Gene Expression and Microarrays

Garrett M. Dancik, Ph.D.

Overview of gene expression

Central Dogma of Molecular Biology:

transcription   translation

DNA → RNA → Protein

T
A
G
C

Thymine
Adenine
Guanine
Cytosine

4-character alphabet   →   20-character alphabet
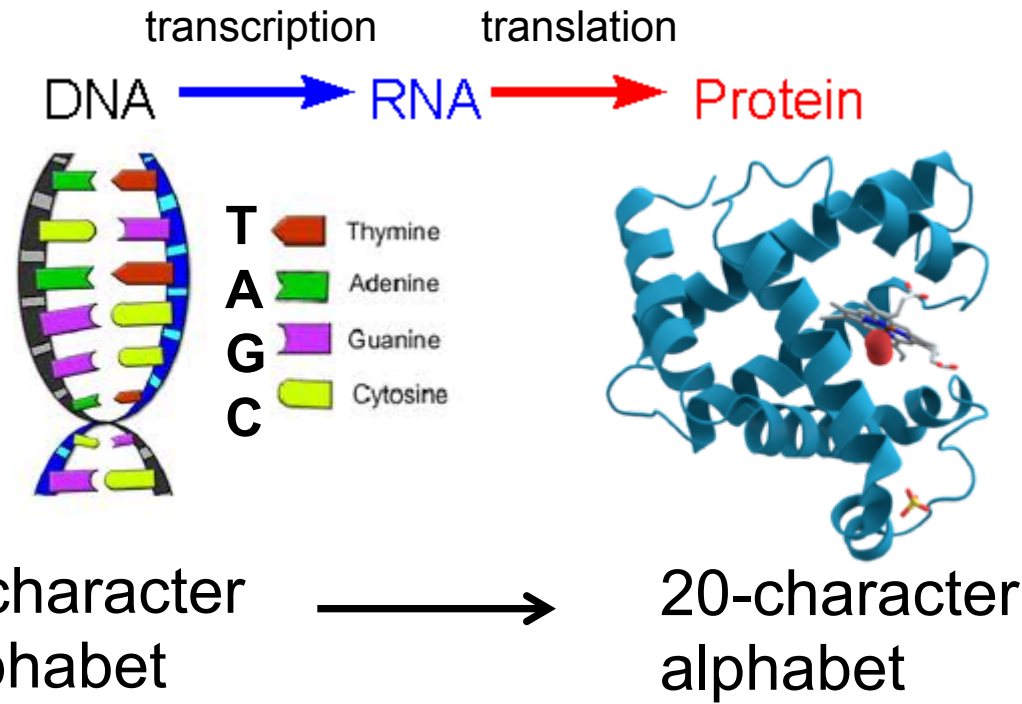
- A *gene* is a unit of hereditary (DNA) that makes a functional RNA or protein
- The human genome is 3 billion characters long
- The human genome contains ~ 20,000 genes

