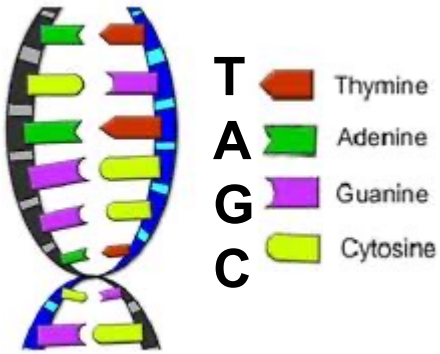
Gene Expression and RNA-Seq

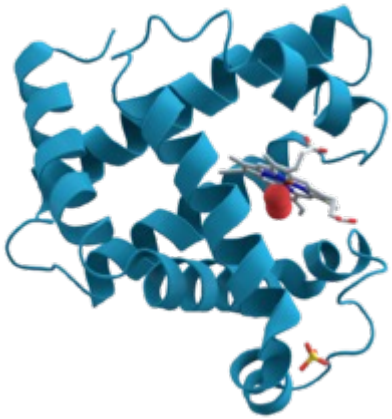
Garrett M. Dancik, Ph.D.

Overview of gene expression

Central Dogma of Molecular Biology:

DNA $\xrightarrow{\text{transcription}}$ RNA $\xrightarrow{\text{translation}}$ Protein





4-character alphabet $\longrightarrow$ 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

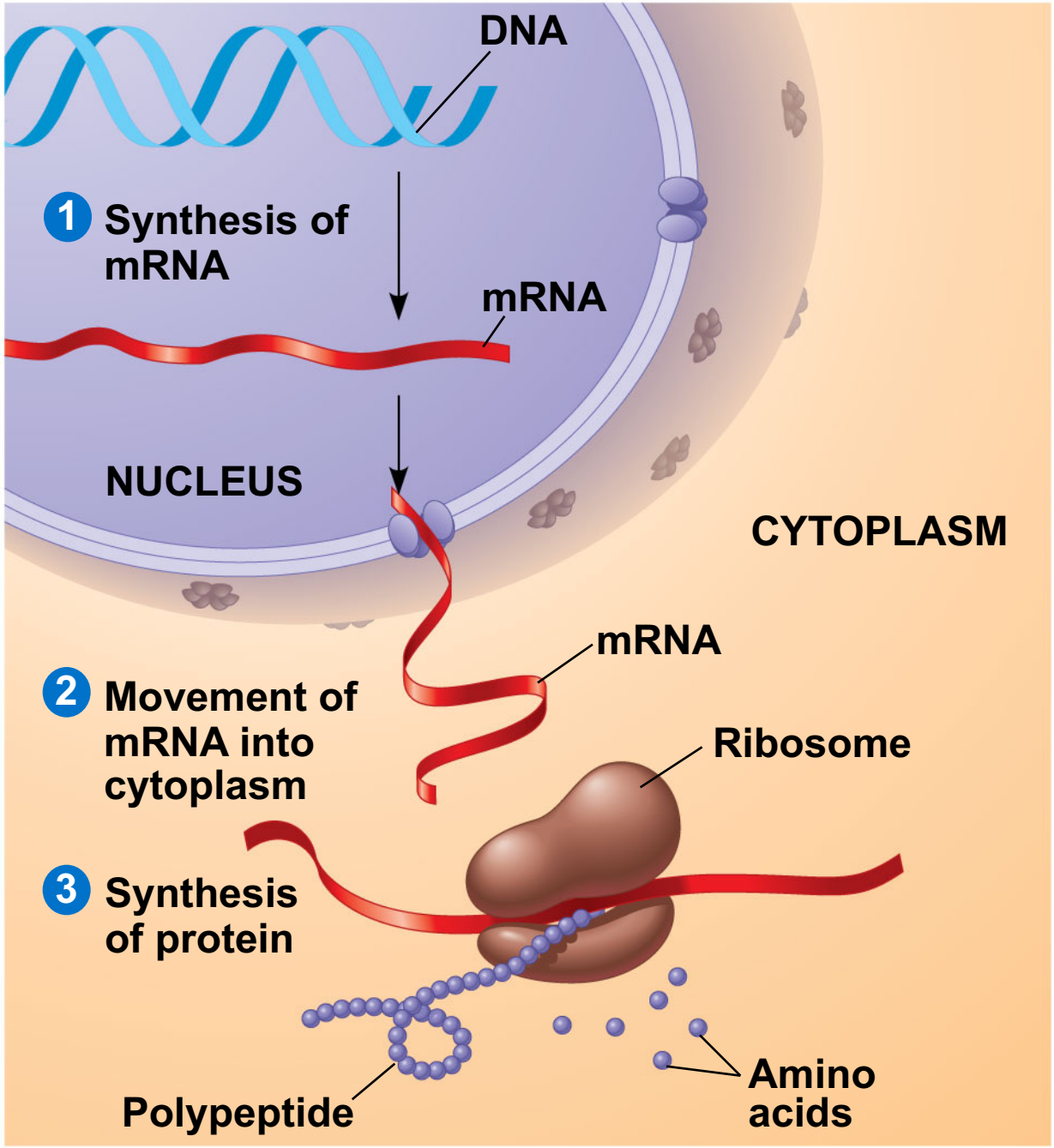
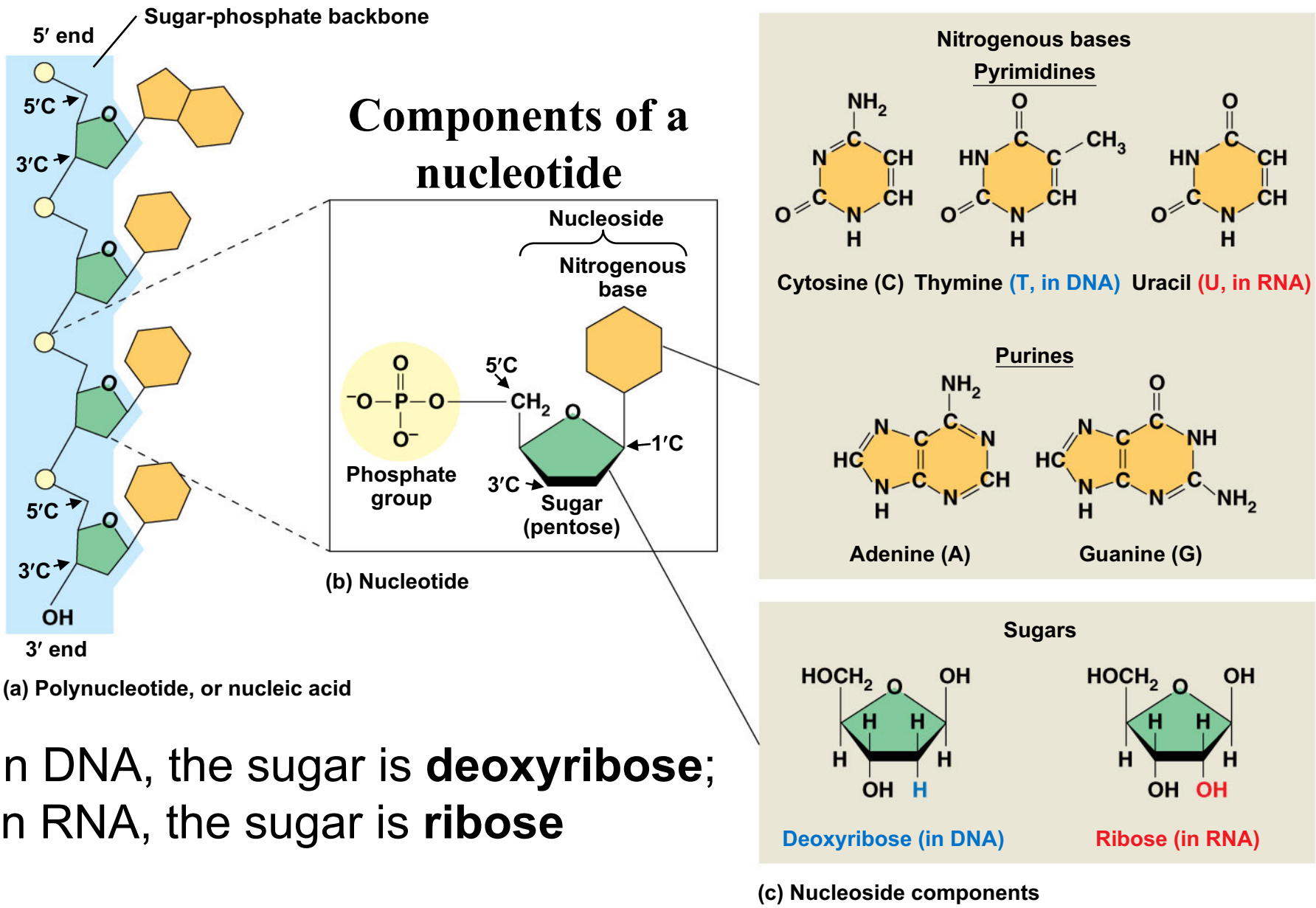


