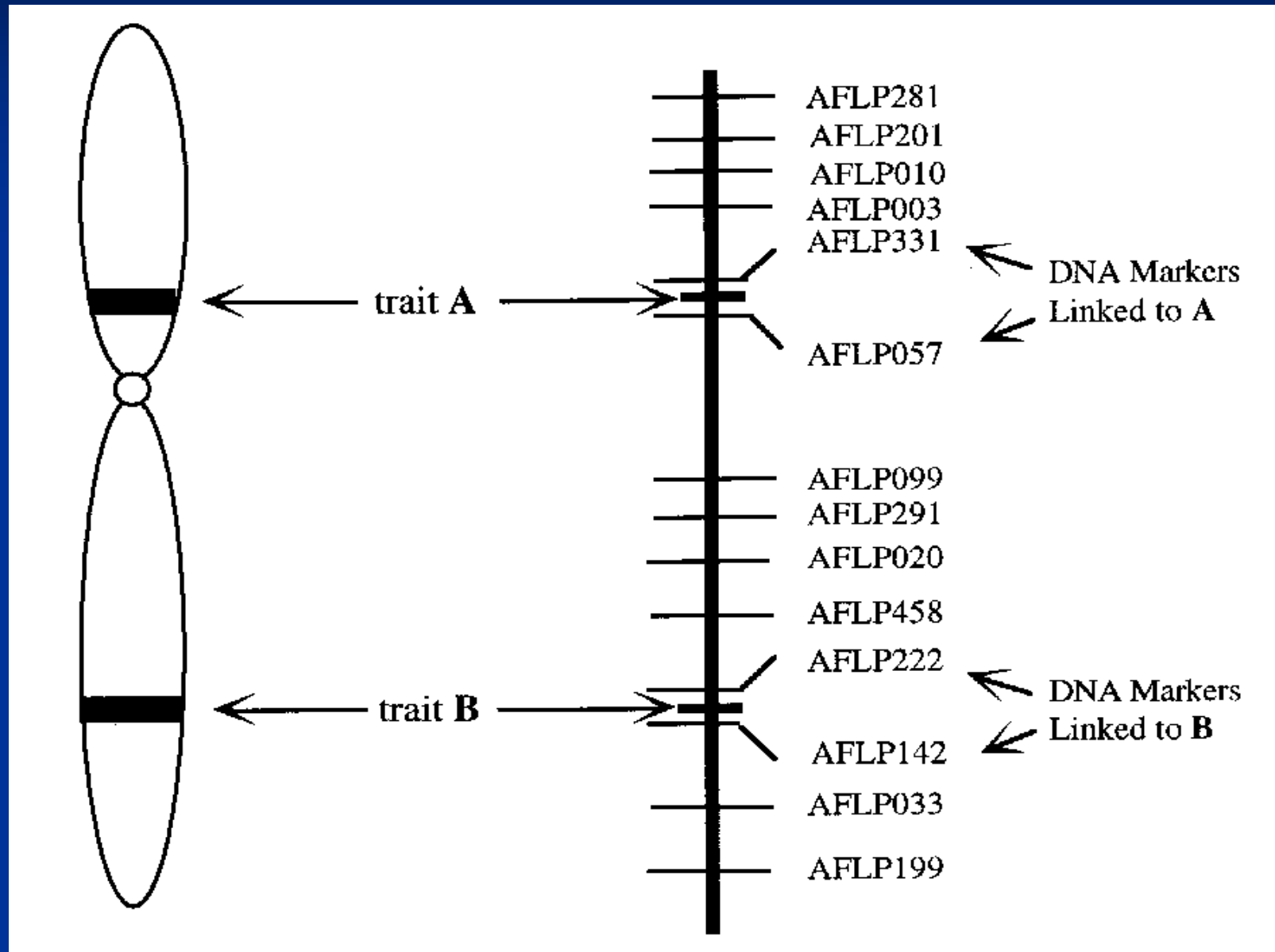
- Programs for identifying matches between a particular sequence and a large population of previously sequenced fragments
- Programs for identifying overlaps of DNA fragments
- Programs for estimating error rates
- Programs for identifying genes in chromosomal sequences

# Vectors used for cloning large inserts for physical mapping

- **YACs** (yeast artificial chromosomes)
  - Insert size 250-2,000 kb
- **BACs** (bacterial artificial chromosomes)
  - Insert size 100 – 300 kb
  - More stable and easier to purify from host DNA than YACs



# Large-scale *linkage maps* serve as guides to whole genomes

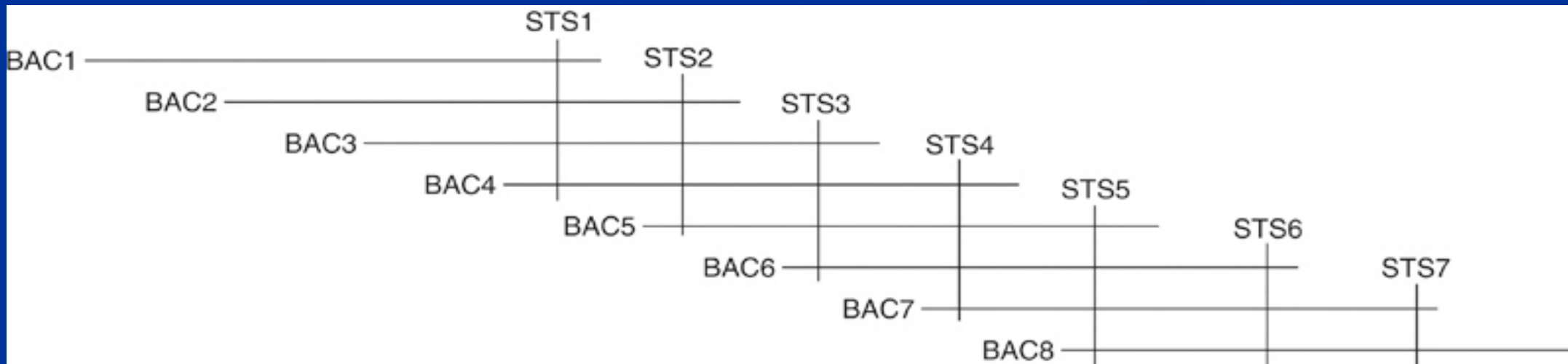




## Long-range *physical maps*

**Physical maps:** a constellation of overlapping DNA fragments that are ordered and oriented and span each of the chromosomes in a genome.

- **STS** (Sequence tagged sites): Short DNA sequences that are already located and amplified by a pair of PCR primers. Each STS represents a unique DNA sequence in the genome.



# Locating cloned genes or markers on the chromosome of a karyotype

- a) Complete set of human chromosomes stained with **Giemsa dye** shows bands.
- b) **Ideograms (染色体G带图)** show idealized banding pattern.