Overview of gene expression: DNA → RNA → Protein

- Genes are made of DNA, a **nucleic acid** made of monomers called nucleotides

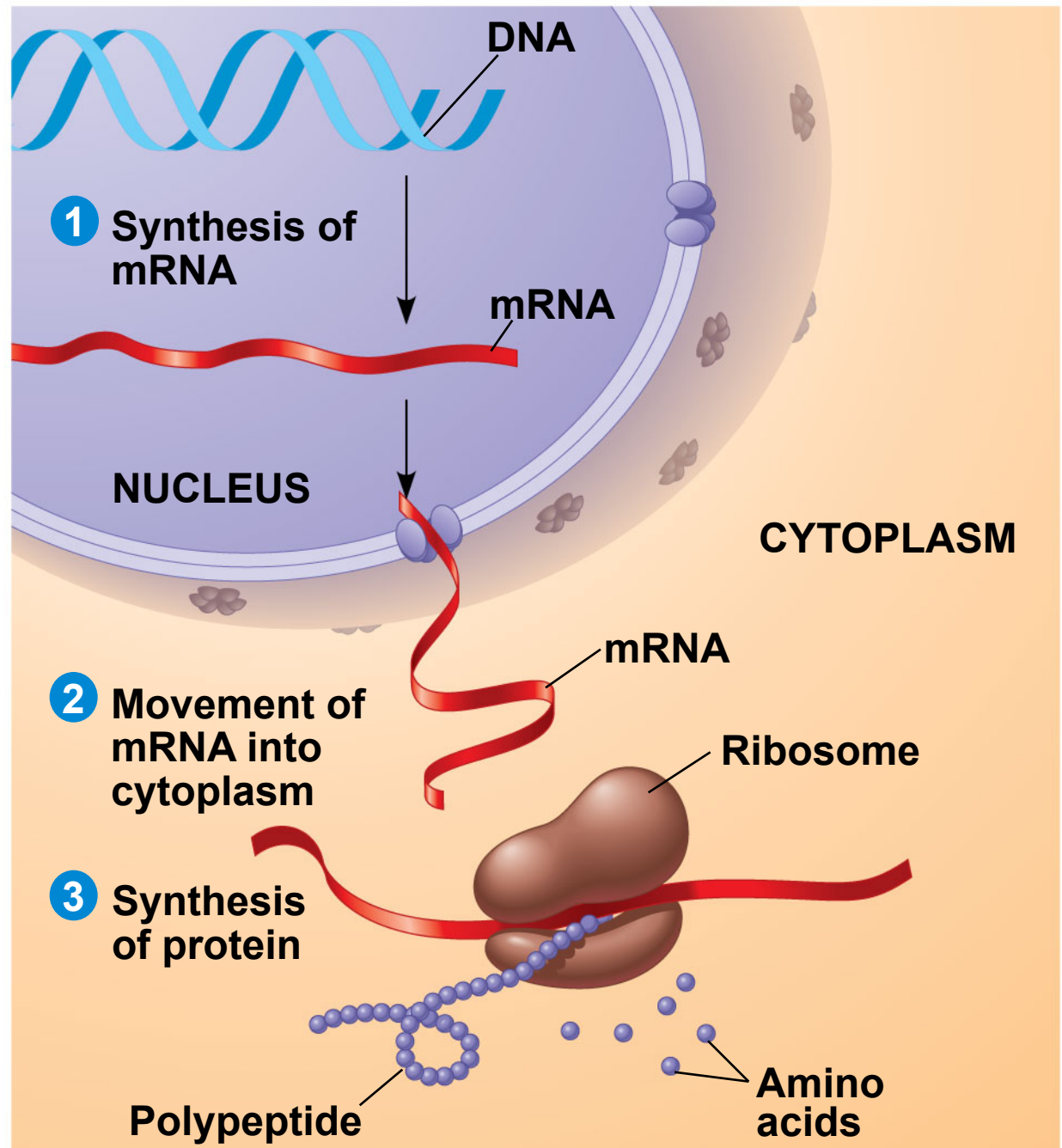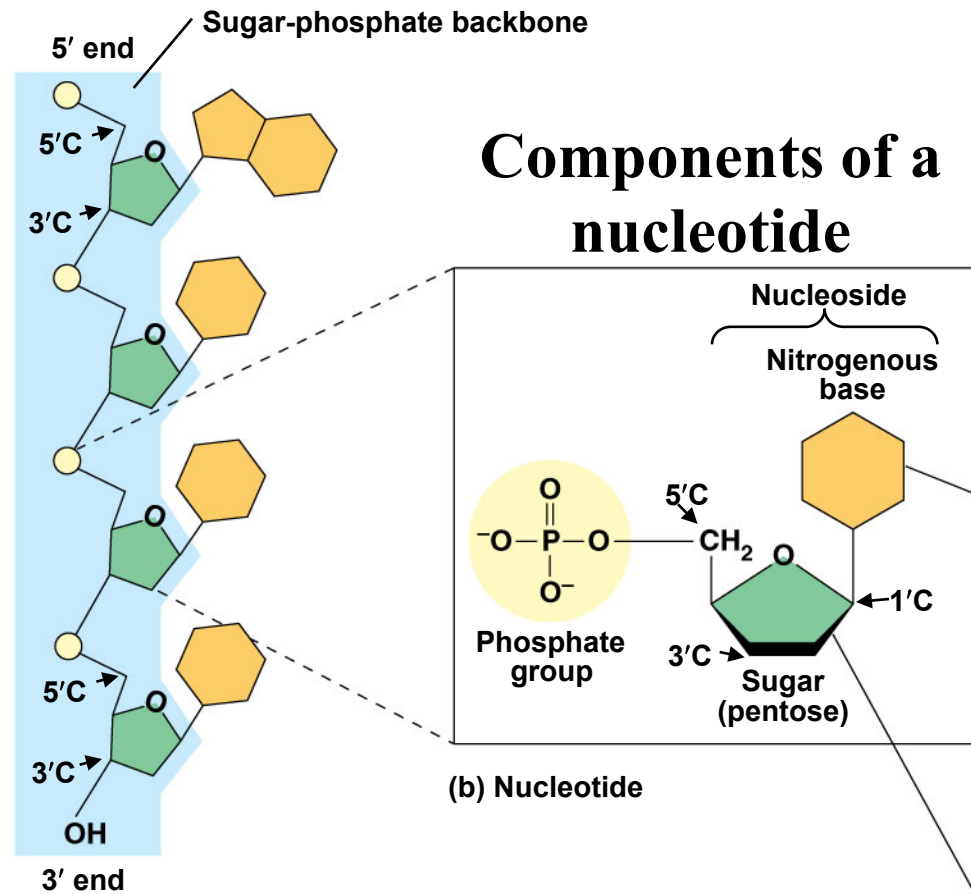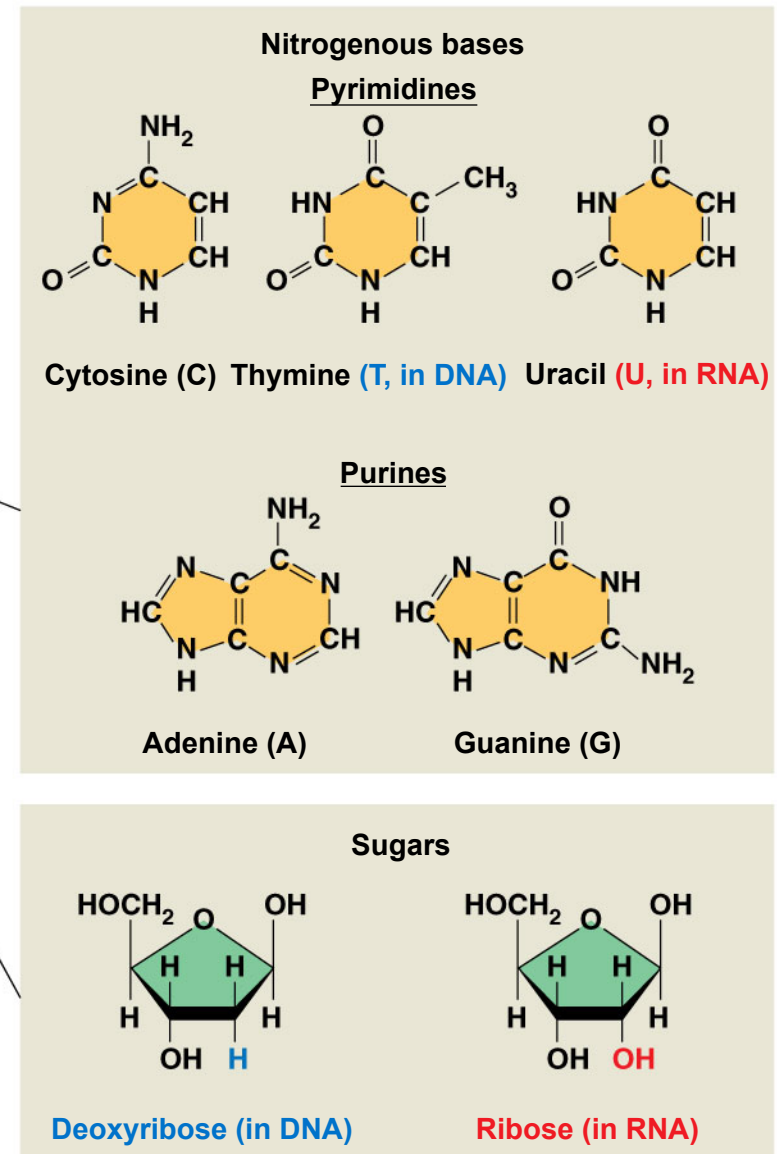