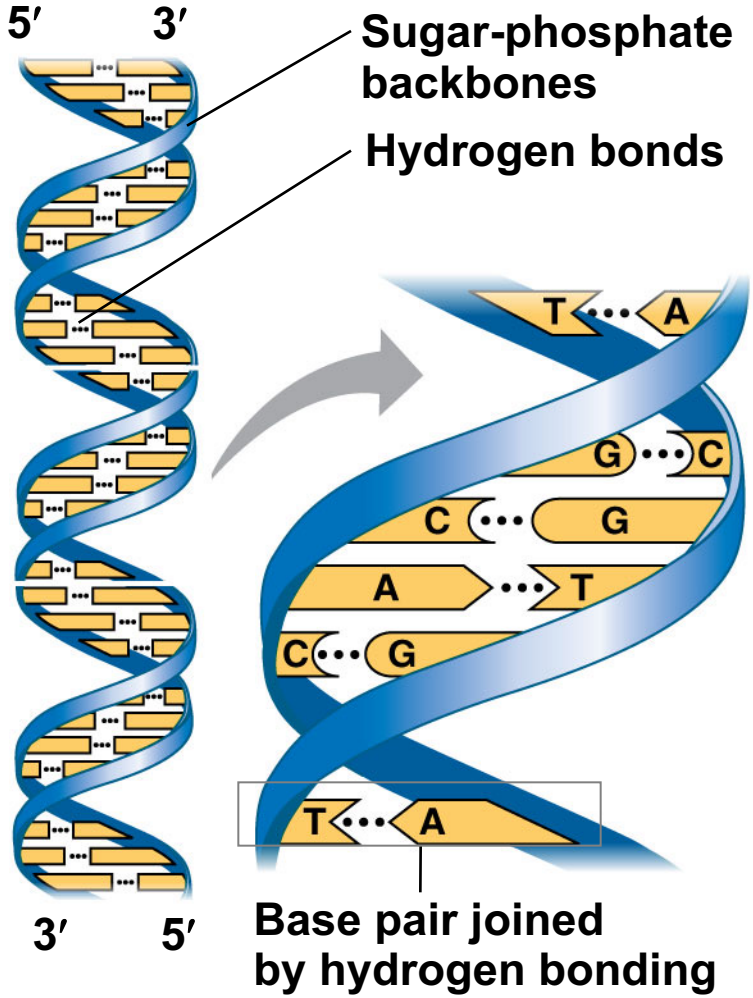
Figure 5.26



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

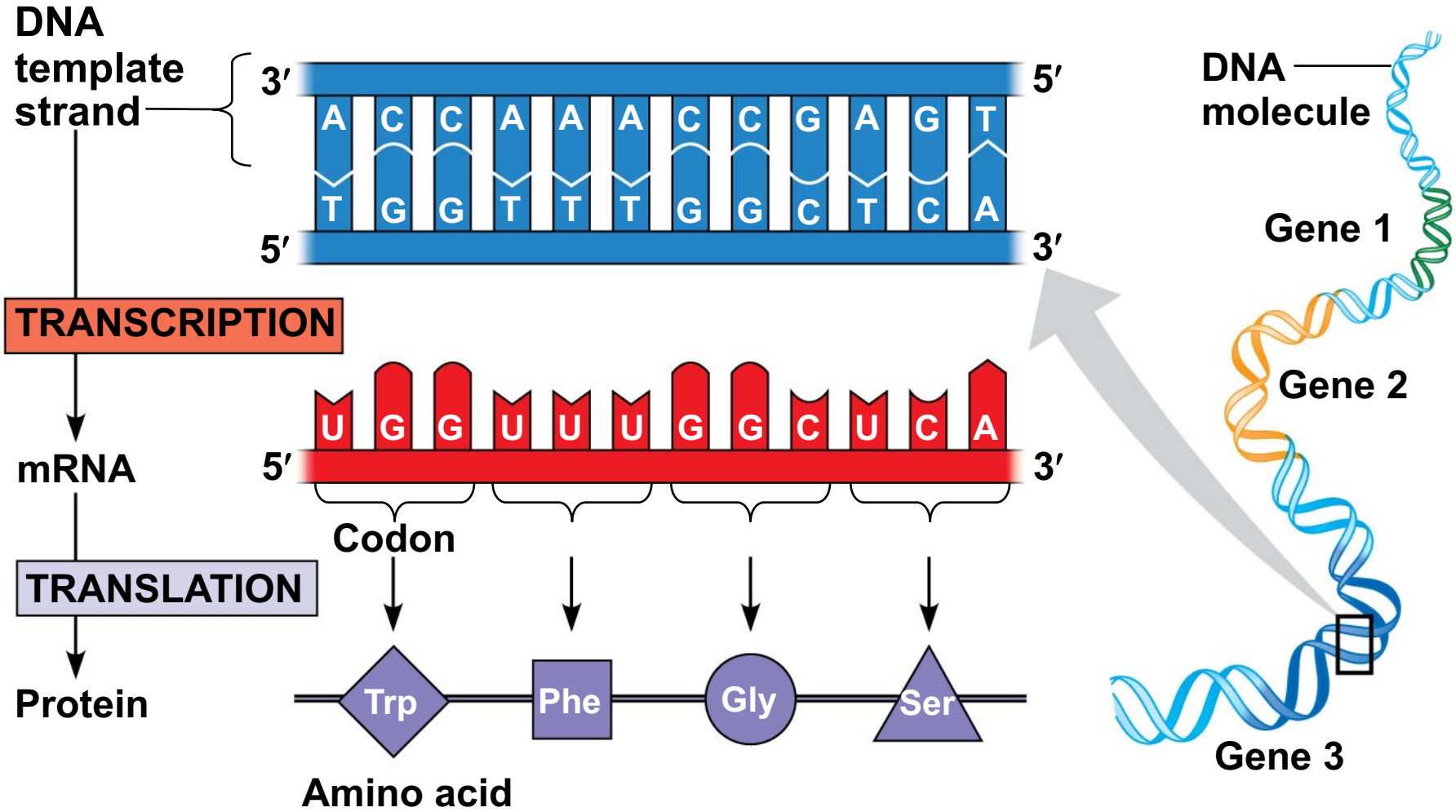
Nucleic Acids are made up of nucleotides