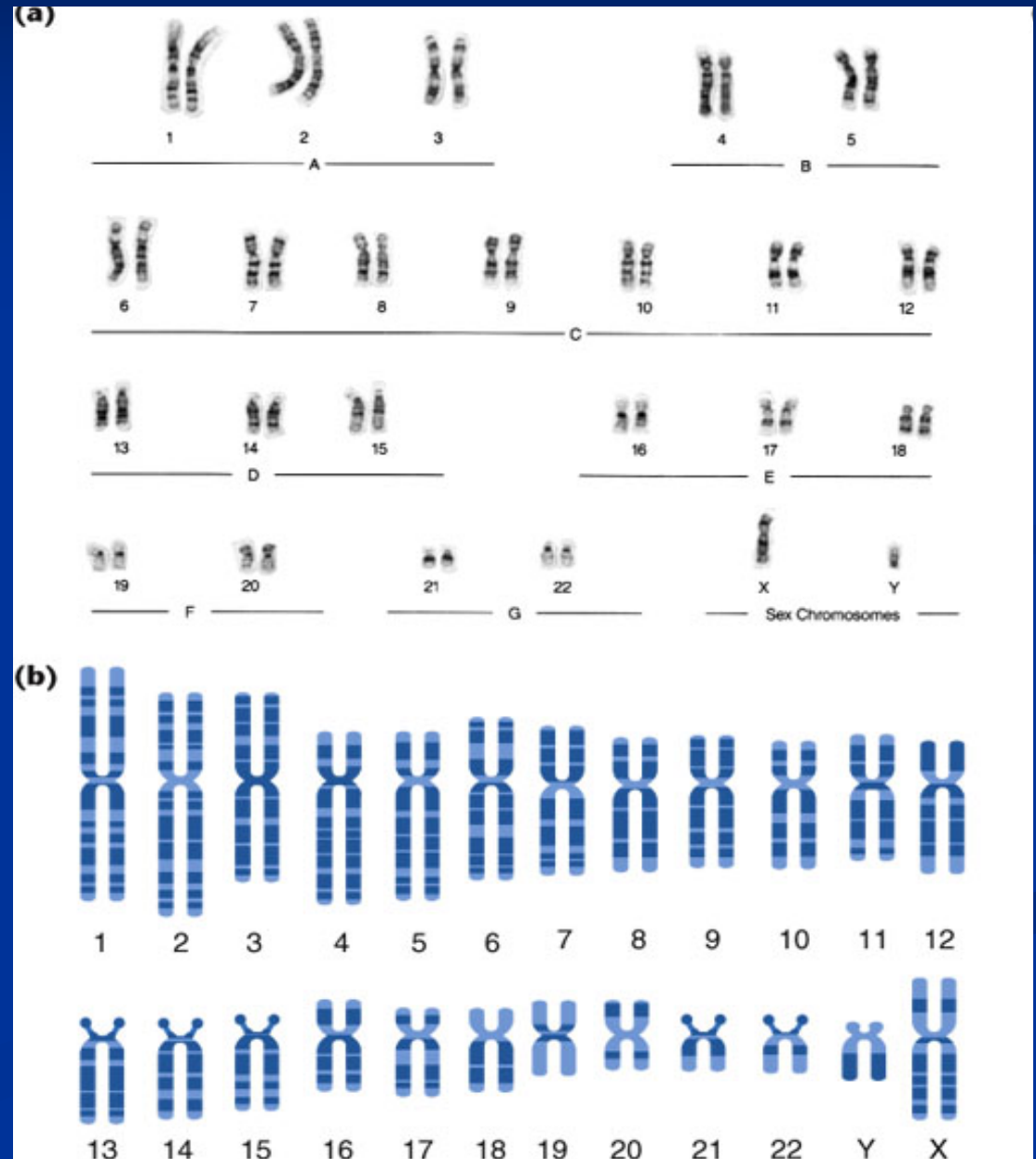


Fig. 10.2 a, b

■ **Chromosome 7 at three different levels of banding resolution.**

p: shorter arm  
q: longer arm

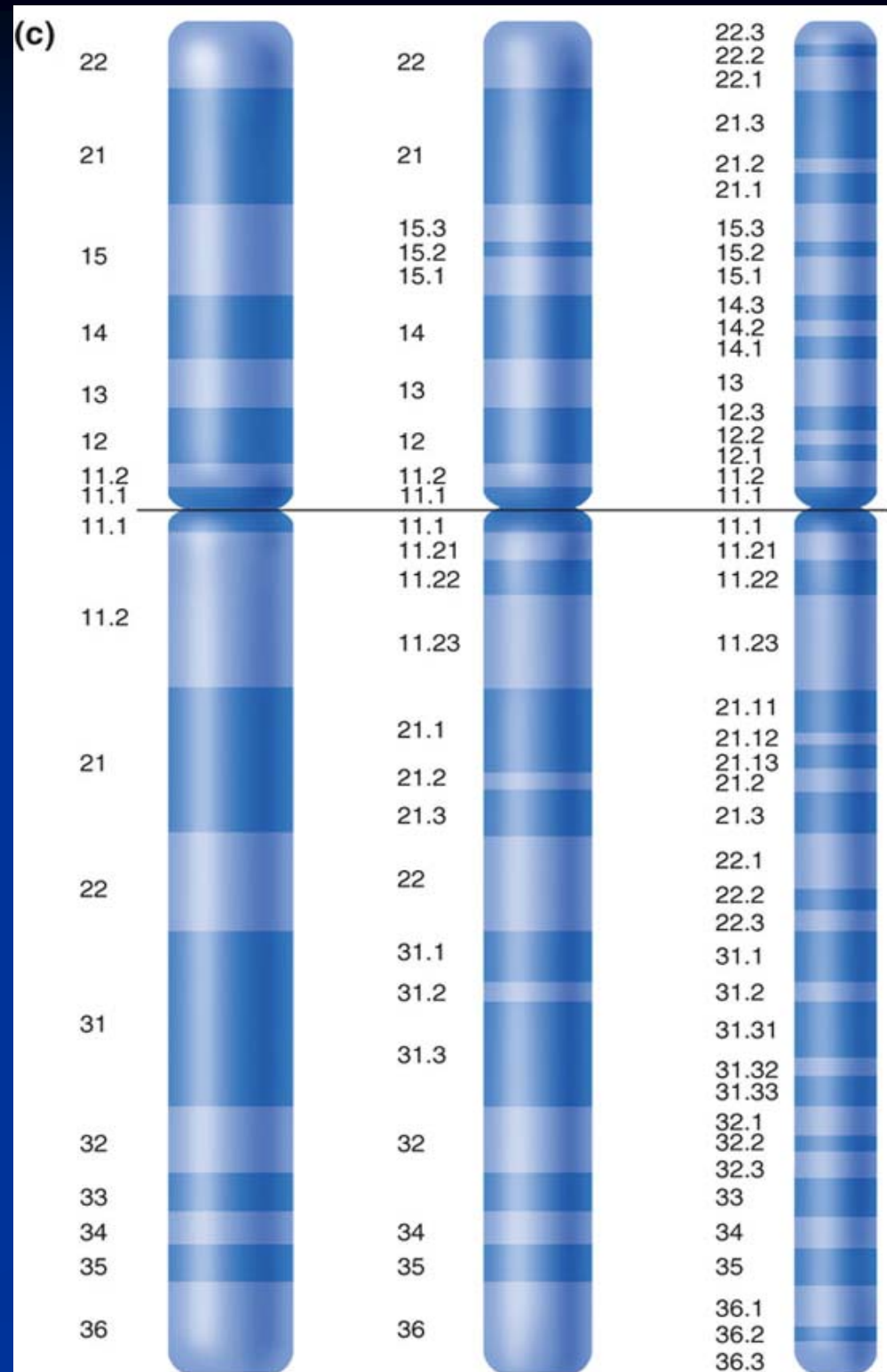
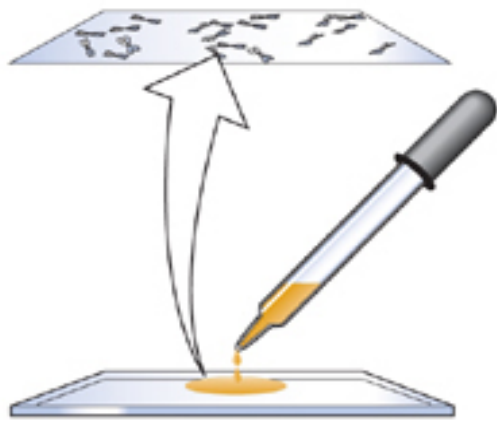


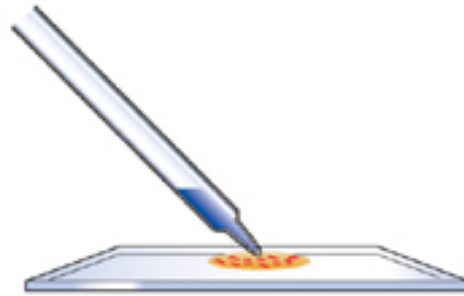
Fig. 10.2 c

# Fluorescent *in situ* hybridization (FISH)

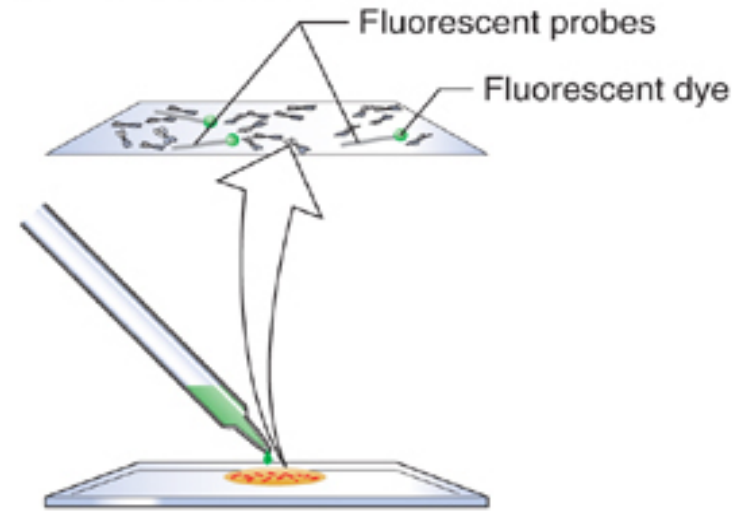
(a)



1. Drop cells onto a glass slide.

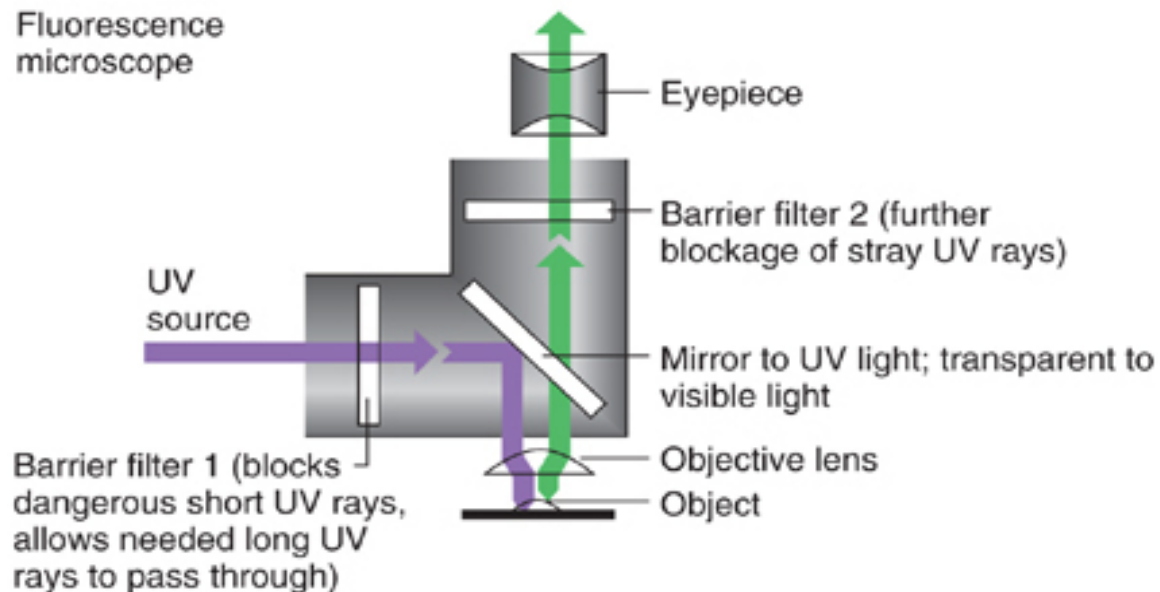


2. Gently denature DNA by treating briefly with DNase.



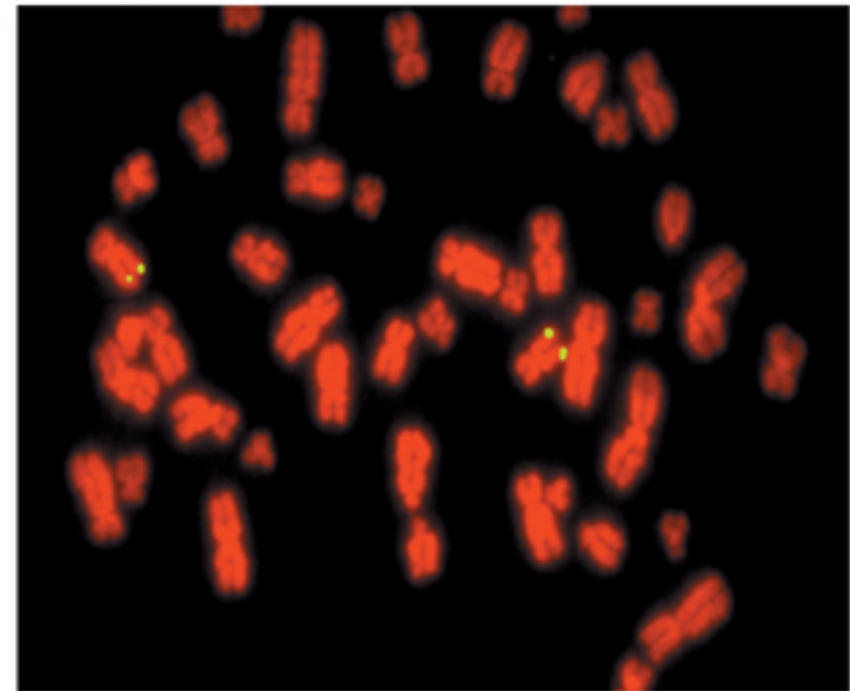
3. Add hybridization probes labeled with fluorescent dye and wash away unhybridized probe.

Fluorescence microscope



4. Expose to ultraviolet (UV) light.  
Take picture of fluorescent chromosomes.

(b)



# The Human Genome Project

- Human genome:  $3 \times 10^9$  bp.
- Proposed in 1985.
- Started in 1990, 15-year time scale, \$ 3 billion dollars U.S. government funding.

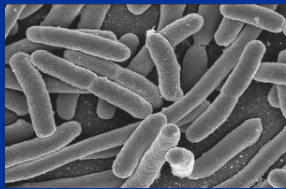




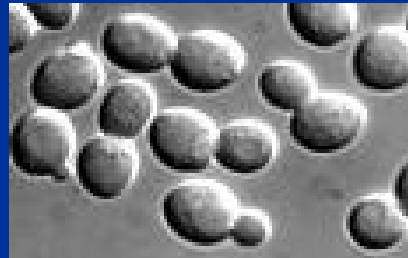
## Six objectives of HGP

1. To generate genetic, physical, and sequence maps of the human genome.
2. To sequence the genomes of **model organisms**.
3. To develop improved technologies for mapping and sequencing.
4. To develop computational tools for capturing, storing, analyzing, displaying, and distributing map and sequence information.
5. To sequence **ESTs** and eventually full-length cDNAs expressed in different cell types of humans and mice.
6. To consider the ethical, legal, and social challenges posed by genomic information.

- **Model organisms:** Organisms used in genomic analysis because they have many genetic mechanisms and cellular pathways in common with each other and with humans. These organisms lend themselves well to classical breeding experiments and direct manipulation of the genome.



Bacteria



Yeast



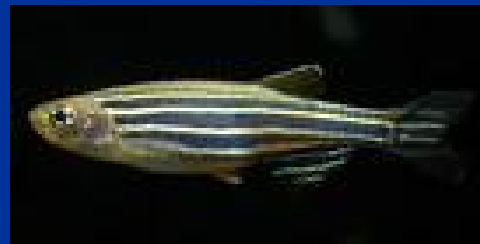
Roundworm



Fruit fly



Plant



Zebrafish

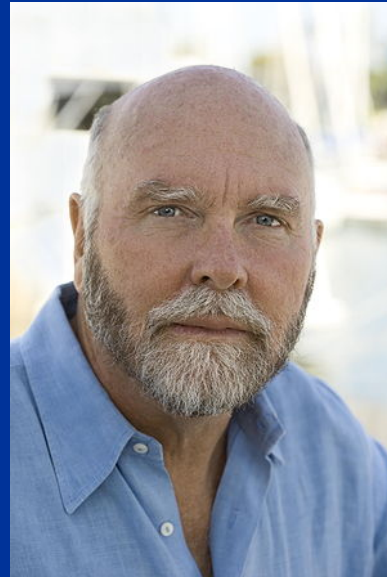


Mouse

- In February 2001, a rough sequence draft was completed. Error rate 1/10,000 with some gaps.
  - **TIGR**, a U.S. government-funded organization
  - **Celera**, a private company
- In 2003, 97% coverage was completed.
- In 2006, 99% coverage was completed.



