Introduction to genome biology and DNA microarray experiments

Sandrine Dudoit and Robert Gentleman

Statistics and Genomics - Lecture 1, Part I

Department of Biostatistics
Harvard School of Public Health

January 23-25, 2002

Outline of lecture 1

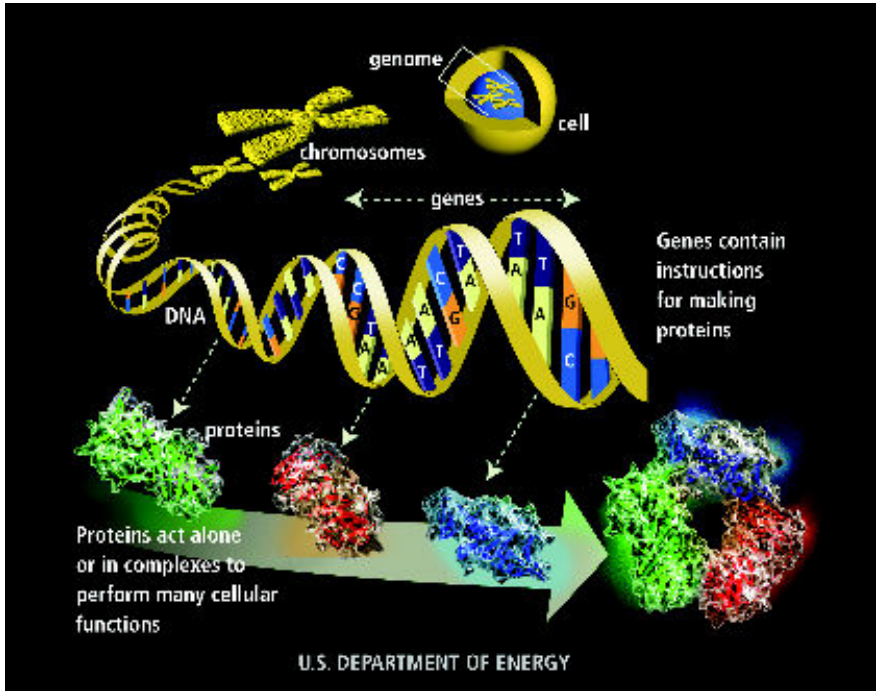
Part I:

- Introduction to genome biology;
- Introduction to microarray experiments.

Part II:

- Image analysis (cDNA microarrays);
- Normalization (cDNA microarrays);
- Experimental design.

Introduction to genome biology



The human genome

- The **cell** is the fundamental working unit of every living organism.
- Humans: trillions of cells (metazoa);
  other organisms like yeast: one cell (protozoa).
- Cells are of many different types (e.g. blood, skin, nerve cells), but all can be traced back to a single cell, the fertilized egg.

The human genome

- The **genome**, or blueprint for all cellular structures and activities in our body, is encoded in **DNA** molecules.

- Each cell contains a complete copy of the organism's **genome.**

The human genome

- The human genome is distributed along 23 pairs of **chromosomes**

  22 autosomal pairs;

  the sex chromosome pair, XX for females and XY for males.

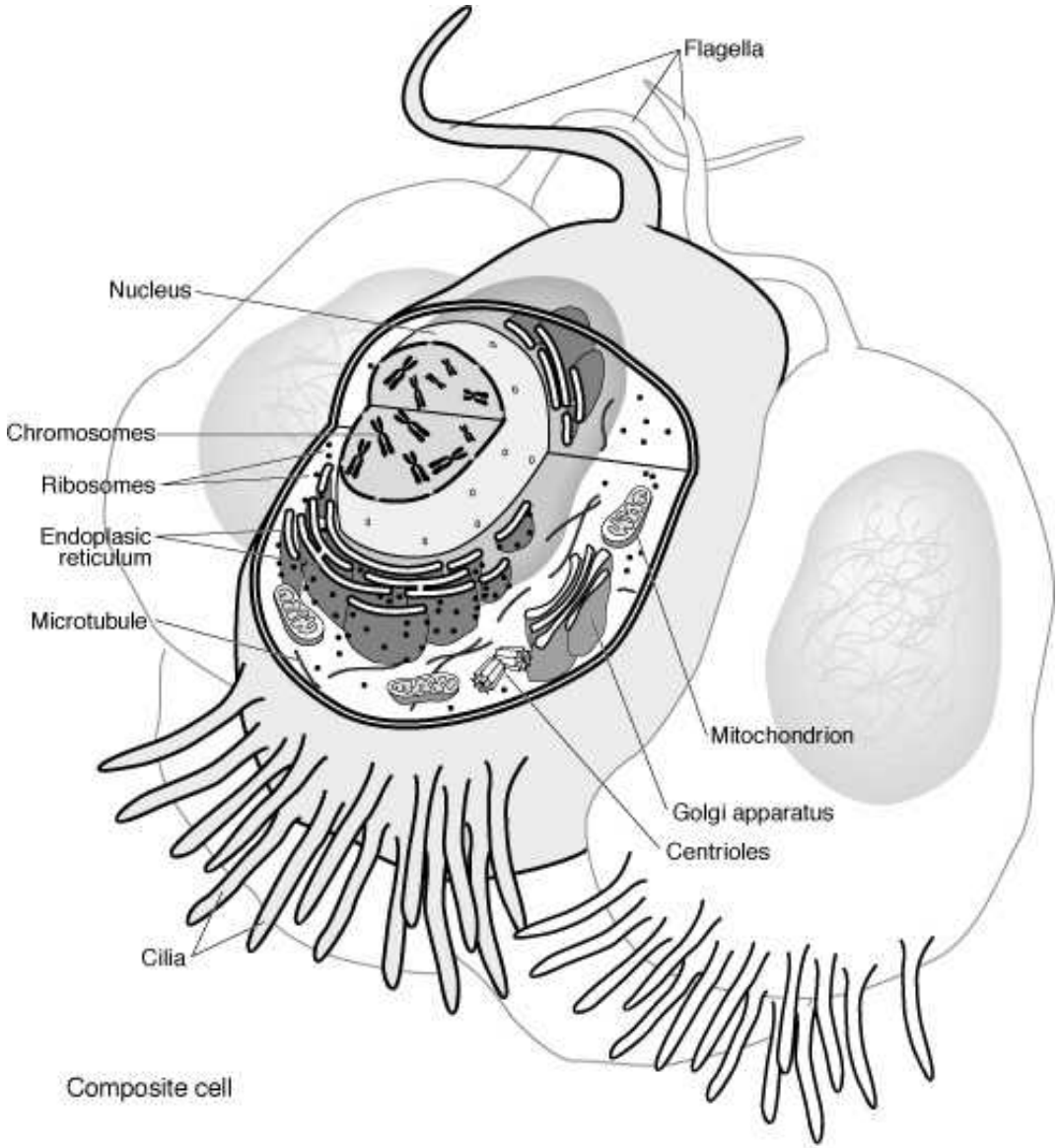
- In each pair, one chromosome is paternally inherited, the other maternally inherited (cf. meiosis).

The human genome

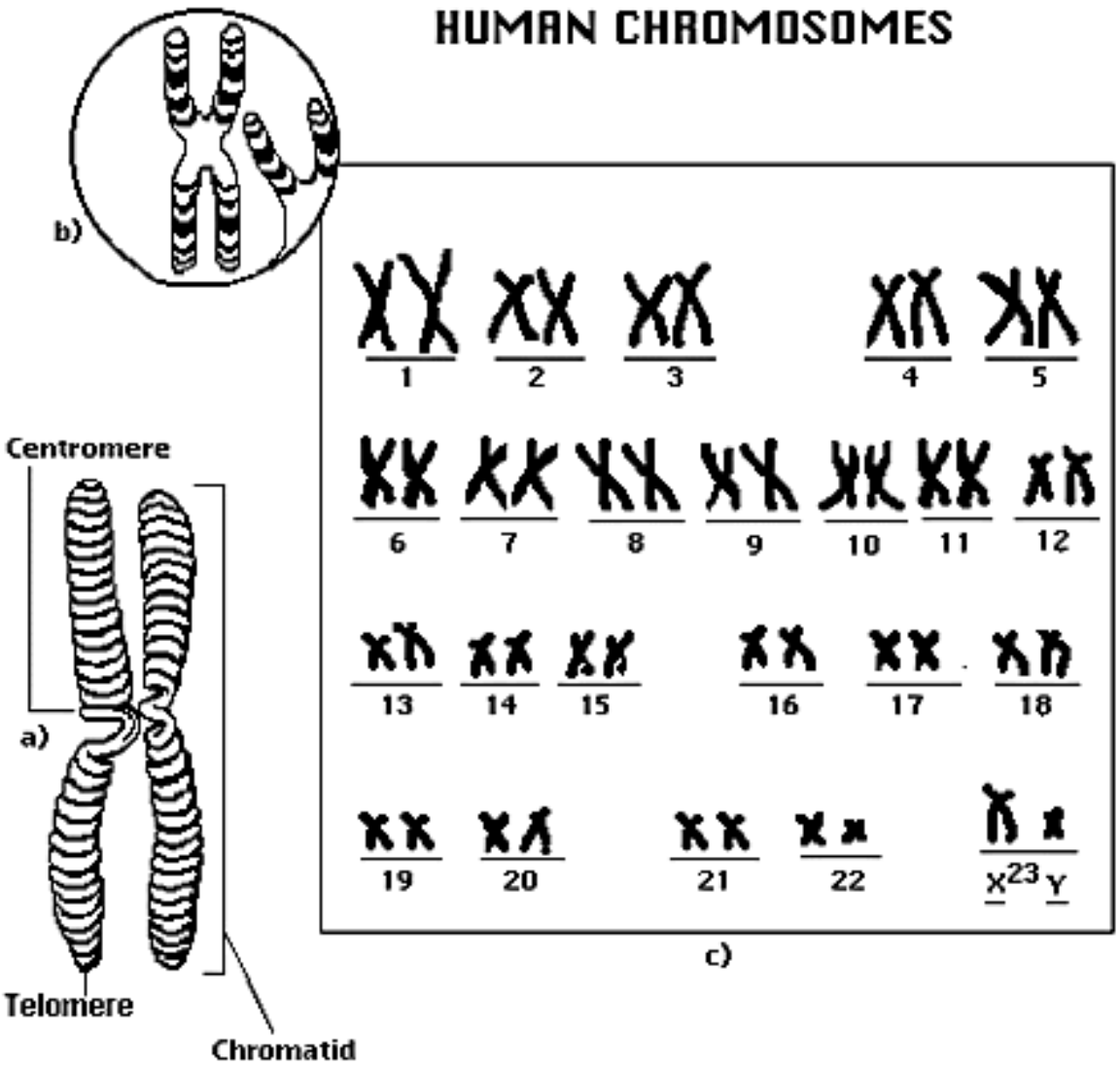
- Chromosomes are made of compressed and entwined **DNA.**

- A (protein-coding) **gene** is a segment of chromosomal **DNA** that directs the synthesis of a **protein**.