Figure 5.27



(a) DNA

- 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



- 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

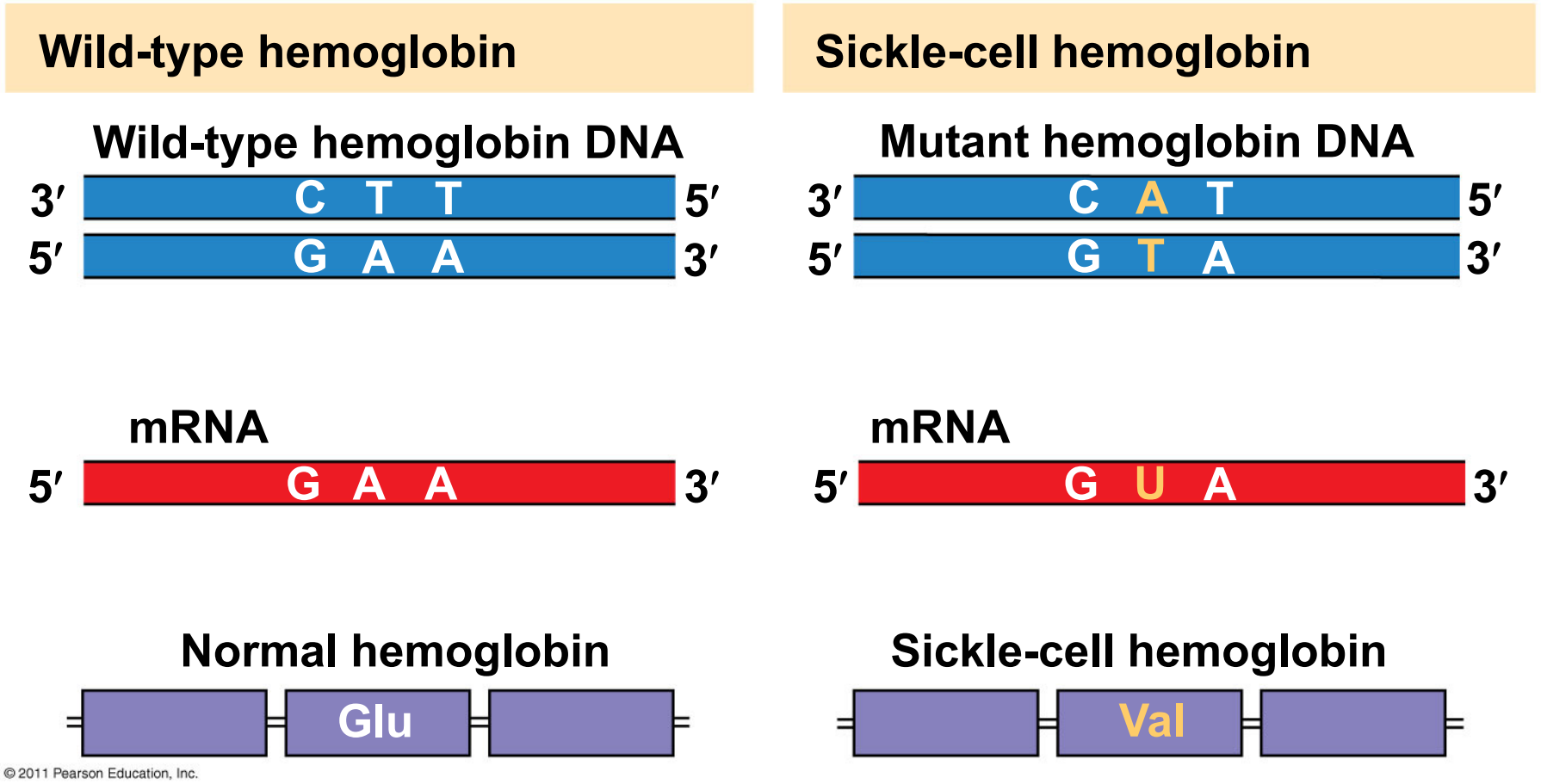
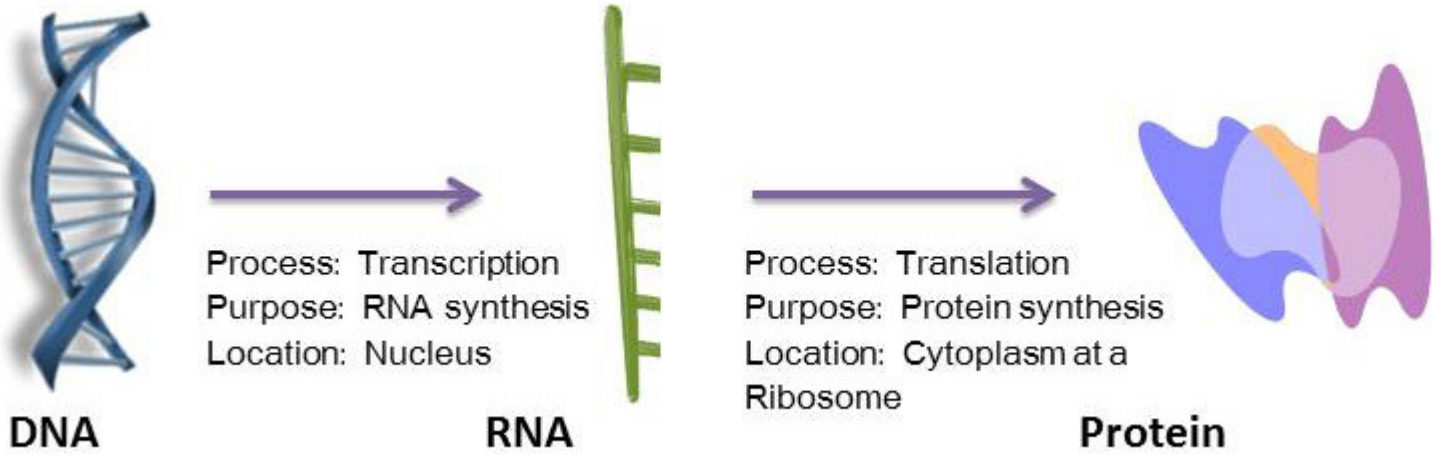


