

# Biological Context for Computational Genomics

Ben Langmead



JOHNS HOPKINS

WHITING SCHOOL  
*of* ENGINEERING

Department of Computer Science

You are free to use these slides. If you do, please sign the guestbook ([www.langmead-lab.org/teaching-materials](http://www.langmead-lab.org/teaching-materials)), or email me ([ben.langmead@gmail.com](mailto:ben.langmead@gmail.com)) and tell me briefly how you're using them. For original Keynote files, email me.

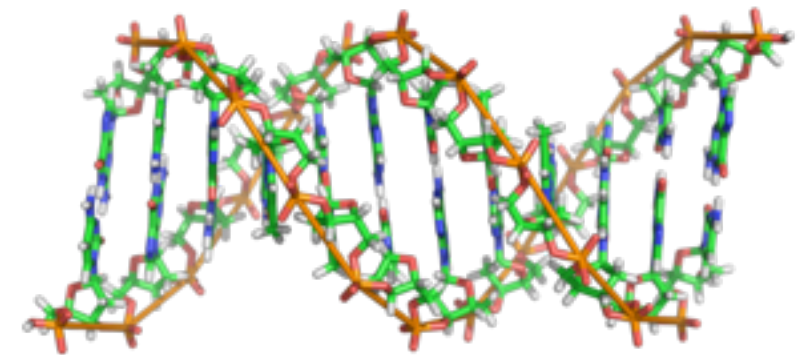
# Genome

“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



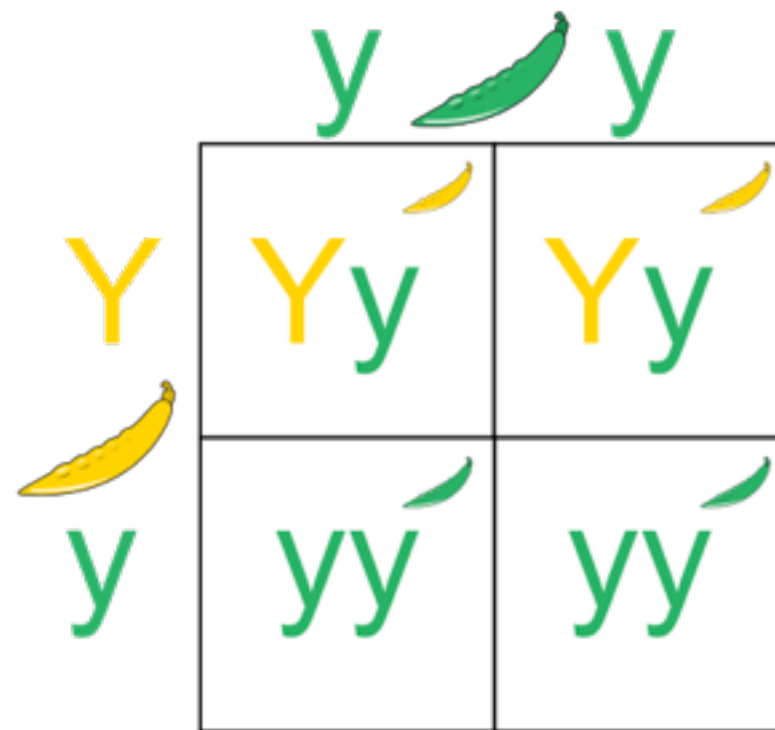
TAGCCCGACTTG



# Genotype & phenotype

*Genotype* is all the inherited information

*Phenotype* is something we observe, like pea pod color



Punnet square

[http://en.wikipedia.org/wiki/Punnett\\_square](http://en.wikipedia.org/wiki/Punnett_square)

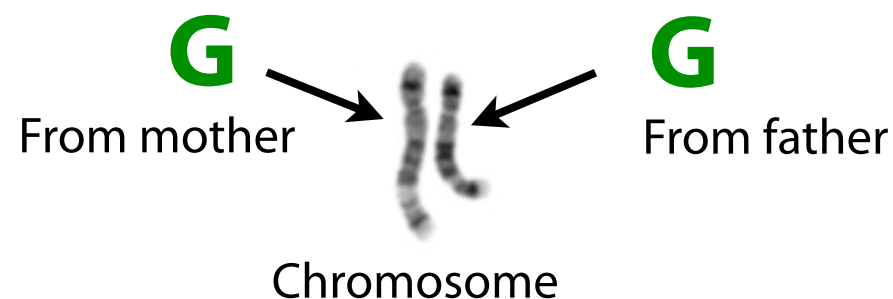
# Genotype & phenotype

Who	Genotype	What It Means
	AA	In Europeans, 85% chance of brown eyes; 14% chance of green eyes; 1% chance of blue eyes.
	AG	In Europeans, 56% chance of brown eyes; 37% chance of green eyes; 7% chance of blue eyes.
Benjamin Langmead	GG	In Europeans, 72% chance of blue eyes; 27% chance of green eyes; 1% chance of brown eyes.

Sources: 23andme.com, Ben's genome

A and G are *alleles*

The variable site is in a gene called HERC2



# Genotype & phenotype

Human Genetics

March 2008, Volume 123, Issue 2, pp 177-187



Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the *HERC2* gene inhibiting *OCA2* expression

Hans Eiberg, Jesper Troelsen, Mette Nielsen, Annemette Mikkelsen, Jonas Mengel-From, Klaus W. Kjaer, Lars Hansen

<http://link.springer.com/article/10.1007%2Fs00439-007-0460-x>

# Genotype, phenotype and environment

Sources: 23andme.com, my genome

Name	Outcome
Alcohol Flush Reaction	Does Not Flush
Bitter Taste Perception	Unlikely to Taste
Earwax Type	Wet
Eye Color	Likely Blue
Hair Curl 	Straighter Hair on Average
Lactose Intolerance	Likely Tolerant
Malaria Resistance (Duffy Antigen)	Possibly Resistant
Male Pattern Baldness 	Decreased Odds
Muscle Performance	Likely Sprinter
Non-ABO Blood Groups	See Report
Norovirus Resistance	Not Resistant

Blood type, height, susceptibility to diseases, ...

