

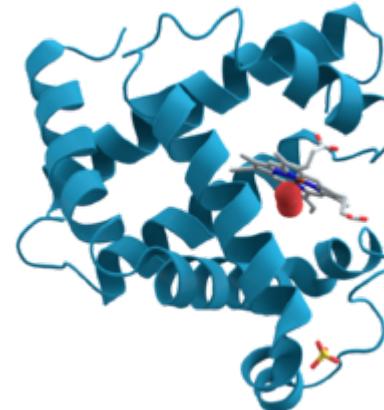
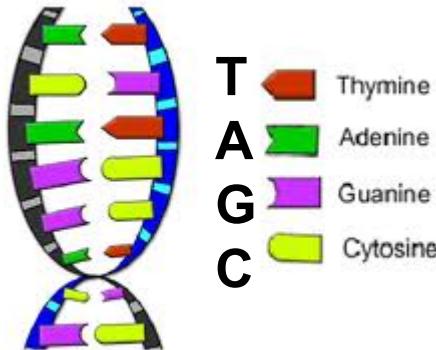
# **Gene Expression and RNA-Seq**

**Garrett M. Dancik, Ph.D.**

Note: All images from slides 3-10 are from Campbell Biology, 9<sup>th</sup> edition,  
© 2011 Pearson Education, Inc.

# Overview of gene expression

Central Dogma of  
Molecular Biology:



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 ~ 25,000 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