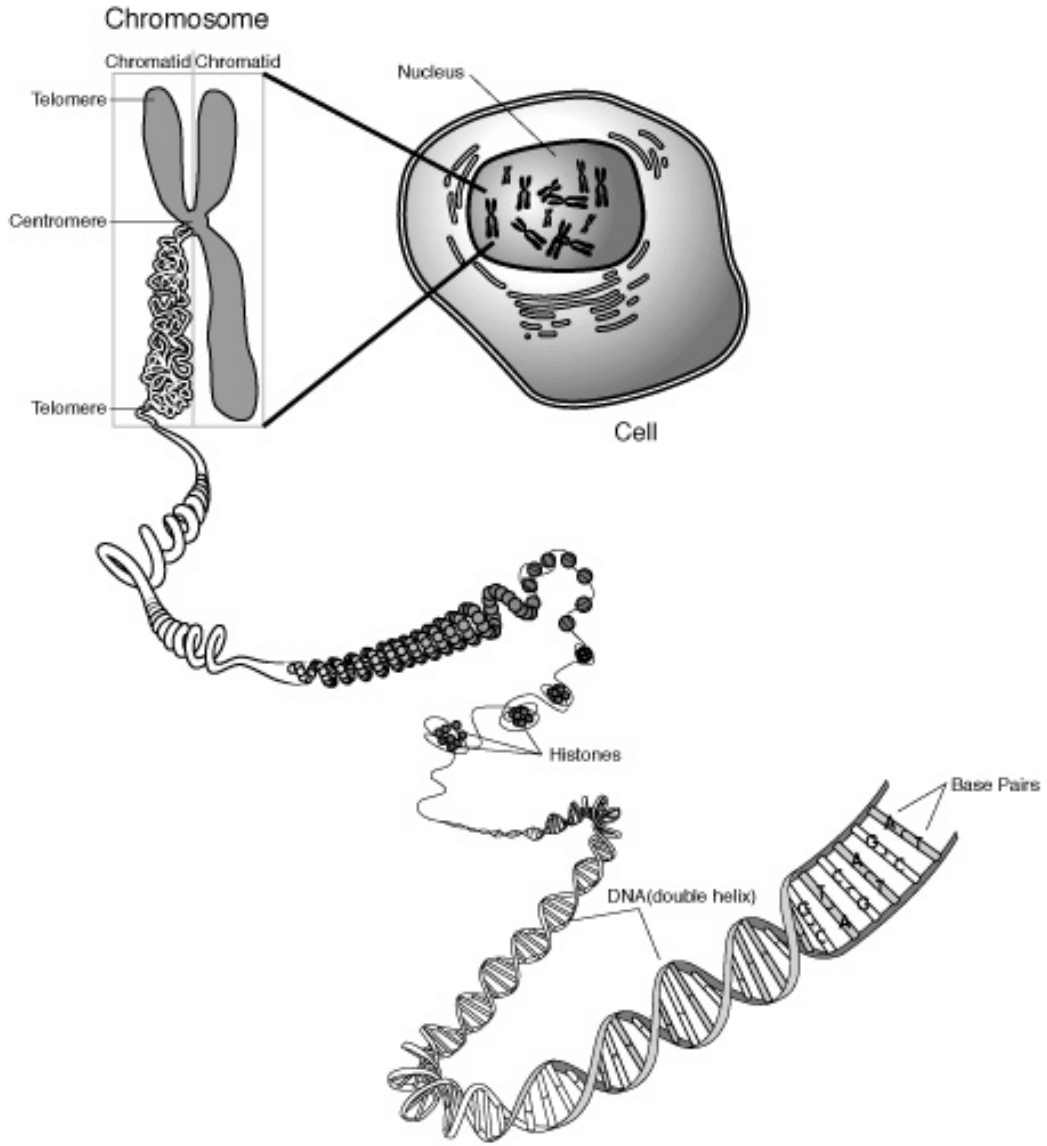
The eukaryotic cell



Chromosomes



Chromosomes and DNA

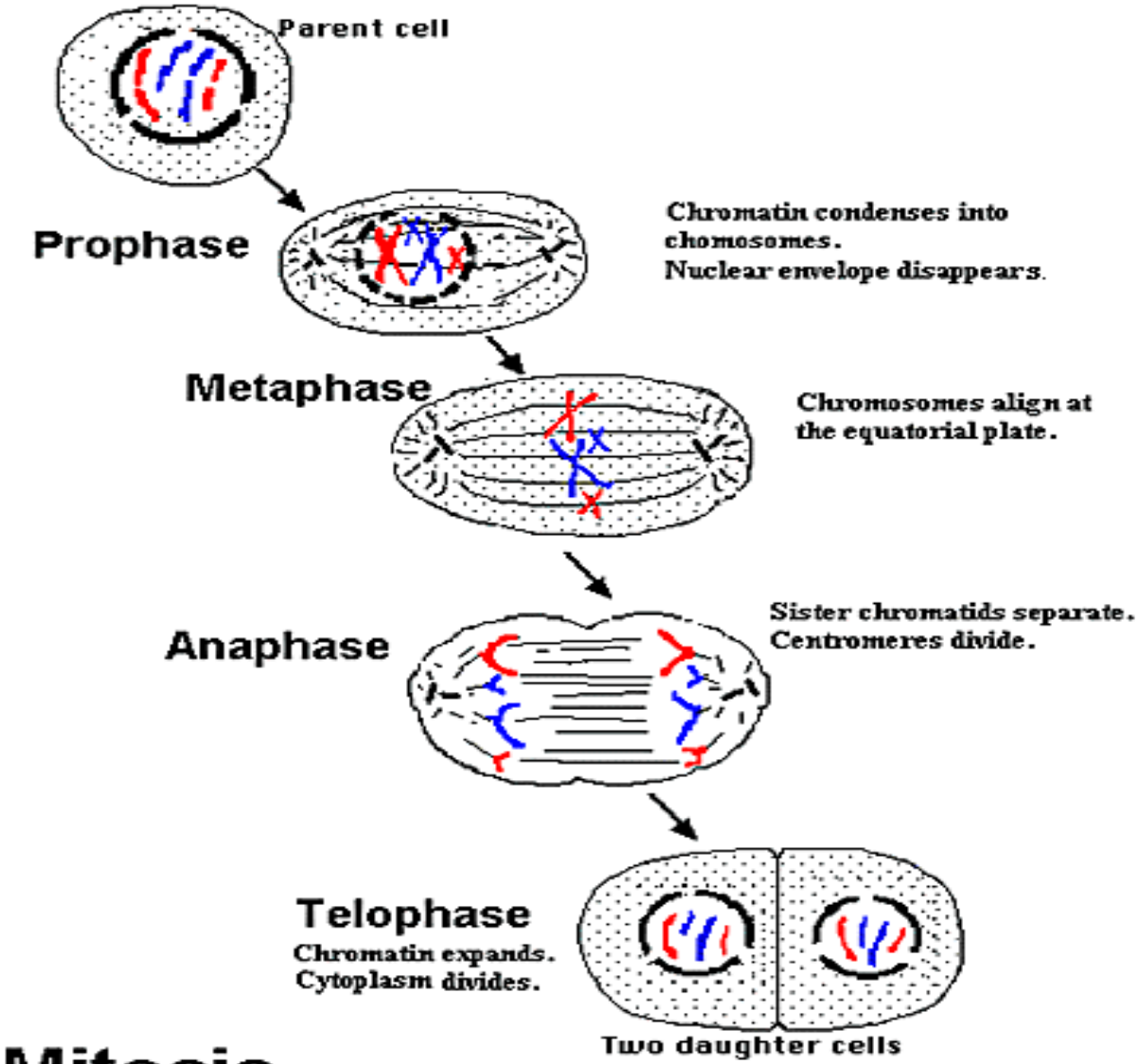


Cell divisions

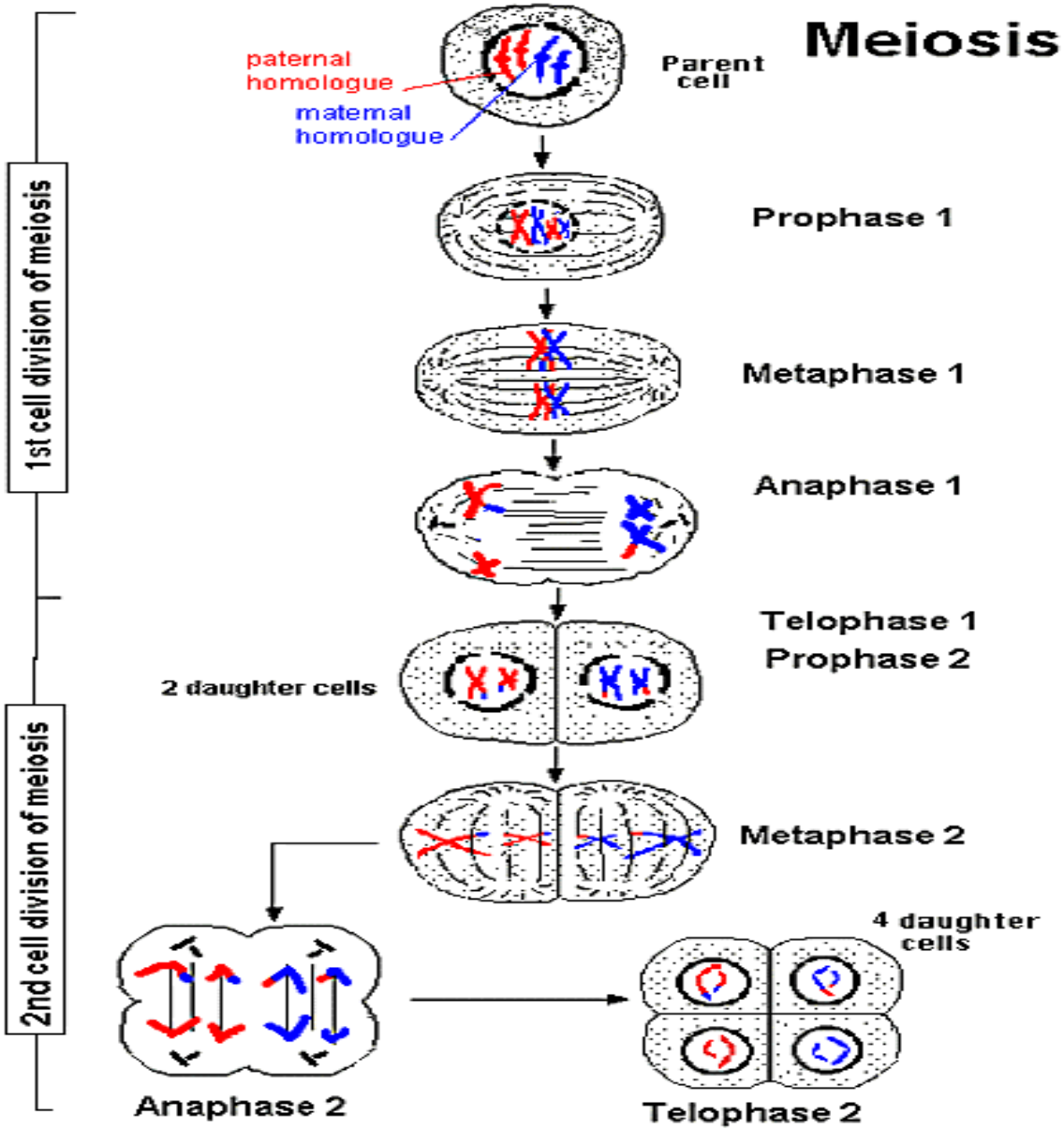
- **Mitosis.** One nuclear division produces two daughter **diploid** nuclei identical to the parent nucleus.
- **Meiosis.** Two successive nuclear divisions produces four daughter **haploid** nuclei, different from original cell.

  Leads to the formation of gametes (egg/sperm).

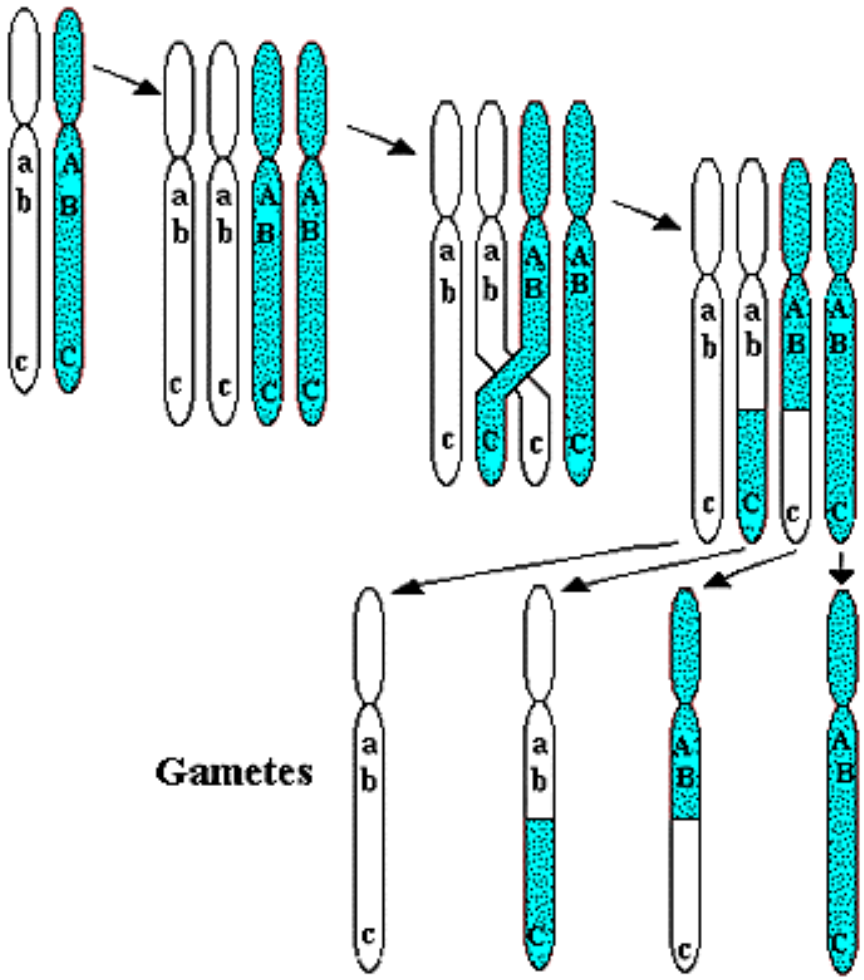
Mitosis


Parent cell
Prophase
Chromatin condenses into chomosomes.
Nuclear envelope disappears.
Metaphase
Chromosomes align at the equatorial plate.
Anaphase
Sister chromatids separate.
Centromeres divide.
Telophase
Chromatin expands.
Cytoplasm divides.
Two daughter cells
Mitosis

Meiosis


Meiosis
paternal homologue
maternal homologue
Parent cell
Prophase 1
Metaphase 1
Anaphase 1
Telophase 1
Prophase 2
2 daughter cells
Metaphase 2
4 daughter cells
Anaphase 2
Telophase 2
1st cell division of meiosis
2nd cell division of meiosis

Recombination

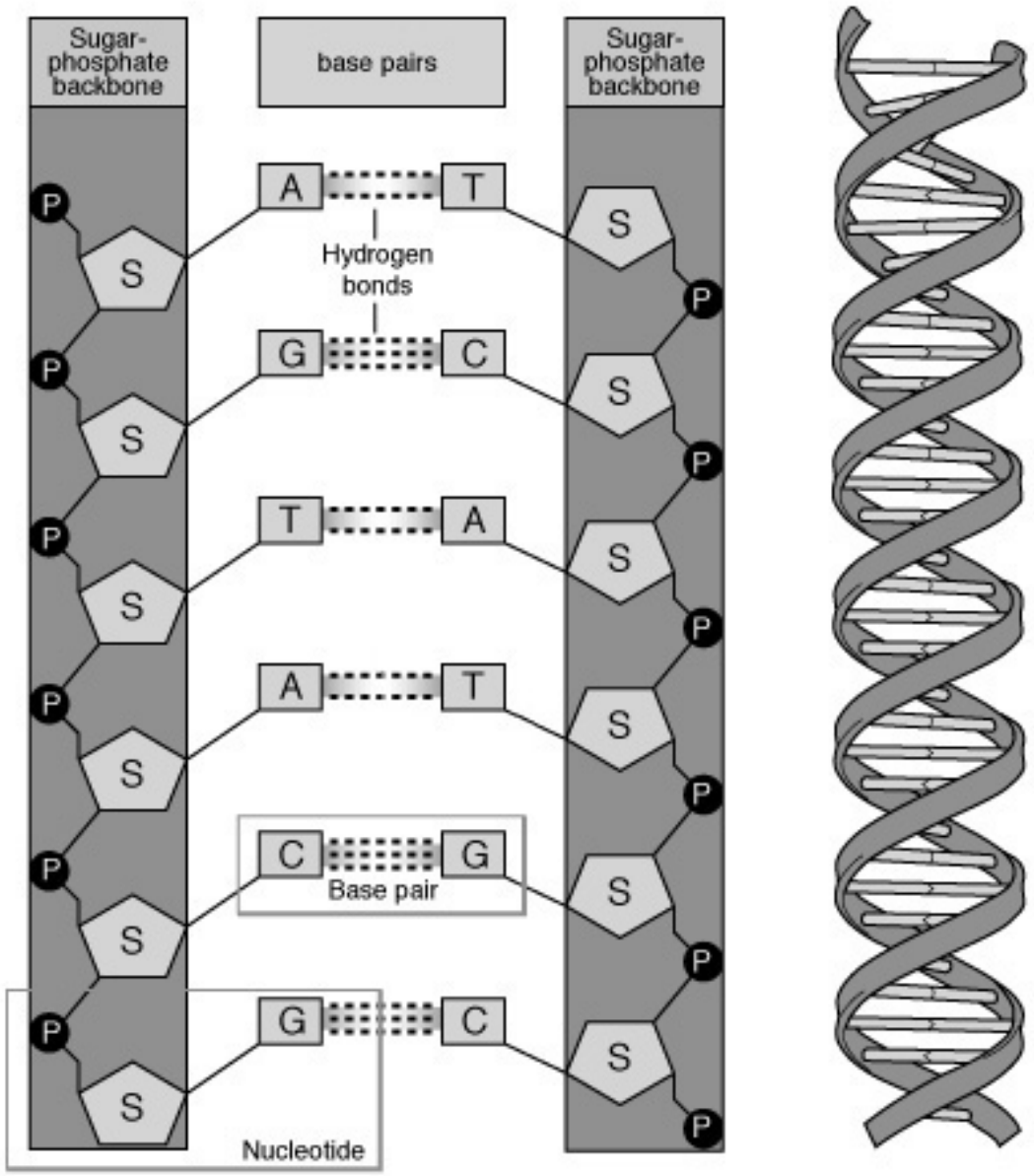


Crossing-over and recombination during meiosis

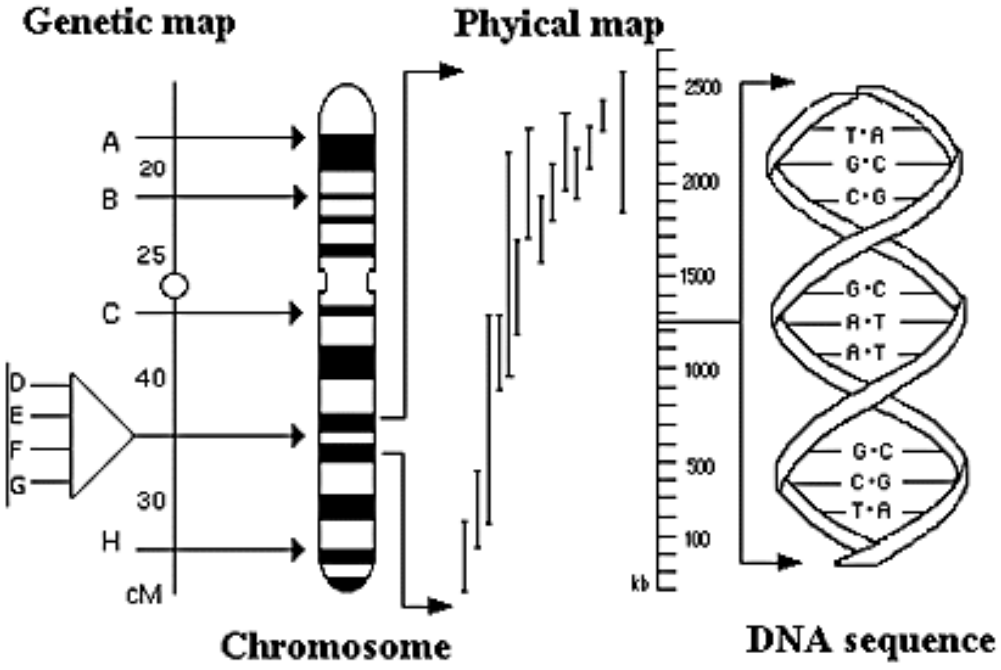
DNA

- A **deoxyribonucleic acid** or **DNA** molecule is a double-stranded polymer composed of four basic molecular units called **nucleotides**.
- Each nucleotide comprises a phosphate group, a deoxyribose sugar, and one of four nitrogen bases: **adenine (A)**, **guanine (G)**, **cytosine (C)**, and **thymine (T)**.
- The two chains are held together by hydrogen bonds between nitrogen bases.
- Base-pairing occurs according to the following rule: **G pairs with C**, and **A pairs with T**.

DNA


Sugar-phosphate backbone
base pairs
Sugar-phosphate backbone
A
T
Hydrogen bonds
G
C
T
A
A
T
C
G
Base pair
G
C
Nucleotide
P
S

Genetic and physical maps



Sequences of base pairs mapping

Genetic and physical maps

- **Physical distance:** number of base pairs (bp).
- **Genetic distance:** expected number of crossovers between two loci, per chromatid, per meiosis.

  Measured in Morgans (M) or centiMorgans (cM).
- 1cM ~ 1 million bp (1Mb).