Sources: 23andme.com, Ben's genome

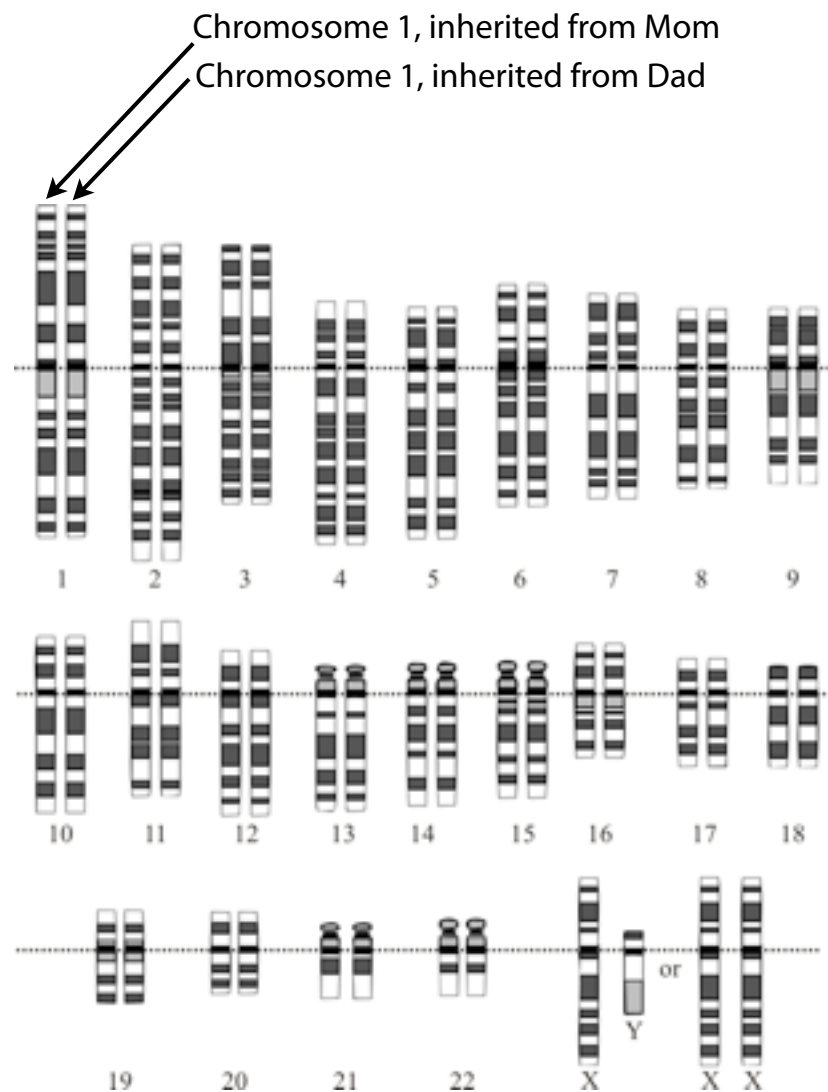
Note qualifiers: *likely, on average, possibly, decreased odds ...*

Outcomes may not be black-and-white since one trait can be affected by many genes or variants (*polygenic* or *quantitative* trait)

Besides genotype, *environment* affects phenotype. Consider muscle size (exercise), skin color (sun exposure), body mass index (diet), baldness (age).



# 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](http://en.wikipedia.org/wiki/Paris_japonica)

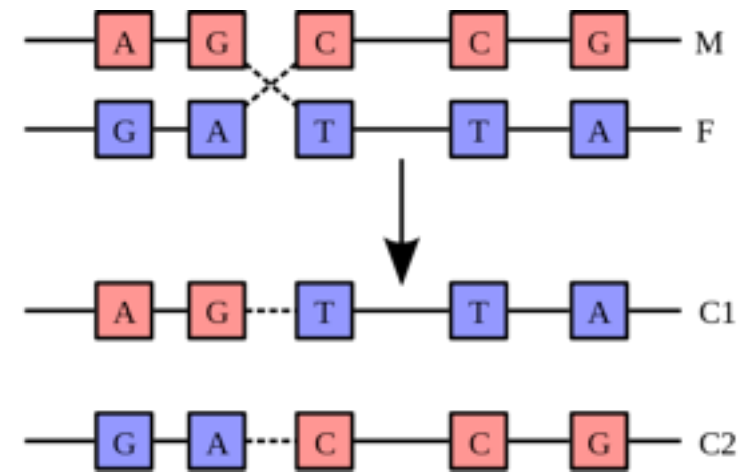
# Evolution: why *these* genotypes?

Organisms reproduce, offspring *inherit* genotype from parents

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

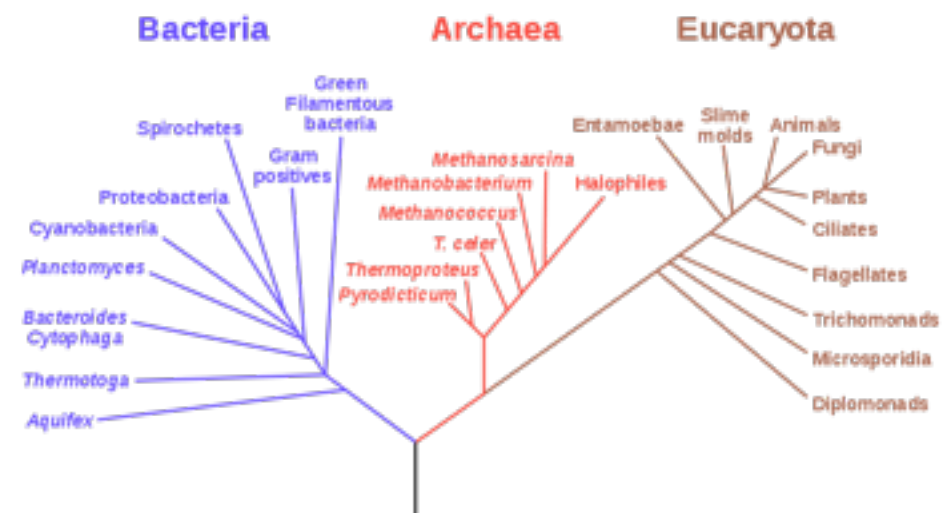
Natural *selection* favors phenotypes that reproduce more

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/Genetic_recombination)