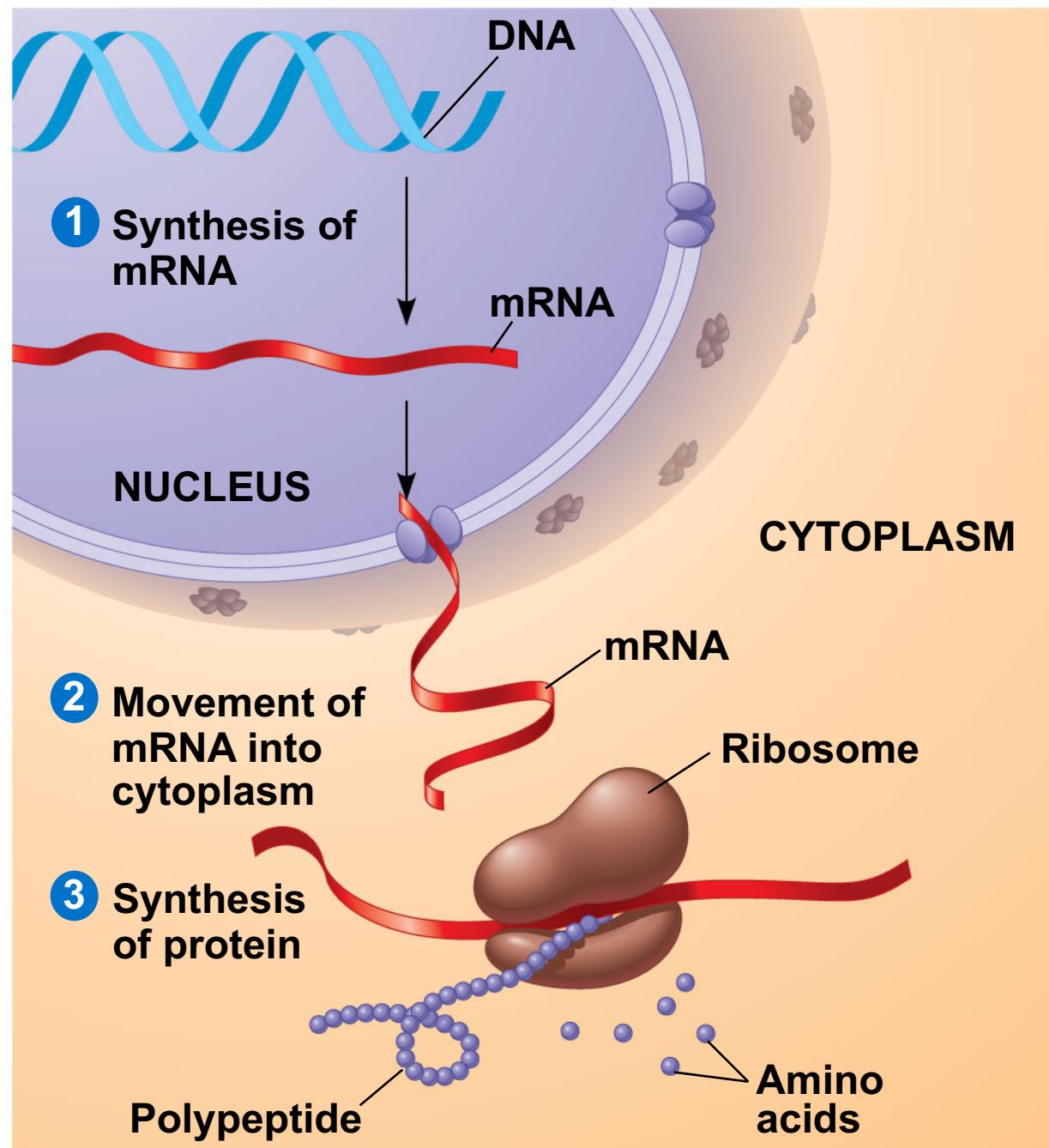
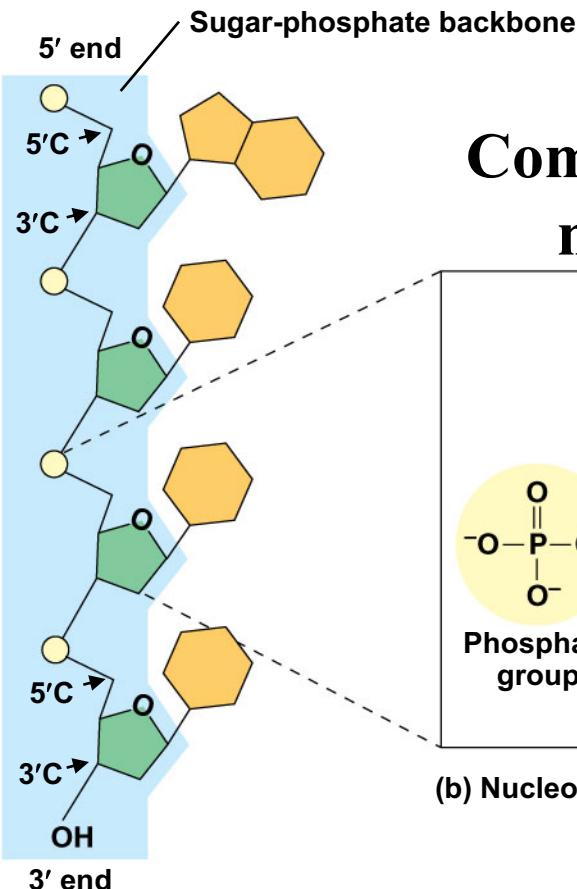