**Francis Collins**



**J. Craig Venter**





- **By 2002, the genomes of 8 model organisms were sequenced.**
- **By mid-2005, more than 50 eukaryotic and 230 prokaryotic genomes had been sequenced.**

**TABLE 10.1****Number of Species with Finished Whole-Genome Sequences Deposited at the National Center for Biotechnology Information as of February 1, 2010**

<b>Organism</b>	<b>Whole Genome</b>	<b>In Progress</b>	<b>Total</b>
Prokaryotes	1058	2354	3412
Mammals	56	69	125
Birds	2	12	14
Fishes	15	17	32
Insects	36	7	43
Flatworms	3	2	5
Roundworms	12	14	26
Amphibians	1	0	1
Reptiles	1	0	1
Other animals	11	18	29
Plants	25	88	113
Fungi	129	91	220
Protists	54	58	112
Total	1403	2730	4133



**TABLE 10.2****A Comparison of the Developmental Complexity and Genome Features\* of Model Organisms**

Type	Organism Species	Developmental Complexity	Genome Size* (Megabases)	Number of Genes	# of Genes per Million bp Sequenced	Date Genome Finished
Bacterium	<i>Escherichia coli</i>	1-cell prokaryote	4.64	3244	905	1997
Yeast	<i>Saccharomyces cerevisiae</i>	1-cell eukaryote	12.07	6201	483	1996
Worm	<i>Caenorhabditis elegans</i>	≈1000 cells	100	21,188	197	1998
Fly	<i>Drosophila melanogaster</i>	≈50,000 cells	180	22,606	117	2000
Mustard weed	<i>Arabidopsis thaliana</i>	10 <sup>10</sup> cells	125	33,500	221	2000
Rice (draft)	<i>Oryza sativa</i>	5 × 10 <sup>10</sup> cells	466/420	29,437	127–155/82–128	2002
Mouse	<i>Mus musculus</i>	10 <sup>11</sup> cells	3200	61,818	10–13	2005
Human	<i>Homo sapiens</i>	10 <sup>14</sup> cells	3200	45,416	18	2003

\*Haploid genome size, including heterochromatic DNA.

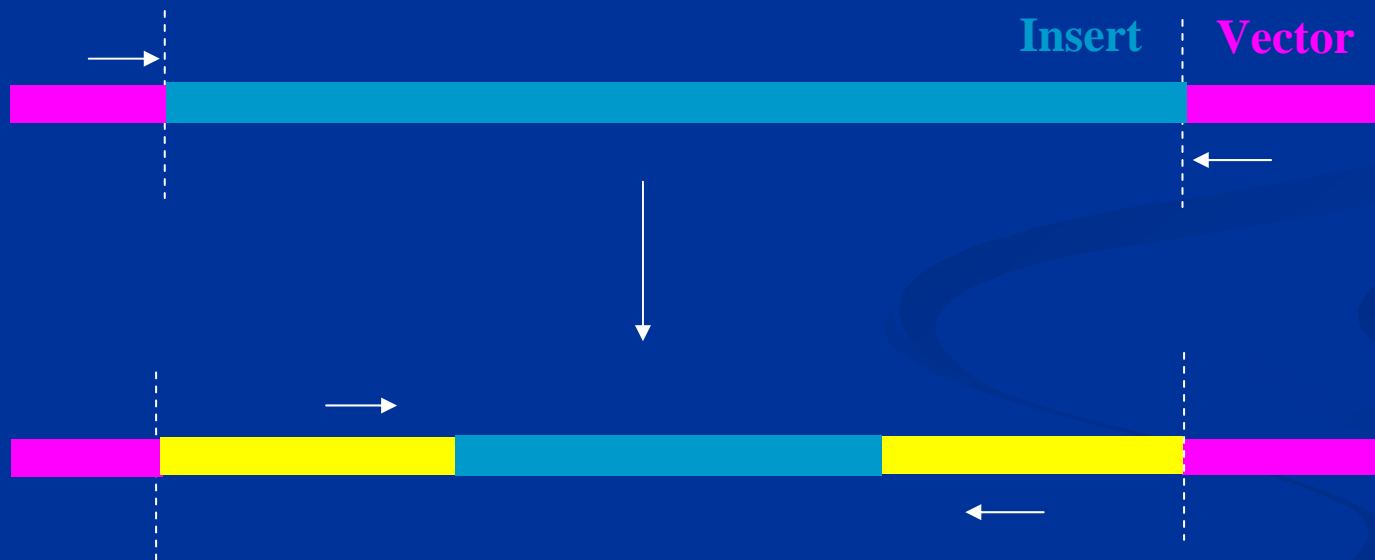
**Note:** For the sequenced genomes of model organisms, gene numbers are taken from the original sequence publications; most numbers have since changed slightly. For rice, two different strains have been sequenced: *Oryza sativa* L. ssp. *Japonica* and *Oryza sativa* L. ssp. *Indica*.

# Sequencing long regions of DNA

- **Primer walking**
- **Shotgun sequencing**

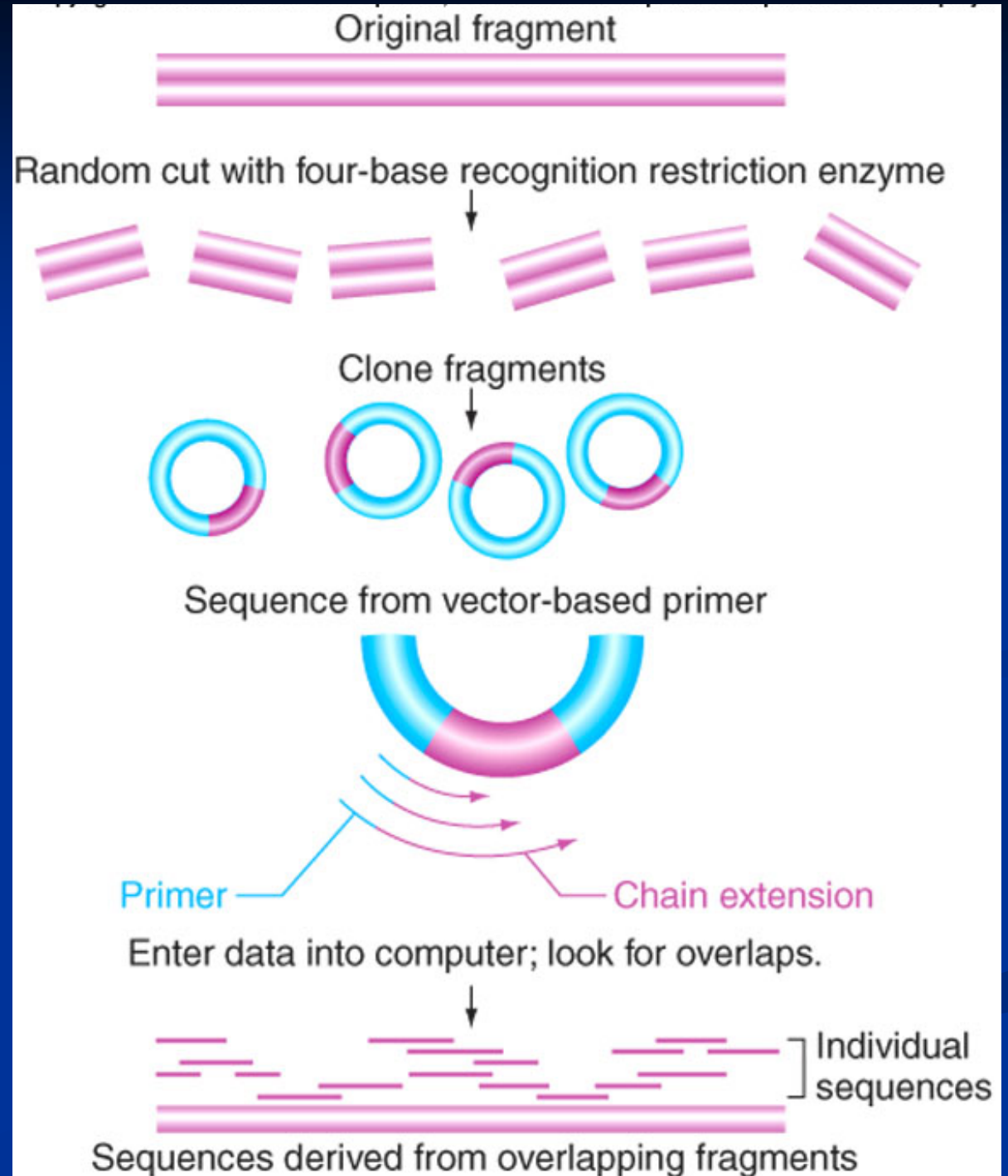
## ■ Primer walking

- Sequence starting from both ends of cloned insert
- New primers derived from sequence obtained in previous round



## ■ Shotgun sequencing

- Long DNA sequence chopped into many small fragments which are cloned individually
- Sequence of all small fragments determined
- Small fragments aligned by computer to generate one long continuous sequence



## ■ **Primer walking**

- **Slower, better for short DNA. Must make primers after each round of sequence.**
- **No redundancy required. No alignment necessary.**
- **Works well in laboratories with few automated sequencers.**

## ■ **Shotgun sequencing**

- **Very fast, better for long DNA.**
- **Relies on redundancy. Must gather sequence information on 3-4 times the actual number of base pairs from the original clone for full coverage.**
- **Sometimes must fill in gaps with primer walking.**
- **Requires many automated sequencers.**



## Two strategies to sequence the human genome:

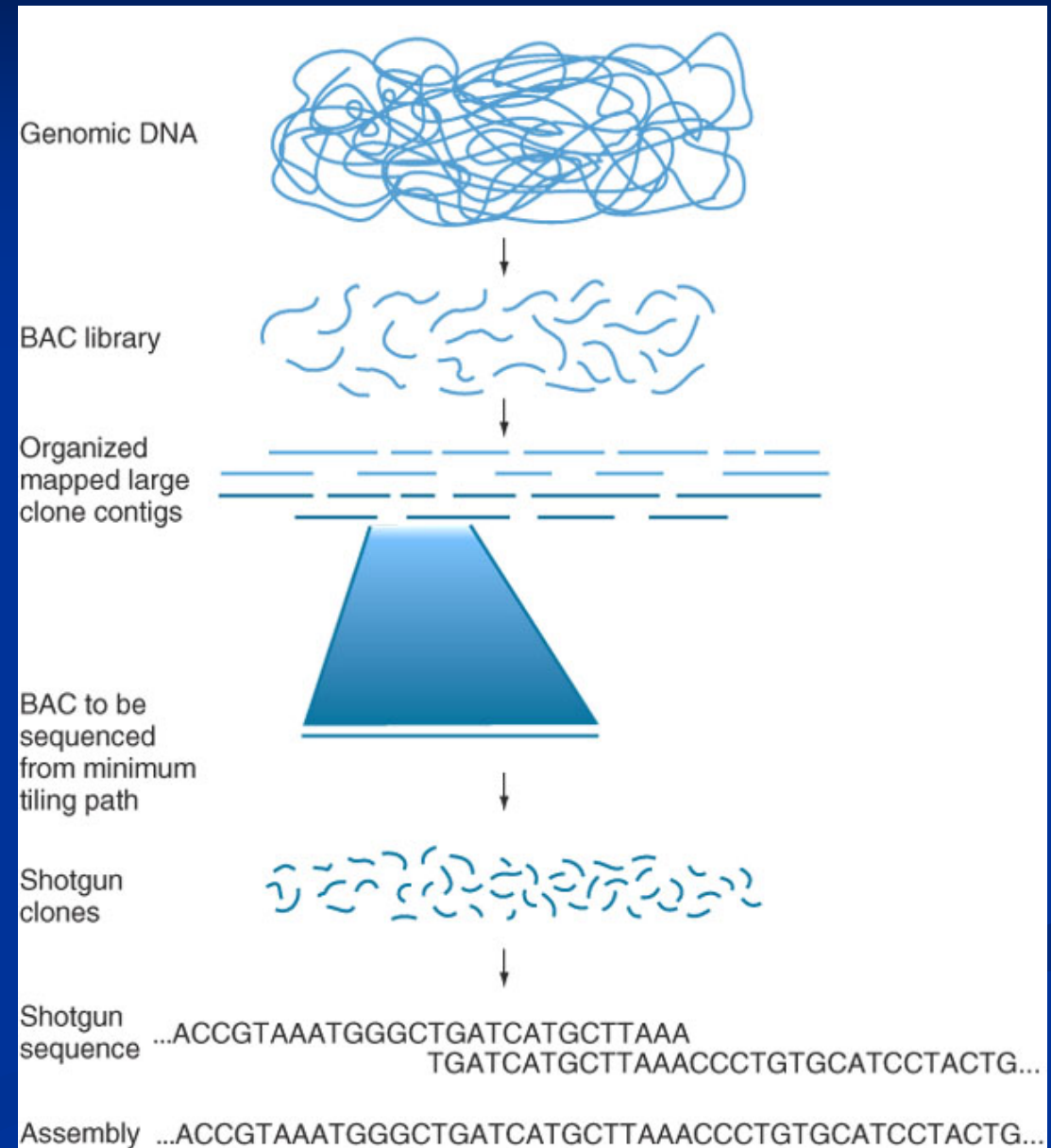
- Hierarchical shotgun (分级鸟枪法)
- Whole-genome shotgun (全基因组鸟枪法)