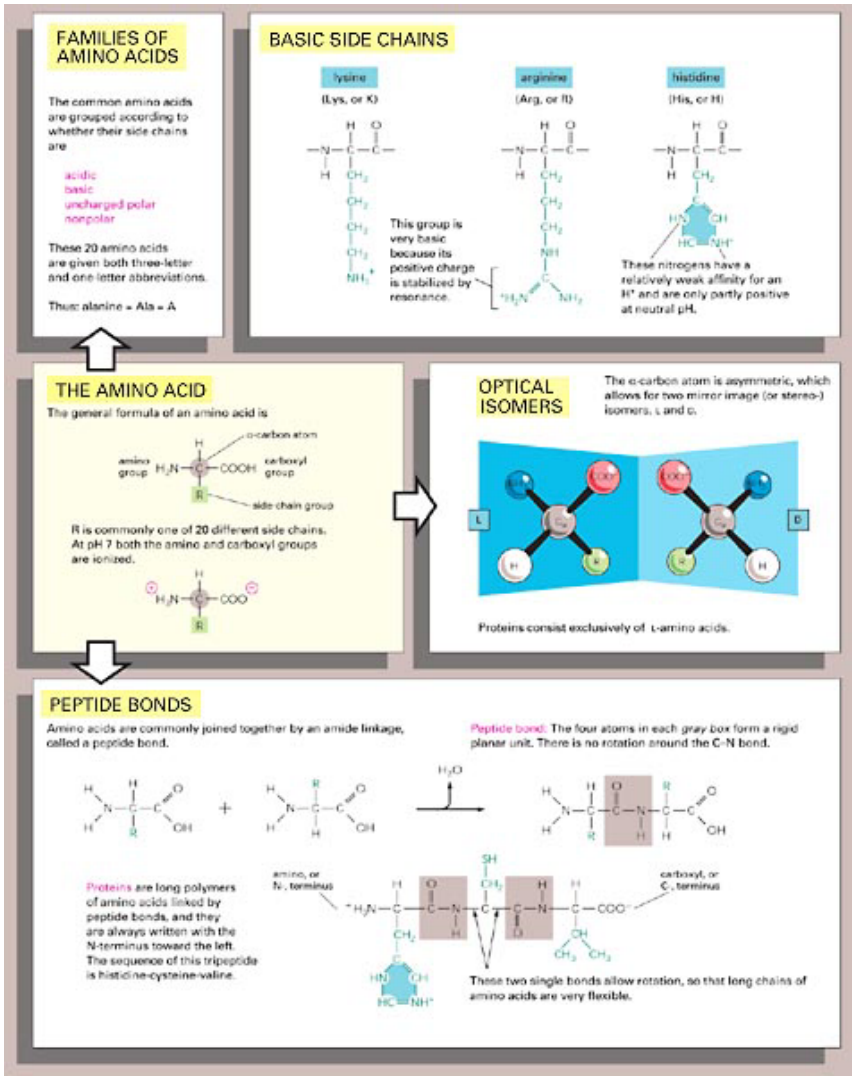
The human genome in numbers

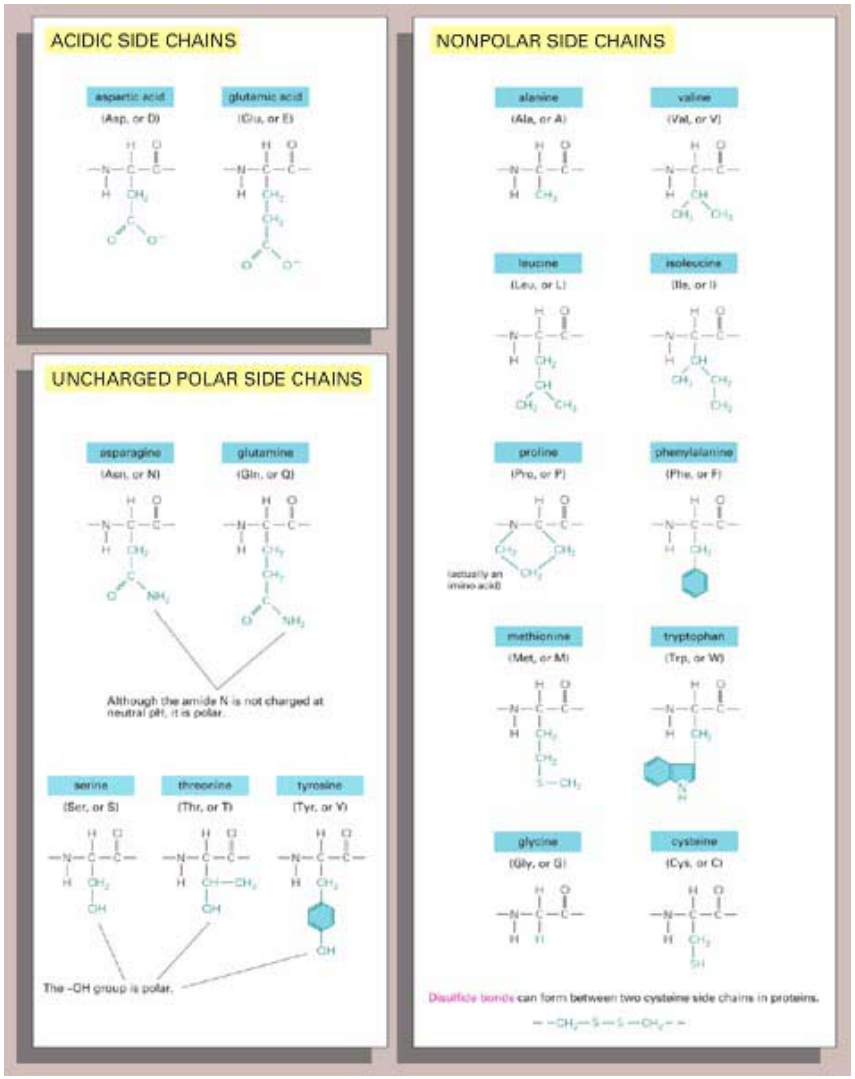
- 23 pairs of chromosomes;
- 2 meters of DNA;
- 3,000,000,000 bp;
- 35 M (males 27M, females 44M);
- 30,000-40,000 genes.

Proteins

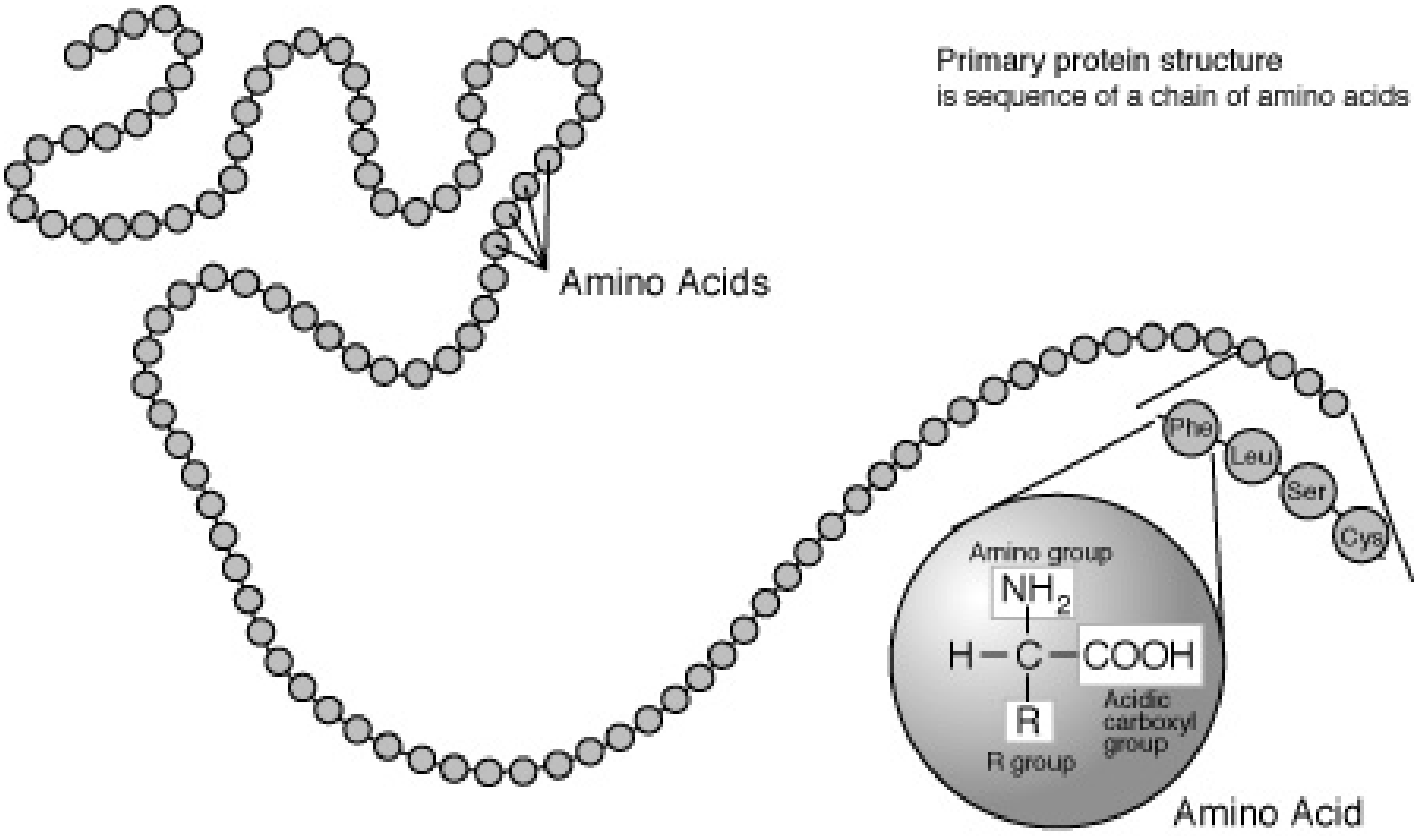
- **Proteins:** large molecules composed of one or more chains of amino acids.
- **Amino acids:** class of 20 different organic compounds containing a basic amino group ($-NH_2$) and an acidic carboxyl group (-COOH).
- The order of the amino acids is determined by the **base sequence** of nucleotides in the **gene** coding for the protein.
- E.g. hormones, enzymes, antibodies.

Amino acids

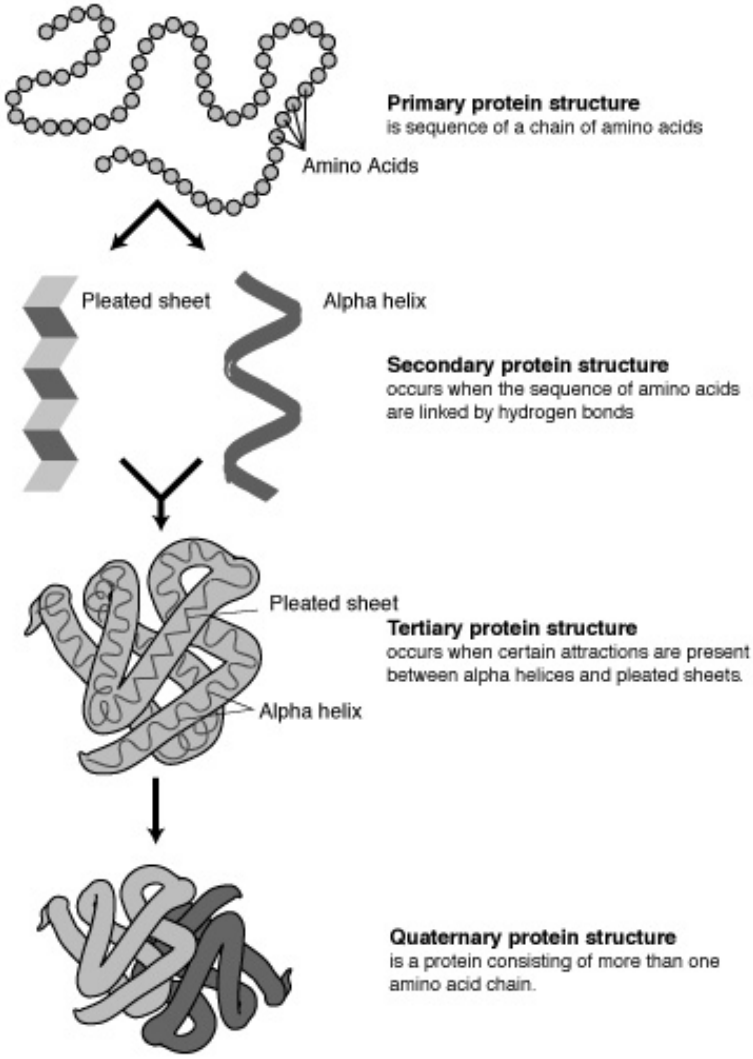

FAMILIES OF AMINO ACIDS
The common amino acids are grouped according to whether their side chains are
acidic
basic
uncharged polar
nonpolar
These 20 amino acids are given both three-letter and one letter abbreviations.
Thus: alanine = Ala = A
BASIC SIDE CHAINS
lysine
(Lys, or K)
arginine
(Arg, or R)
histidine
(His, or H)
This group is very basic because its positive charge is stabilized by resonance.
These nitrogens have a relatively weak affinity for an H+ and are only partly positive at neutral pH.
THE AMINO ACID
The general formula of an amino acid is
α-carbon atom
amino group
carboxyl group
side chain group
R is commonly one of 20 different side chains. At pH 7 both the amino and carboxyl groups are ionized.
OPTICAL ISOMERS
The α-carbon atom is asymmetric, which allows for two mirror image (or stereo-) isomers, L and D.
L
D
Proteins consist exclusively of L-amino acids.
PEPTIDE BONDS
Amino acids are commonly joined together by an amide linkage, called a peptide bond.
Peptide bond: The four atoms in each gray box form a rigid planar unit. There is no rotation around the C-N bond.
Proteins are long polymers of amino acids linked by peptide bonds, and they are always written with the N-terminus toward the left. The sequence of this tripeptide is histidine-cysteine-valine.
amino, or N-, terminus
carboxyl, or C-, terminus
These two single bonds allow rotation, so that long chains of amino acids are very flexible.


ACIDIC SIDE CHAINS
aspartic acid
(Asp, or D)
glutamic acid
(Glu, or E)
UNCHARGED POLAR SIDE CHAINS
asparagine
(Asn, or N)
glutamine
(Gln, or Q)
Although the amide N is not charged at neutral pH, it is polar.
serine
(Ser, or S)
threonine
(Thr, or T)
tyrosine
(Tyr, or Y)
The –OH group is polar.
NONPOLAR SIDE CHAINS
alanine
(Ala, or A)
valine
(Val, or V)
leucine
(Leu, or L)
isoleucine
(Ile, or I)
proline
(Pro, or P)
(actually an imino acid)
phenylalanine
(Phe, or F)
methionine
(Met, or M)
tryptophan
(Trp, or W)
glycine
(Gly, or G)
cysteine
(Cys, or C)
Disulfide bonds can form between two cysteine side chains in proteins.