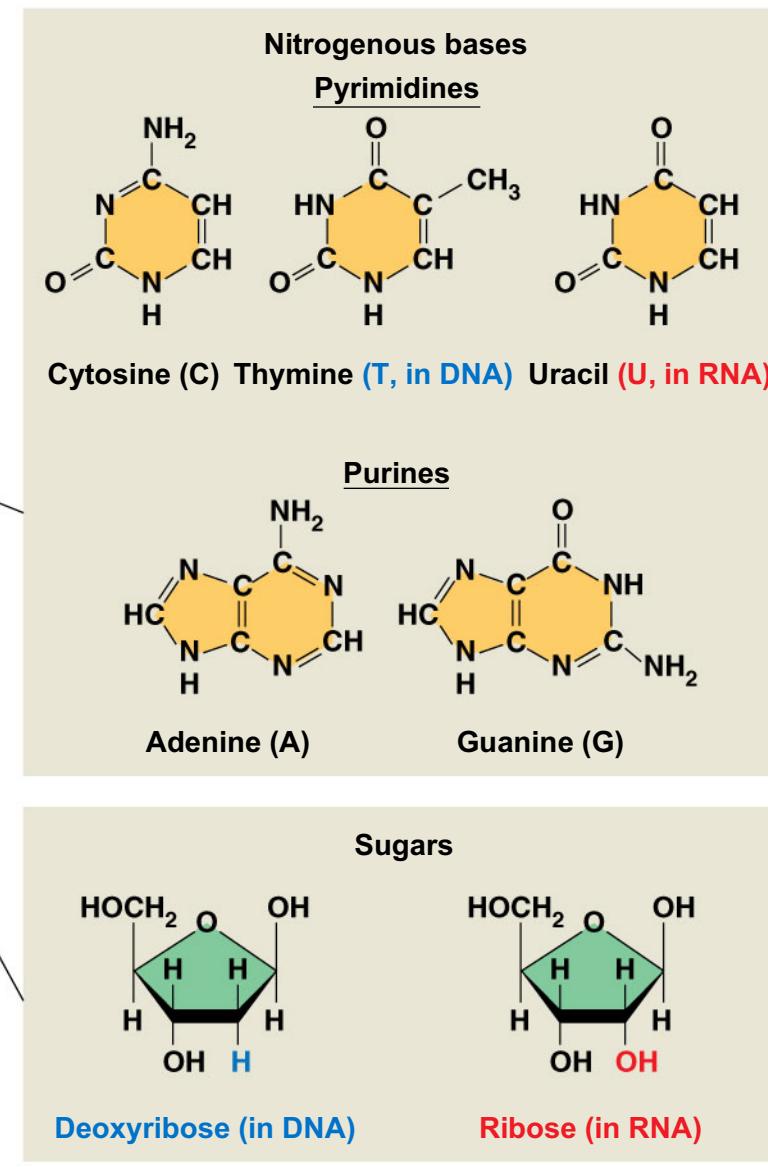
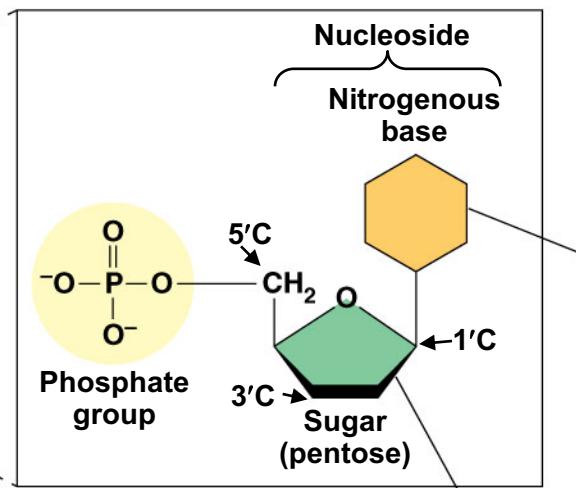