# Hierarchical shotgun strategy (分级鸟枪法)

Used in publicly funded effort to sequence human genome

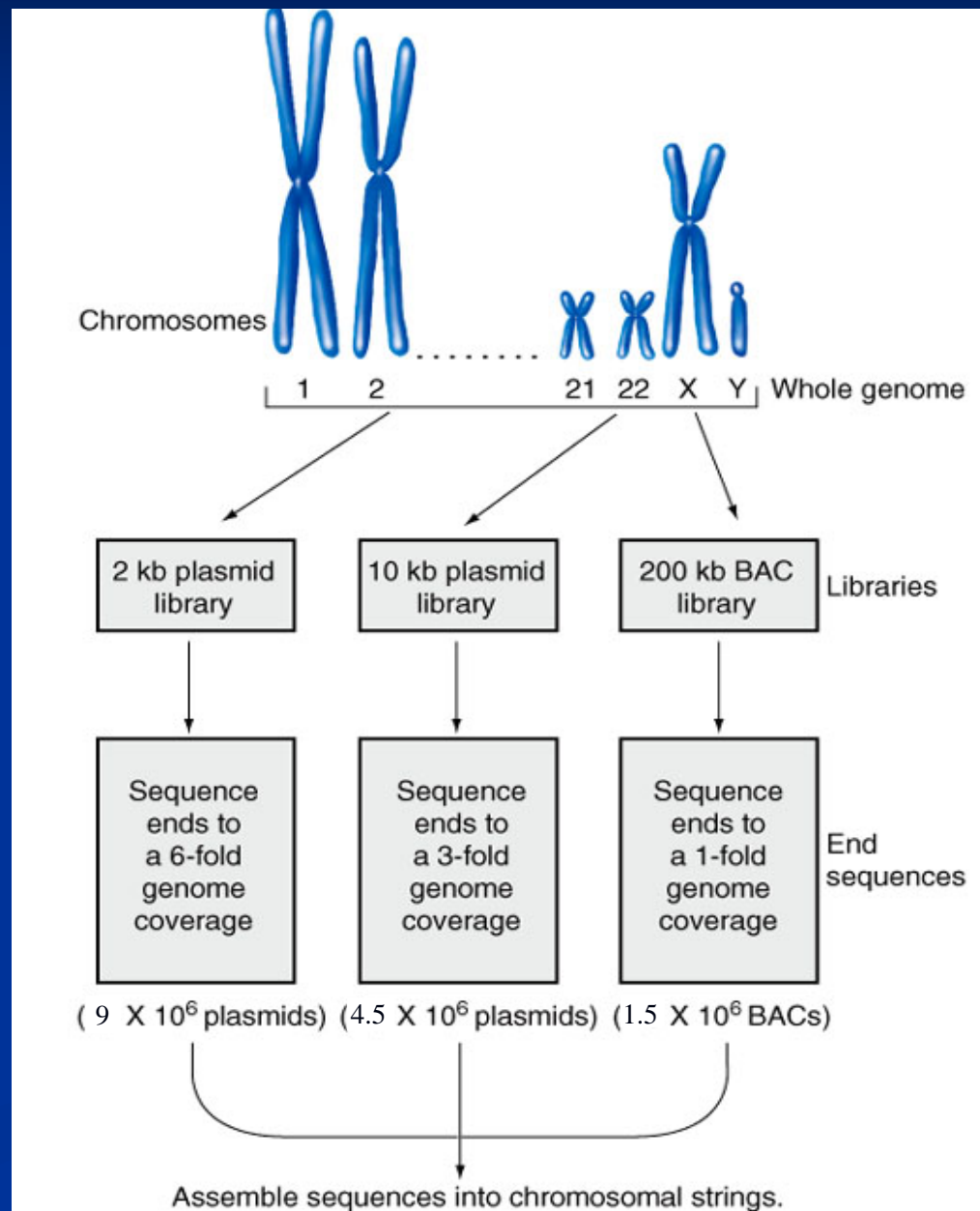
- Make **BAC library**.
- Select a set of minimal overlapping **BAC clones**.
- Shear 200 kb BAC clone into ~2 kb fragments and sequence ends at 10-fold coverage.
- Need ~20,000 BACs and ~1000 2 kb plasmids per BAC to cover genome.
- Data from linkage and physical maps used to assemble sequence maps of chromosomes.



# Whole-genome shotgun strategy

Private company Celera used to sequence whole human genome

- Whole genome randomly sheared three times
  - Plasmid library constructed with ~ 2 kb inserts
  - Plasmid library with ~10 kb inserts
  - BAC library with ~ 200 kb inserts
- Computer program assembles sequences into chromosomes.



## **Advantages of whole-genome shotgun strategy**

- 1. No physical map construction required.**
- 2. Require only one BAC library and two plasmid libraries.  
Three libraries of different DNA length overcome problems of repeat sequences.**
- 3. Rely on only DNA sequencing.**

# Limitations of large-scale sequencing

- **Some DNA can not be cloned.**
  - e.g., heterochromatin
- **Some sequences rearrange or sustain deletions when cloned.**
- **Some sequences are contaminated sequences from other species !!!**

## 2 Major insights from the human and model organism genome sequences

The human genome project has advanced gene finding and analysis.

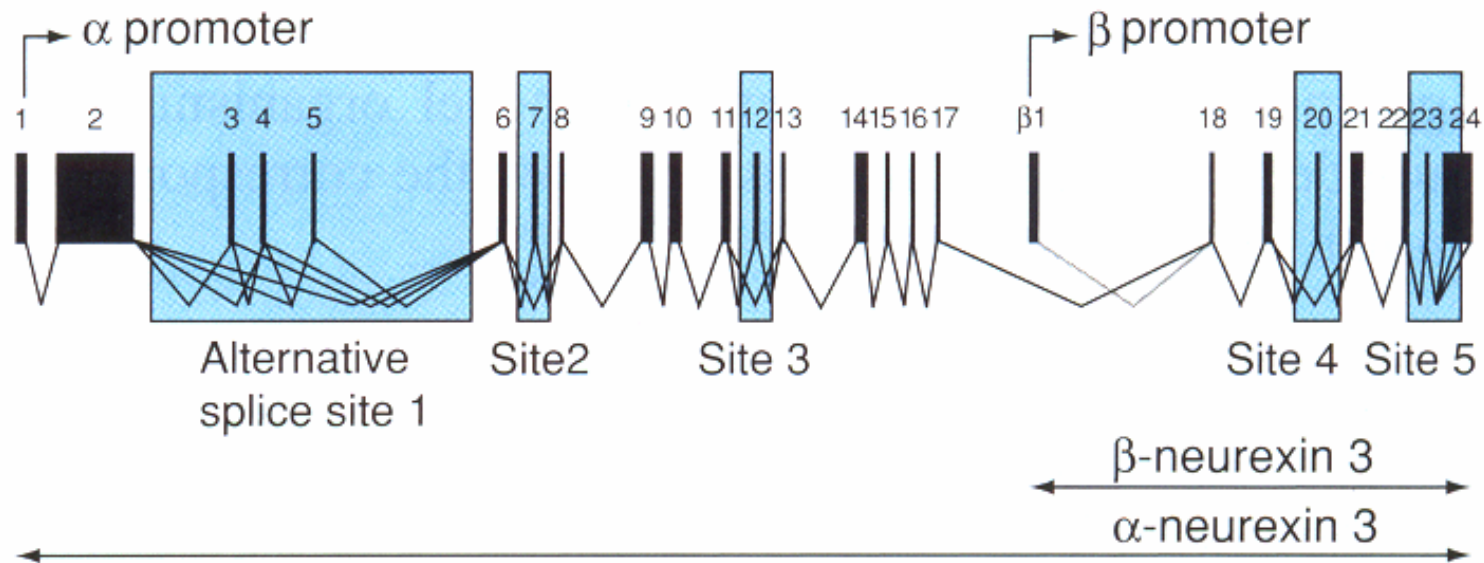
1. The identification of genes in one organism facilitates the identification of homologous genes in another organism.
  - **Orthologs (直系同源)**: Genes arose from the same gene in the common ancestor.
  - **Paralogs (旁系同源)**: Genes arise by duplication within a single species.
2. Comparisons between sequenced genomes help the assembling of a newly sequenced into its chromosomal strings.
3. Many genes are composed of distinct exons that encode discrete protein domains.
4. The individual genomes show **polymorphisms (多态性)**.

# The human genome contains approximately 25,000 genes

- The gene number is lower than expected (100,000 genes).
- The human genome could generate more complexity in protein function.
  - New arrangements of domain architecture generate new proteins.
  - **Alternative RNA splicing** (可变剪接) could increase the number of proteins produced.
  - More than 400 types of **post-translational modifications** on proteins can change a protein's functions.



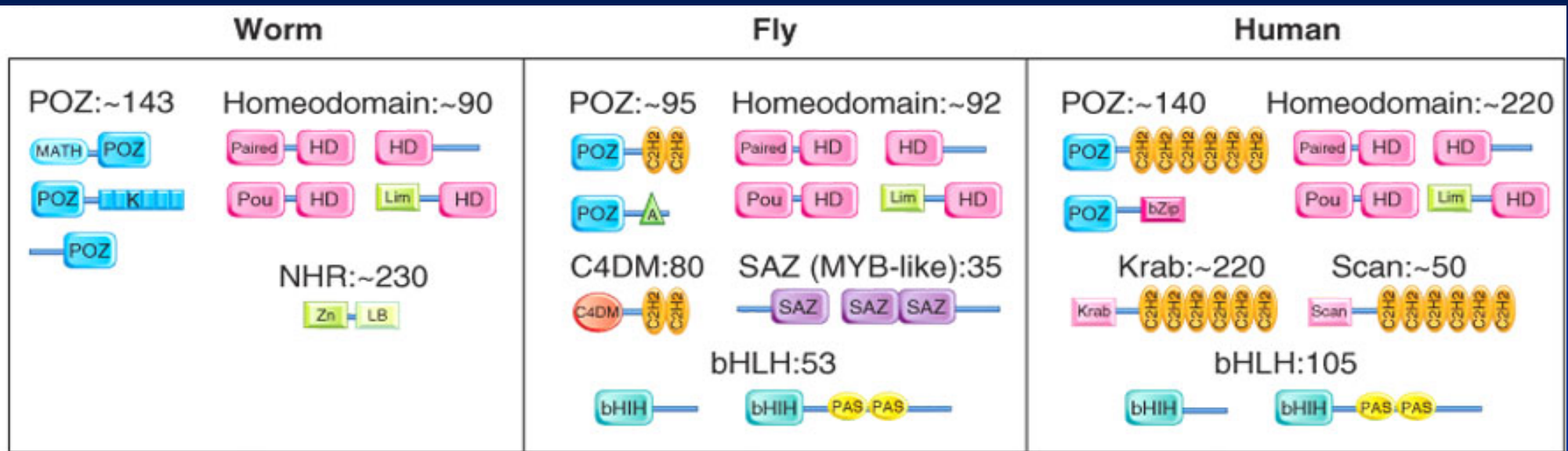
# Neurexins: Three genes and >1000 products



Neurexin Statistics			
Gene	Length in Human	Length in Puffer Fish	Number of Potential Alternative Splice Variants in Human
NRXN1	1112 kb	>163 kb	292
NRXN2	117 kb	unknown	194
NRXN3	1692 kb	>181 kb	1764