## Phylogenetic Tree of Life

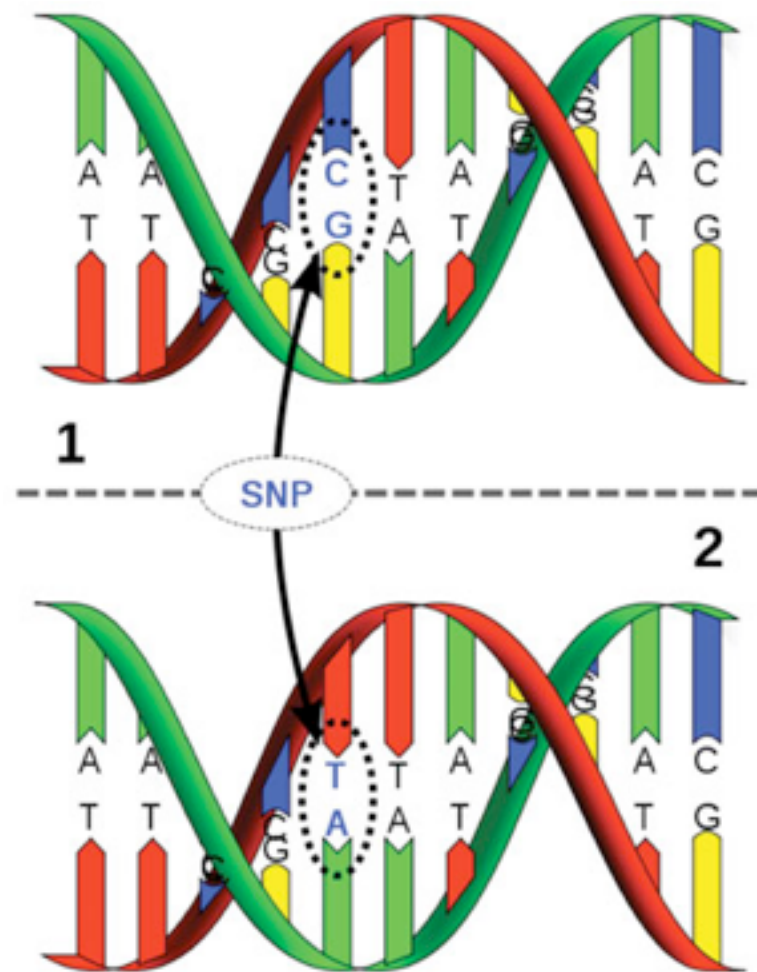


[http://en.wikipedia.org/wiki/Evolutionary\\_tree](http://en.wikipedia.org/wiki/Evolutionary_tree)



# 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).

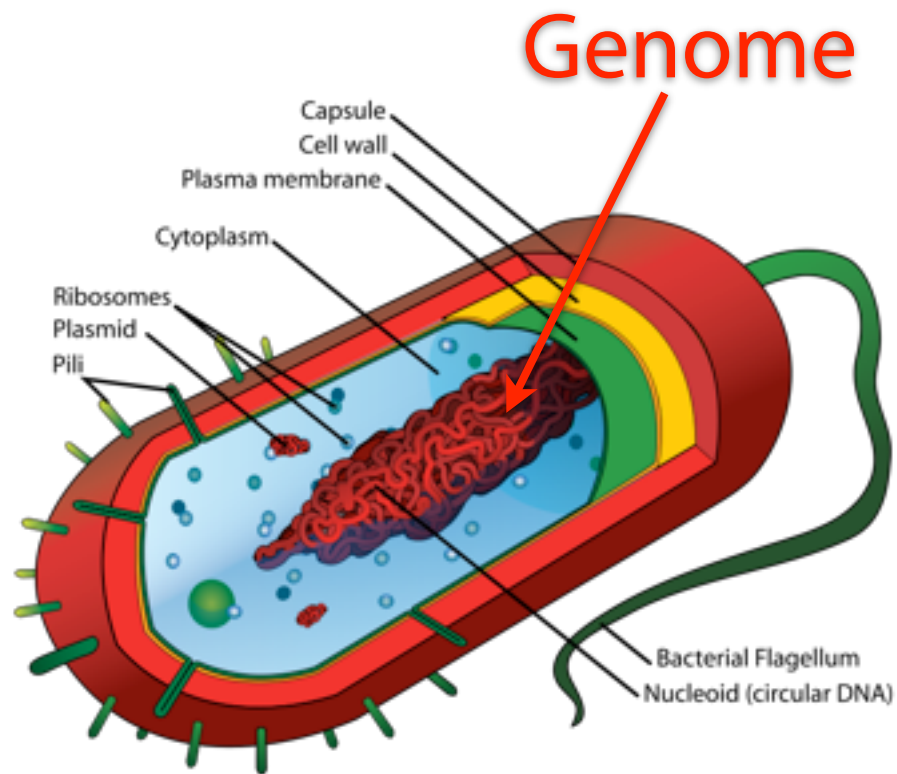


Human and chimpanzee genomes are about 96% similar



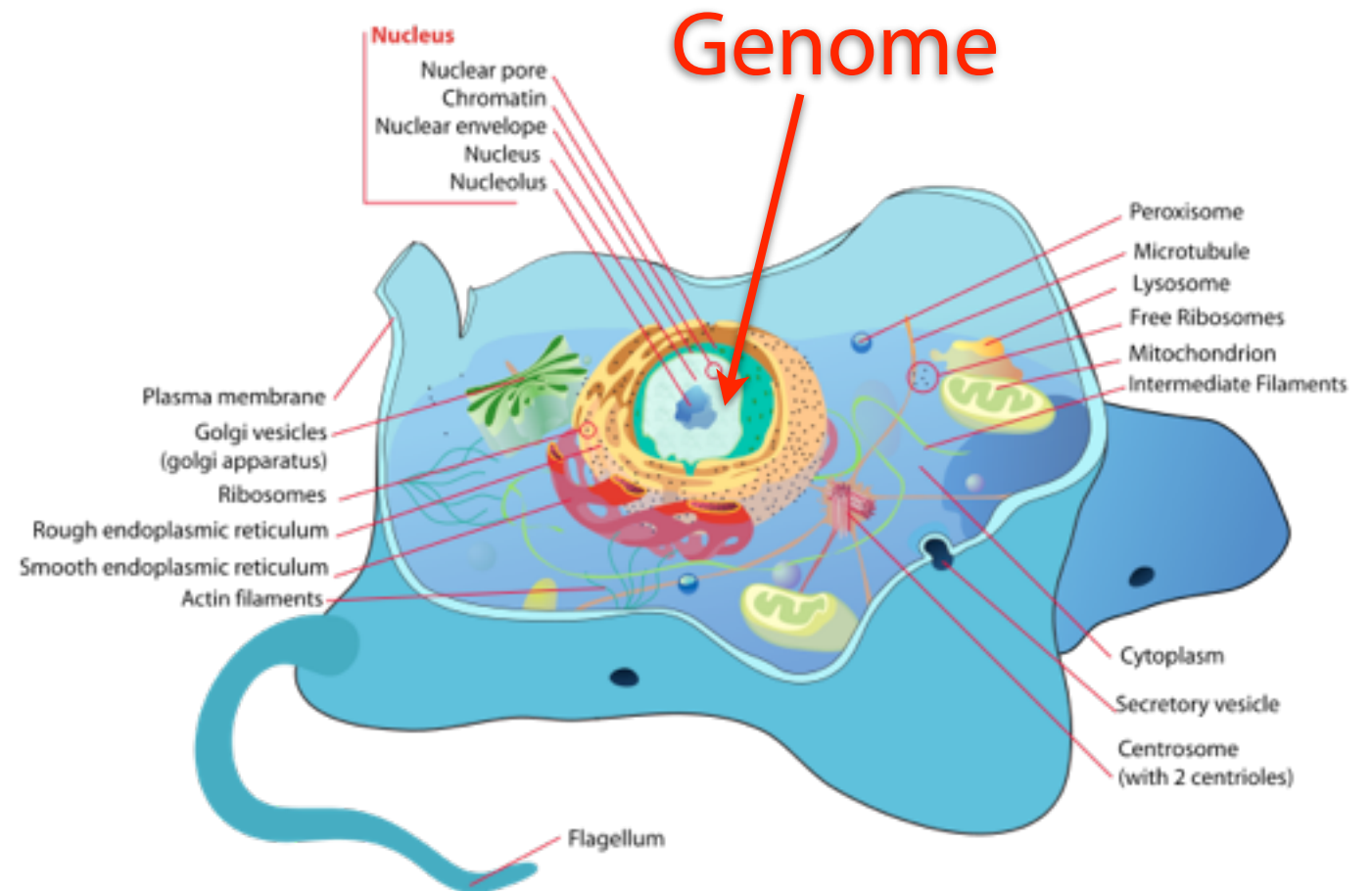
Pictures: <http://www.dana.org/news/publications/detail.aspx?id=24536>,  
<http://en.wikipedia.org/wiki/Chimpanzee>