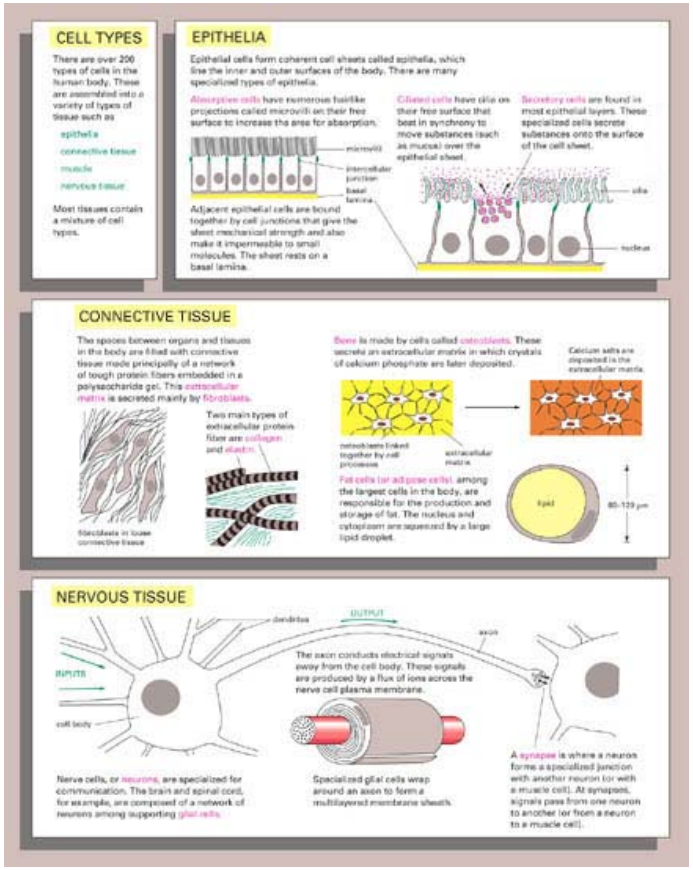
Proteins

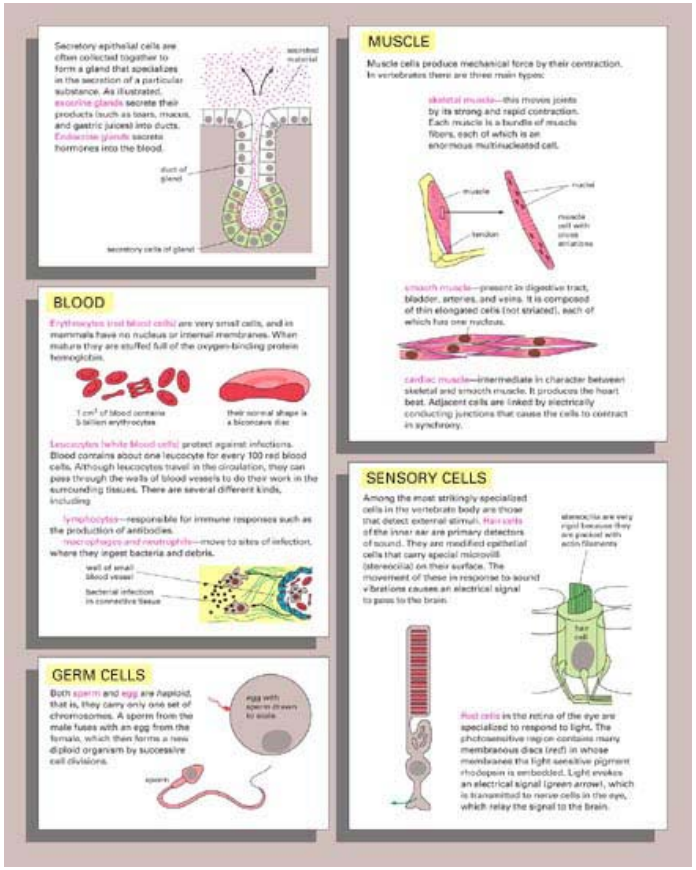

Primary protein structure
is sequence of a chain of amino acids
Amino Acids
Phe
Leu
Ser
Cys
Amino group
NH2
H—C—COOH
R
Acidic carboxyl group
R group
Amino Acid

Proteins


Primary protein structure
is sequence of a chain of amino acids
Amino Acids
Pleated sheet
Alpha helix
Secondary protein structure
occurs when the sequence of amino acids
are linked by hydrogen bonds
Pleated sheet
Tertiary protein structure
occurs when certain attractions are present
between alpha helices and pleated sheets.
Alpha helix
Quaternary protein structure
is a protein consisting of more than one
amino acid chain.

Cell types


CELL TYPES
There are over 200 types of cells in the human body. These are assembled into a variety of types of tissue such as
epithelia
connective tissue
muscle
nervous tissue
Most tissues contain a mixture of cell types.
EPITHELIA
Epithelial cells form coherent cell sheets called epithelia, which line the inner and outer surfaces of the body. There are many specialized types of epithelia.
Absorptive cells have numerous hairlike projections called microvilli on their free surface to increase the area for absorption.
Ciliated cells have cilia on their free surface that beat in synchrony to move substances (such as mucus) over the epithelial sheet.
Secretory cells are found in most epithelial layers. These specialized cells secrete substances onto the surface of the cell sheet.
microvilli
intercellular junction
basal lamina
cilia
nucleus
Adjacent epithelial cells are bound together by cell junctions that give the sheet mechanical strength and also make it impermeable to small molecules. The sheet rests on a basal lamina.
CONNECTIVE TISSUE
The spaces between organs and tissues in the body are filled with connective tissue made principally of a network of tough protein fibers embedded in a polysaccharide gel. This extracellular matrix is secreted mainly by fibroblasts.
Two main types of extracellular protein fiber are collagen and elastin.
fibroblasts in loose connective tissue
Bone is made by cells called osteoblasts. These secrete an extracellular matrix in which crystals of calcium phosphate are later deposited.
Calcium salts are deposited in the extracellular matrix
osteoblasts linked together by cell processes
extracellular matrix
Fat cells (or adipose cells), among the largest cells in the body, are responsible for the production and storage of fat. The nucleus and cytoplasm are squeezed by a large lipid droplet.
lipid
80–120 μm
NERVOUS TISSUE
dendrites
OUTPUT
axon
INPUTS
cell body
The axon conducts electrical signals away from the cell body. These signals are produced by a flux of ions across the nerve cell plasma membrane.
Nerve cells, or neurons, are specialized for communication. The brain and spinal cord, for example, are composed of a network of neurons among supporting glial cells.
Specialized glial cells wrap around an axon to form a multilayered membrane sheath.
A synapse is where a neuron forms a specialized junction with another neuron (or with a muscle cell). At synapses, signals pass from one neuron to another (or from a neuron to a muscle cell).


Secretory epithelial cells are often collected together to form a gland that specializes in the secretion of a particular substance. As illustrated, exocrine glands secrete their products (such as tears, mucus, and gastric juices) into ducts. Endocrine glands secrete hormones into the blood.
secreted material
duct of gland
secretory cells of gland
MUSCLE
Muscle cells produce mechanical force by their contraction. In vertebrates there are three main types:
skeletal muscle—this moves joints by its strong and rapid contraction. Each muscle is a bundle of muscle fibers, each of which is an enormous multinucleated cell.
muscle
nuclei
tendon
muscle cell with cross striations
smooth muscle—present in digestive tract, bladder, arteries, and veins. It is composed of thin elongated cells (not striated), each of which has one nucleus.
cardiac muscle—intermediate in character between skeletal and smooth muscle. It produces the heart beat. Adjacent cells are linked by electrically conducting junctions that cause the cells to contract in synchrony.
BLOOD
Erythrocytes (red blood cells) are very small cells, and in mammals have no nucleus or internal membranes. When mature they are stuffed full of the oxygen-binding protein hemoglobin.
1 cm³ of blood contains 5 billion erythrocytes
their normal shape is a biconcave disc
Leucocytes (white blood cells) protect against infections. Blood contains about one leucocyte for every 100 red blood cells. Although leucocytes travel in the circulation, they can pass through the walls of blood vessels to do their work in the surrounding tissues. There are several different kinds, including:
lymphocytes—responsible for immune responses such as the production of antibodies.
macrophages and neutrophils—move to sites of infection, where they ingest bacteria and debris.
wall of small blood vessel
bacterial infection in connective tissue
GERM CELLS
Both sperm and egg are haploid, that is, they carry only one set of chromosomes. A sperm from the male fuses with an egg from the female, which then forms a new diploid organism by successive cell divisions.
egg with sperm drawn to scale
sperm
SENSORY CELLS
Among the most strikingly specialized cells in the vertebrate body are those that detect external stimuli. Hair cells of the inner ear are primary detectors of sound. They are modified epithelial cells that carry special microvilli (stereocilia) on their surface. The movement of these in response to sound vibrations causes an electrical signal to pass to the brain.
stereocilia are very rigid because they are packed with actin filaments
hair cell
Rod cells in the retina of the eye are specialized to respond to light. The photosensitive region contains many membranous discs (red) in whose membranes the light-sensitive pigment rhodopsin is embedded. Light evokes an electrical signal (green arrow), which is transmitted to nerve cells in the eye, which relay the signal to the brain.

Differential expression

- Each cell contains a complete copy of the organism's genome.
- Cells are of many different types and states

  E.g. blood, nerve, and skin cells, dividing cells, cancerous cells, etc.
- What makes the cells different?
- **Differential gene expression**, i.e., **when**, **where**, and in **what quantity** each gene is expressed.
- On average, 40% of our genes are expressed at any given time.

Central dogma

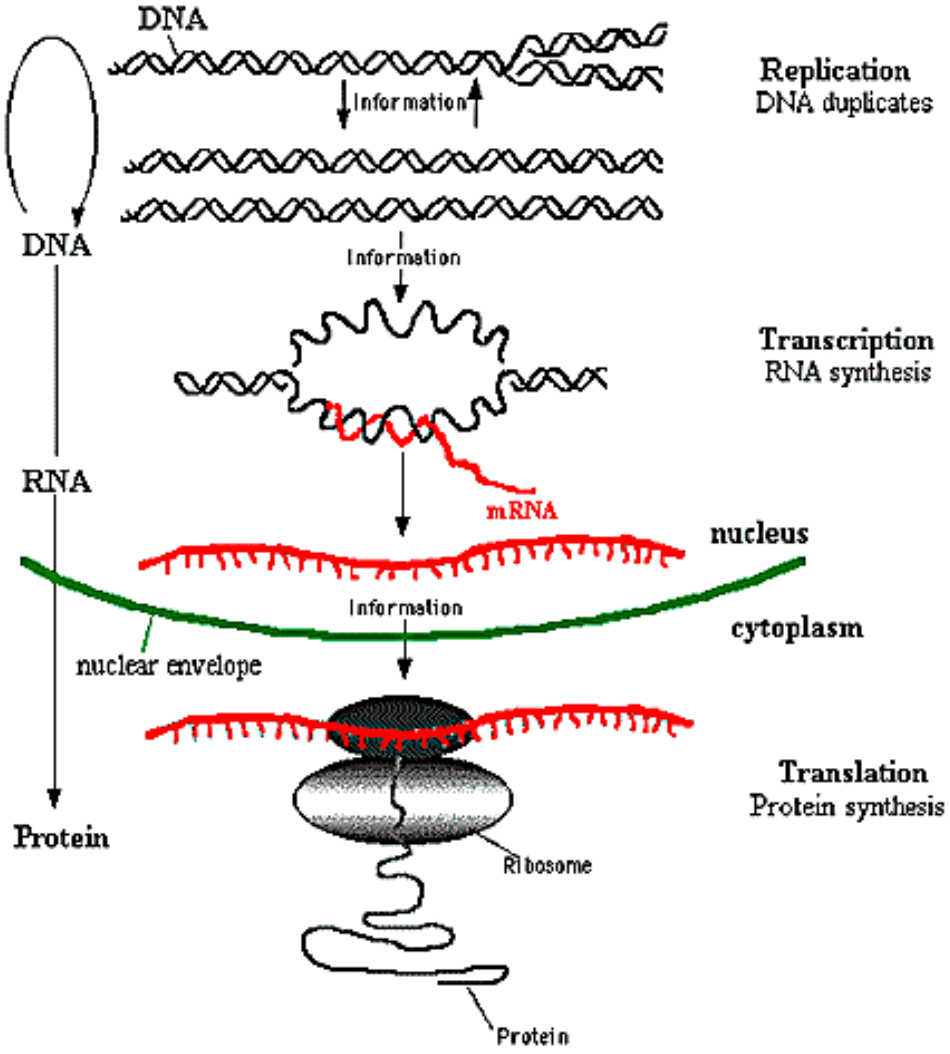
The **expression** of the genetic information stored in the DNA molecule occurs in two stages:

- (i) **transcription**, during which DNA is transcribed into mRNA;
- (ii) **translation**, during which mRNA is translated to produce a protein.

DNA → mRNA → protein

Other important aspects of regulation: methylation, alternative splicing, etc.

Central dogma



The Central Dogma of Molecular Biology

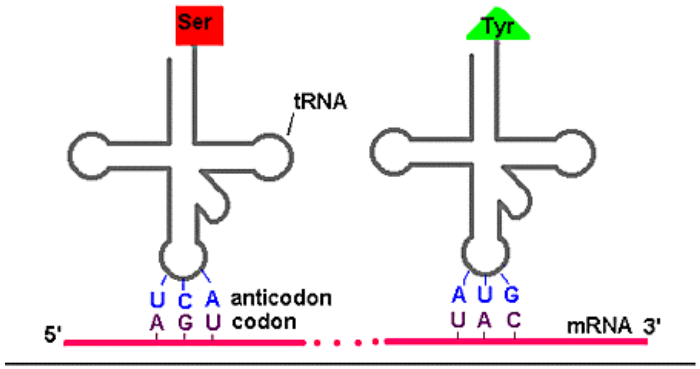
RNA

- A **ribonucleic acid** or **RNA** molecule is a nucleic acid similar to DNA, but
  - single-stranded;
  - ribose sugar rather than deoxyribose sugar;
  - **uracil (U)** replaces thymine (T) as one of the bases.
- RNA plays an important role in protein synthesis and other chemical activities of the cell.
- Several classes of RNA molecules, including **messenger RNA** (**mRNA**), transfer RNA (tRNA), ribosomal RNA (rRNA), and other small RNAs.

The genetic code

- **DNA:** sequence of **four** different nucleotides.
- **Proteins:** sequence of **twenty** different amino acids.
- The correspondence between DNA's four-letter alphabet and a protein's twenty-letter alphabet is specified by the **genetic code**, which relates nucleotide triplets or **codons** to **amino acids**.

The genetic code



2nd base in codon

1st base in codon	U	C	A	G	3rd base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

The Genetic Code

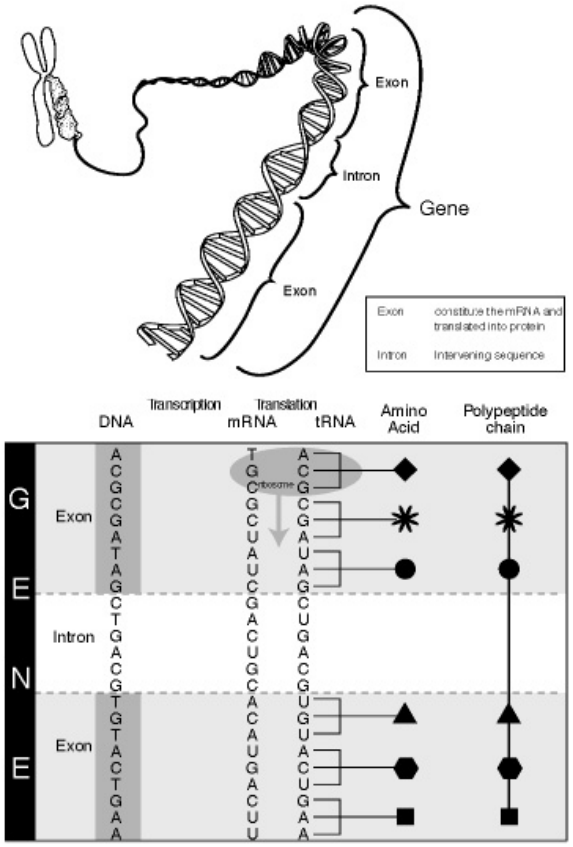
Mapping between codons and amino acids is many-to-one: 64 codons but only 20 a.a..

Third base in codon is often redundant,
e.g., stop codons.