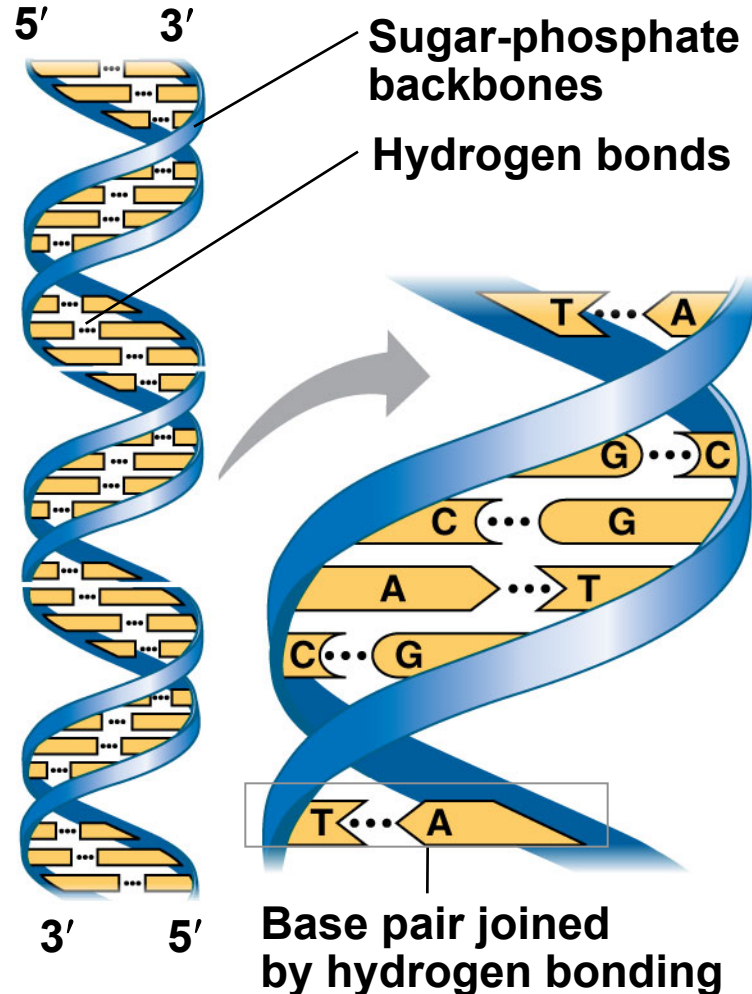- A gene is a unit of inheritance that codes for the amino acid sequence of a polypeptide