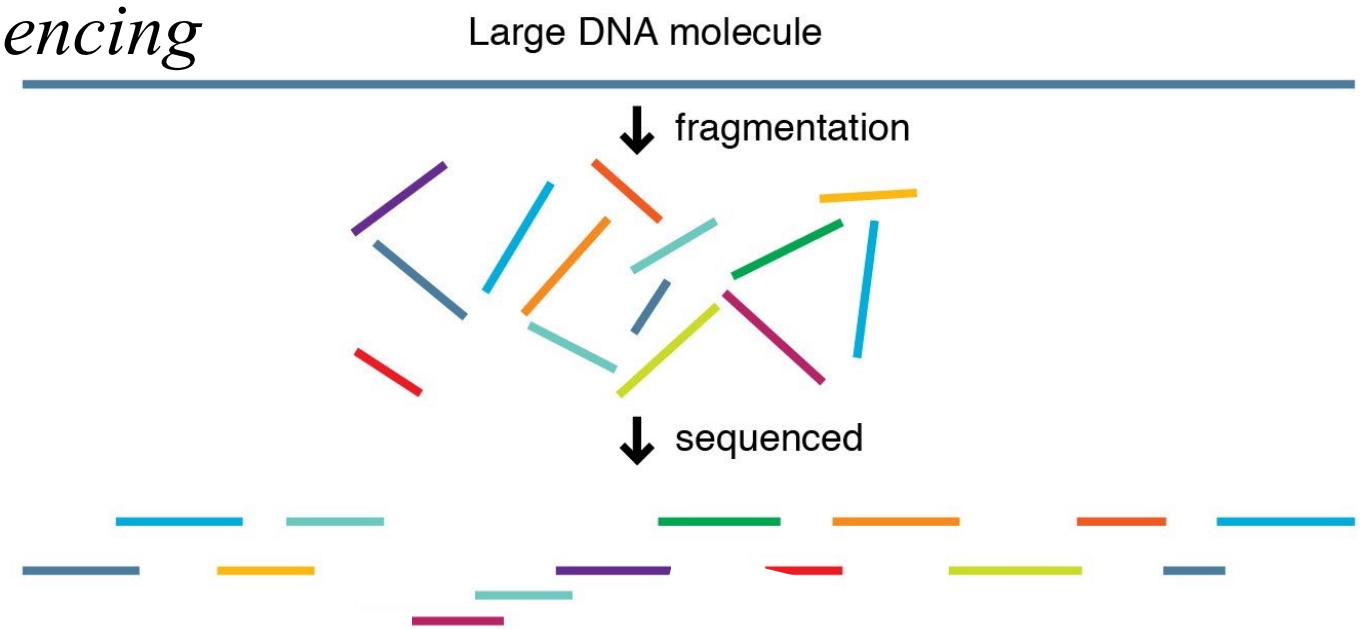
Figure 5.21

	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 α β	Molecules crystallize into a fiber; capacity to carry oxygen is reduced.	10 μm

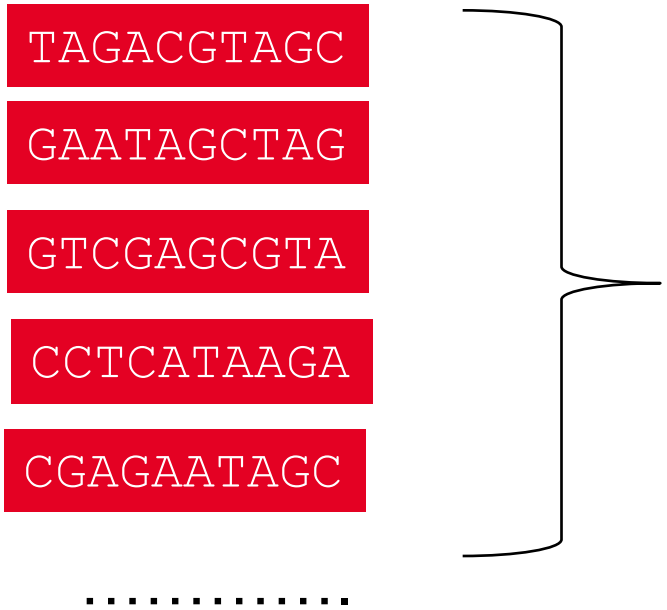
Gene Expression



Genomic sequencing



http://knowgenetics.org



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

Genome assembly when a reference genome is available

Reference Genome Sequence (~3 billion bp for humans)

----ACGTCGAGCGTAGACGTAGCGAGAATAGCTAGCTATAAAGGCCTCGTAAGA---

Align fragments to reference genome; must allow for variation

TAGACGTAGC

GAATAGCTAG

GTCGAGCGTA

CCTCATAAGA

CGAGAATAGC

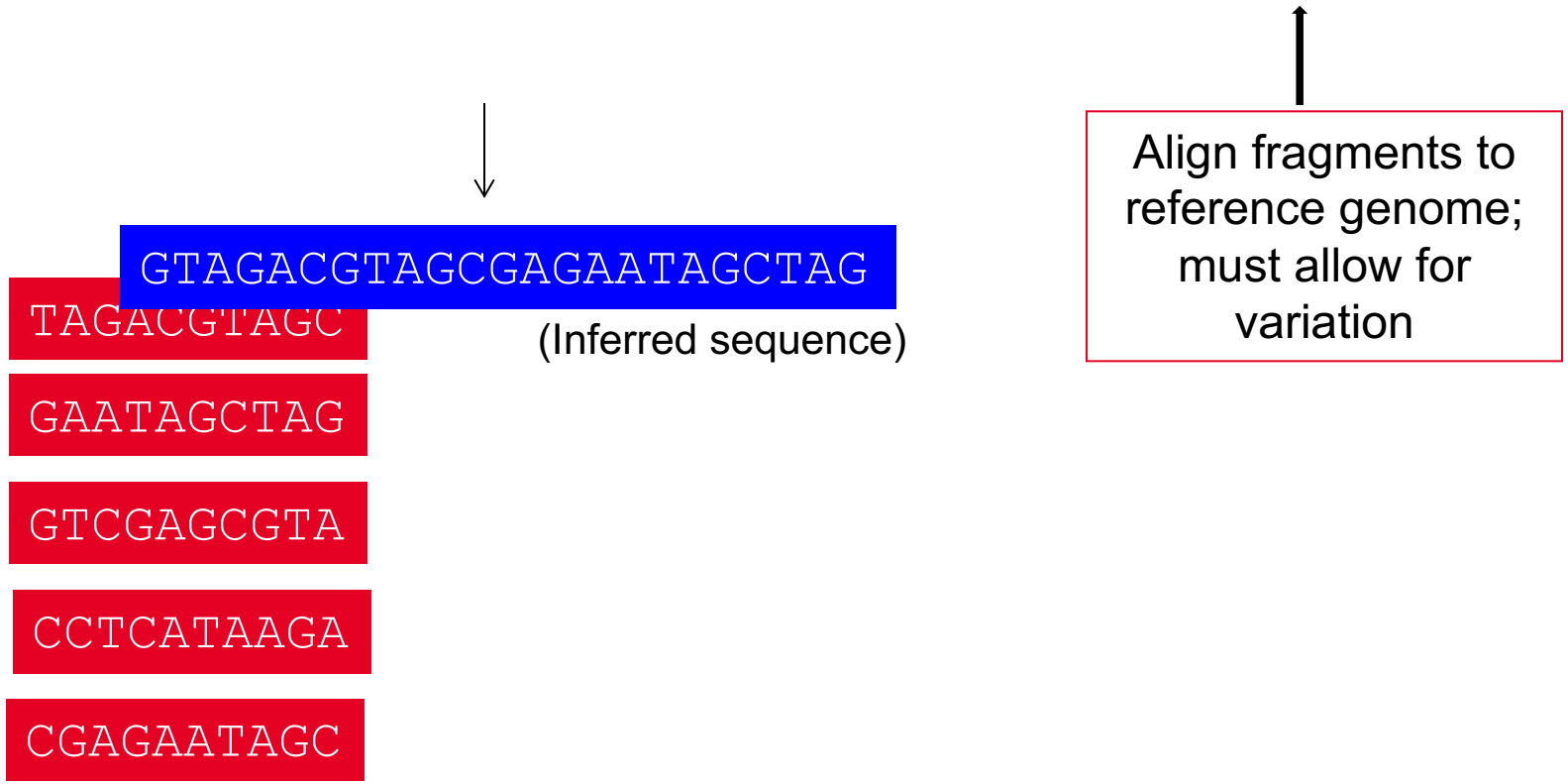
.....