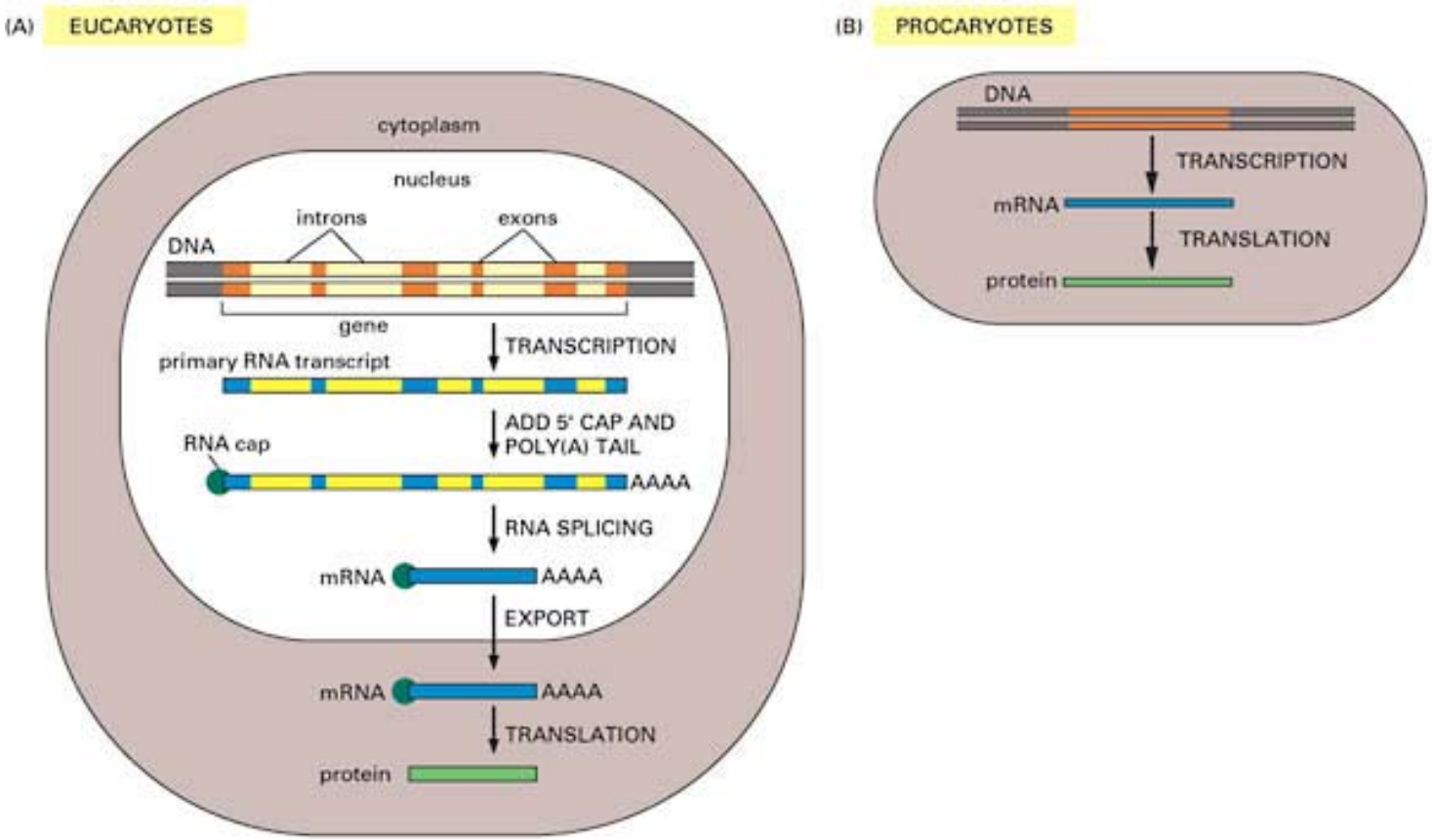
Exons and introns

- Genes comprise only about 2% of the human genome; the rest consists of non-coding regions, whose functions may include providing chromosomal structural integrity and regulating when, where, and in what quantity proteins are made (regulatory regions).
- The terms **exon** and **intron** refer to coding (translated into a protein) and non-coding DNA, respectively.

Exons and introns



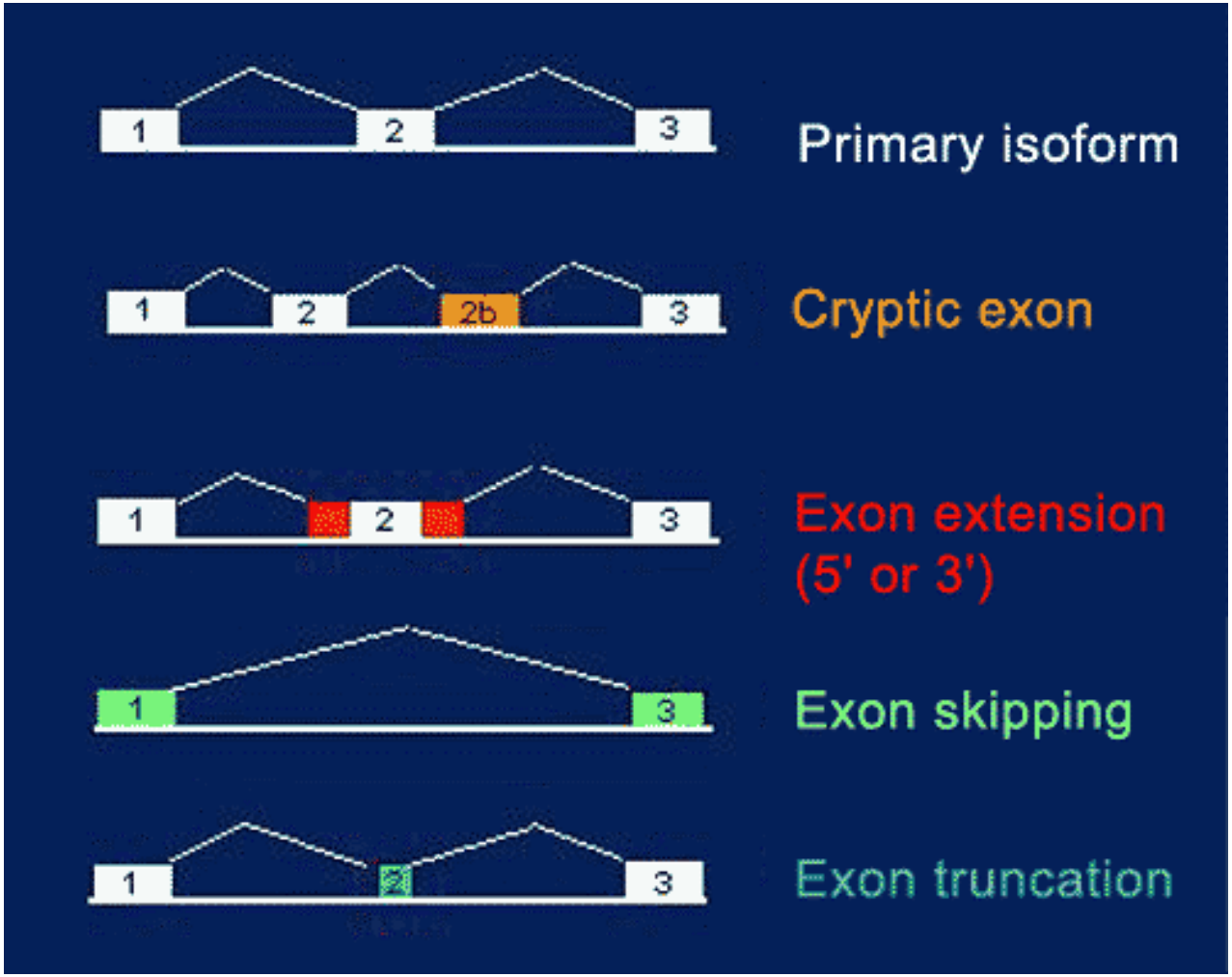
Splicing


(A) EUCARYOTES
cytoplasm
nucleus
introns
exons
DNA
gene
TRANSCRIPTION
primary RNA transcript
ADD 5' CAP AND POLY(A) TAIL
RNA cap
AAAA
RNA SPLICING
mRNA
AAAA
EXPORT
mRNA
AAAA
TRANSLATION
protein
(B) PROCARYOTES
DNA
TRANSCRIPTION
mRNA
TRANSLATION
protein

Alternative splicing

- There are more than 1,000,000 different human antibodies. How is this possible with only ~30,000 genes?
- **Alternative splicing** refers to the different ways of combining a gene's exons. This can produce different forms of a protein for the same gene,
- Alternative pre-mRNA splicing is an important mechanism for regulating gene expression in higher eukaryotes.
- E.g. in humans, it is estimated that approximately 30% genes are subject to alternative splicing.

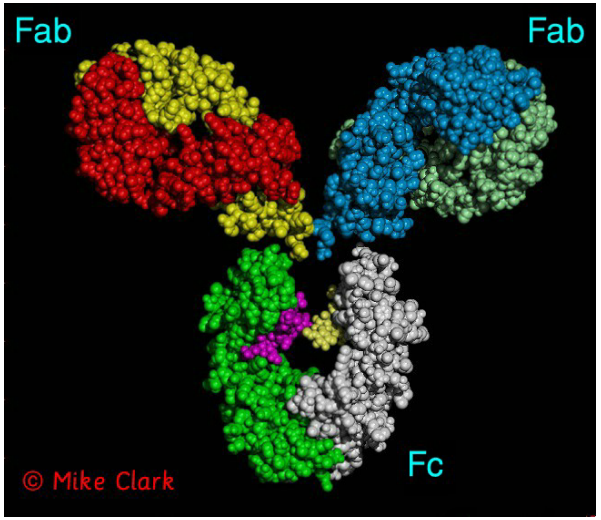
Alternative splicing



Immunoglobulin

- B cells produce antibody molecules called immunoglobulins (Ig) which fall in five broad classes.

- Diversity of Ig molecules
  - DNA sequence: recombination, mutation.
  - mRNA sequence: alternative splicing.
  - Protein structure: post-translational proteolysis, glycosylation.



IgG1

Functional genomics

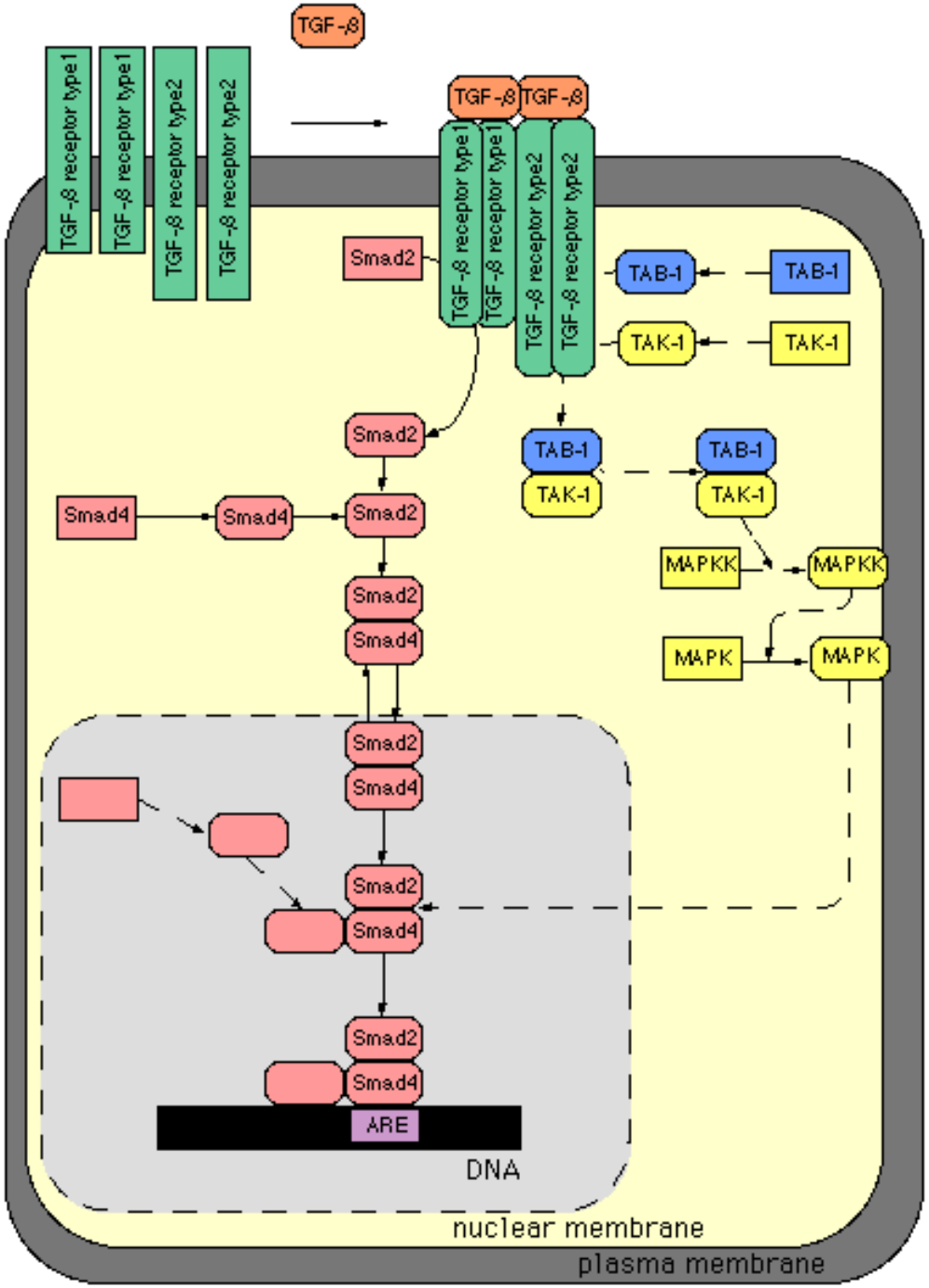
- The various **genome projects** have yielded the complete DNA sequences of many organisms.
  - E.g. human, mouse, yeast, fruitfly, etc.
  - Human: 3 billion base-pairs, 30-40 thousand genes.
- Challenge: **go from sequence to function**, i.e., define the role of each gene and understand how the genome functions as a whole.

Pathways

- The complete genome sequence doesn't tell us much about how the organism functions as a biological system.
- We need to study how different gene products function to produce various components.
- Most important activities are not the result of a single molecule but depend on the coordinated effects of multiple molecules.

TFG-β pathway

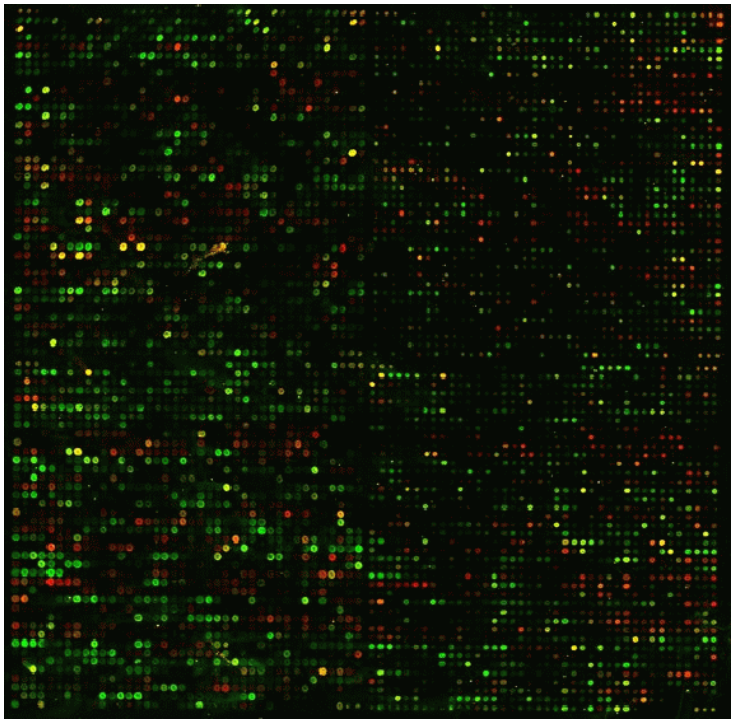
- TGF-β (transforming growth factor beta) plays an essential role in the control of development and morphogenesis in multicellular organisms.
- This is done through SMADS, a family of signal transducers and transcriptional activators.


TGF-β
TGF-β receptor type1
TGF-β receptor type1
TGF-β receptor type2
TGF-β receptor type2
TGF-β
TGF-β
TGF-β receptor type1
TGF-β receptor type1
TGF-β receptor type2
TGF-β receptor type2
Smad2
TAB-1
TAB-1
TAK-1
TAK-1
Smad2
TAB-1
TAB-1
TAK-1
TAK-1
Smad4
Smad4
Smad2
MAPKK
MAPKK
Smad2
Smad4
MAPK
MAPK
Smad2
Smad4
Smad2
Smad4
Smad2
Smad4
ARE
DNA
nuclear membrane
plasma membrane

Pathways

- http://www.grt.kyushu-u.ac.jp/spad/
- There are many open questions regarding the relationship between expression level and pathways.
- It is not clear whether expression level data will be informative.

DNA microarrays

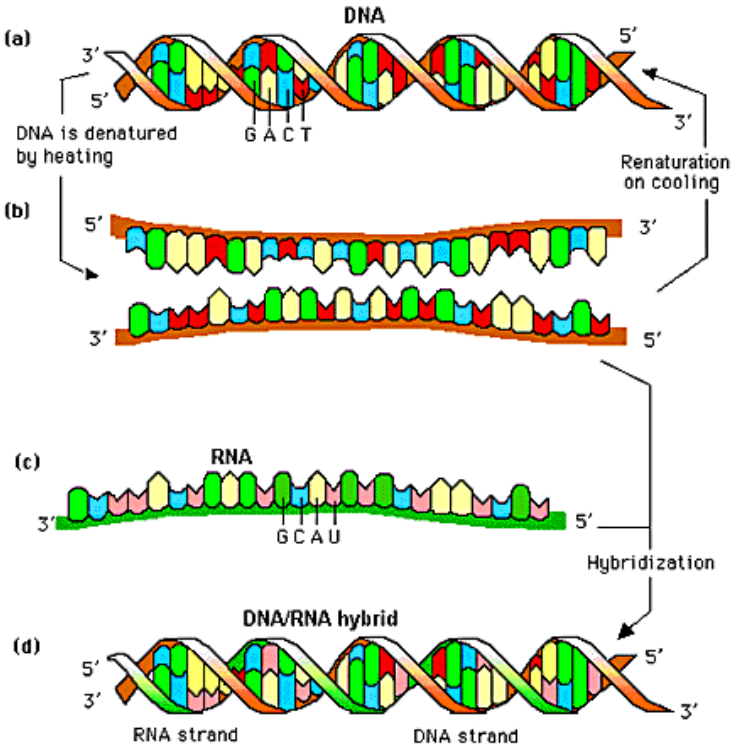


DNA microarrays

DNA microarrays rely on the **hybridization** properties of nucleic acids to monitor DNA or RNA abundance on a genomic scale in different types of cells.

The ancestor of microarrays: the Northern blot.

Nucleic acid hybridization



Nucleic Acid Hybridization

Gene expression assays

The main types of gene expression assays:

- Serial analysis of gene expression (SAGE);
- **Short oligonucleotide arrays (Affymetrix);**
- Long oligonucleotide arrays (Agilent Inkjet);
- Fibre optic arrays (Illumina);
- **cDNA arrays (Brown/Botstein).**

Applications of microarrays

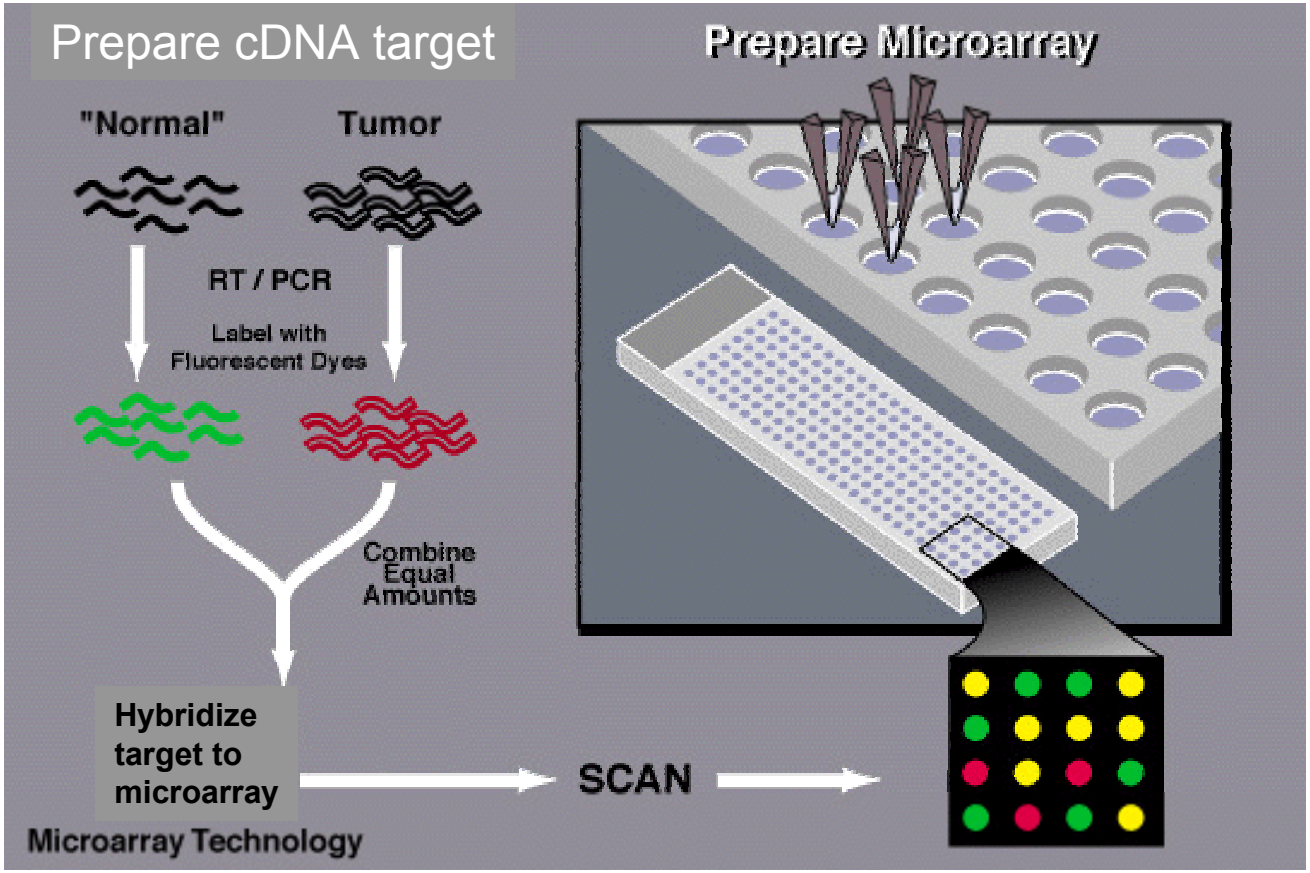
- **Measuring transcript abundance (cDNA arrays);**
- Genotyping;
- Estimating DNA copy number (CGH);
- Determining identity by descent (GMS);
- Measuring mRNA decay rates;
- Identifying protein binding sites;
- Determining sub-cellular localization of gene products;
- …

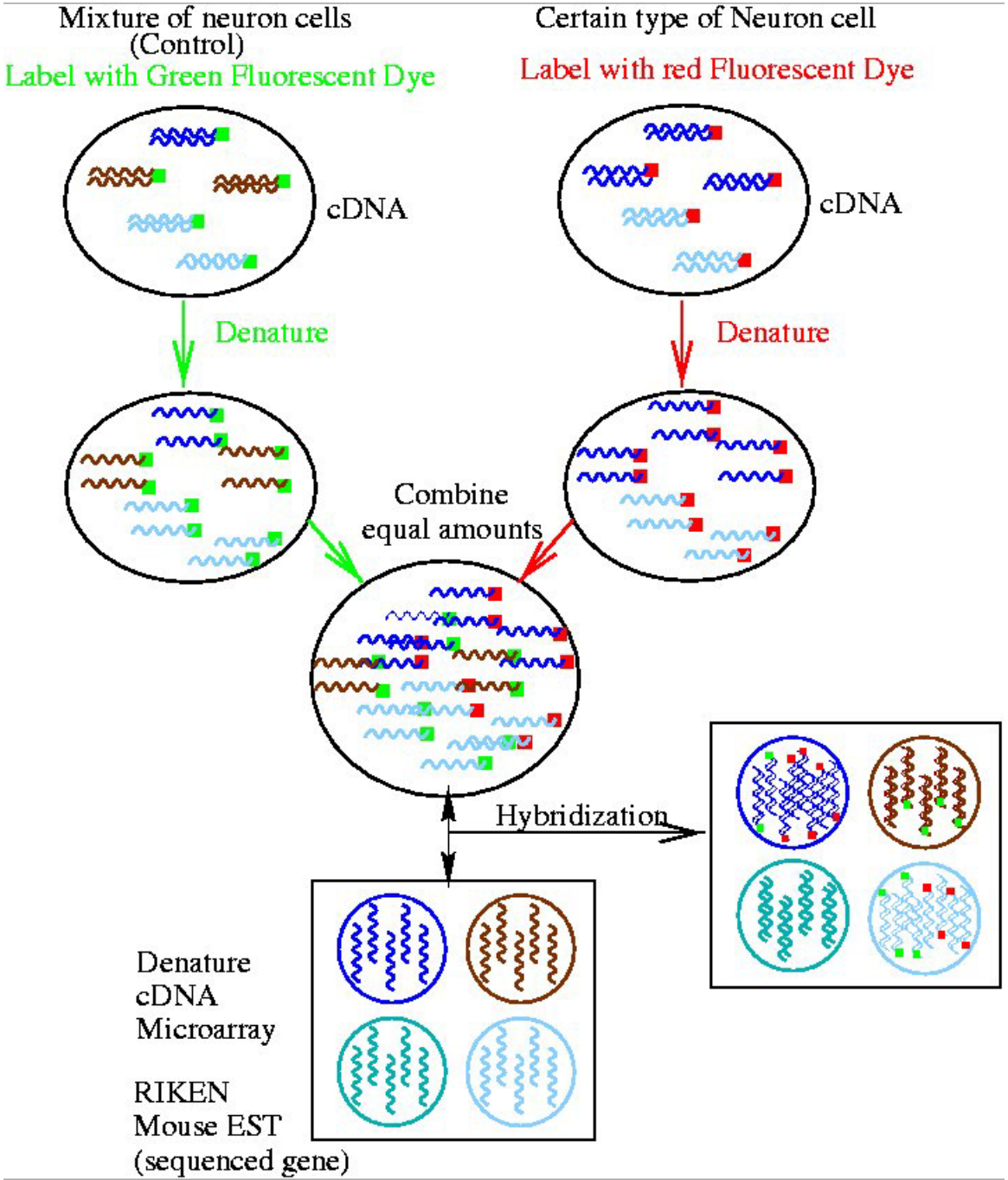
Applications of microarrays

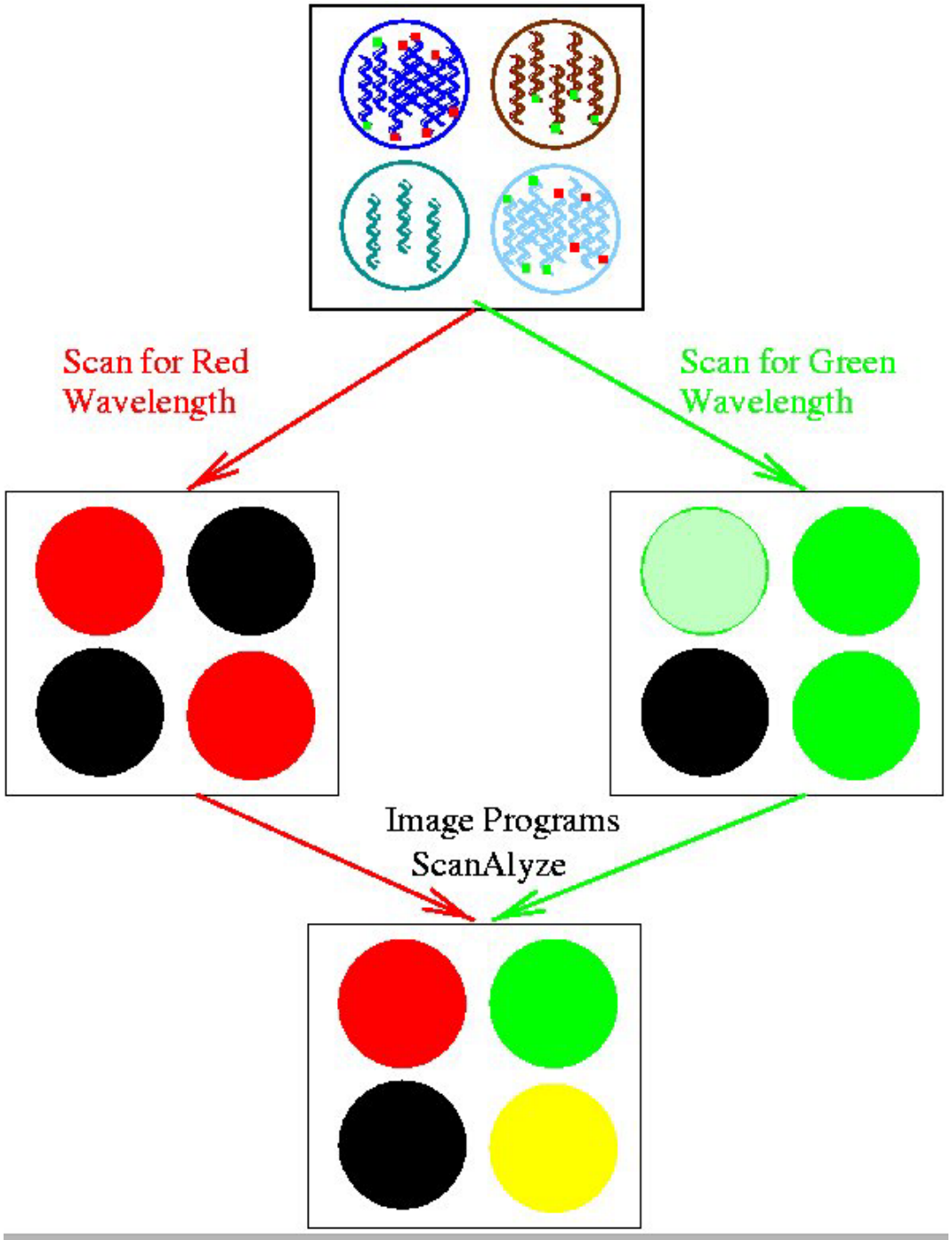
- **Cancer research:** Molecular characterization of tumors on a genomic scale

  → more reliable diagnosis and effective treatment of cancer.
- **Immunology:** Study of host genomic responses to bacterial infections; reversing immunity.
- …

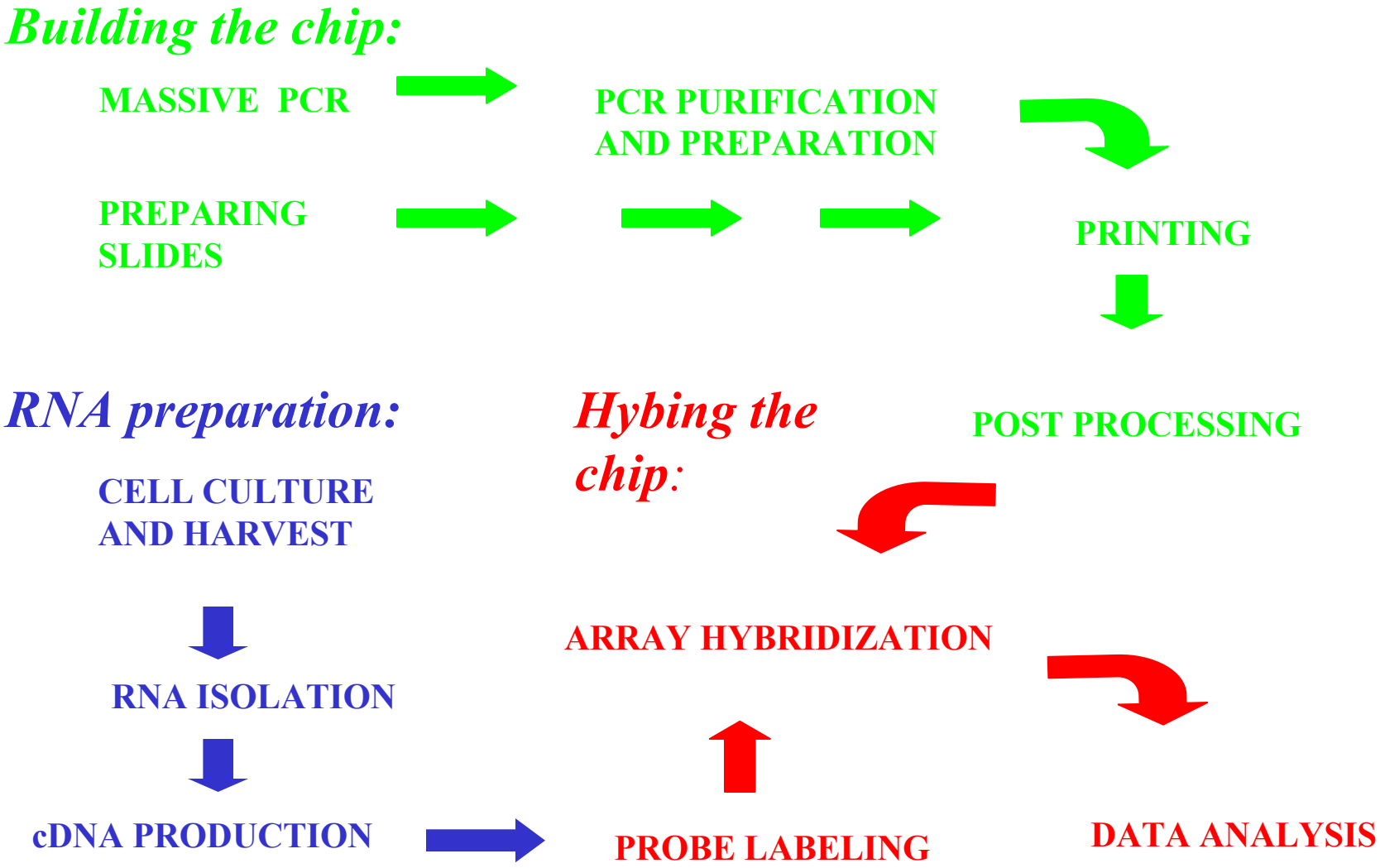
cDNA microarray experiment



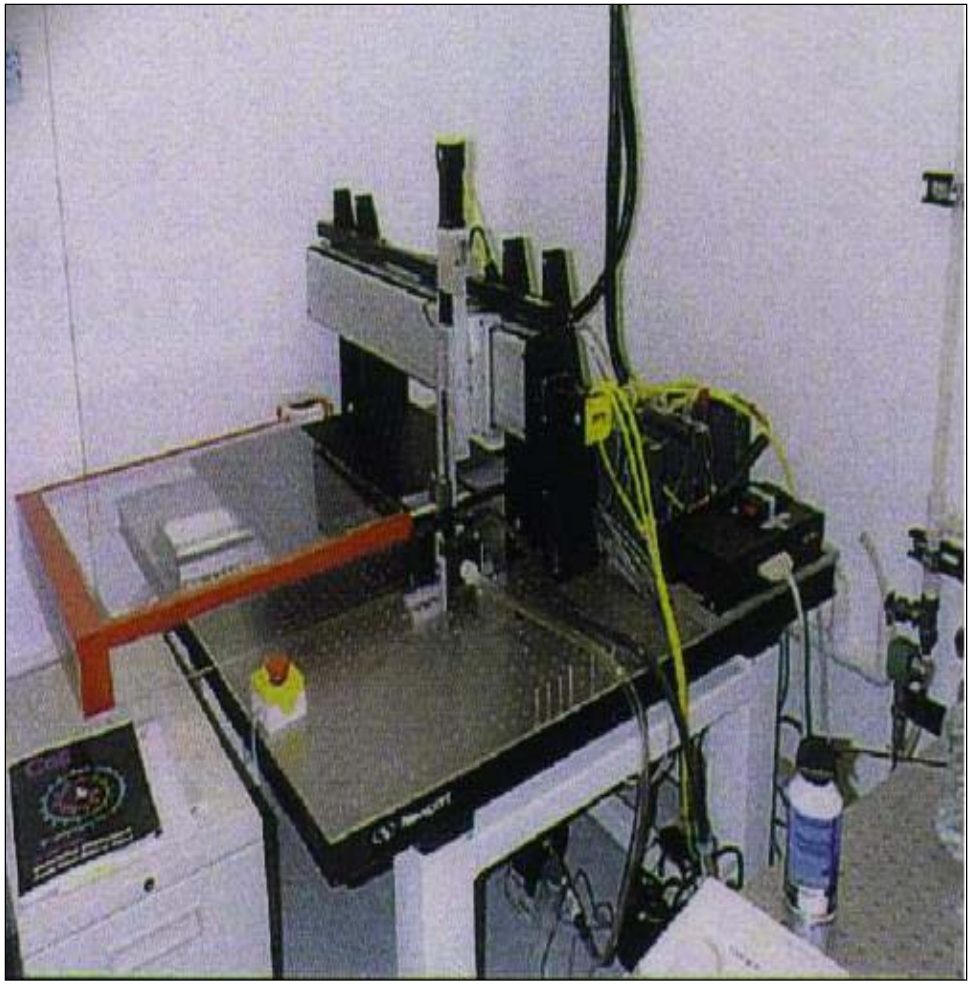

Mixture of neuron cells
(Control)
Label with Green Fluorescent Dye
Certain type of Neuron cell
Label with red Fluorescent Dye
cDNA
cDNA
Denature
Denature
Combine
equal amounts
Hybridization
Denature
cDNA
Microarray
RIKEN
Mouse EST
(sequenced gene)


Scan for Red
Wavelength
Scan for Green
Wavelength
Image Programs
ScanAlyze

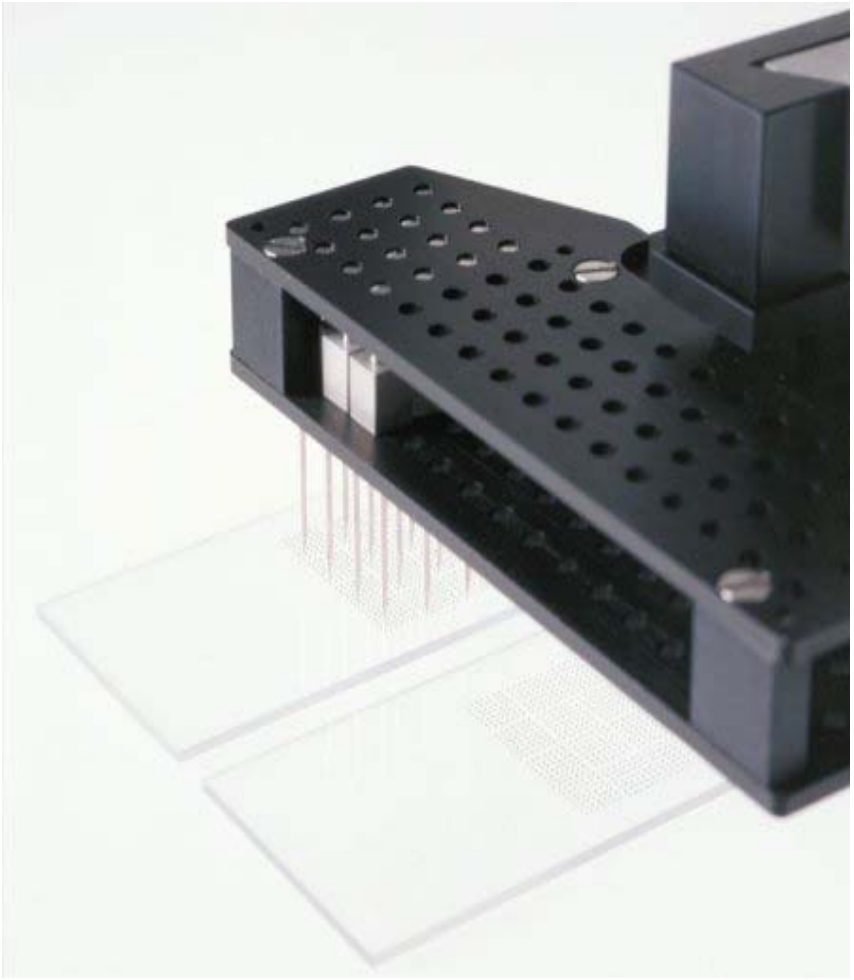
The process


Building the chip:
MASSIVE PCR
PCR PURIFICATION AND PREPARATION
PREPARING SLIDES
PRINTING
POST PROCESSING
RNA preparation:
Hybing the chip:
CELL CULTURE AND HARVEST
RNA ISOLATION
cDNA PRODUCTION
PROBE LABELING
ARRAY HYBRIDIZATION
DATA ANALYSIS

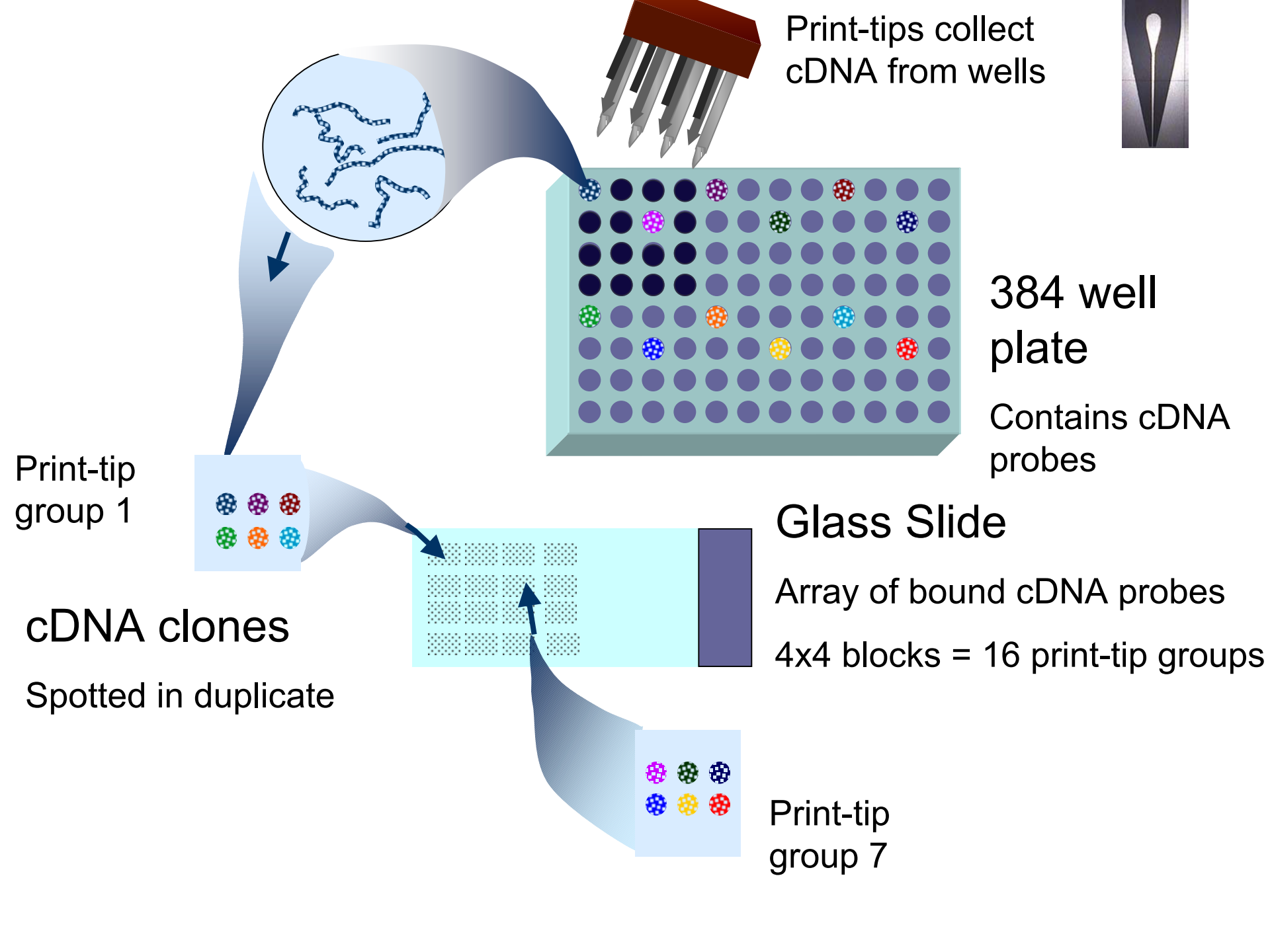
The arrayer



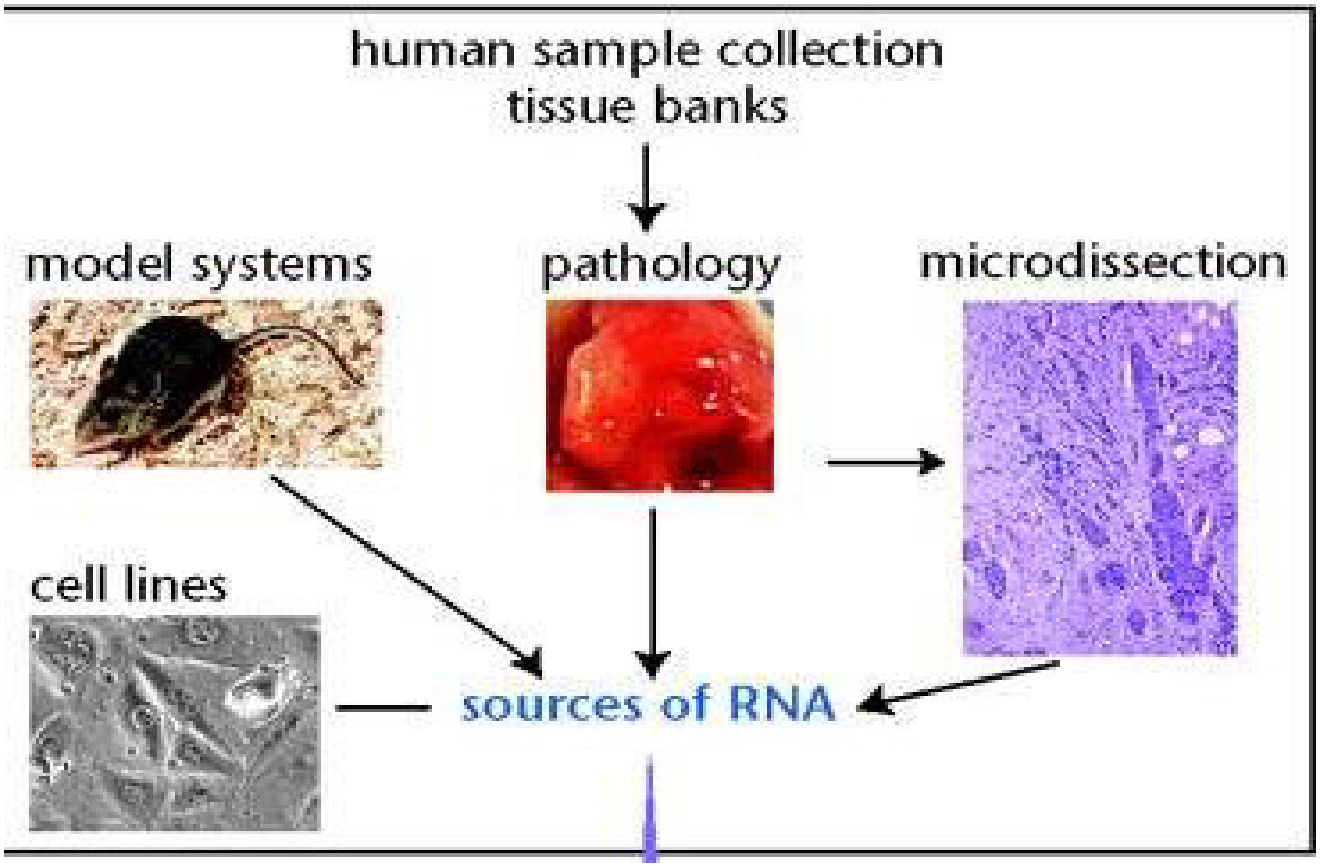
Ngai Lab arrayer , UC Berkeley



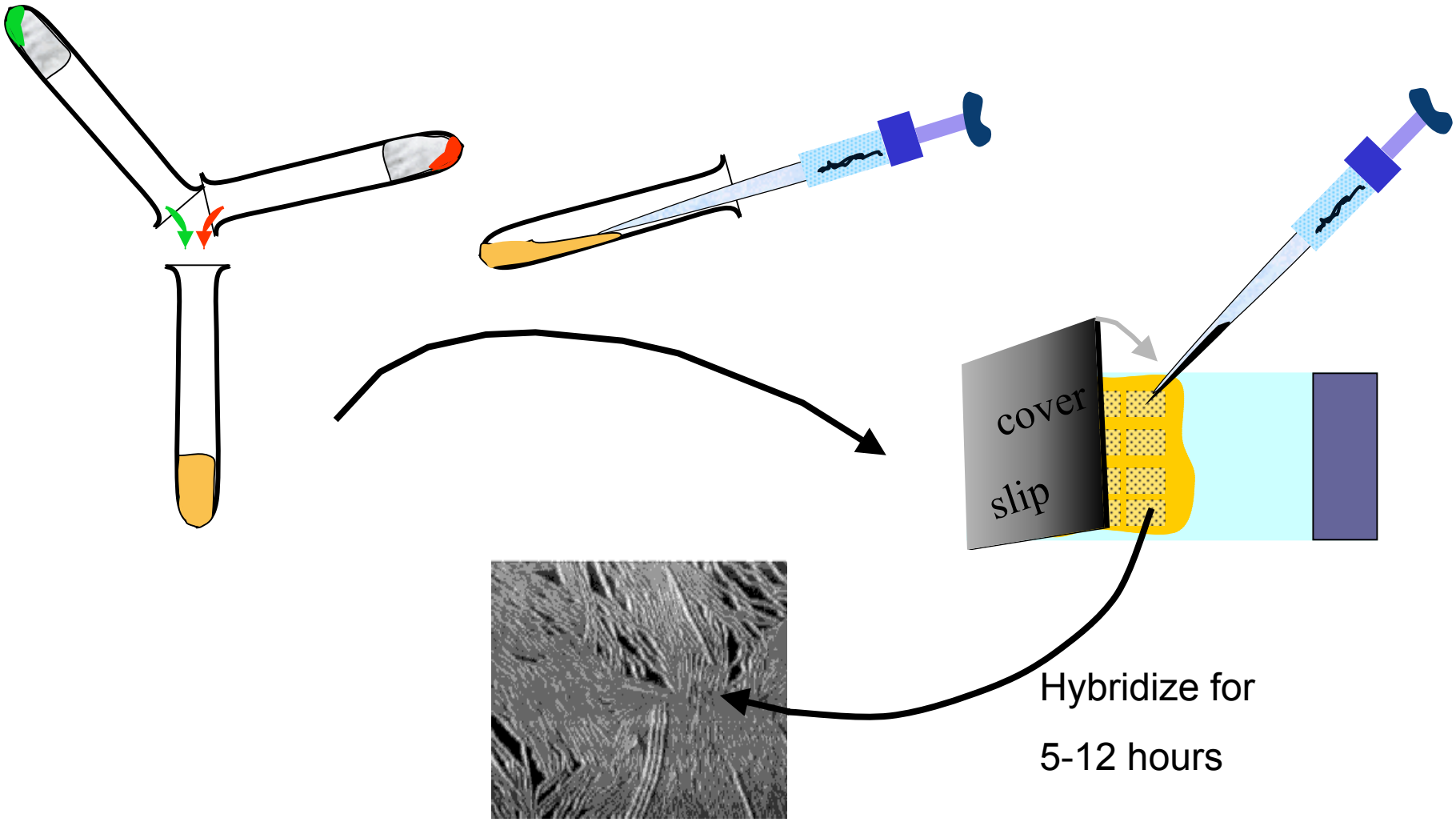
Print-tip head


Print-tips collect cDNA from wells
384 well plate
Contains cDNA probes
Print-tip group 1
Glass Slide
Array of bound cDNA probes
4x4 blocks = 16 print-tip groups
cDNA clones
Spotted in duplicate
Print-tip group 7