DNA

1 **Synthesis of mRNA**

mRNA

NUCLEUS

CYTOPLASM

mRNA

2 **Movement of mRNA into cytoplasm**

Ribosome

3 **Synthesis of protein**

Polypeptide

Amino acids

Figure 5.26

Components of a nucleotide

Sugar-phosphate backbone

5′ end
5′C
3′C

3′C
OH

3′ end

(a) Polynucleotide, or nucleic acid

Nucleoside

Nitrogenous base

5′C

CH₂

Phosphate group

3′C

1′C

Sugar (pentose)

(b) Nucleotide

Nitrogenous bases

Pyrimidines

Cytosine (C) Thymine (T, in DNA) Uracil (U, in RNA)

Purines

NH_2

Adenine (A) Guanine (G)

Sugars

$HOCH_2$ OH $HOCH_2$ OH

OH H   OH OH

Deoxyribose (in DNA)   Ribose (in RNA)

(c) Nucleoside components

In DNA, the sugar is **deoxyribose**; in RNA, the sugar is **ribose**

4

Nucleic Acids are made up of nucleotides

Figure 5.27

5′ **3′**

Sugar-phosphate backbones

Hydrogen bonds

T ···· A

G ···· C

C ···· G

A ···· T

C ···· G

T ···· A

3′ **5′**

Base pair joined by hydrogen bonding

(a) DNA

© 2011 Pearson Education, Inc.

- Complementary base pairing
  - The nitrogenous bases in DNA pair up and form hydrogen bonds: adenine (A) always with thymine (T), and guanine (G) always with cytosine (C)
  - Complementary pairing can also occur between two RNA molecules or between parts of the same molecule
- In RNA, thymine is replaced by uracil (U) so A and U pair

5
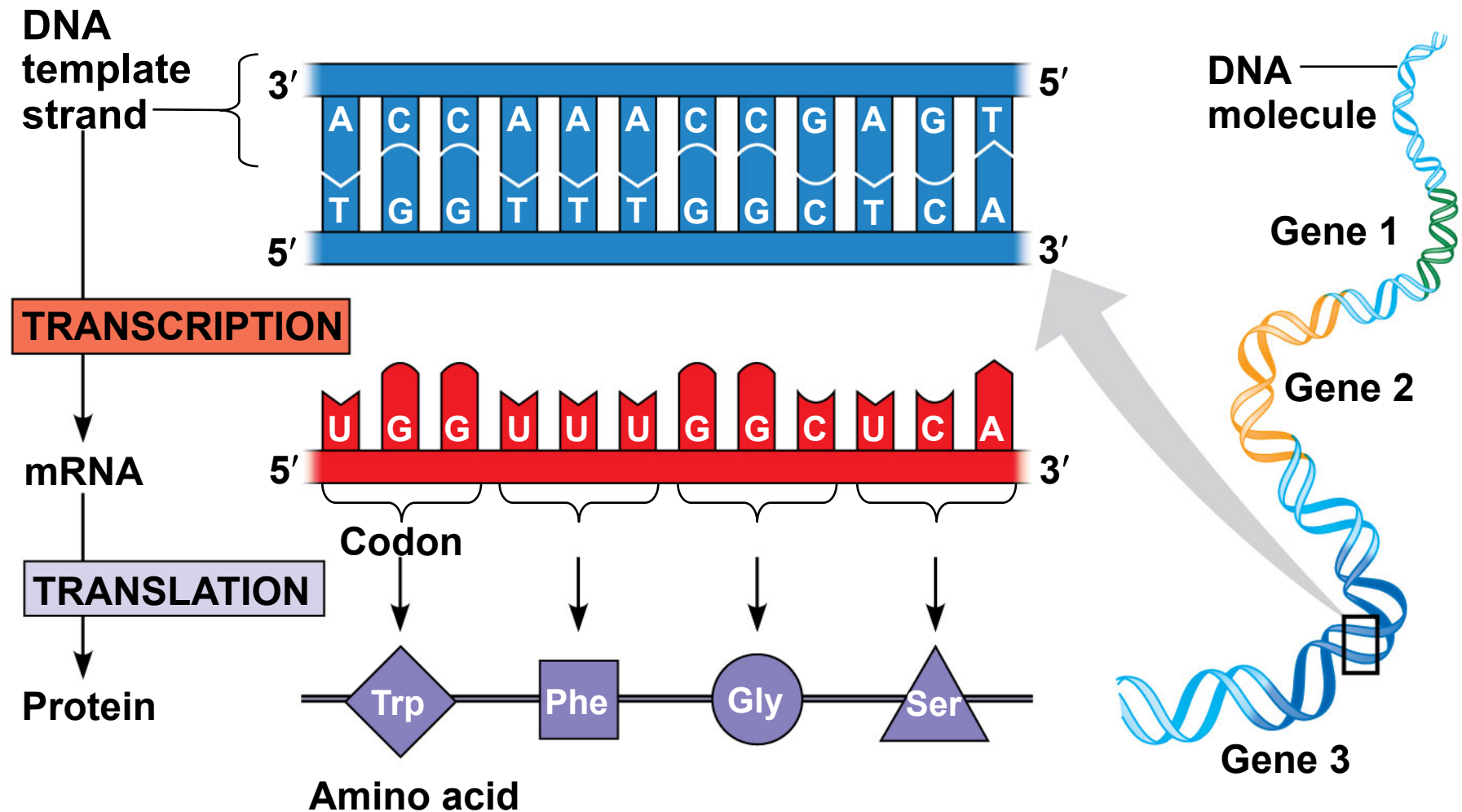
- The **genetic code** is a triplet code where a 3-nucleotide DNA word codes for a 3-nucleotide mRNA word (a **codon**) which codes for an amino acid