# Cells: where genomes live



## Prokaryotic cell

A bacterium consists of a single prokaryotic cell

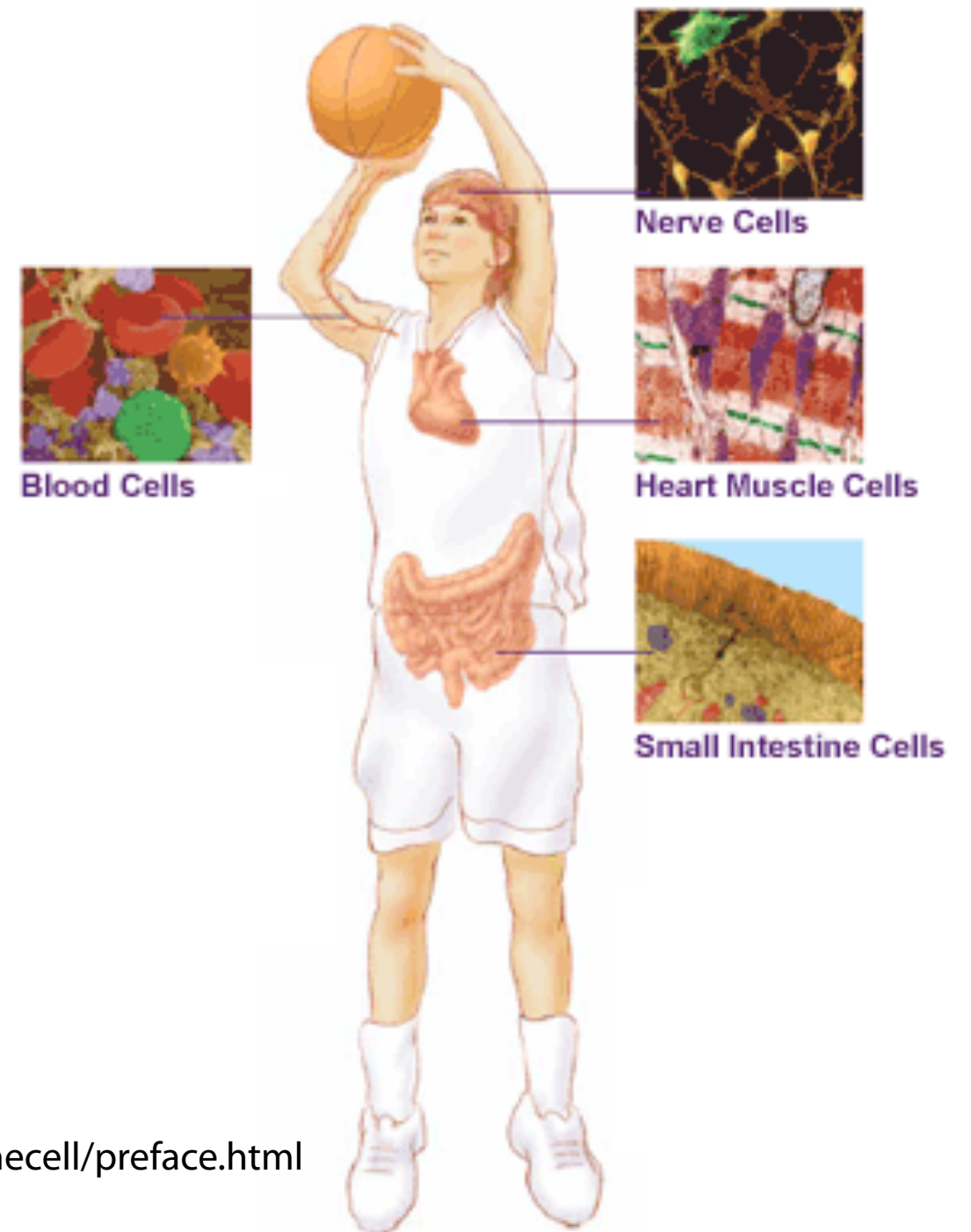


## Eukaryotic cell (pictured: animal cell)

Make up animals, plants, fungi, other eukaryotes

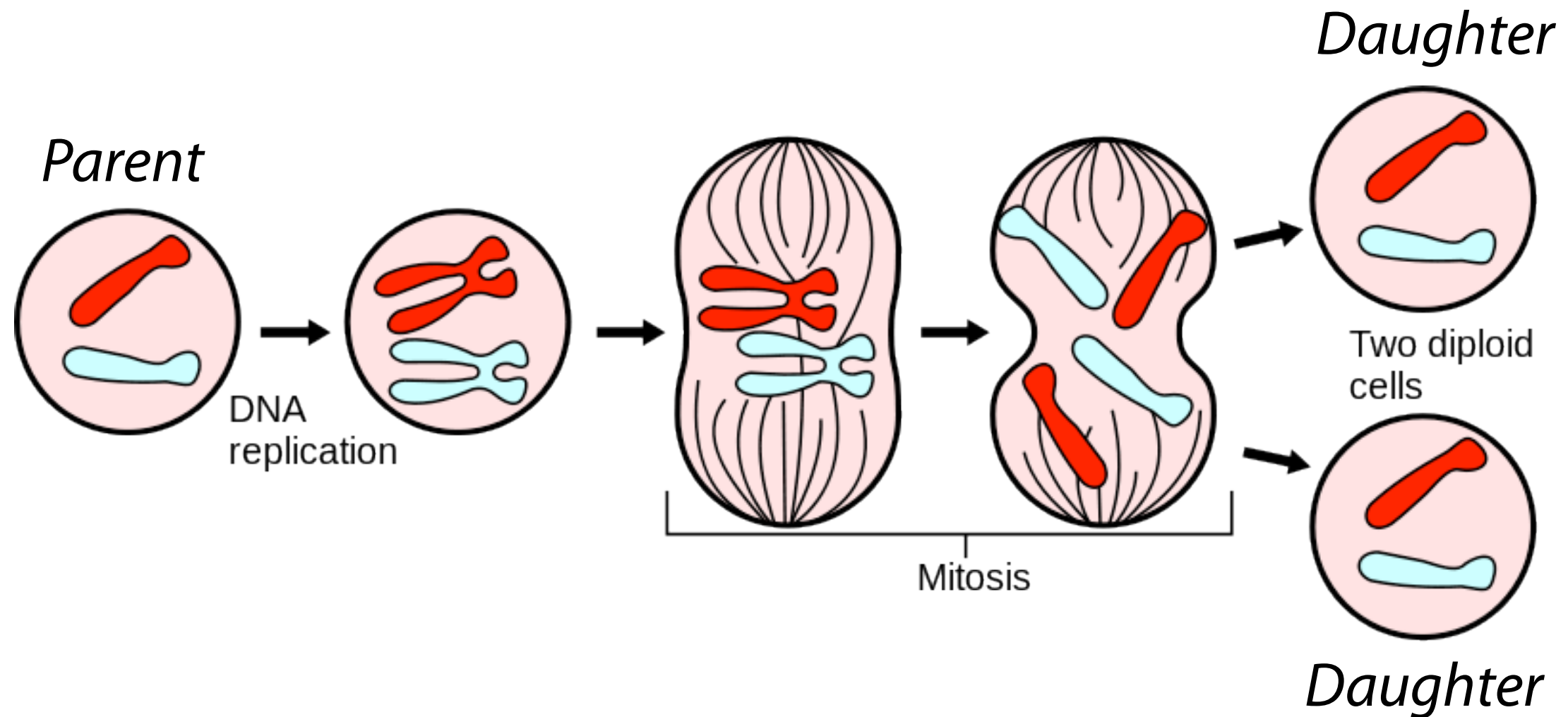
# Cells: where genomes live

All the trillions of cells in a person have same genomic DNA in the nucleus



Picture: <http://publications.nigms.nih.gov/insidethecell/preface.html>

# Cells: division

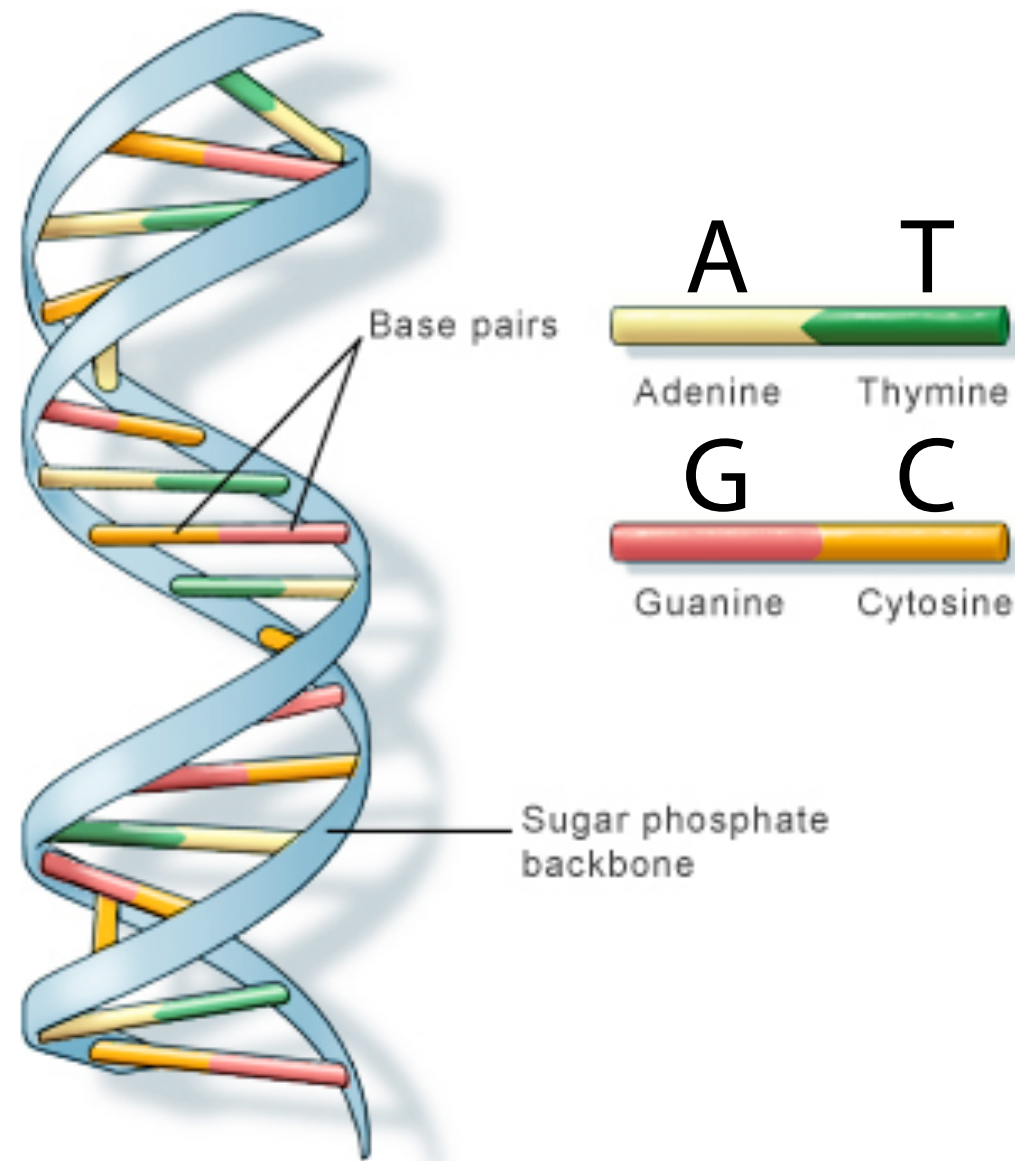


During cell division (*mitosis*), the genome is copied

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



# DNA: the genome's molecule



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)