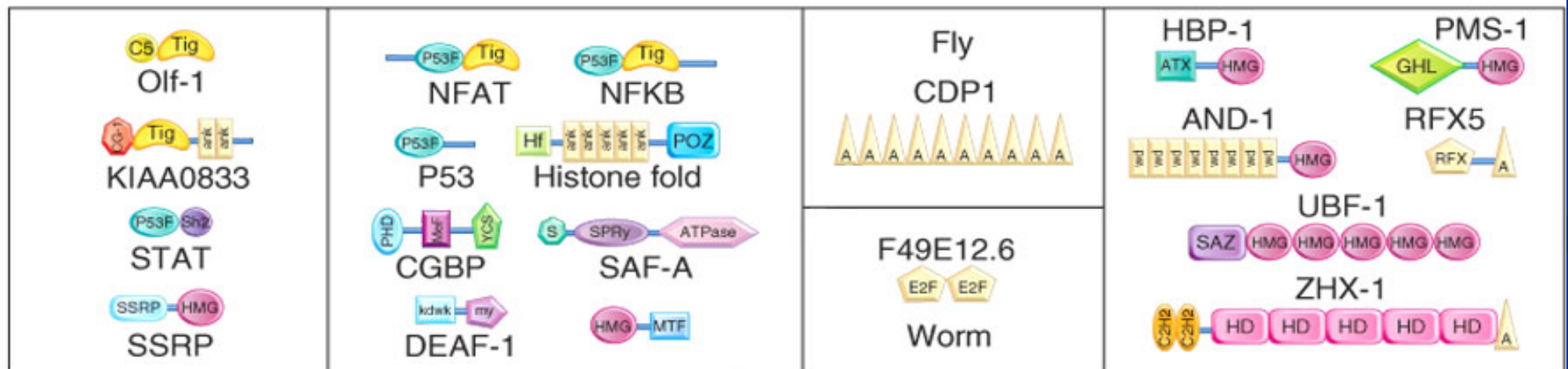
Fig. 10.14

# Transcription factor protein domains have expanded in specific lineages



(b)

## Unique and shared domain organizations in animals



Ancient architectures conserved in all animals

Shared by fly and human

Unique to fly or worm

Unique to human



# The genome contain distinct types of gene organization

## ■ Gene families

- Closely related genes clustered or dispersed

## ■ Gene-rich regions

- Functional or chance events?

## ■ Gene deserts

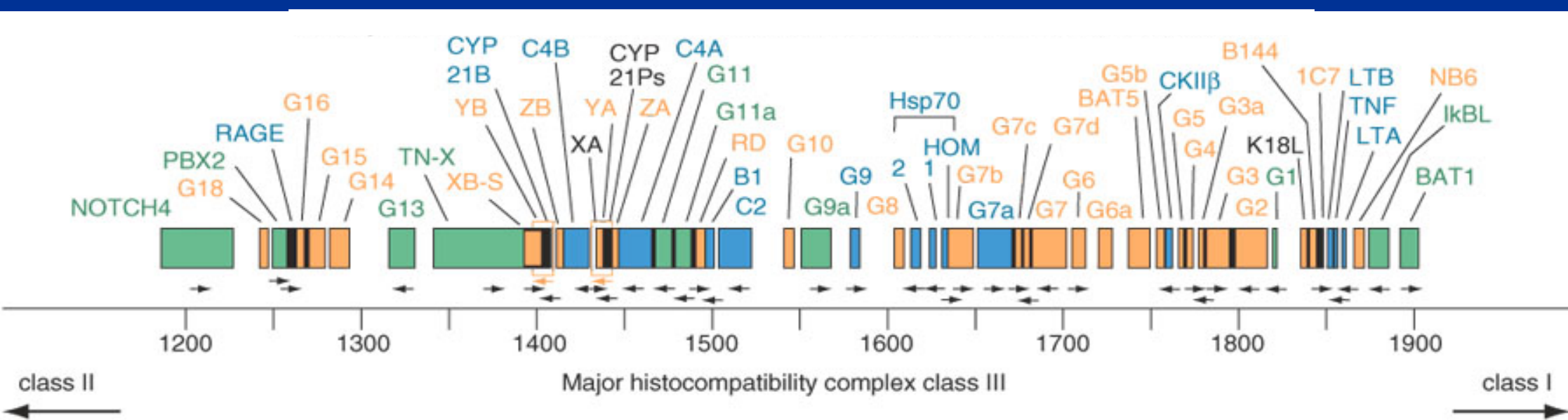
- Span 144 Mb or 3% of genome

- Contain regions difficult to identify?

- e.g., big genes – nuclear transcript spans 500 kb or more with very large introns (exons < 1% of DNA)

# Gene-rich regions



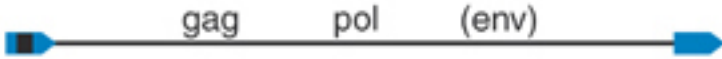



Class III region of human major histocompatibility complex contains 60 genes in 700 kb



# **Repeat sequences constitute more than 50% of the human genome**

- **Transposon-derived repeats: more than 45% of the genome.**
- **Processed pseudogenes**
- **Segmental duplications of 10-300 kb: more than 5%.**
- **Blocks of repeated sequences at centromeres, telomeres and other chromosomal features.**

# Transposon-derived repeats

			Length	Copy number	Fraction of genome
LINES	Autonomous		6–8 kb	850,000	21%
SINEs	Nonautonomous		100–300 bp	1,500,000	13%
Retrovirus-like elements	Autonomous		6–11 kb	450,000	8%
	Nonautonomous		1.5–3 kb		
DNA transposon fossils	Autonomous		2–3 kb	300,000	3%
	Nonautonomous		80–2000 bp		

- **Complexity of proteome increases from yeast to humans.**
  - More genes: ~ 25,000 genes in human genome
  - More **paralogous genes**
  - Shuffling, increase, or decrease of functional modules
  - Alternative RNA splicing – humans exhibit significantly more
  - Chemical modification of proteins is higher in humans.

# 3 Global analysis of genes and their mRNAs

- DNA microarrays
- DNA chips



An oligonucleotide array

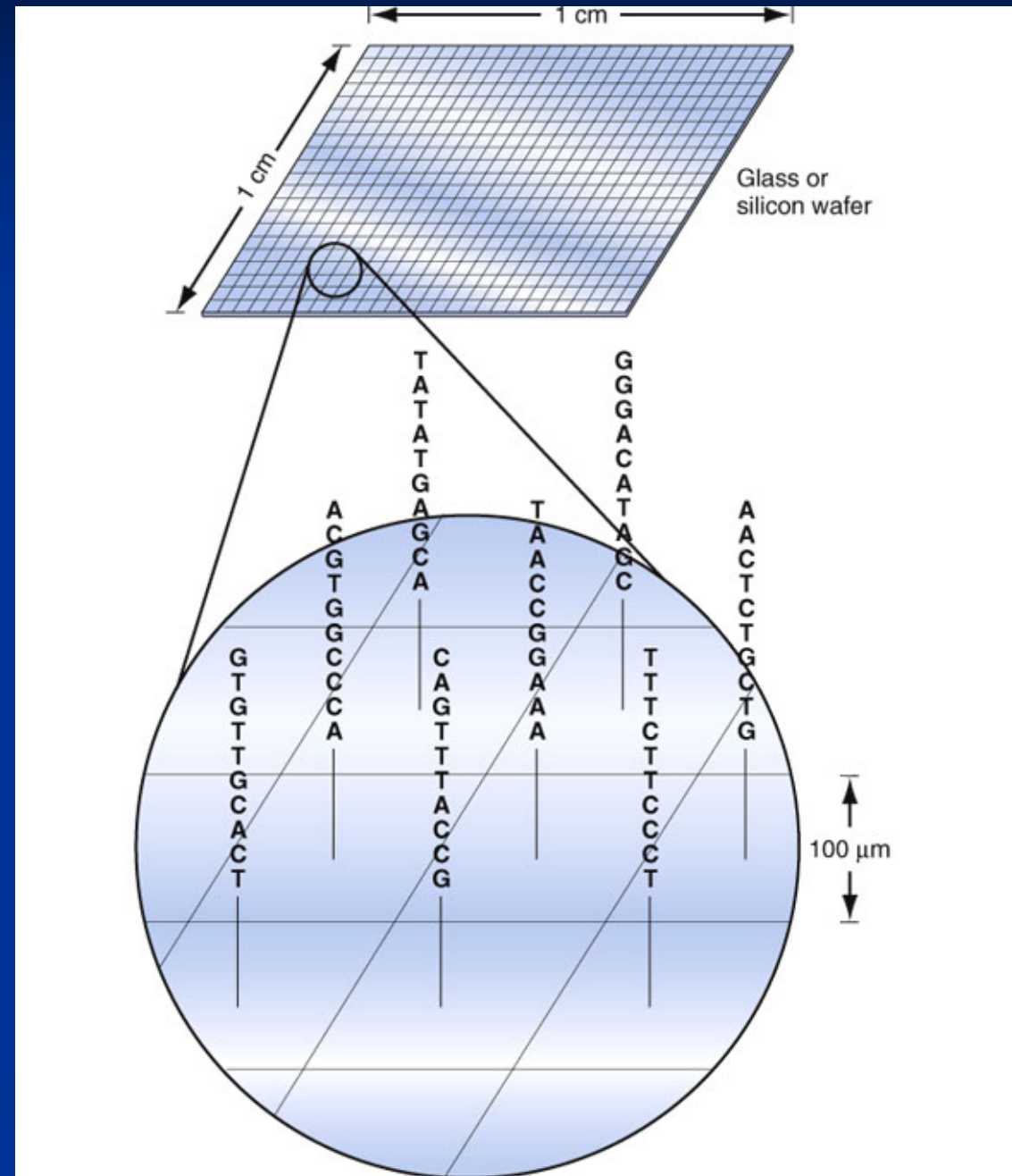
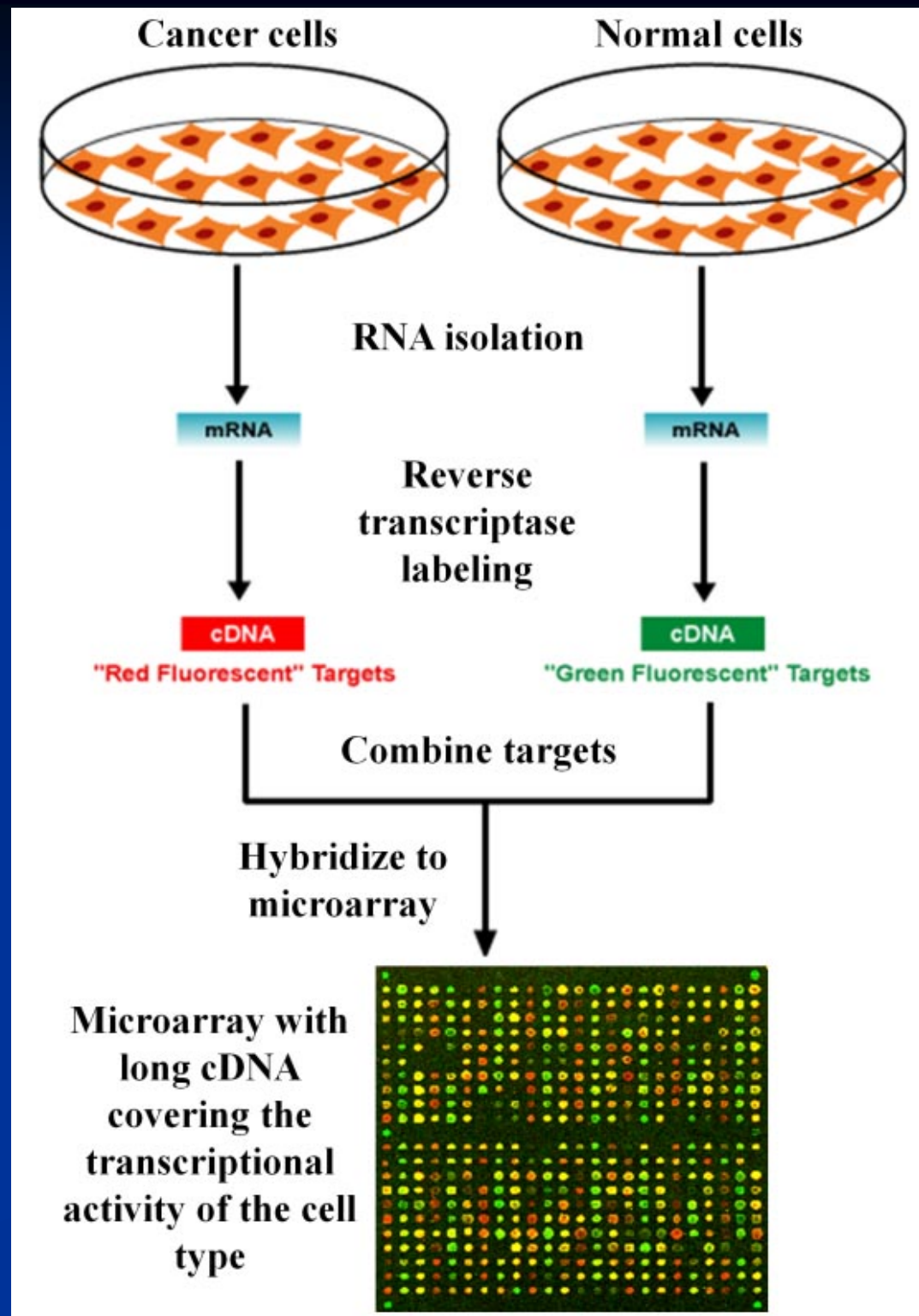