Mutations of one or a few nucleotides can affect protein structure and function

- **Mutations** are changes in the genetic material of a cell or virus

- **Point mutations** are chemical changes in just one base pair of a gene

  – May or may not change the protein

- **Insertions/deletions** may cause **frameshift** mutations that have a disasterous effect on the protein

Sickle-Cell Disease: A Change in Primary Structure

- A slight change in the amino acid (primary structure) can affect a protein's structure and ability to function
  - What causes a change in the primary structure?
- **Sickle-cell disease**, an inherited blood disorder, results from a single amino acid substitution in the protein hemoglobin

Point mutation that causes sickle cell disease

Wild-type hemoglobin	Sickle-cell hemoglobin

Wild-type hemoglobin DNA

3′ C T T 5′

5′ G A A 3′

Mutant hemoglobin DNA

3′ C A T 5′

5′ G T A 3′

mRNA

5′ G A A 3′

mRNA

5′ G U A 3′
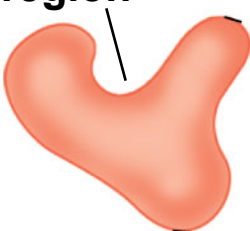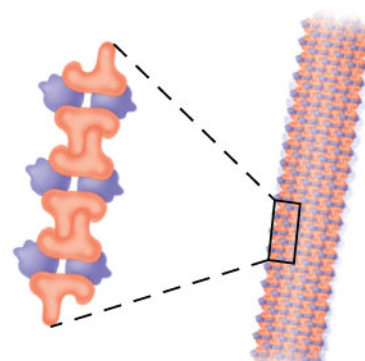
Normal hemoglobin

Glu

Sickle-cell hemoglobin

Val
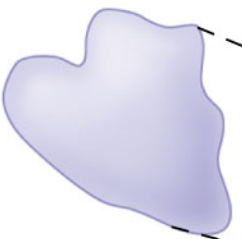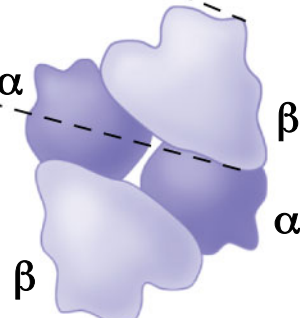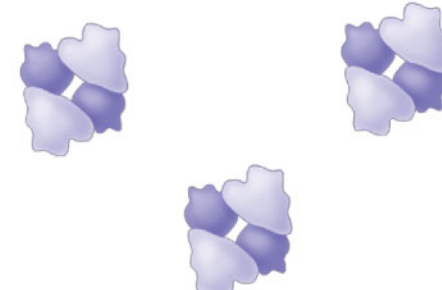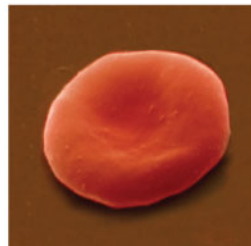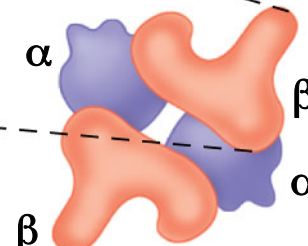
Figure 5.21