Figure 5.26

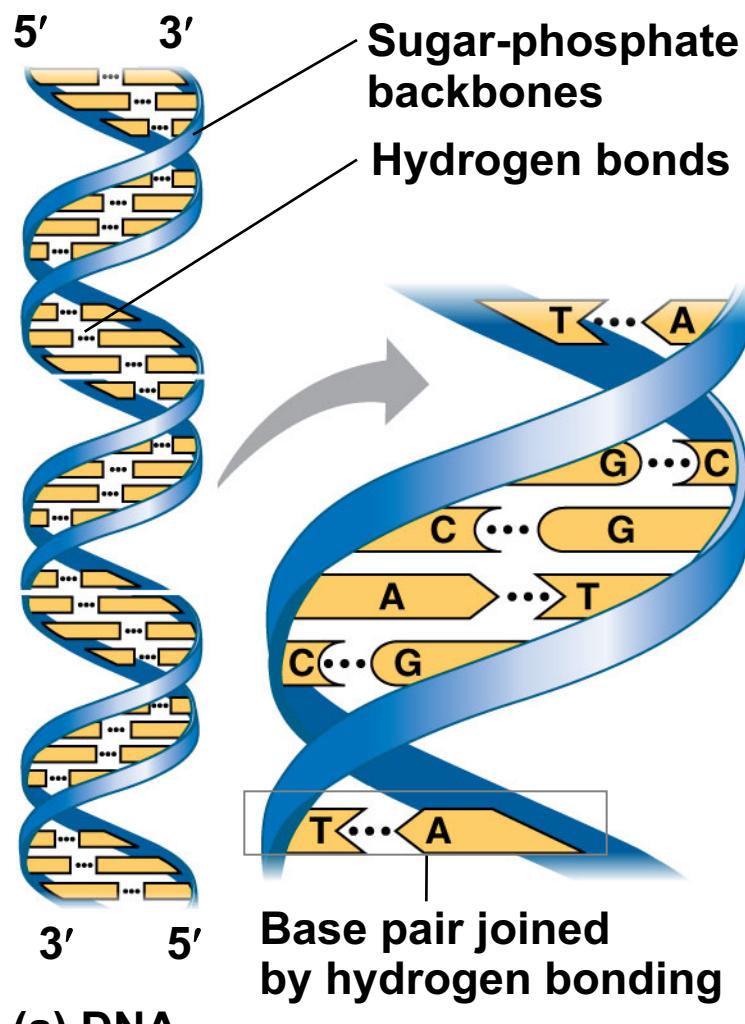


## Components of a nucleotide



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

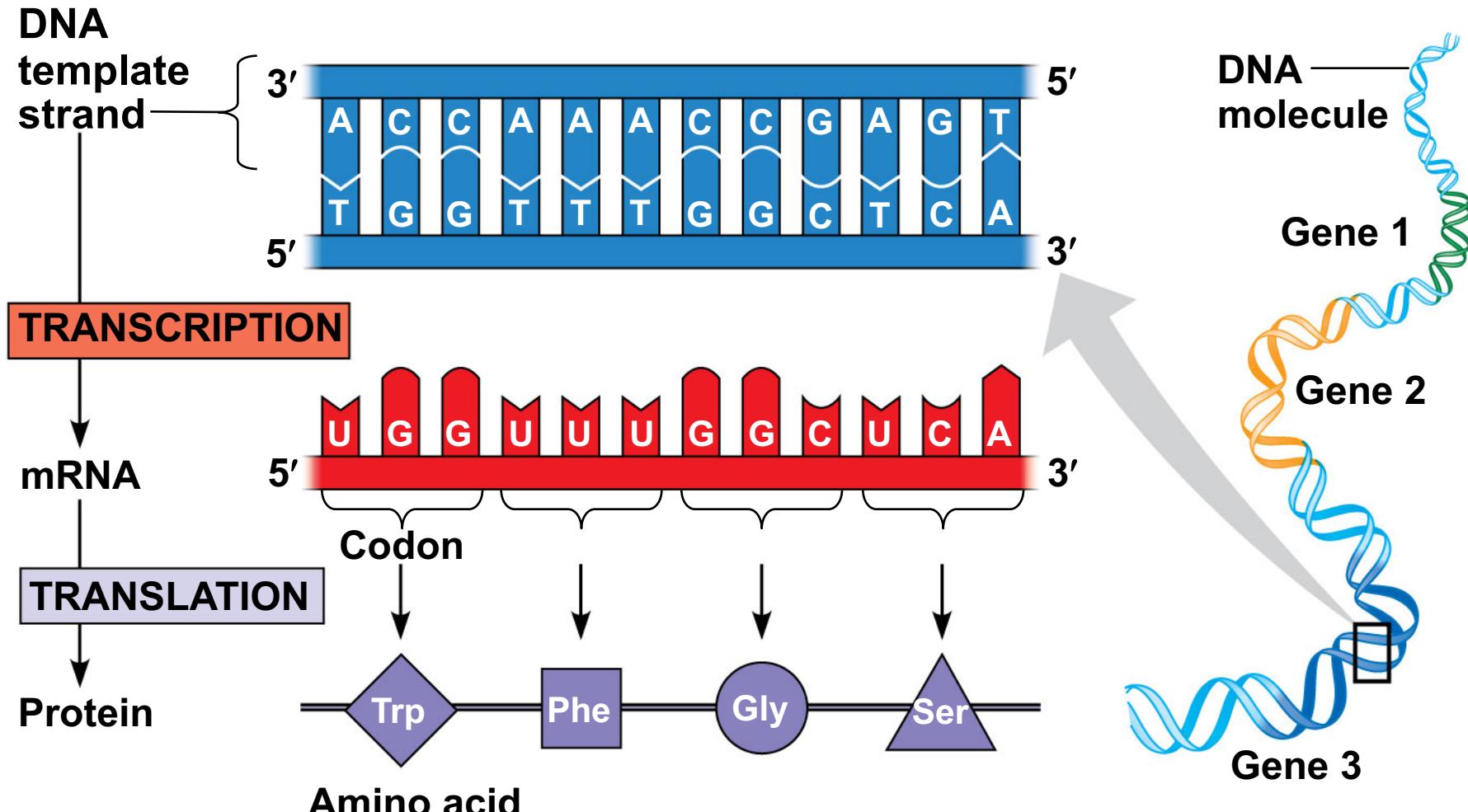
Figure 5.27



(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



© 2011 Pearson Education, Inc.

- 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 disastrous 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

### Wild-type hemoglobin DNA

3' C T T 5'  
5' G A A 3'

## Sickle-cell hemoglobin

### 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'

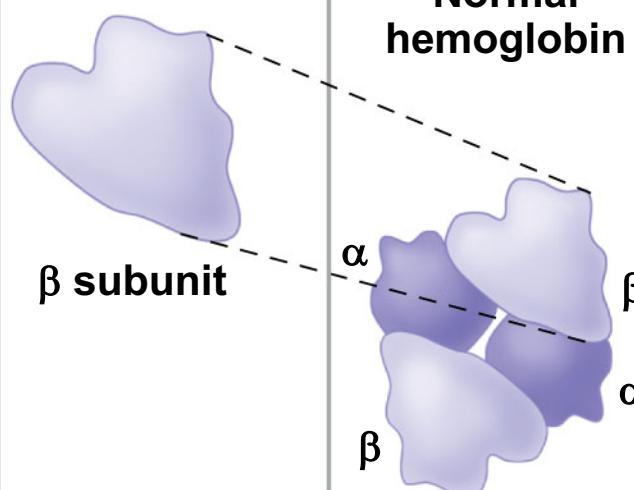
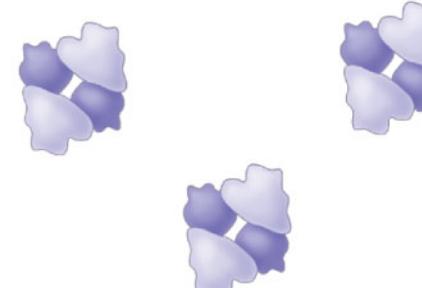
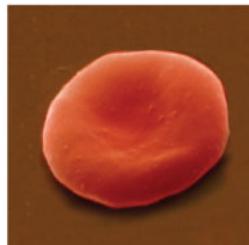
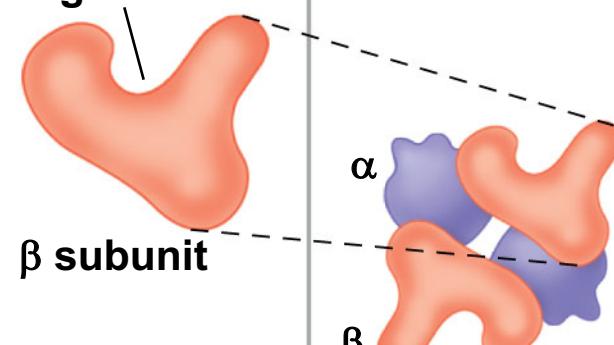
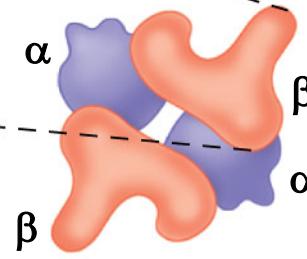
### Normal hemoglobin

= [purple box] = Glu = [purple box] =

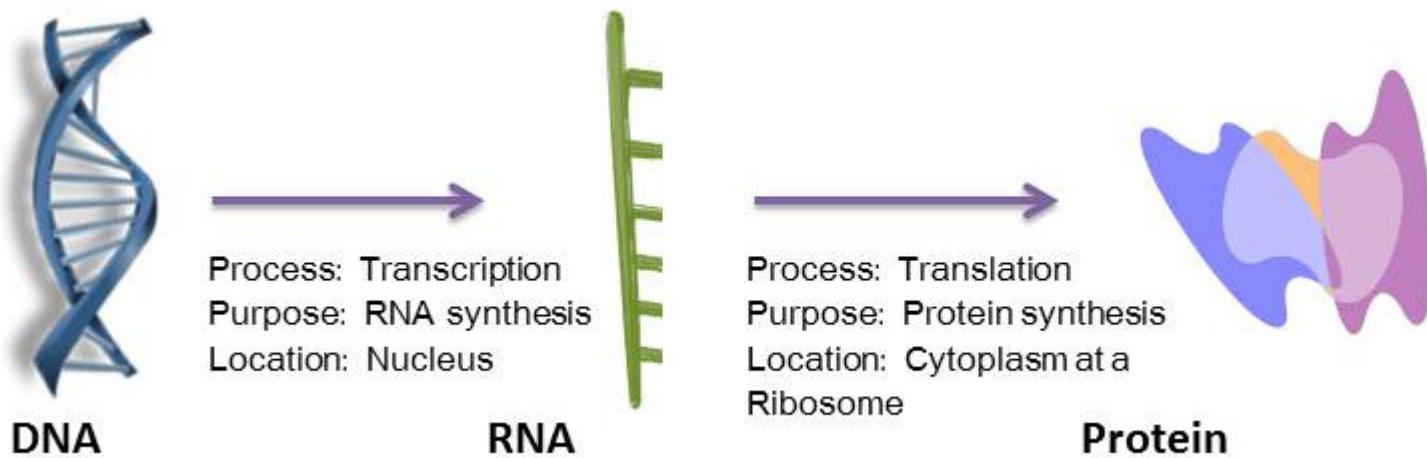
### Sickle-cell hemoglobin

= [purple box] = Val = [purple box] =

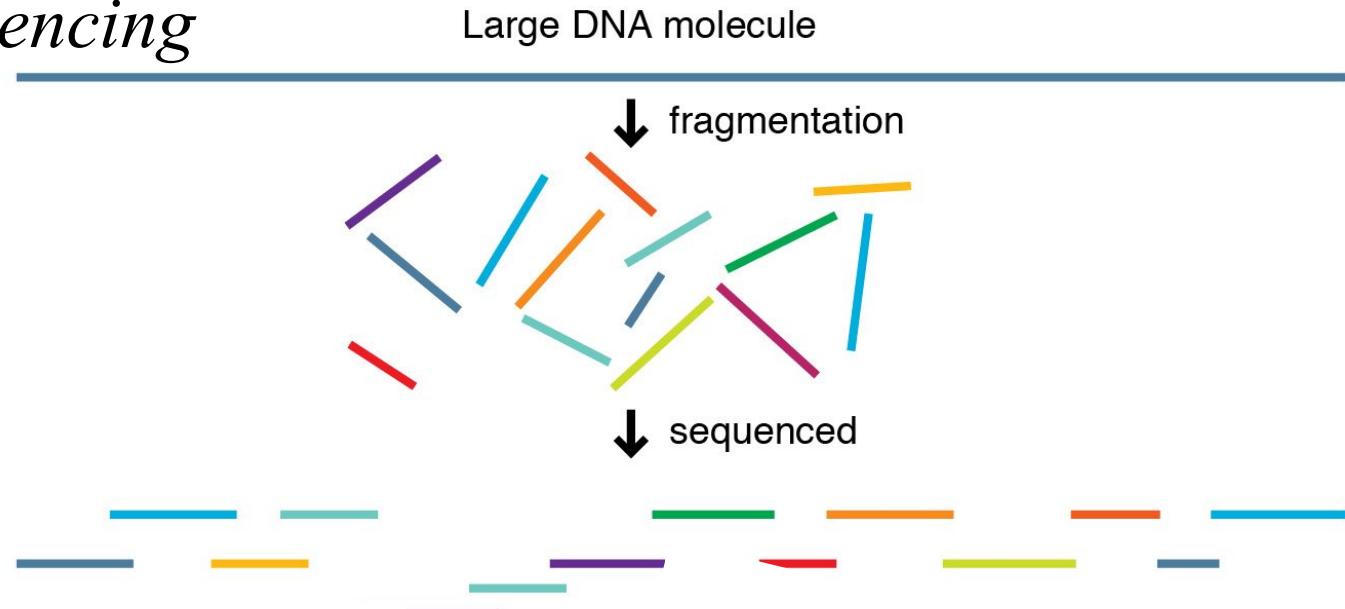
Figure 5.21

	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	 <p><math>\beta</math> subunit</p> <p>Normal hemoglobin</p>	 <p><math>\alpha</math></p> <p><math>\beta</math></p> <p><math>\alpha</math></p> <p><math>\beta</math></p>	Molecules do not associate with one another; each carries oxygen.	 <p>10 <math>\mu\text{m}</math></p>
Sickle-cell hemoglobin	1 Val 2 His 3 Leu 4 Thr 5 Pro 6 Val 7 Glu	 <p>Exposed hydrophobic region</p> <p><math>\beta</math> subunit</p>	 <p><math>\alpha</math></p> <p><math>\beta</math></p> <p><math>\alpha</math></p> <p><math>\beta</math></p>	Molecules crystallize into a fiber; capacity to carry oxygen is reduced.	 <p>10 <math>\mu\text{m}</math></p>

# Gene Expression

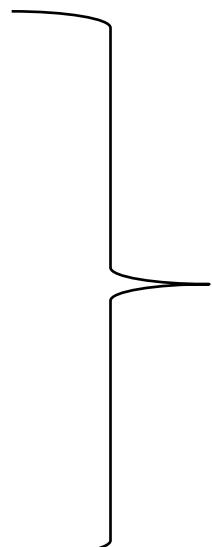


## *Genomic sequencing*



TAGACGTAGC  
GAATAGCTAG  
GTCGAGCGTA  
CCTCATAAGA  
CGAGAATAGC

.....

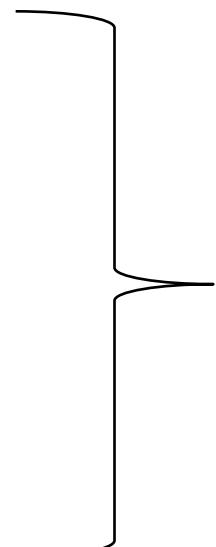


- ~ 1 billion reads
- Each read is ~ 100 bp

# Reference Genome Sequence (~3 billion bp for humans)

-----ACGTCGAGCGTAGACGTAGCGAGAATAGCTAGCTATAAAGGCCTCGTAAGA-----

TAGACGTAGC  
GAATAGCTAG  
GTCGAGCGTA  
CCTCATAAGA  
CGAGAATAGC  
.....

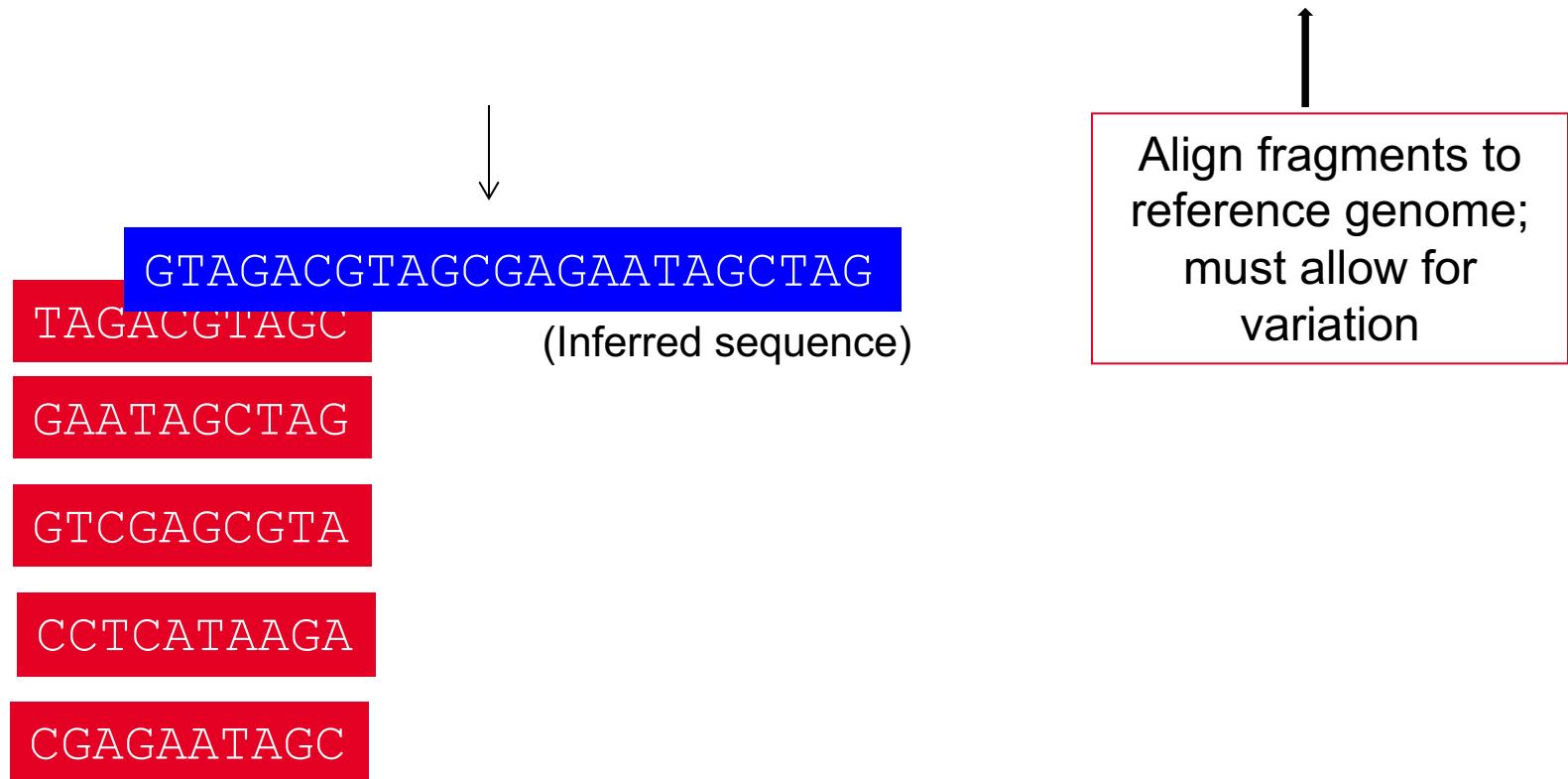


- ~ 1 billion reads
- Each read is ~ 100 bp

Align fragments to reference genome;  
must allow for variation

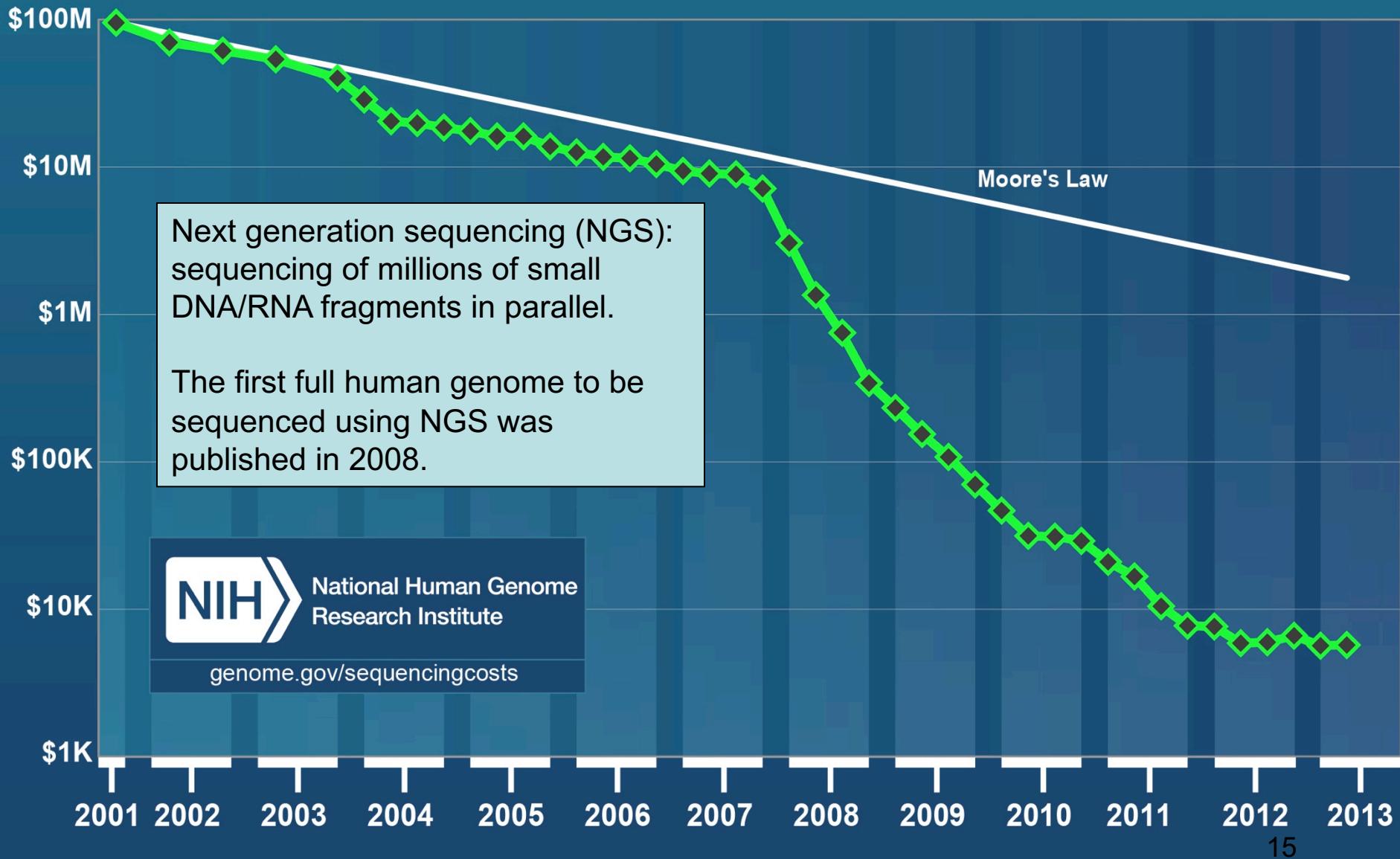
# Reference Genome Sequence (~3 billion bp for humans)