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

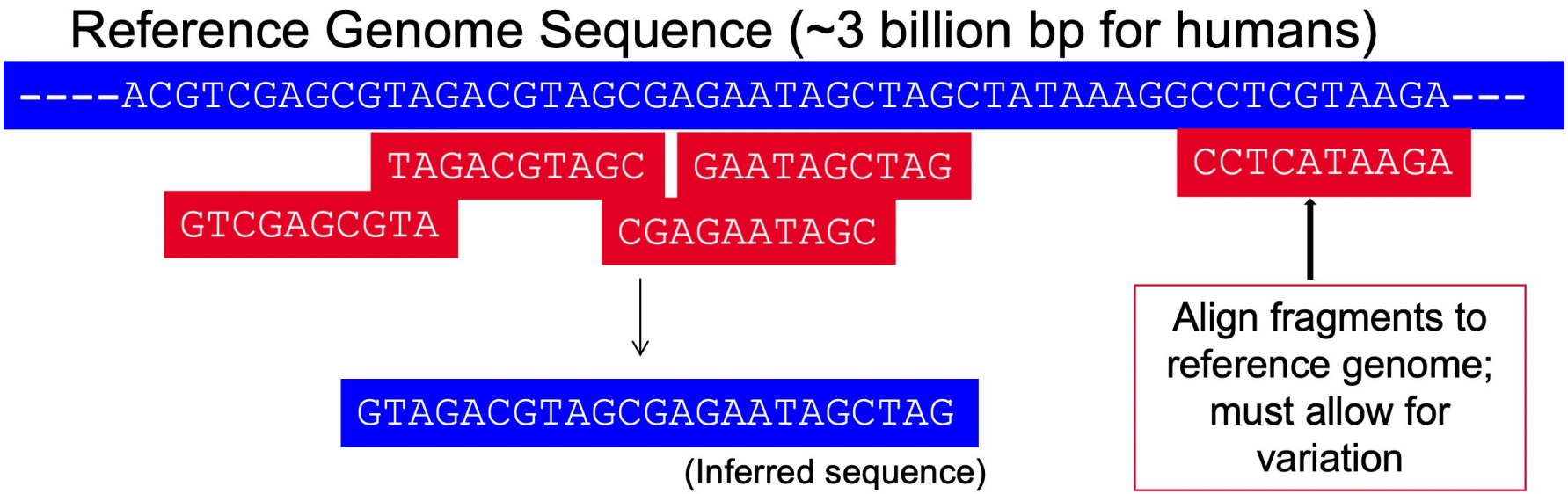
Genome assembly when a reference genome is available

Reference Genome Sequence (~3 billion bp for humans)