	Primary Structure	Secondary and Tertiary Structures	Quaternary Structure	Function	Red Blood Cell Shape
Normal hemoglobin	1 Val 2 His 3 Leu 4 Thr 5 Pro 6 Glu 7 Glu	β subunit	Normal hemoglobin α β α β	Molecules do not associate with one another; each carries oxygen.	10 μm
Sickle-cell hemoglobin	1 Val 2 His 3 Leu 4 Thr 5 Pro 6 Val 7 Glu	Exposed hydrophobic region β subunit	Sickle-cell hemoglobin α β β α	Molecules crystallize into a fiber; capacity to carry oxygen is reduced.	10 μm

10

© 2011 Pearson Education, Inc.

Oligonucleotide microarray

Cells of person 1/condition 1
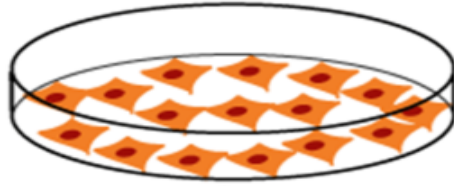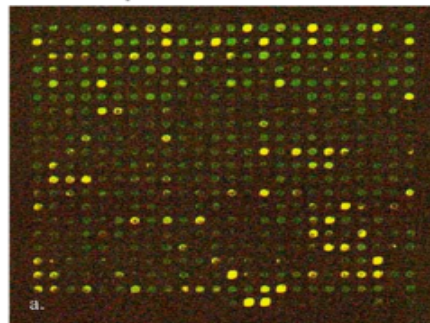
RNA isolation

mRNA

Reverse transcriptase labeling

cDNA
"Green Fluorescent" Targets

Hybridize to microarray

Microarray with short ssDNA spanning the entire genome

----ACTGA----

----ACTGA----

----GAGAT----
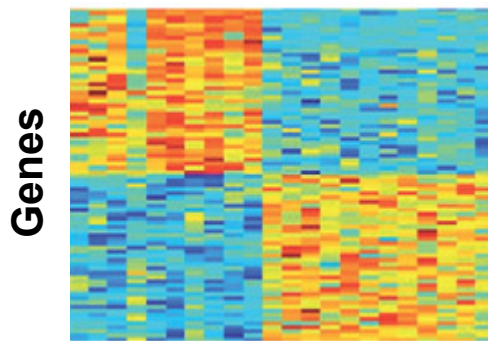
Probe 1: TGACT

Probe 2: CTCTA

…

Probe 20000: TTTAG

Biomarkers and personalized medicine

Gene expression profiles

Samples

Genes

- Bioinformatics challenges
  - Identification of genes or gene signature
  - Choice of classification method or gene model

Possible comparisons

A	B	Biomarker identification (gene or gene signature)
Tumor	Normal	*Diagnostic*: predictive of a clinical variable
High risk	Low risk	*Prognostic*: predictive of disease outcome
Responder	Non-responder	*Predictive*: predictive of therapeutic response