Sample preparation

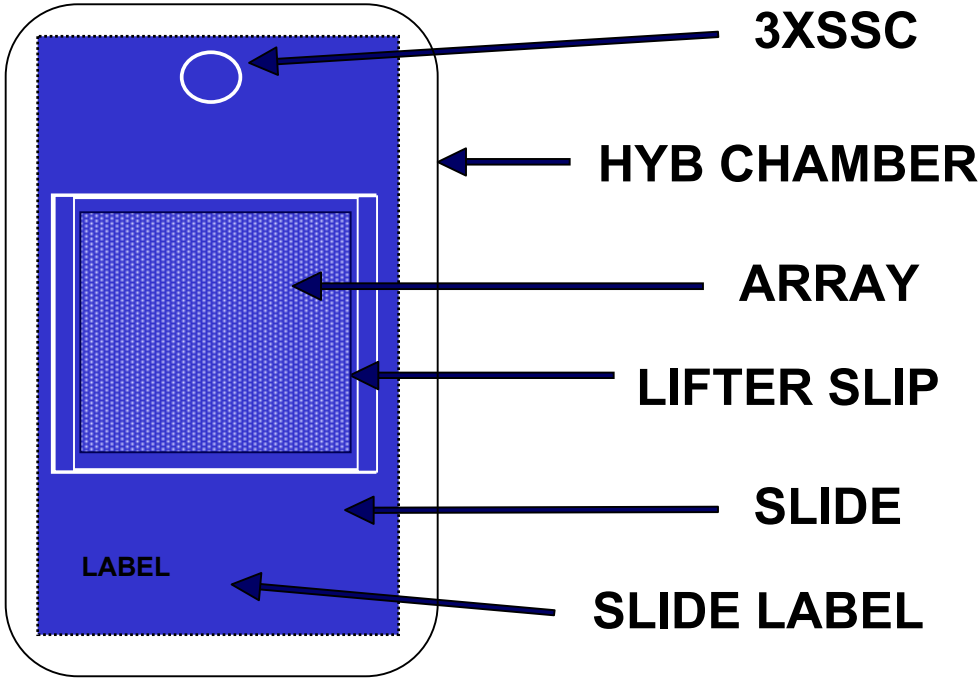


Hybridization



Binding of cDNA target samples to cDNA probes on the slide

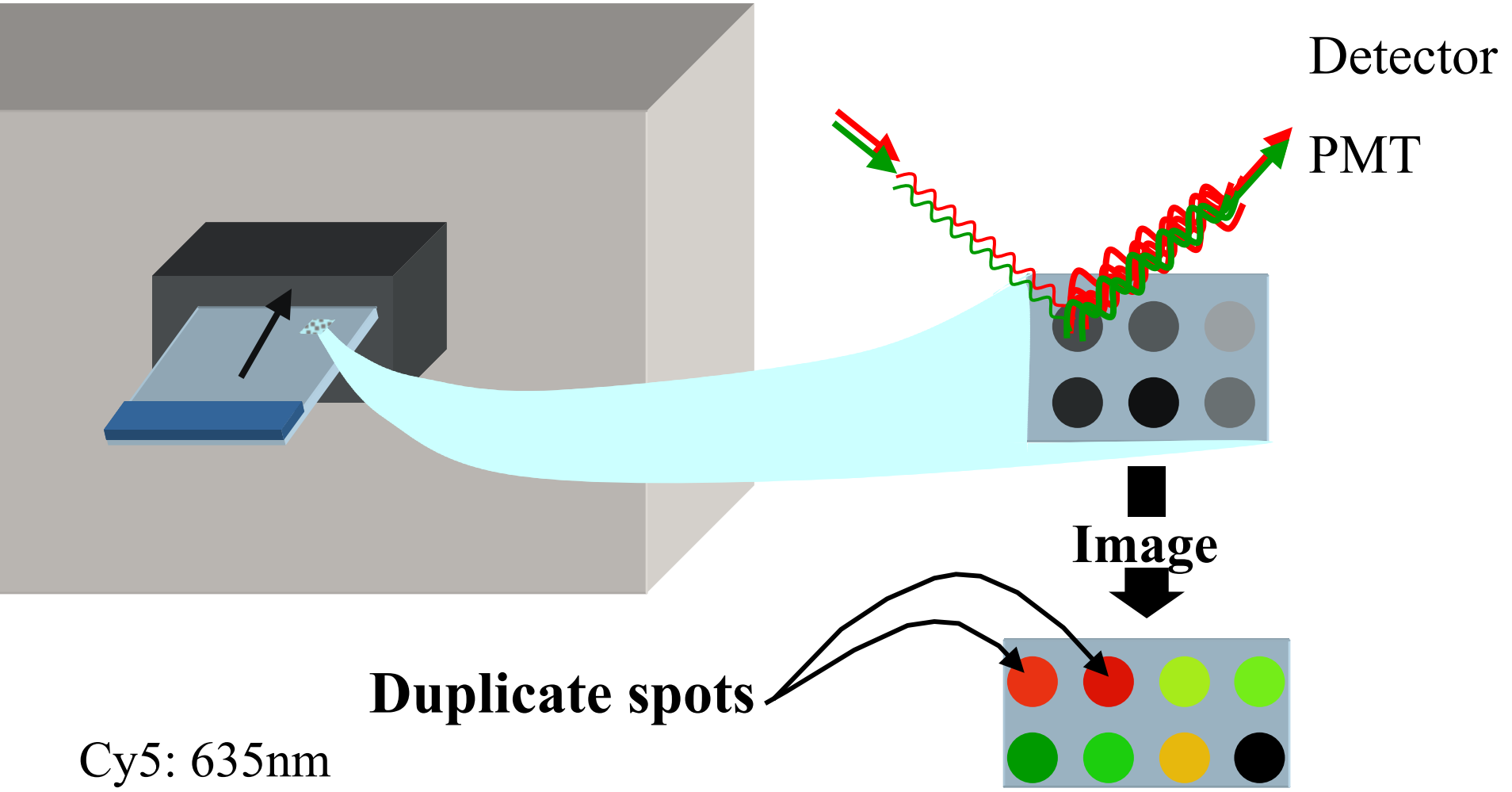
Hybridization chamber



- Humidity
- Temperature
- Formamide

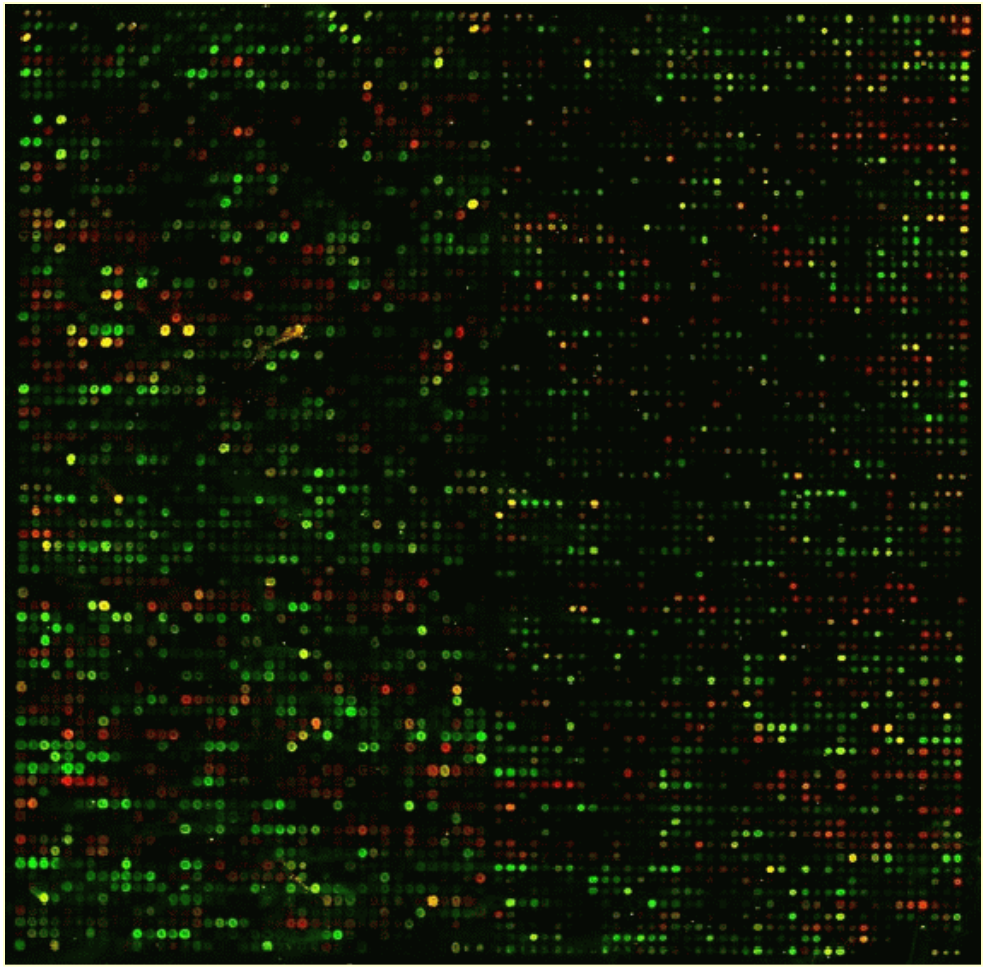
(Lowers the Tmp)

Scanning



Cy5: 635nm
Cy3: 532nm

RGB overlay of Cy3 and Cy5 images



Raw data

- Human cDNA arrays
  - ~43K spots;
  - 16–bit TIFFs: ~ 20Mb per channel;
  - ~ 2,000 x 5,500 pixels per image;
  - Spot separation: ~ 136um;
  - For a “typical” array:

    Mean = 43, med = 32, SD = 26 pixels per spots

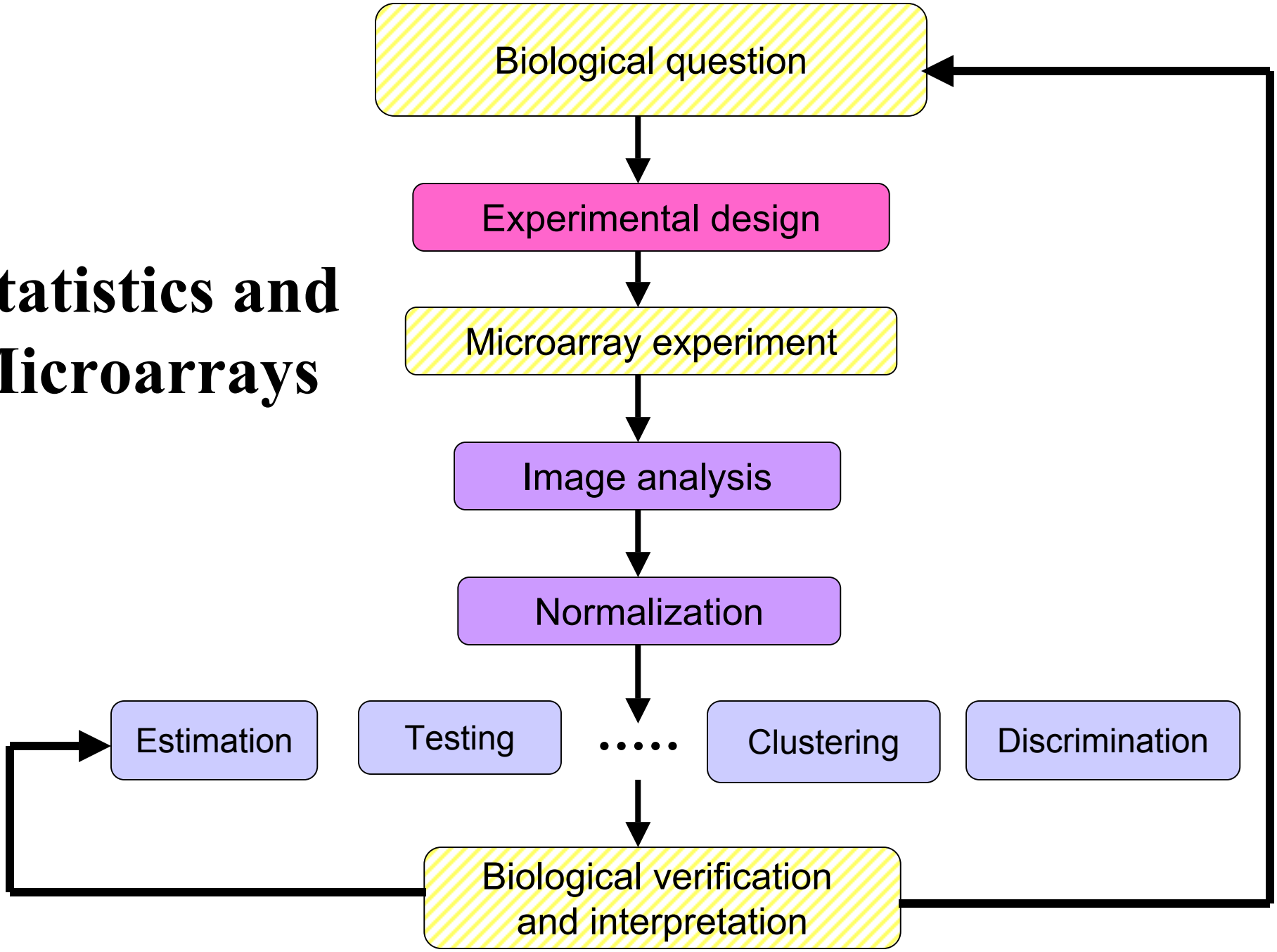
WWW resources

- Complete guide to “microarraying”
  http://cmgm.stanford.edu/pbrown/mguide/
  http://www.microarrays.org
  - Parts and assembly instructions for printer and scanner;
  - Protocols for sample prep;
  - Software;
  - Forum, etc.
- Animation:
  http://www.bio.davidson.edu/courses/genomics/chip/chip.html

Integration of biological data

- Expression, sequence, structure, annotation.
- Integration will depend on our using a common language and will rely on database methodology as well as statistical analyses.
- This area is largely unexplored.

Statistics and Microarrays



Statistical computing

Everywhere …

- for statistical design and analysis:

  pre-processing, estimation, pattern discovery and recognition, etc.

- for integration with biological information resources (in-house and external databases).

Road map

- Lecture1, Part II: cDNA arrays
  - Pre-processing: Image analysis;
  - Pre-processing: Normalization;
  - Experimental design.

Road map

- Lecture 2: Differential expression.
- Lecture 3: Applications of HMMs to sequence analysis.
- Lecture 4: Affymetrix chips.
- Lecture 5: Classification.