-----ACGTCGAGCGTAGACGTAGCGAGAATAGCTAGCTATAAAGGCCTCGTAAGA-----

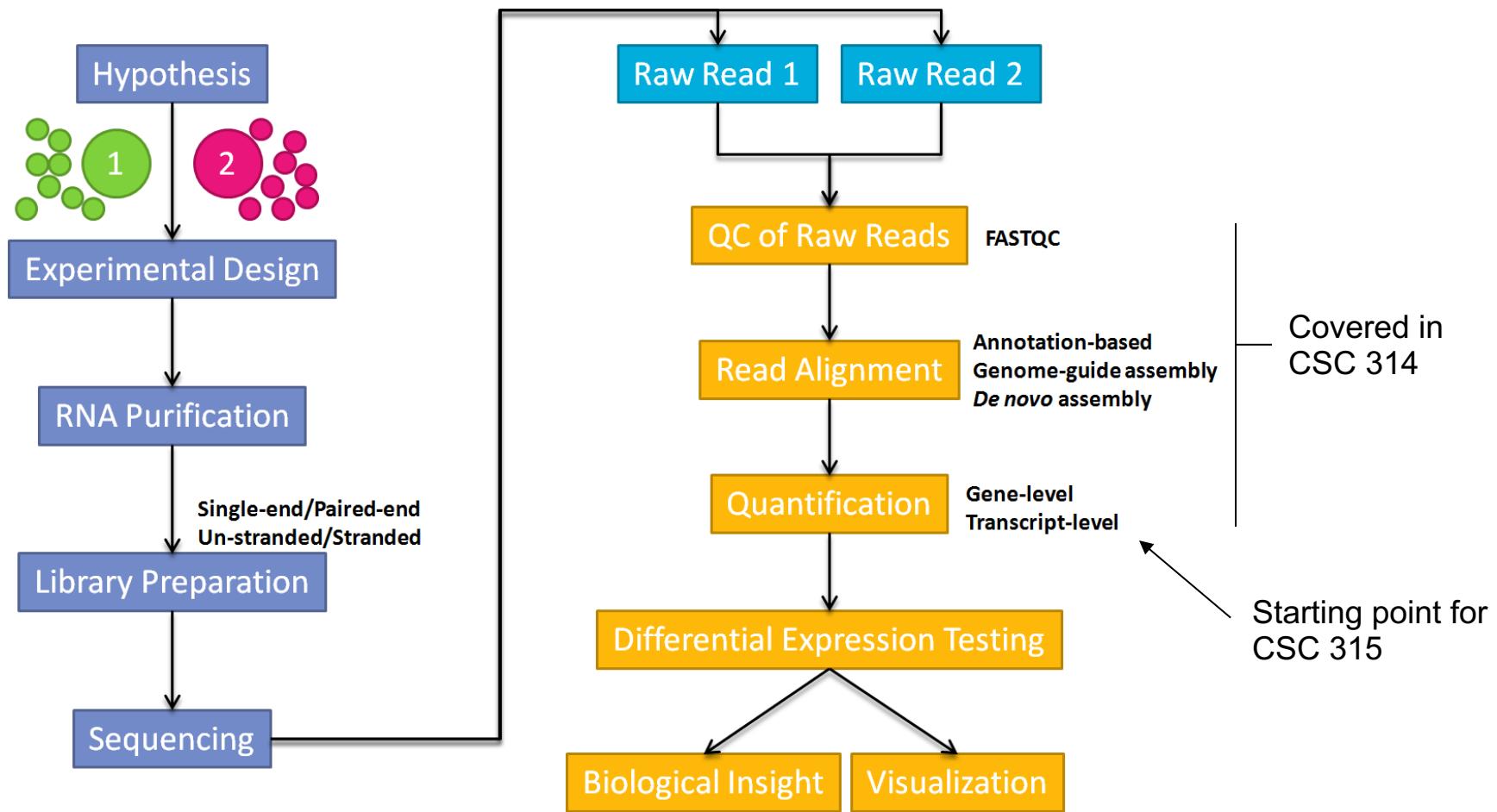


Genome assembly when a reference genome is available

# *Cost per Genome*

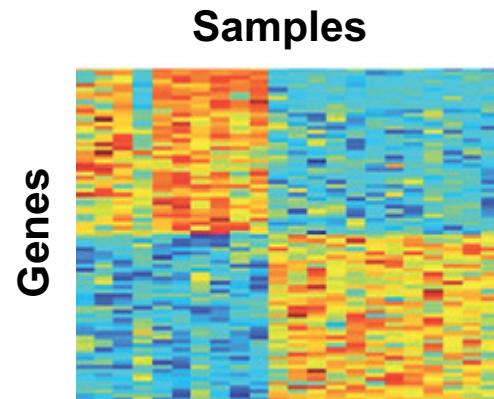


# RNA-Seq overview



# Biomarkers and personalized medicine

## Gene expression profiles



- Bioinformatics challenges
  - Accurate quantification of gene expression
  - **Identification of genes or gene signature**
  - **Choice of classification method or gene model (time permitting)**

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

# Example: Development of a RNA-Seq Based Prognostic Signature in Lung Adenocarcinoma

**Figure 2.** Four-gene prognostic signature biomarker characteristics in The Cancer Genome Atlas (TCGA) cohort.

