



# Molecular Biology Primer



# Starting 19<sup>th</sup> century...

---

- Cellular biology:

- Cell as a fundamental building block

- 1850s+:

- “DNA” was discovered by Friedrich Miescher and Richard Altmann

- Mendel’s experiments with garden pea plants

- Laws of inheritance, “Alleles”, “genotype” vs. “Phenotype”

- 1909: Wilhelm Johannsen coined the word “gene”

- