Biological Context for Computational Genomics

Ben Langmead

JOHNS HOPKINS
WHITING SCHOOL *of* ENGINEERING

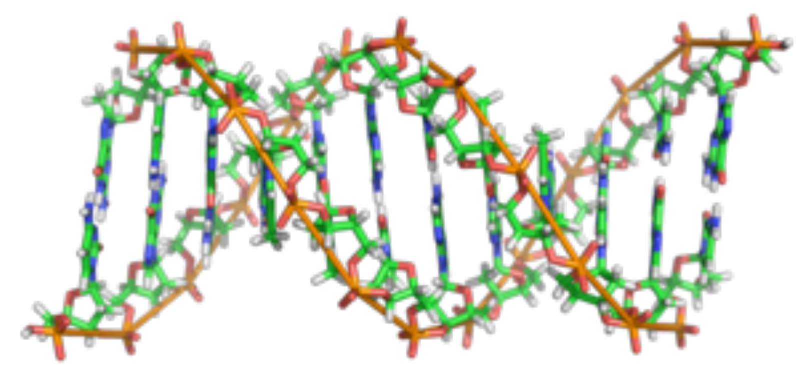
Department of Computer Science

Genome

"The complete set of genes or genetic material present in a cell or organism."

Oxford dictionaries

"Blueprint" or "recipe" of life

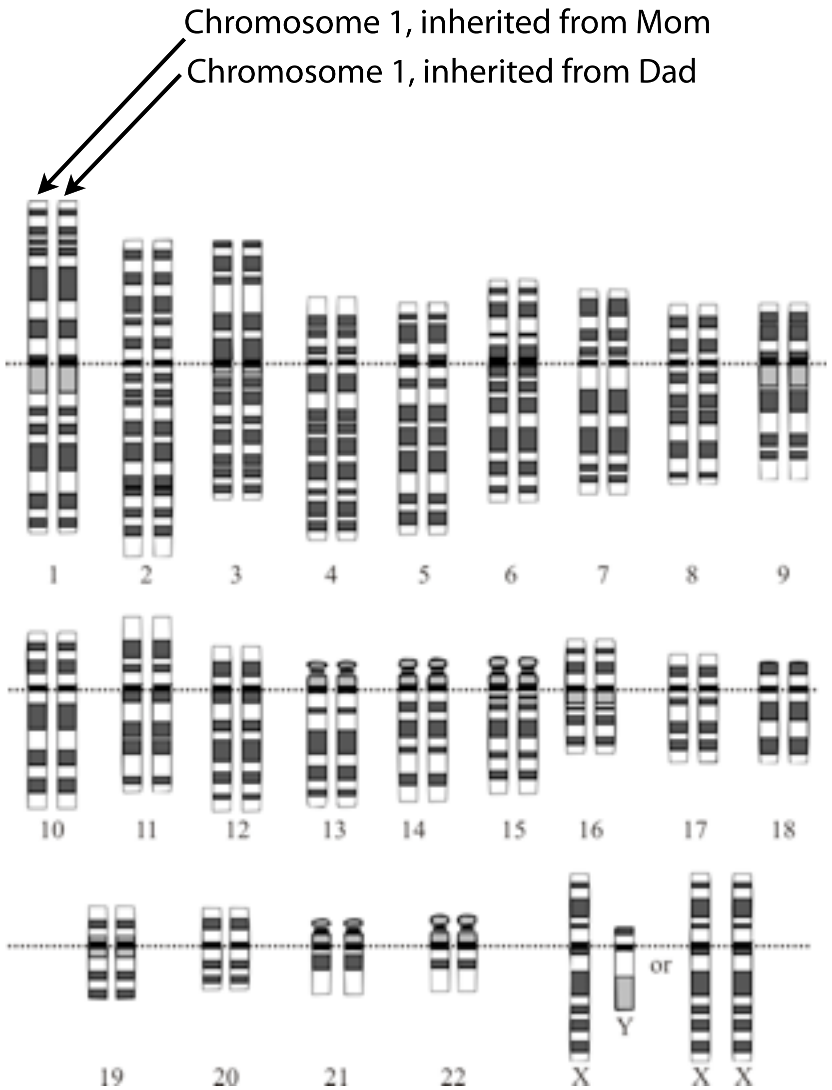


Self-copying store of read-only information about how to develop and maintain an organism

TAGCCCGACTTG

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

The genome: where genotypes live



Human chromosomes

23 pairs, 46 total
22 pairs are "autosomes"
1 pair are "sex chromosomes"

Genome is the entire DNA sequence of an individual; all chromosomes

Human genome is 3 billion nt

"nt" = nucleotides long
similarly: "bp"

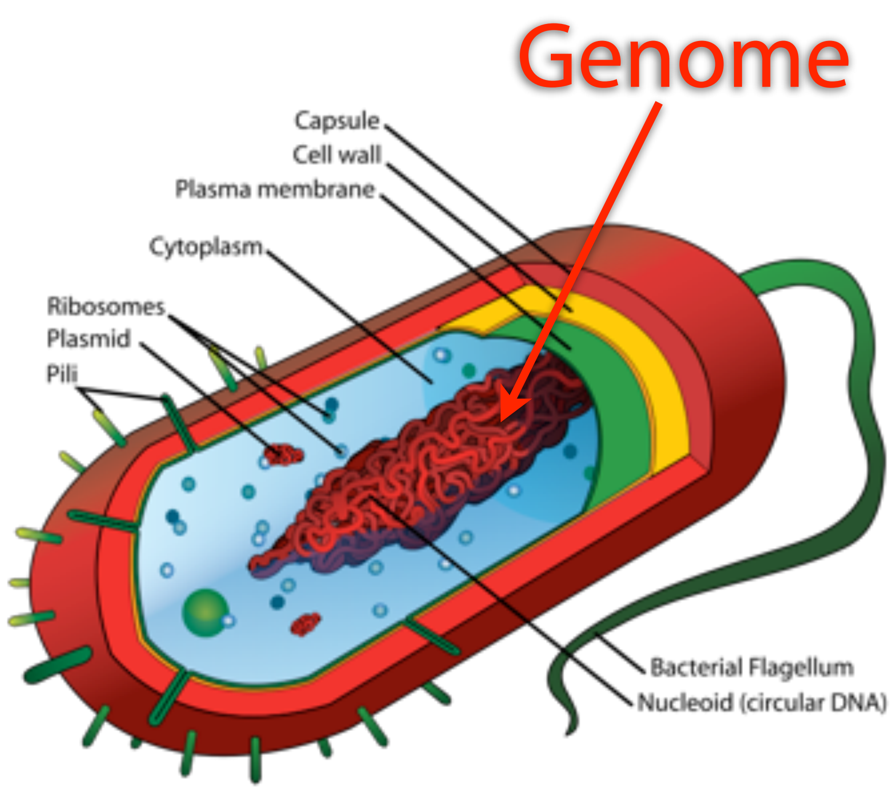
Most bacterial genomes are a few million nt. Most viral genomes are tens of thousands of nt. This plant's genome is about 150 billion nt. ⟶

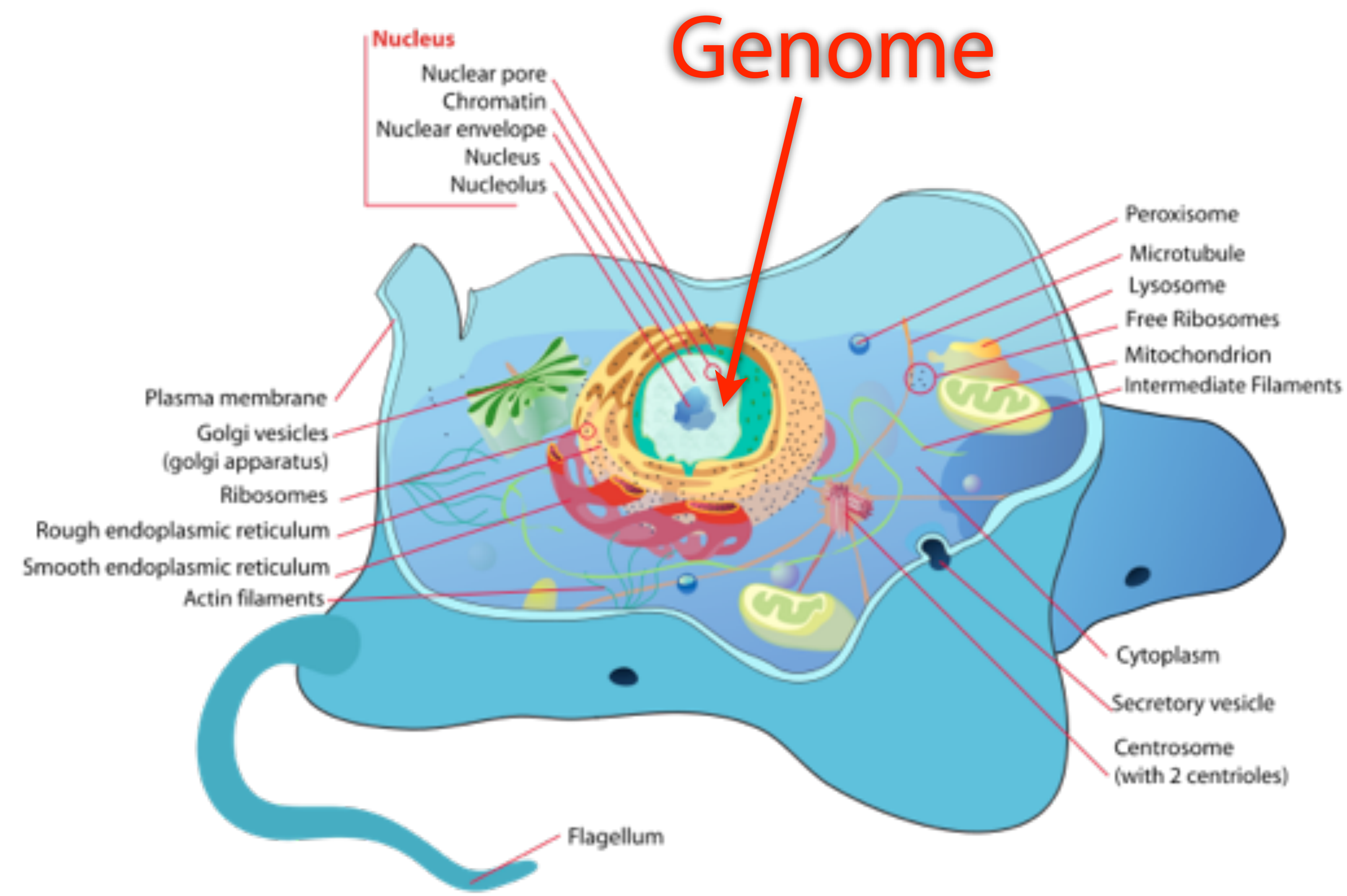
Paris japonica

Pictures: http://en.wikipedia.org/wiki/Chromosome,
http://en.wikipedia.org/wiki/Paris_japonica

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

Cells: where genomes live





Prokaryotic cell

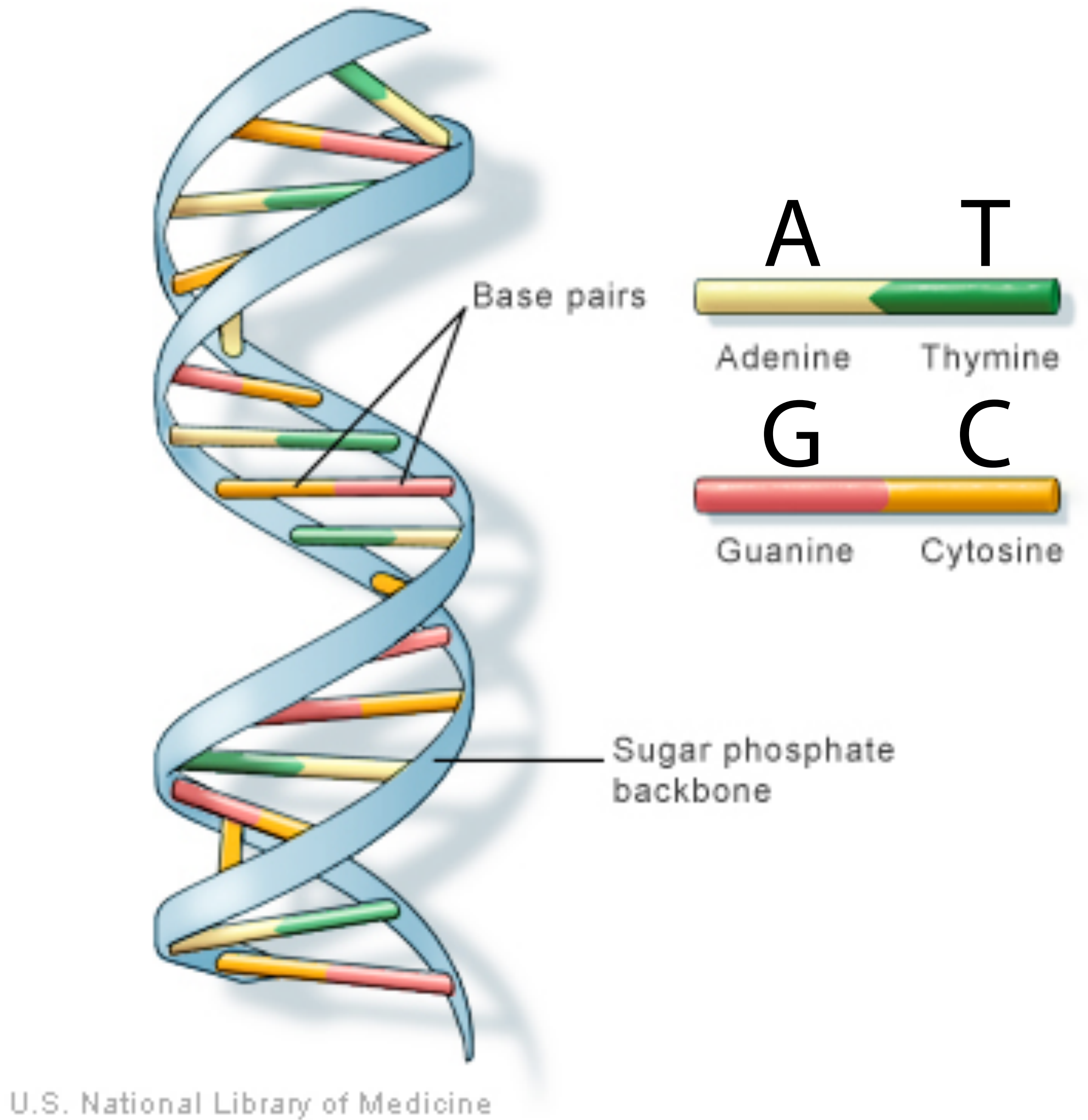
A bacterium consists of a single prokaryotic cell

Eukaryotic cell
(pictured: animal cell)

Make up animals, plants, fungi, other eukaryotes

Pictures: http://en.wikipedia.org/wiki/Cell_(biology)

DNA: the genome's molecule



Picture: http://ghr.nlm.nih.gov/handbook/basics/dna

Deoxyribonucleic acid

"Rungs" of DNA double-helix are base pairs. Pair combines two complementary bases.

Complementary pairings: A-T, C-G

Single base also called a "nucleotide"

DNA fragment lengths are measured in "base pairs" (abbreviated bp), "bases" (b) or "nucleotides" (nt)

Stringizing DNA

DNA has *direction* (a *5′ head* and a *3′ tail*). When we write a DNA *string*, we follow this convention.

When we write a DNA string, we write just one strand. The other strand is its *reverse complement*.

To get reverse complement, reverse then complement nucleotides (i.e. interchange A/T and C/G)

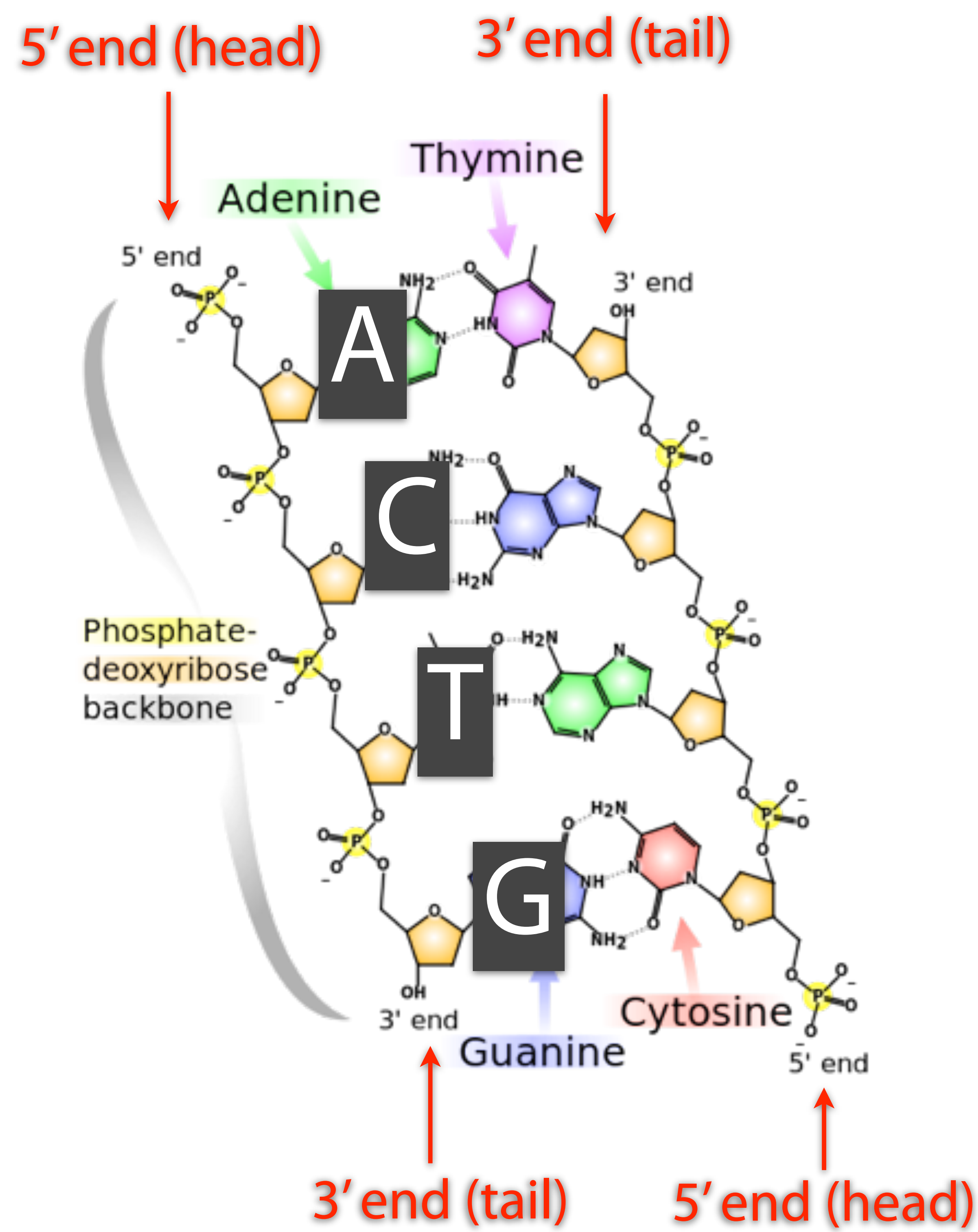
5′ end **A C T G** *3′ end*

↑

reverse complement

↓

5′ end **C A G T** *3′ end*



Picture: http://en.wikipedia.org/wiki/DNA

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

RNA

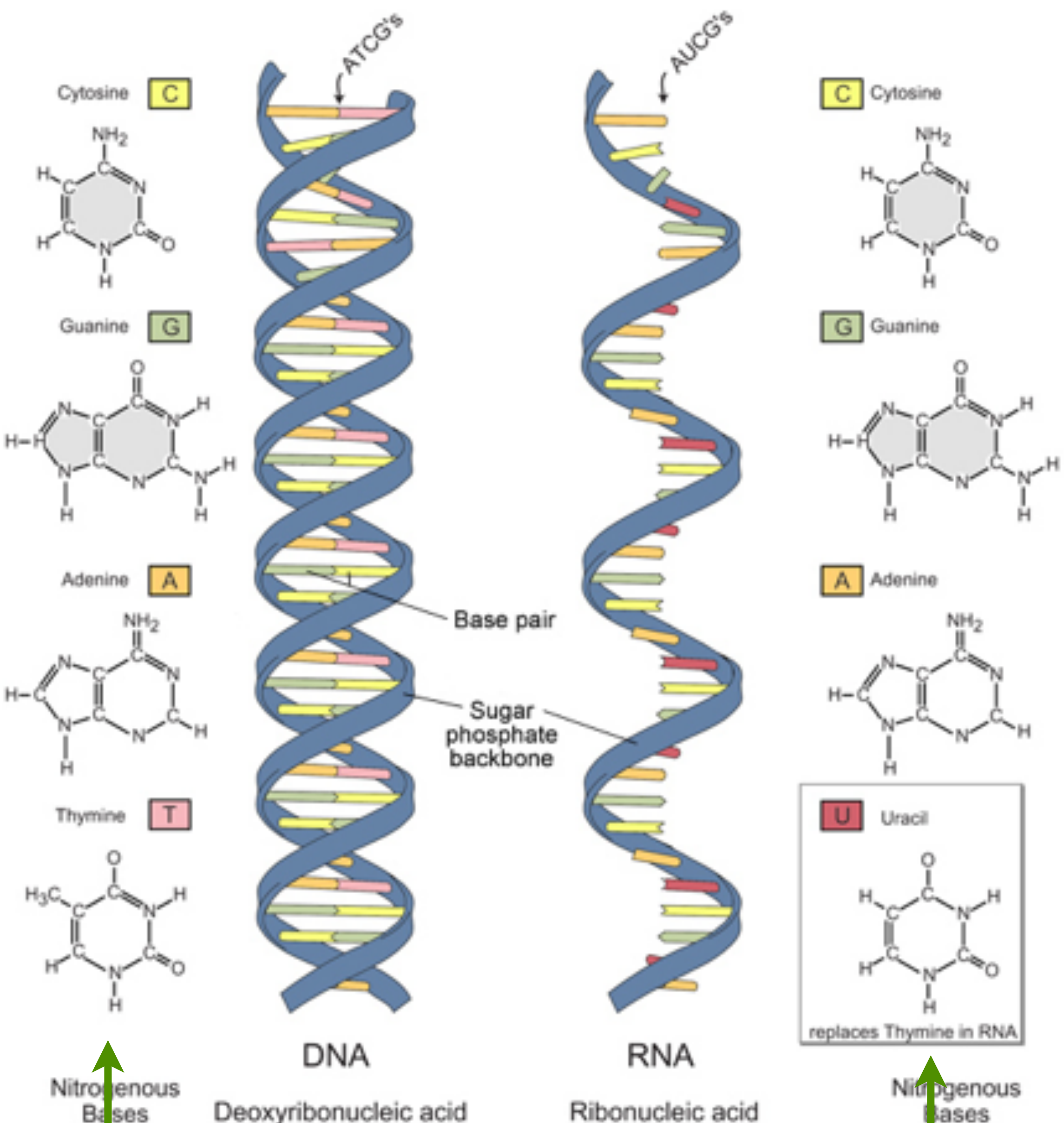


Image adapted from: National Human Genome Research Institute.

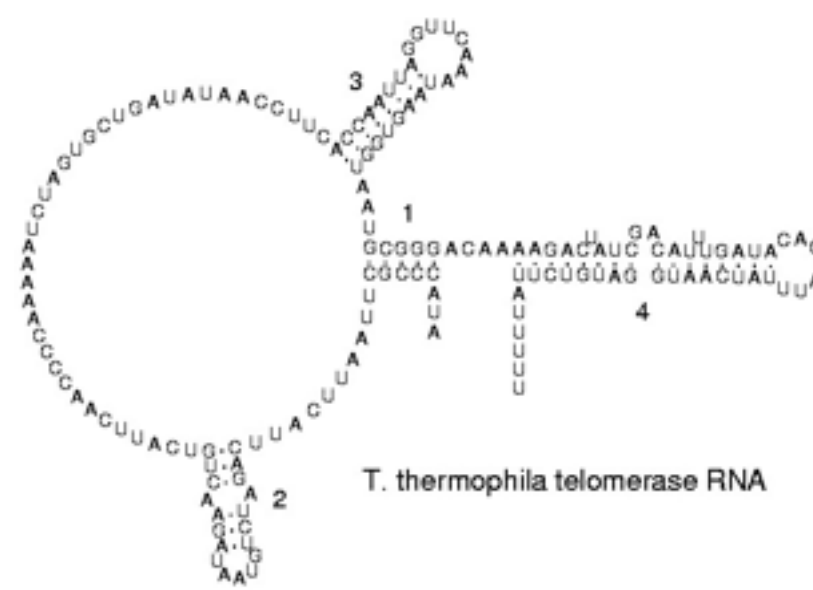
U instead of T

Like DNA but:

Single-stranded

Uses Uracil (U) instead of Thymine (T)

Sugar in the backbone is ribose instead of deoxyribose



T. thermophila telomerase RNA

Picture: http://en.wikipedia.org/wiki/Rna

The central dogma of molecular biology

Short version:

DNA -> RNA -> Protein

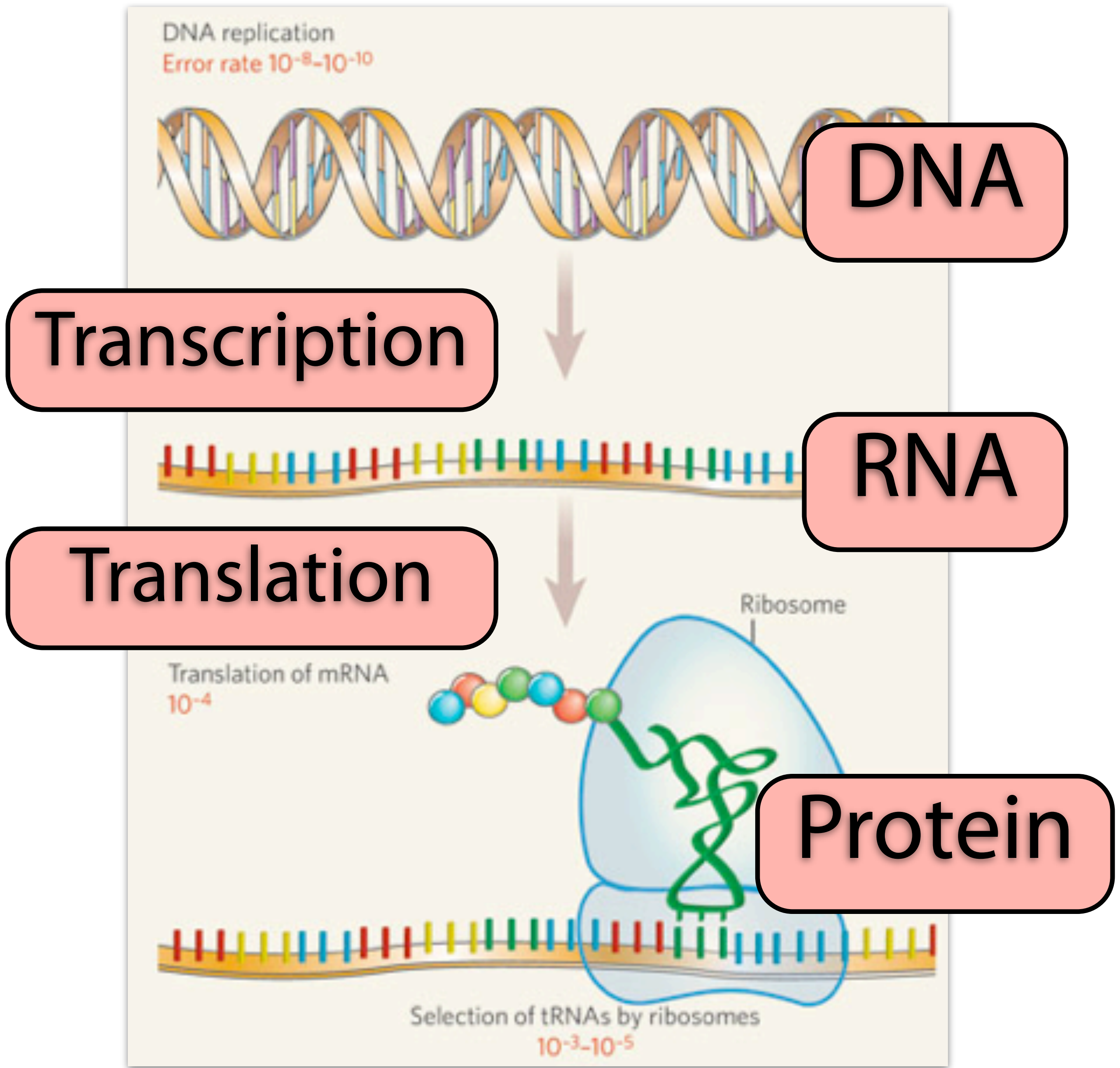
Long version:

DNA molecules contain information about how to create proteins; this information is *transcribed* into RNA molecules, which, in turn, direct chemical machinery which *translates* the nucleic acid message into a protein.

Hunter, Lawrence. "Life and its molecules: A brief introduction." *AI Magazine* 25.1 (2004): 9.

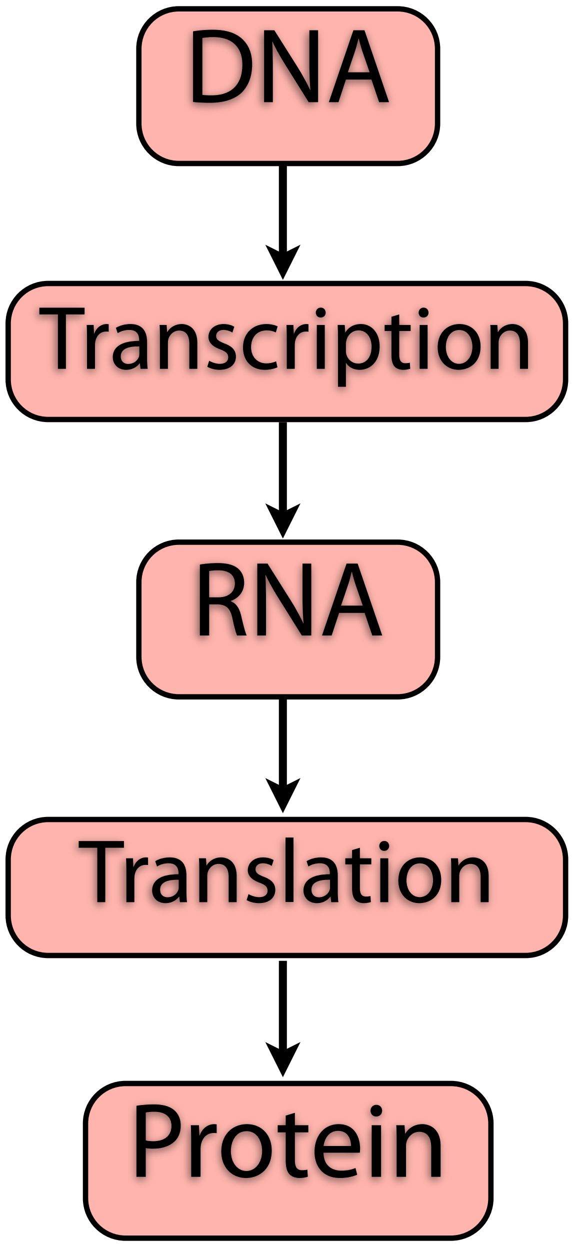
Links genotype and phenotype

First stated by Francis Crick in 1958



Picture from: Roy H, Ibba M. Molecular biology: sticky end in protein synthesis. Nature. 2006 Sep 7;443(7107):41-2.

The central dogma of molecular biology



Transcription: process whereby protein-coding stretches of DNA are **transcribed** into messenger RNA molecules

Translation: process whereby messenger RNAs are fed into the ribosome, which **translates** RNA nucleic acids into protein amino acids

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

The Central Dogma: Genetic code

DNA codes for protein, but DNA alphabet has 4 nucleic acids, whereas protein alphabet has ~20 amino acids

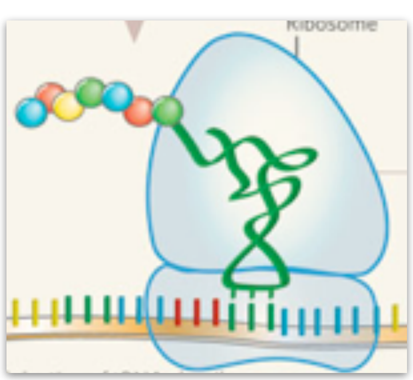
A *triplet* of nucleic acids (*codon*) codes for one amino acid

The code is *redundant*. E.g., both GGC and GGA code for Gly (Glycine)

Second letter

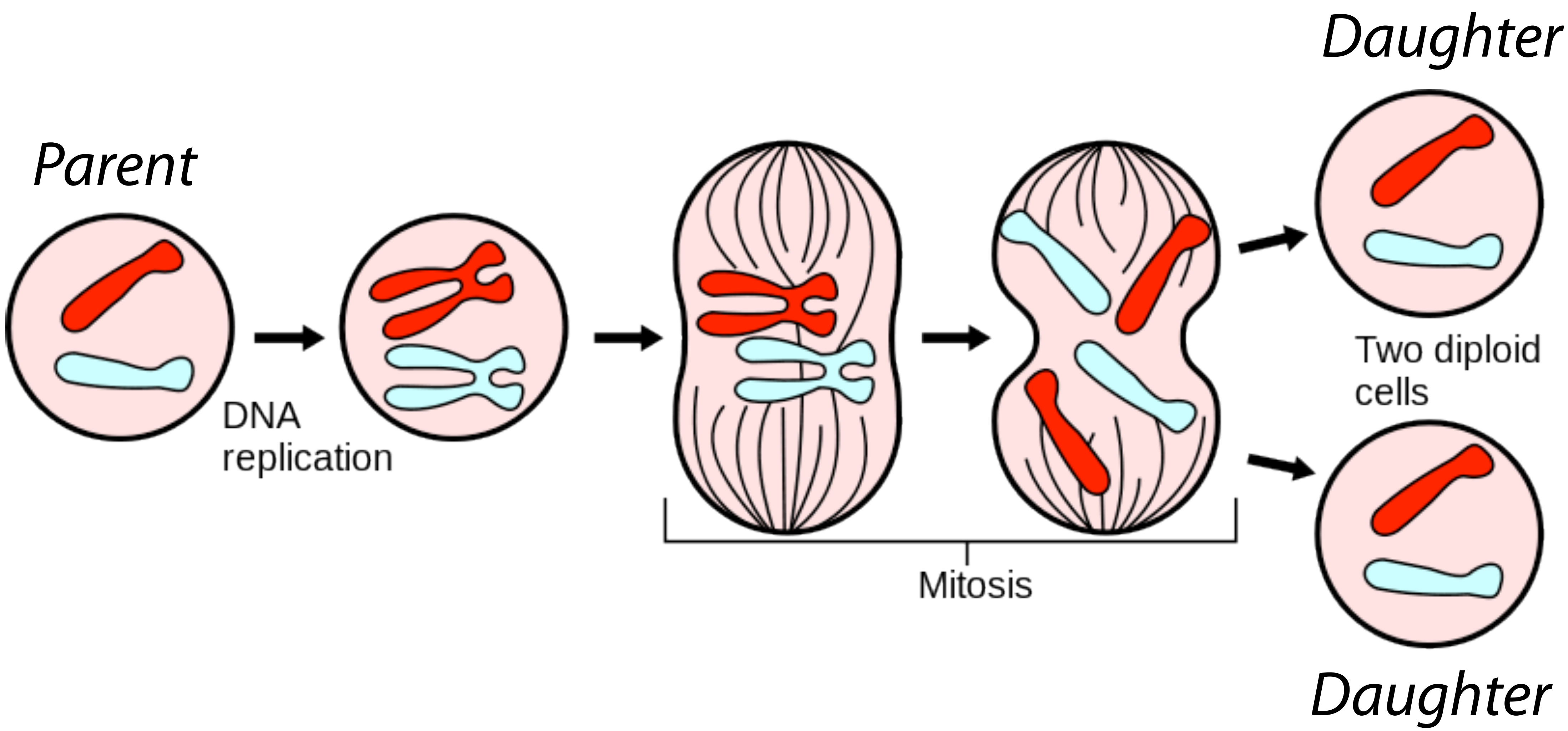
First letter	U	C	A	G	Third letter
U	UUU Phe UUC Phe UUA Leu UUG Leu	UCU Ser UCC Ser UCA Ser UCG Ser	UAU Tyr UAC Tyr UAA Stop UAG Stop	UGU Cys UGC Cys UGA Stop UGG Trp	U C A G
C	CUU Leu CUC Leu CUA Leu CUG Leu	CCU Pro CCC Pro CCA Pro CCG Pro	CAU His CAC His CAA Gln CAG Gln	CGU Arg CGC Arg CGA Arg CGG Arg	U C A G
A	AUU Ile AUC Ile AUA Ile AUG Met	ACU Thr ACC Thr ACA Thr ACG Thr	AAU Asn AAC Asn AAA Lys AAG Lys	AGU Ser AGC Ser AGA Arg AGG Arg	U C A G
G	GUU Val GUC Val GUA Val GUG Val	GCU Ala GCC Ala GCA Ala GCG Ala	GAU Asp GAC Asp GAA Glu GAG Glu	GGU Gly GGC Gly GGA Gly GGG Gly	U C A G

Picture: http://www.mun.ca/biology/scarr/MGA2_03-20.html



JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Cells: division



During cell division (*mitosis*), the genome is copied

Picture: http://en.wikipedia.org/wiki/Mitosis

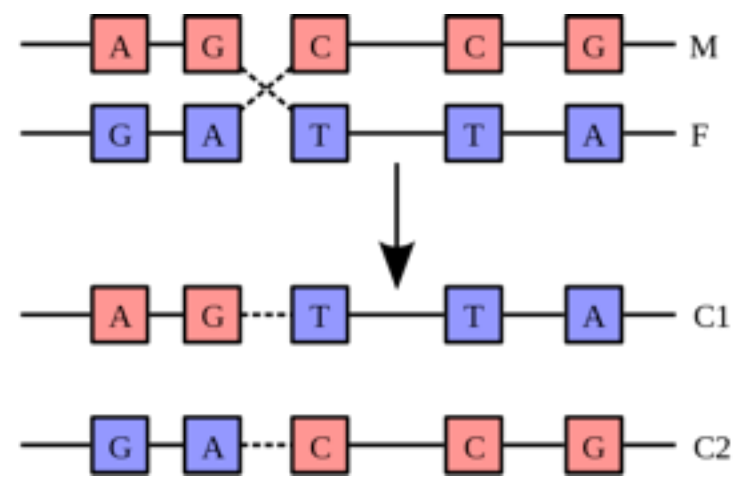
Evolution: why *these* genotypes?

Organisms reproduce, offspring *inherit* genotype from parents

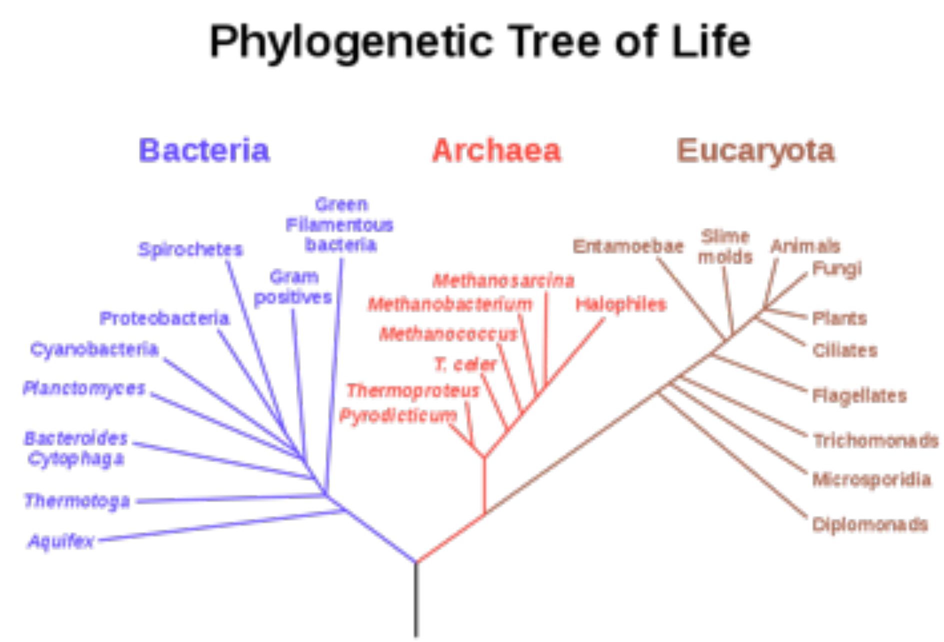
Random *mutation* changes genotypes and *recombination* shuffles chunks of genotypes together in new combinations

Natural *selection* favors phenotypes that reproduce more

Over time, this yields the variety of life on Earth. Incredibly, all organisms share a common ancestor.



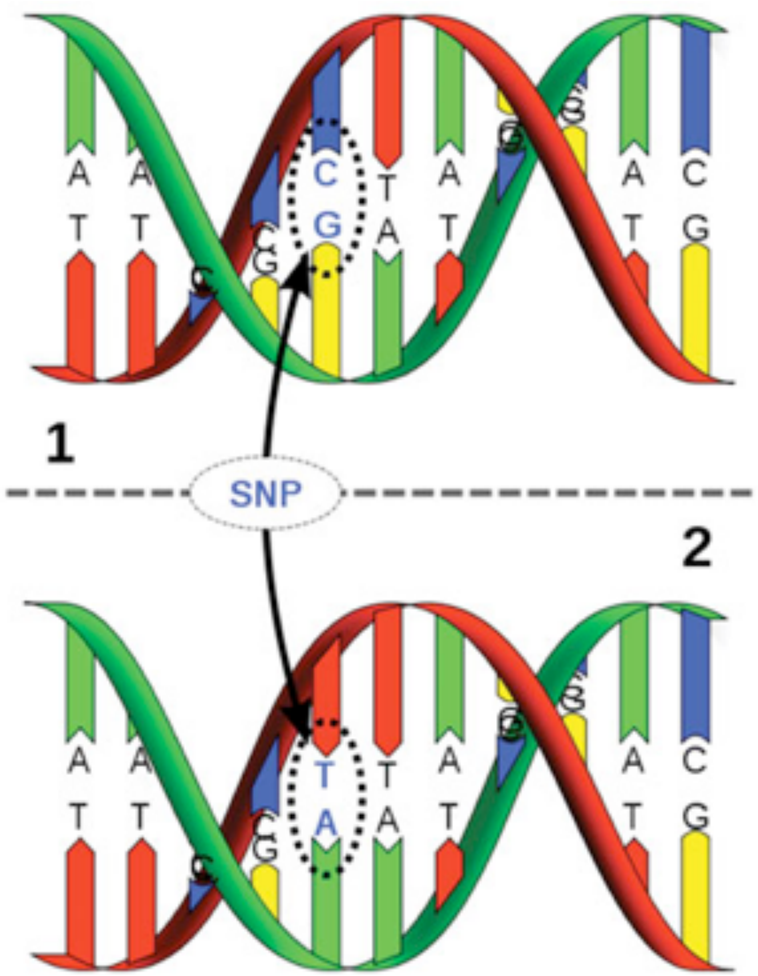
http://en.wikipedia.org/wiki/Genetic_recombination



http://en.wikipedia.org/wiki/Evolutionary_tree

The genome: variation

Two unrelated humans have genomes that are ~99.8% similar by sequence. There are about 3-4 million differences. Most are small, e.g. Single Nucleotide Polymorphisms (SNPs).



Human and chimpanzee genomes are about 96% similar

Pictures: http://www.dana.org/news/publications/detail.aspx?id=24536, http://en.wikipedia.org/wiki/Chimpanzee