Fig. 10.16

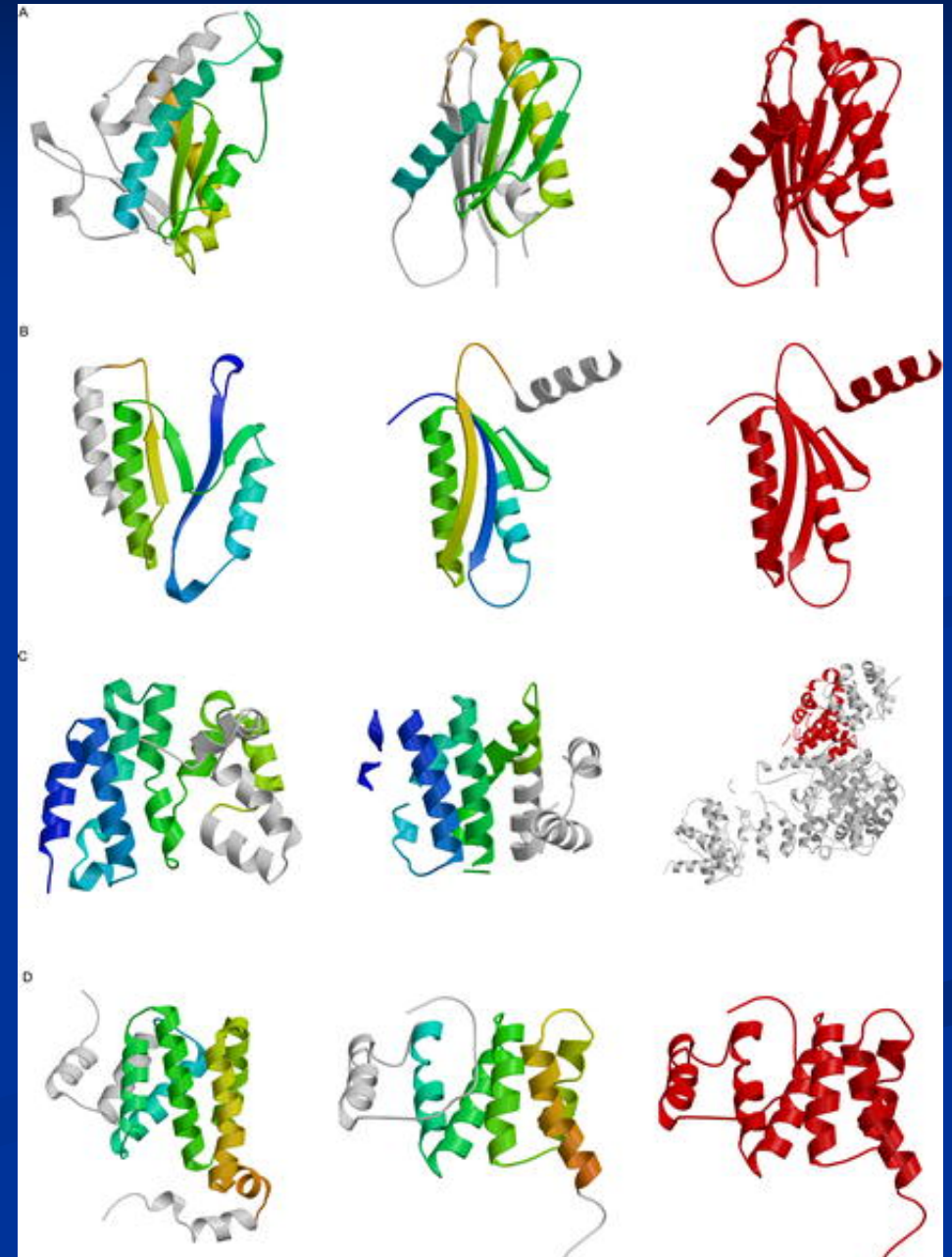
- **Two-color DNA microarray analysis** of gene expression profile in two samples.





## 4 Global analysis of proteomes

**Proteome:** The complete set of proteins in a particular cell or organism.





# **Analysis of proteomes is complex and challenging**

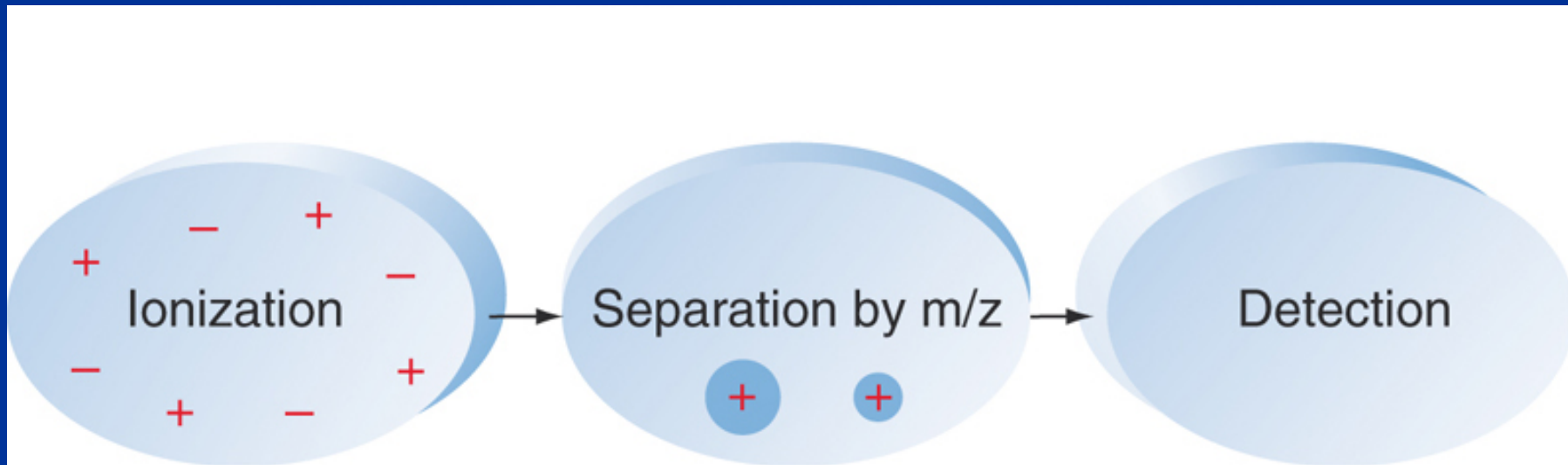
- **The range of protein expression in cells is enormous.**
- **Proteins in complex mixtures have many different features that must be identified and characterized.**
  - **Gene(s) encoding each protein**
  - **Levels of expression in particular cell types**
  - **Chemical modifications**
  - **Interactions with other proteins**
  - **Compartmentalization within the cell**
  - **State of activation**
  - **Half-lives**
  - **Three dimensional structure**
  - **Relationship between structure and function**

# Identifying proteins in complex mixtures

- **Human Genome Project** provides information to make **mass spectrometry** a powerful tool for identifying protein components of proteomes.
- **25,000** gene sequences in human genome.
  - **Mass spectrometer** can identify corresponding protein sequences.
  - Computationally determine mass of each protein and peptides derived from it.

# The mass spectrometer

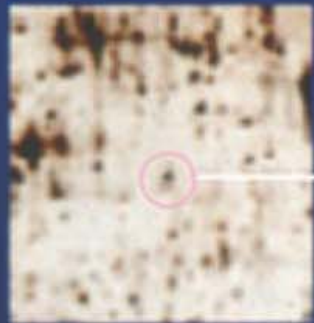
Measures masses of a wide variety of molecules including proteins.



## **Steps involved in identifying proteins in sequenced genomes**

- **Purify cells of a particular type.**
- **Extract proteins from cells.**
- **Digest with trypsin to split proteins.**
- **Partially fractionate peptides on a reverse phase column.**
- **Aliquots fed into mass spectrometer to obtain mass.**

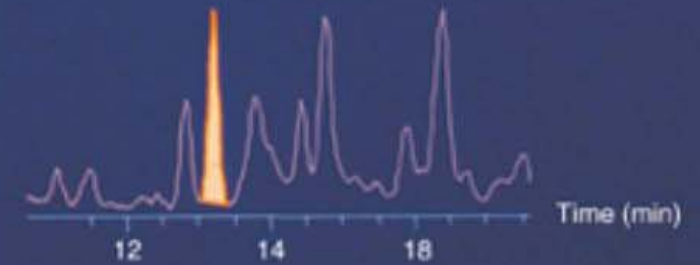
# Two-dimensional gels and mass spectrometry



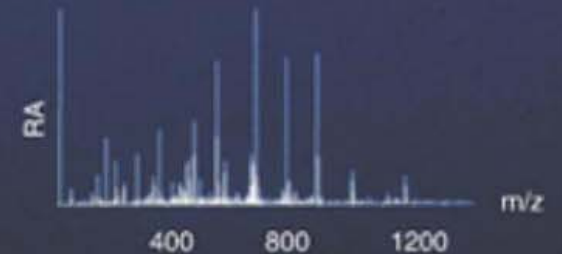
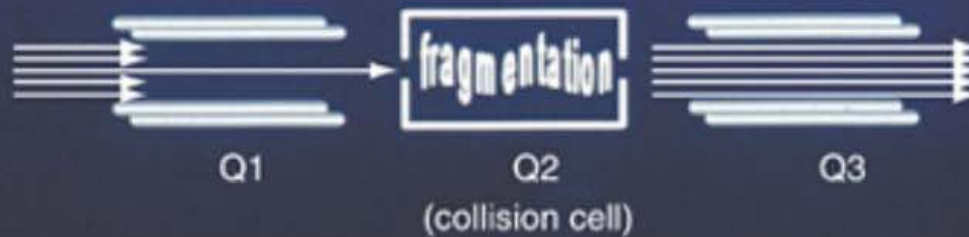
Protein