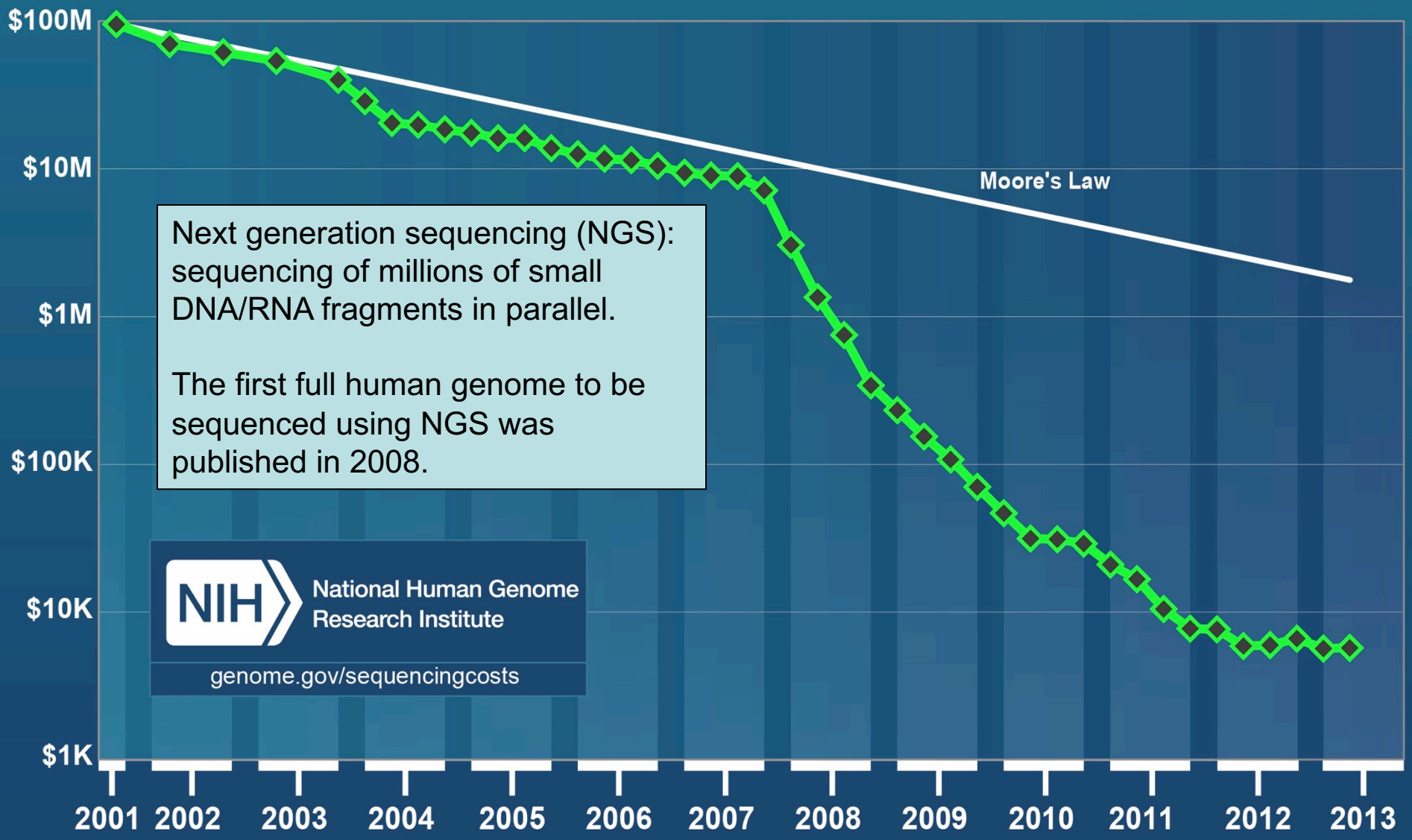
----ACGTCGAGCGTAGACGTAGCGAGAATAGCTAGCTATAAAGGCCTCGTAAGA---



Genome assembly when a reference genome is available


Reference Genome Sequence (~3 billion bp for humans)
----ACGTCGAGCGTAGACGTAGCGAGAATAGCTAGCTATAAAGGCCTCGTAAGA---
TAGACGTAGC
GAATAGCTAG
CCTCATAAGA
GTCGAGCGTA
CGAGAATAGC
GTAGACGTAGCGAGAATAGCTAG
(Inferred sequence)
Align fragments to reference genome; must allow for variation

Cost per Genome



Next generation sequencing (NGS): sequencing of millions of small DNA/RNA fragments in parallel.

The first full human genome to be sequenced using NGS was published in 2008.