U.S. National Library of Medicine

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

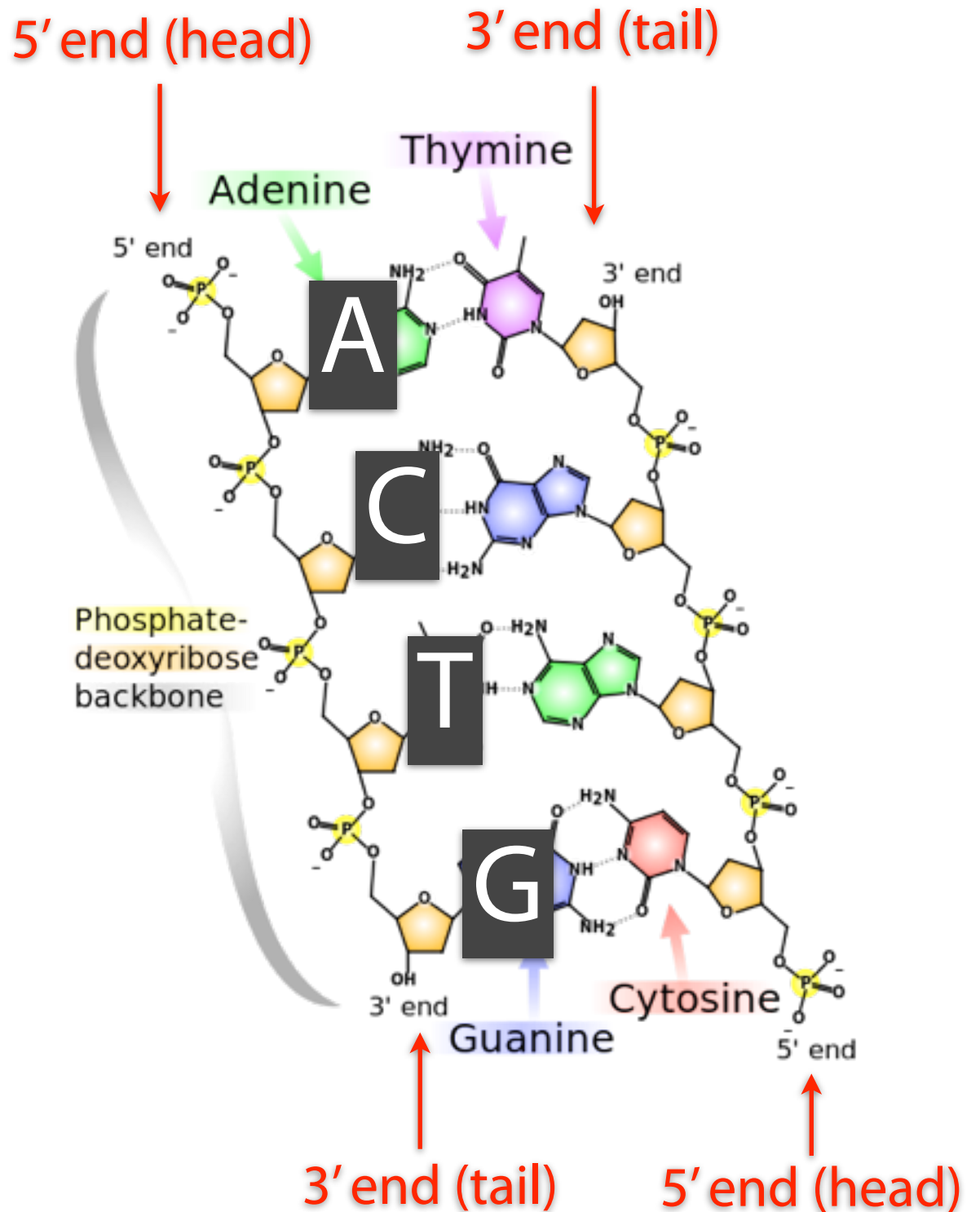
# Stringizing DNA

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>



# The central dogma of molecular biology

Short version:

DNA → RNA → Protein

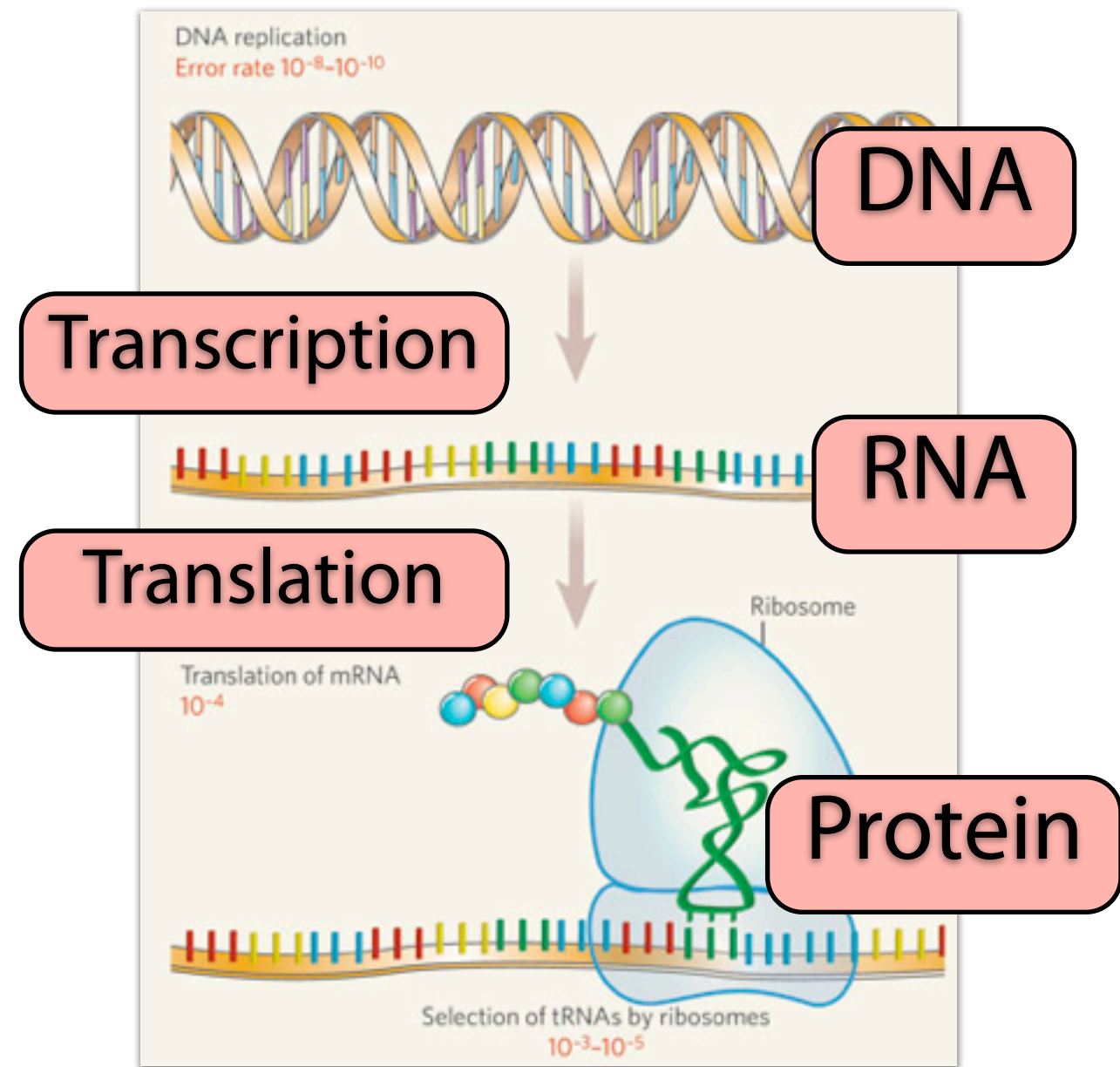
Long version:

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.

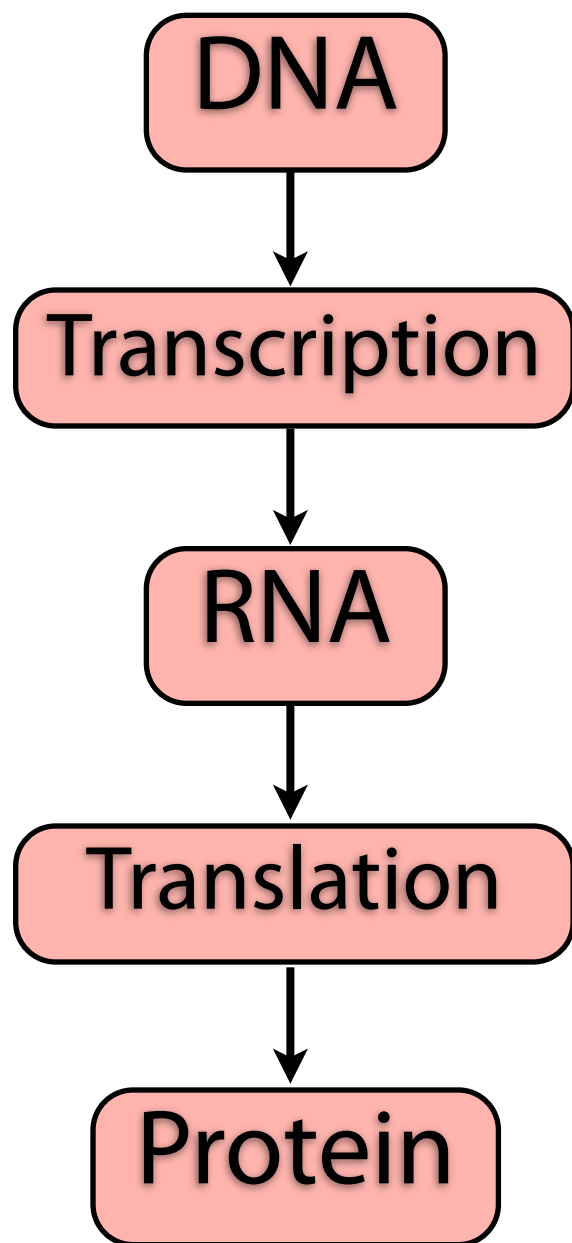
Links genotype and phenotype

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.

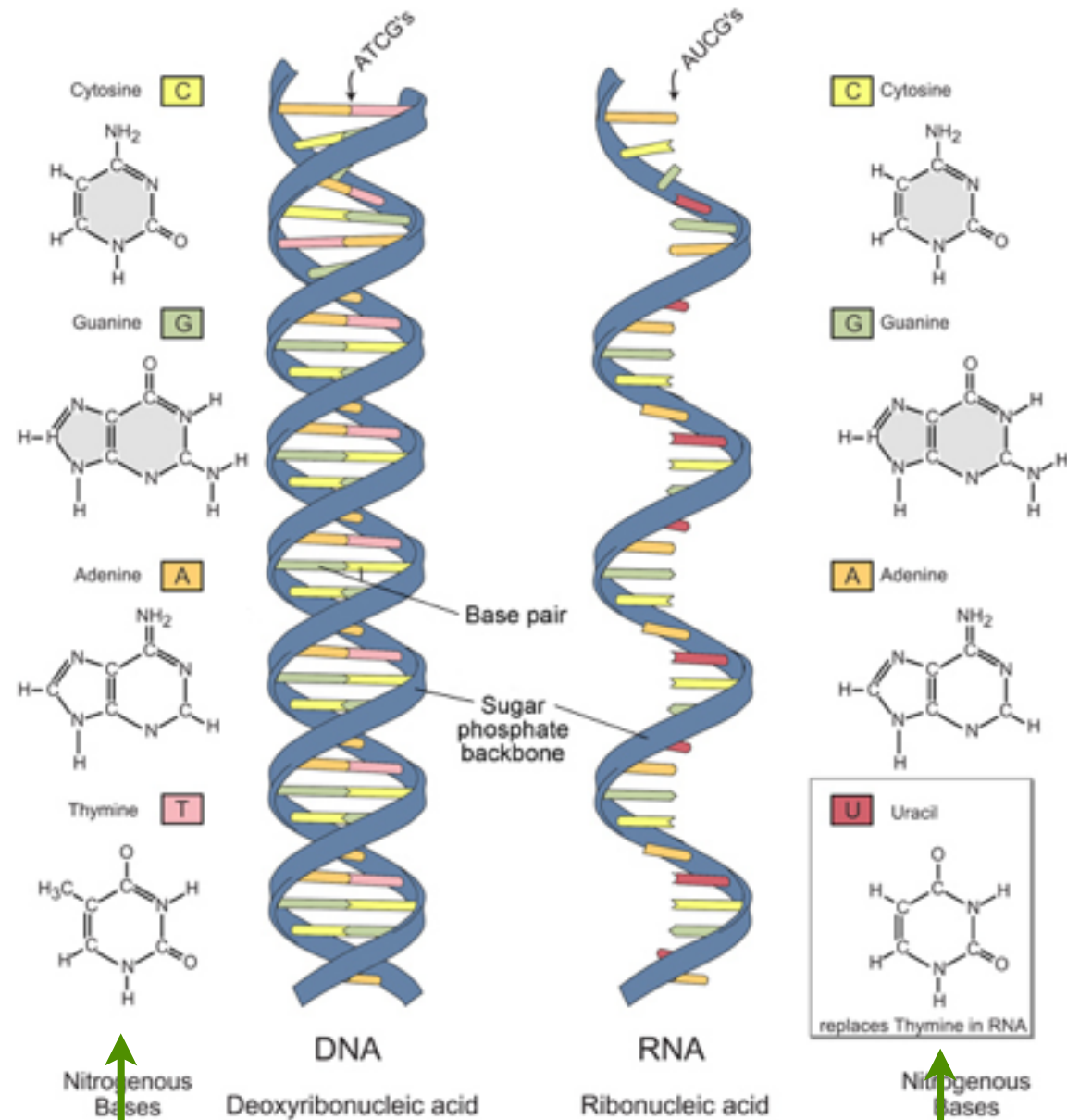
# The central dogma of molecular biology



**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

# RNA



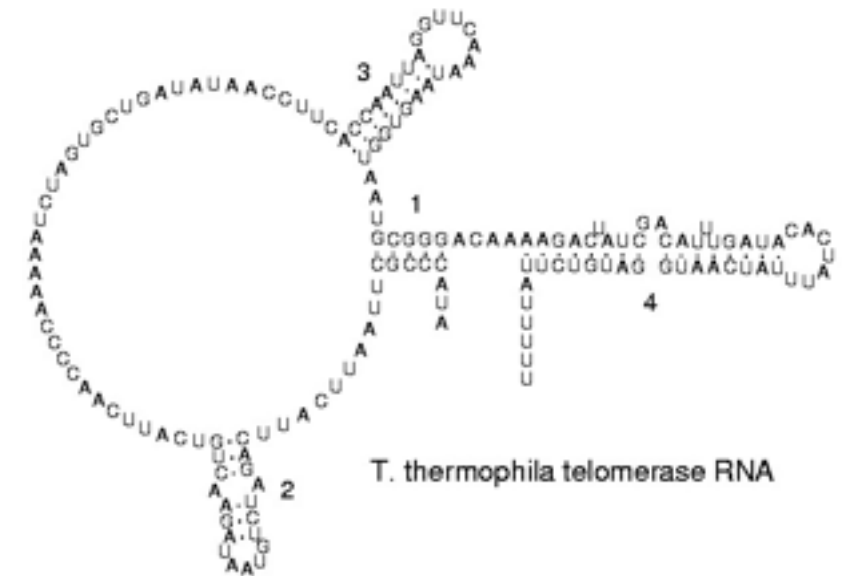
U instead of T

Like DNA but:

Single-stranded

Uses Uracil (U) instead of Thymine (T)

Sugar in the backbone is ribose instead of deoxyribose



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



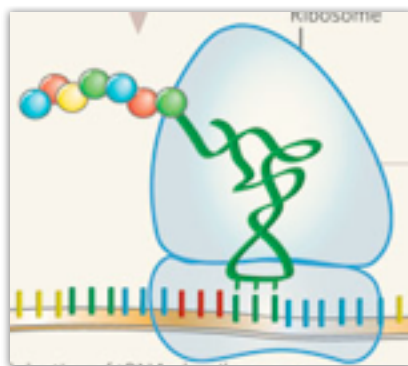
JOHNS HOPKINS  
WHITING SCHOOL  
of ENGINEERING

# 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)



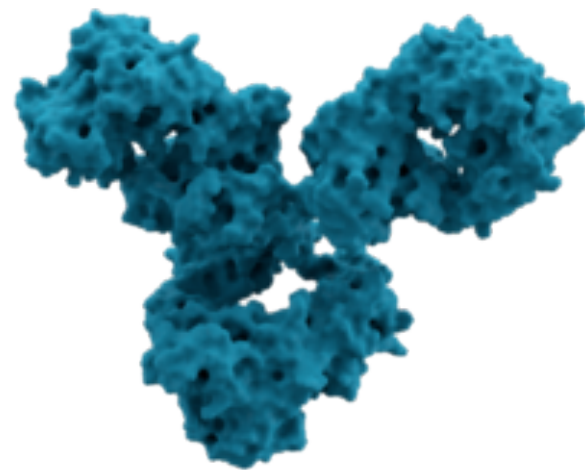
		Second letter				
		U	C	A	G	
First letter	U	UUU } Phe UUC } UUA } Leu UUG }	UCU } UCC } Ser UCA } UCG }	UAU } Tyr UAC } UAA Stop UAG Stop	UGU } Cys UGC } UGA Stop UGG Trp	U C A G
	C	CUU } CUC } Leu CUA } CUG }	CCU } CCC } Pro CCA } CCG }	CAU } His CAC } CAA } Gln CAG }	CGU } CGC } Arg CGA } CGG }	U C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

Picture: [http://www.mun.ca/biology/scarr/MGA2\\_03-20.html](http://www.mun.ca/biology/scarr/MGA2_03-20.html)

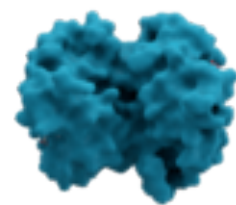


# The Central Dogma: Proteins

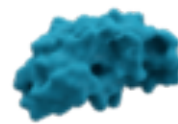
Proteins are typically 100s or 1000s of amino acids long, and fold into exquisitely complicated shapes



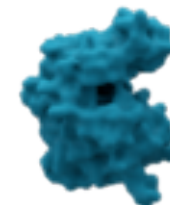
Immunoglobulin



Hemoglobin



Insulin



Adenylate  
Kinase

Proteins perform a vast array of functions within living organisms: catalyzing metabolic reactions, replicating DNA, transporting molecules from one location to another, etc

Sources: <http://en.wikipedia.org/wiki/Protein>, [http://en.wikipedia.org/wiki/Protein\\_structure](http://en.wikipedia.org/wiki/Protein_structure)