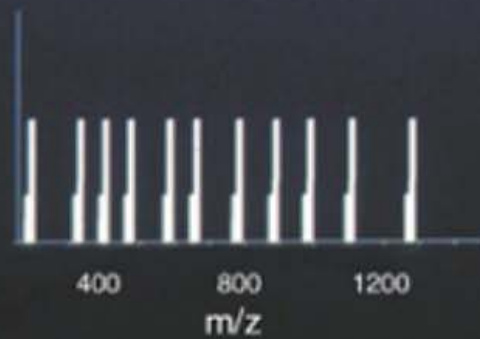
Peptides



MS base peak



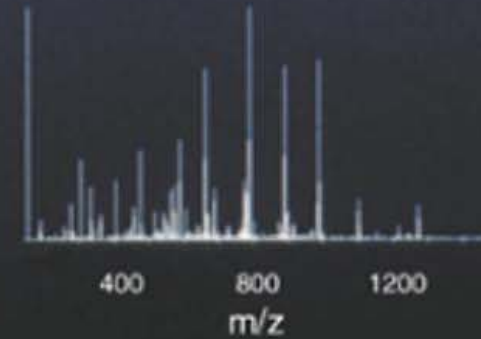
Tandem mass spectrum



Theoretical

Computer sequence  
database searching

Peptide identification  
and  
protein identification



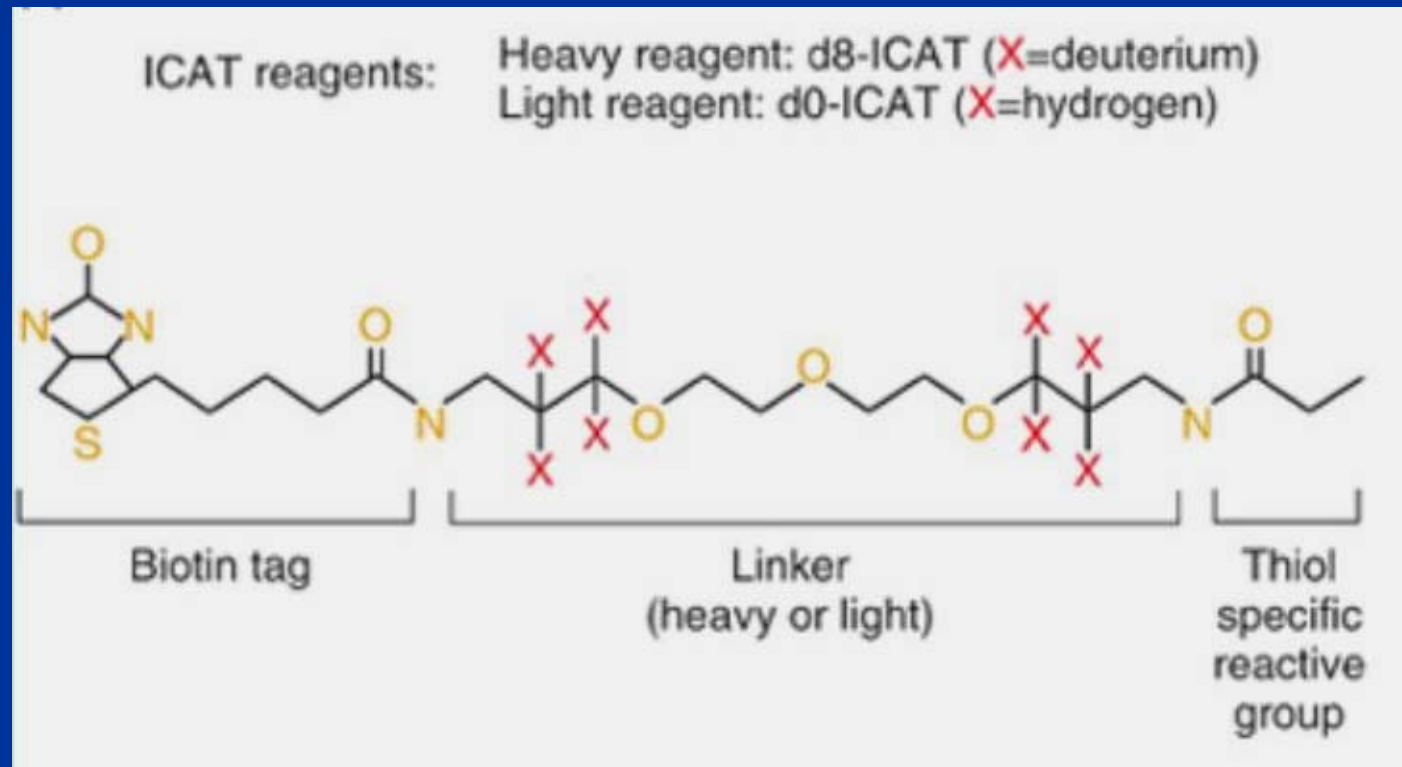
Acquired

# Quantifying changes in protein concentration in different cell or tissue states

- Isotope analysis using **isotope-coded affinity tags (ICATs)**
  - Analyze protein expression in two different cellular states.



- **Three components of ICATs:**
  - A biotin tag
  - A linker (can be labeled with eight hydrogen or deuterium isotopes)
  - A chemical group that reacts with the thiol (-SH) group of cysteine amino acids



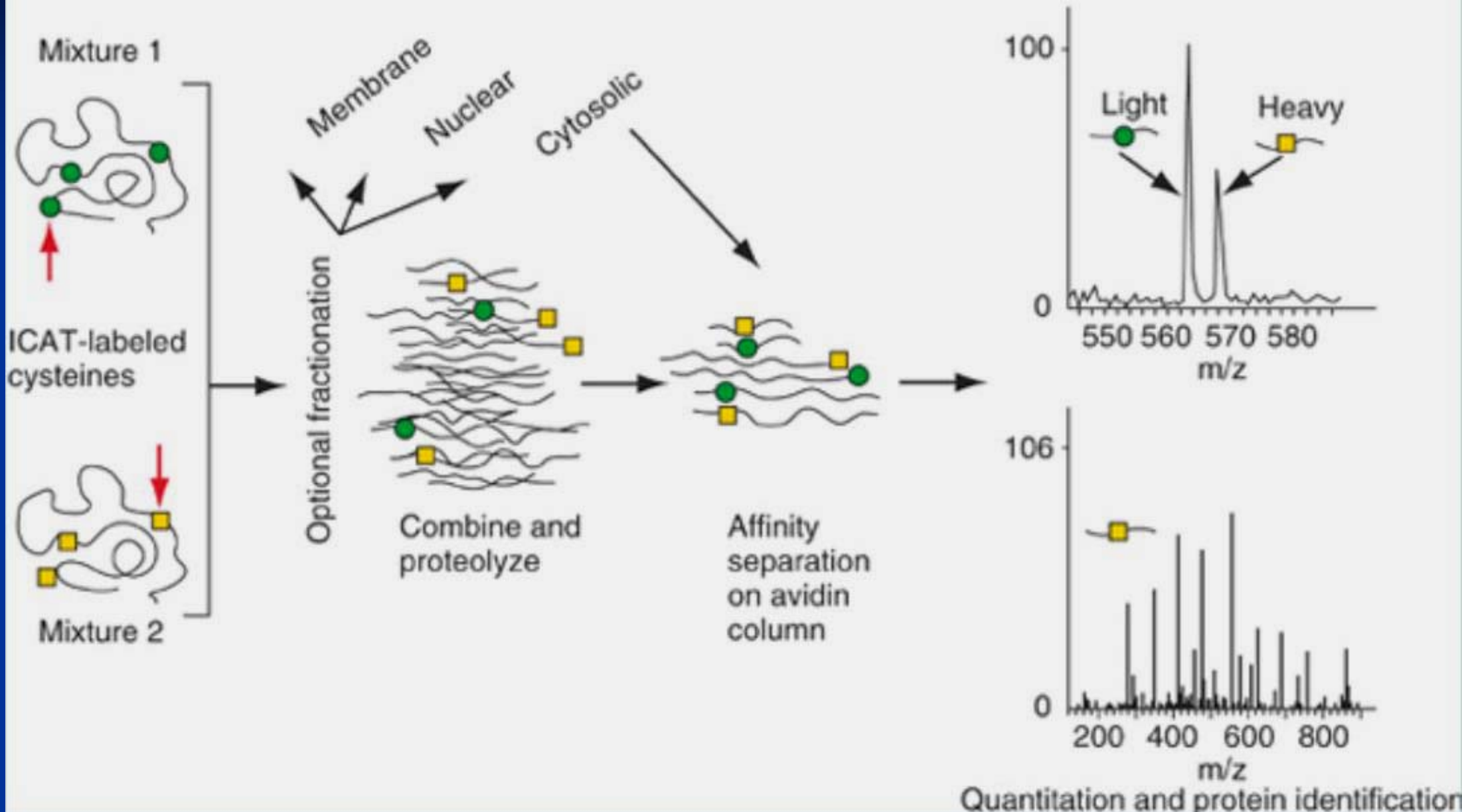


Fig. 10.22 b

# Identification of protein-protein interactions

- **Affinity purification and mass spectrometry**
- **Protein arrays**
- **Yeast two-hybrid**

# Affinity purification and mass spectrometry

- Antibody “pulls down” a protein from complex mixture.
- Identified in the mass spectrometer

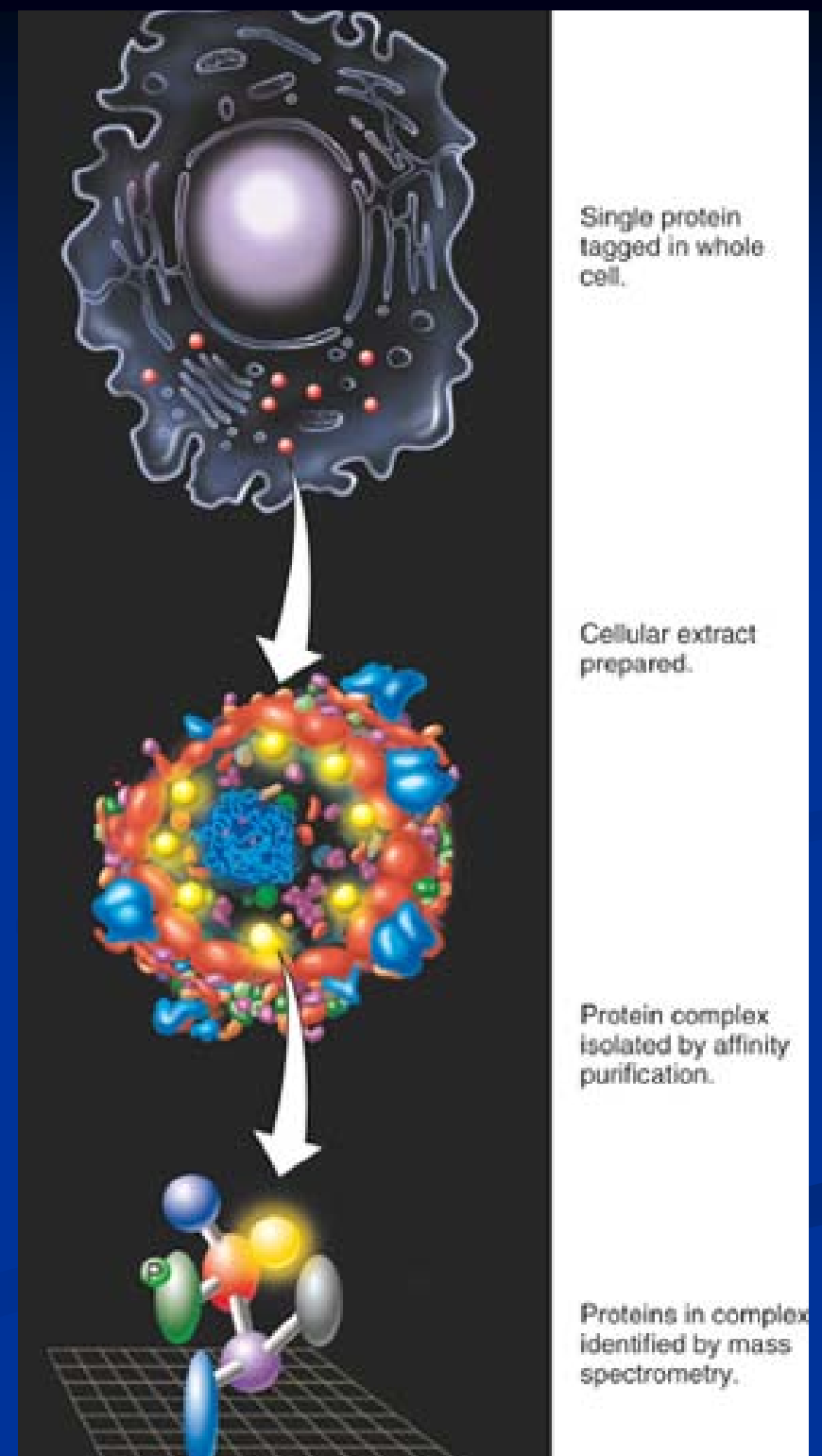
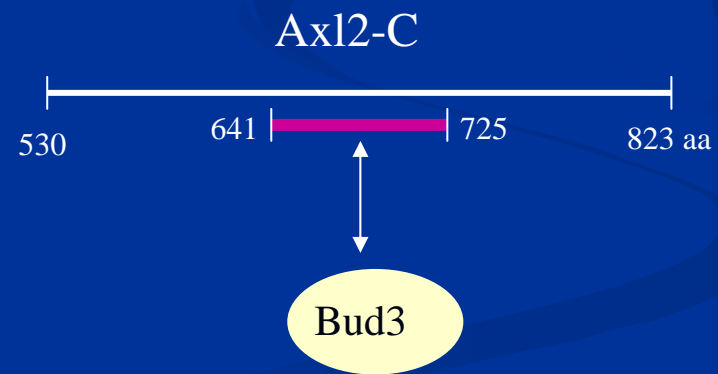
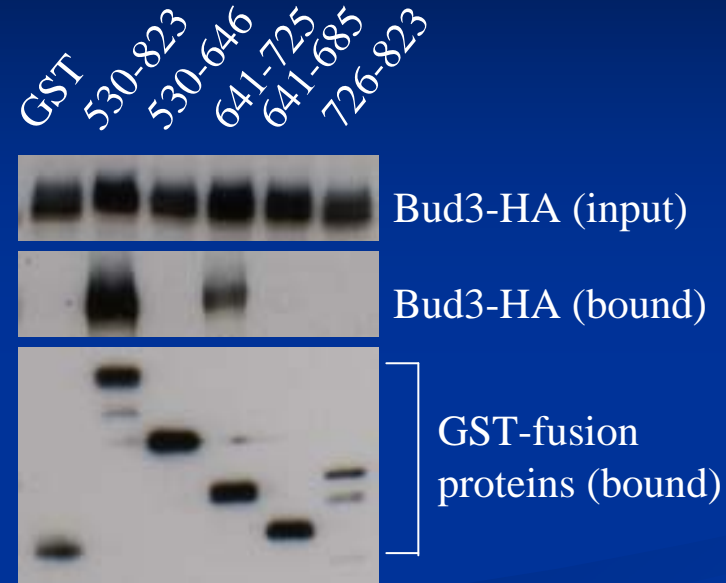