Microarrays in more detail
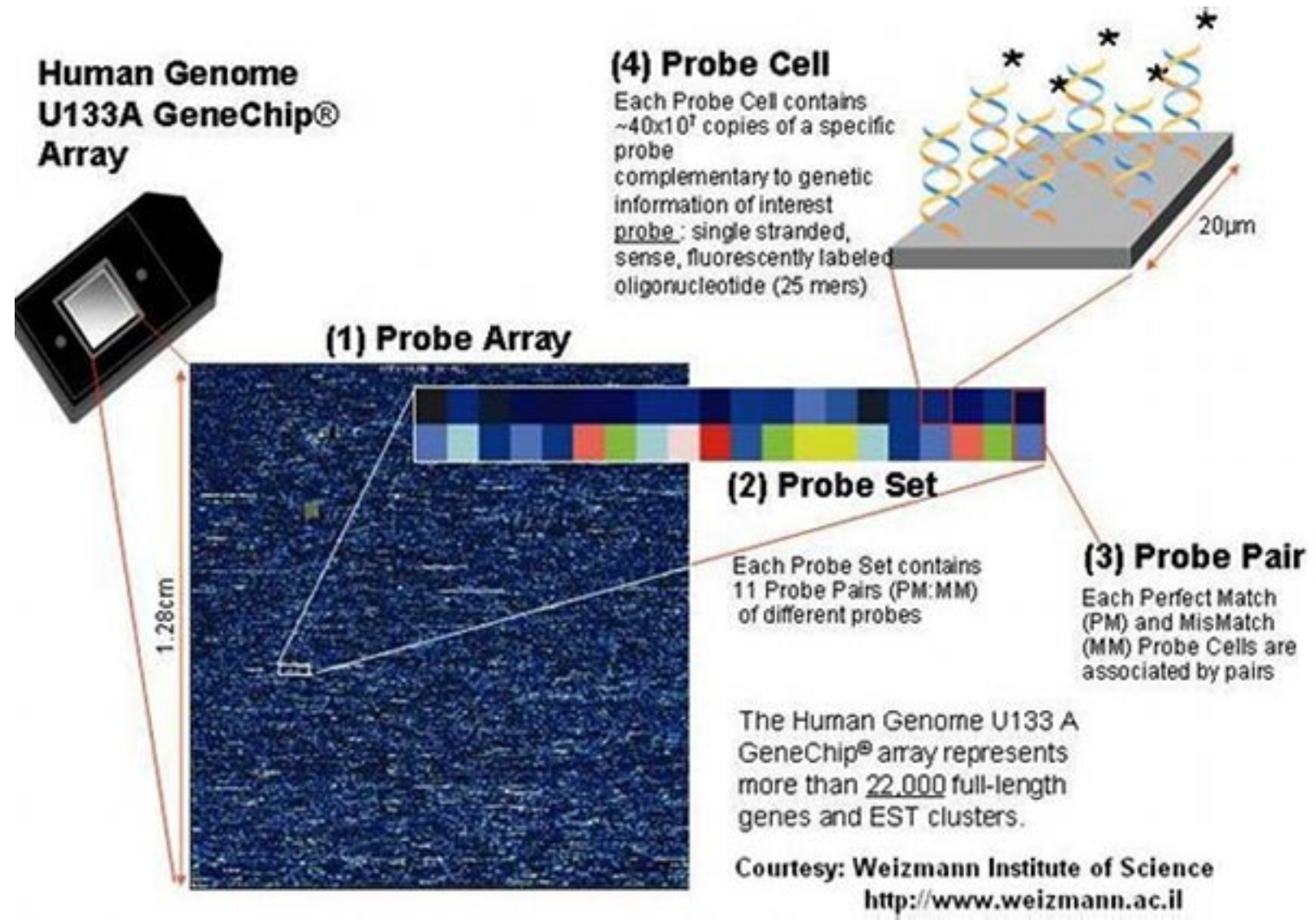
Microarray Analysis

- Analysis will be performed using several Bioconductor packages (http://bioconductor.org)

- Data is available from the Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/)

  – We will look at how to download raw and processed data from GEO

Gene Expression Omnibus (GEO)

- GEO (http://www.ncbi.nlm.nih.gov/geo/) is a public functional genomics data repository for gene expression (microarray) and sequence-based data.

- There are four kinds of records on GEO (http://www.ncbi.nlm.nih.gov/geo/info/overview.html)

Gene Expression Omnibus (GEO)

- A **GEO sample** (GSM*) describes an individual sample, including the experimentally conditions in which it was collected, and the gene expression value for each element on the array.

- A **GEO platform** (GPL*) is a summary of the array used, and links the array probe to a gene

- A **GEO series** (GSE*) links together a collection of samples with one or more platforms for a particular experiment or study (such as profiling gene expression from 100 patients with lung cancer)

- A **GEO dataset** is a curated collection of samples that allows for user-friendly analysis. Not all series exist as datasets.