RNA-Seq overview

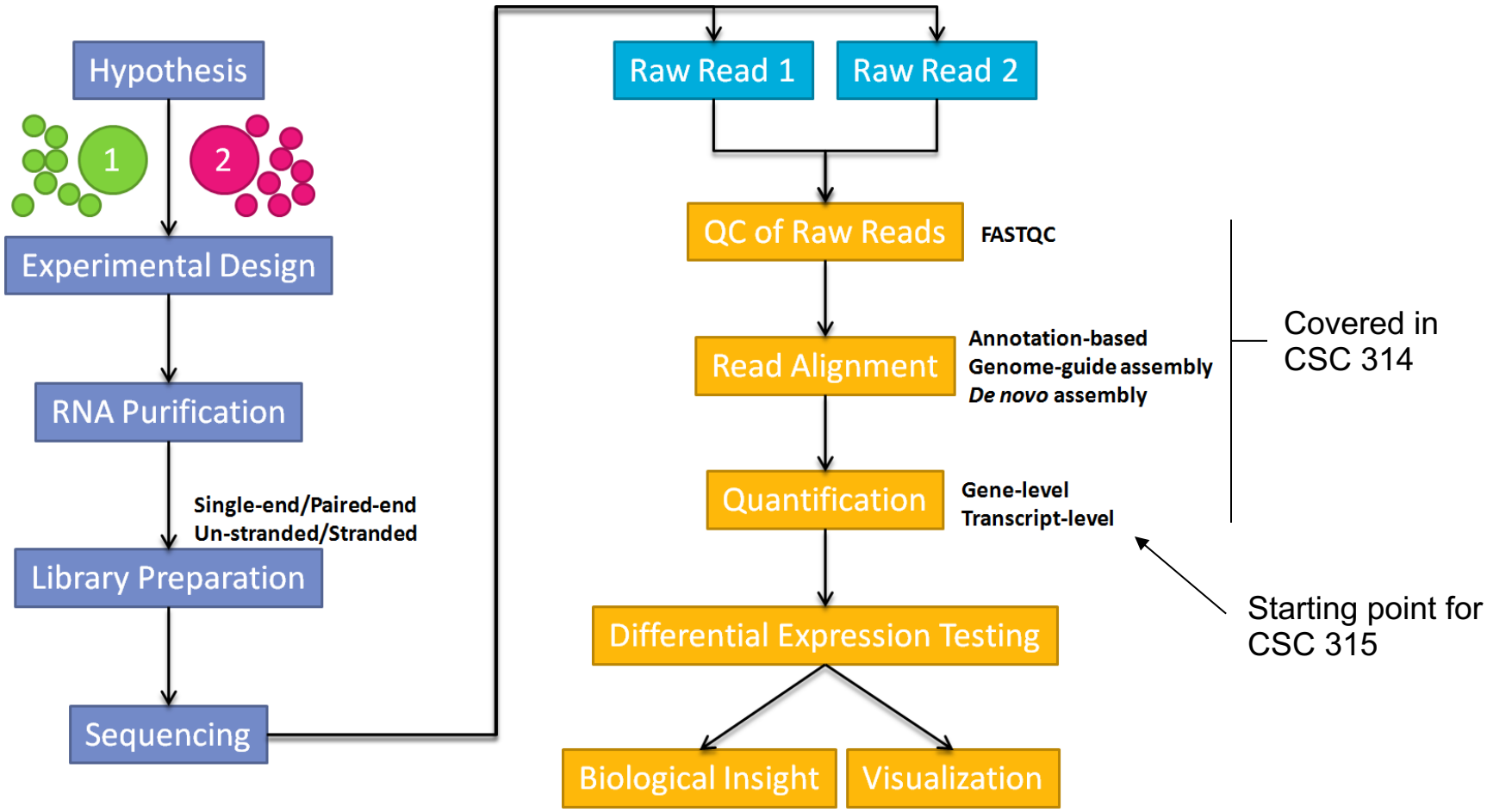


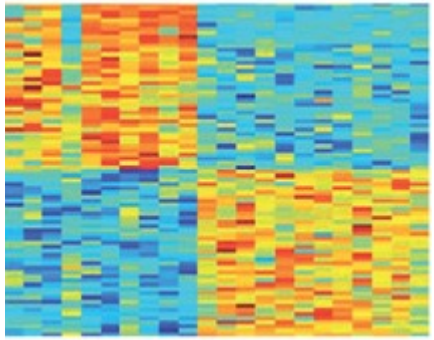
Image from: https://databeauty.com/blog/tutorial/2016/09/13/RNA-seq-analysis.html

Biomarkers and personalized medicine

Gene expression profiles

Samples

Genes



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

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

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

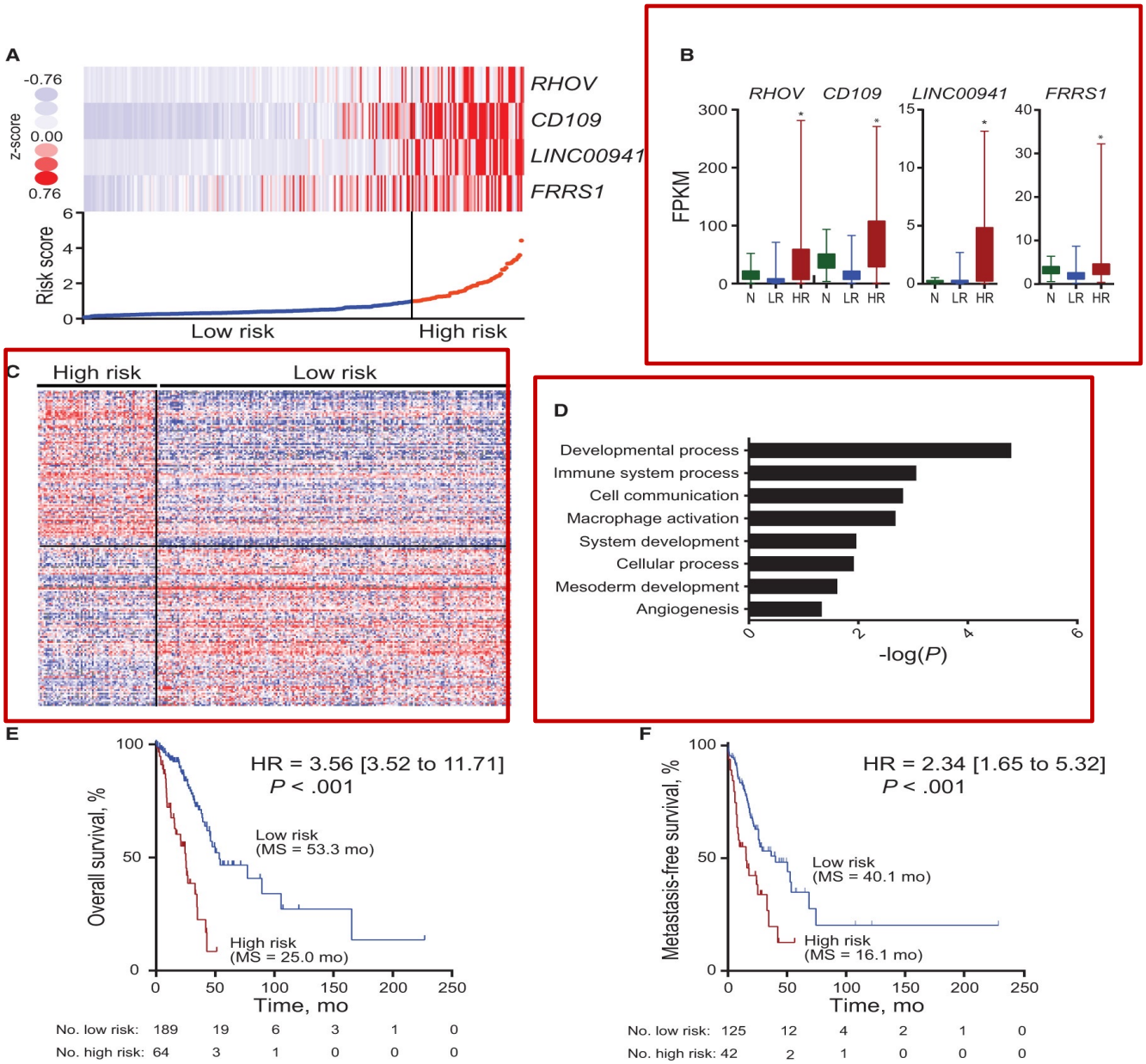


Figure 2. Four-gene prognostic signature biomarker characteristics in The Cancer Genoma Atlast (TCGA) cohort.

J Natl Cancer Inst, Volume 109, Issue 1, January 2017, djw200, https://doi.org/10.1093/jnci/djw200