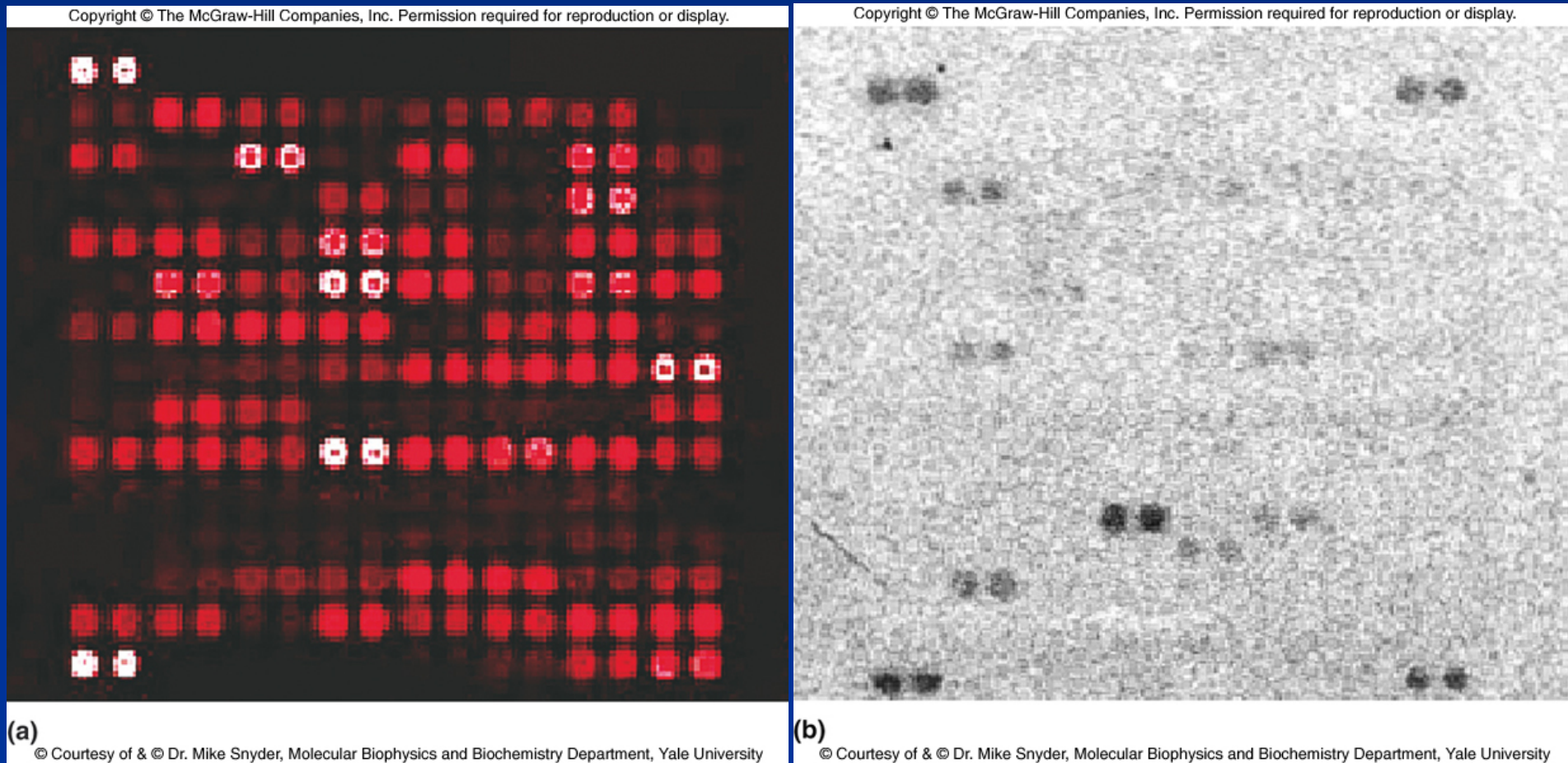
Fig. 10.23

# GST pull-down assay

## GST-Axl2-C fragments



# Identification of protein interactions and potential chemical modification by protein arrays

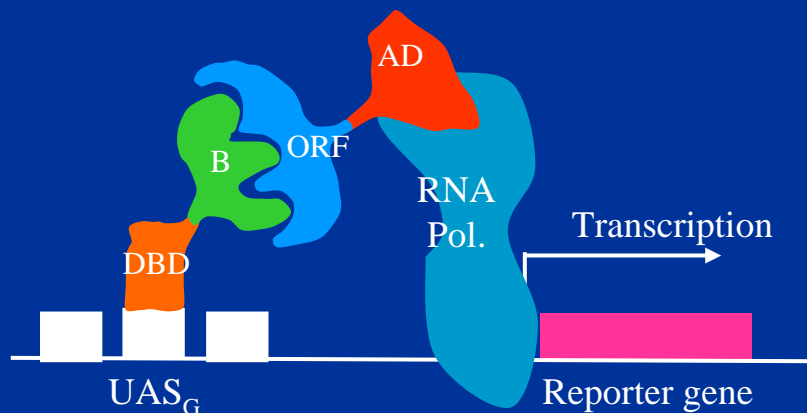


A protein array of different types of protein kinases



# Yeast two-hybrid

- **DBD**: Gal4 DNA-binding domain.
- Bait (**B**): protein of interest.
- **AD**: activator domain of Gal4.
- Prey (**ORF**): protein tested against bait.



- *HIS3* (Growth on SC-Histidine).
- *ADE2* (Growth on SC-Adenine)
- *LEU2* (Growth on SC-Leucine)
- *lacZ* (Blue color on X-GAL)

# Identification of protein-DNA interactions by ChIP/Chip

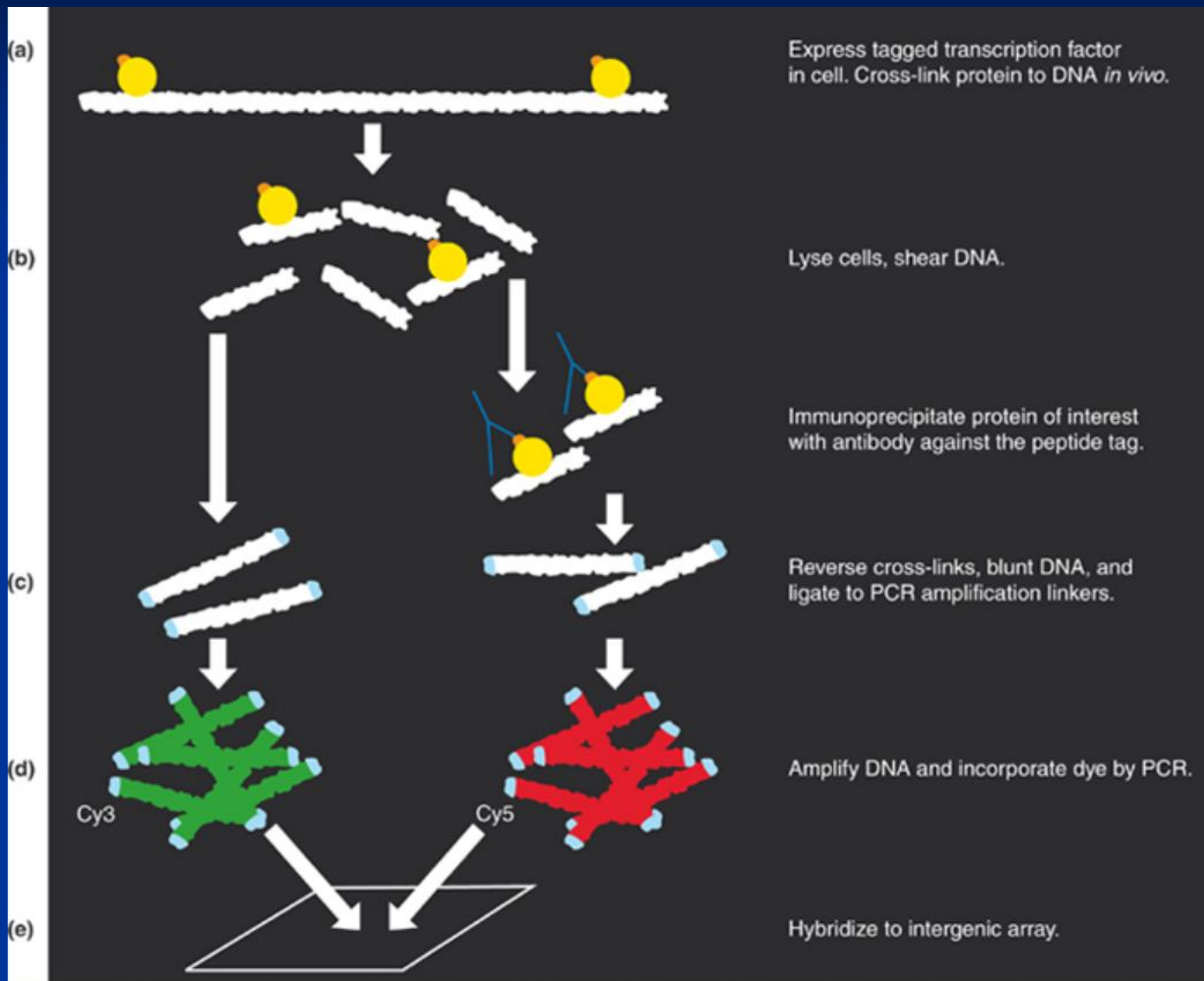


Fig. 10.25

# **5 Repercussions of the human genome project and high-throughput technology**

**Global proteomics strategies and high-throughput platforms** make it possible to gather and analyze system-wide protein data.

**For each organism, the tasks include:**

- Sequence the organism's genome.
- Define quantitatively **transcriptomes** and proteomes in each cell type.
- Delineate the nature of the proteome interactions in various cell types.
- Analyze other features of the proteome: localization, half-life, 3D structure.

# **Human Genome Project has changed the potential for predictive/preventive medicine**

- **Provided access to DNA polymorphisms underlying human variability.**
- **Help to identify genes predisposing to disease.**
- **Help to understand defective genes in context of biological systems.**
- **Help to circumvent limitations of defective genes.**
  - **Novel drugs**
  - **Environmental controls**
  - **Approaches such as stem-cell transplants or gene therapy**

## Social, ethical, and legal issues

- Privacy of genetic information.
- Limitations on the use of genetic testing.
- Patentability of DNA sequences.
- Society's view of growing number of older people.
- **Gene therapy** – inserting replacement genes to cure disease.
  - Somatic gene therapy
  - Germ-line gene therapy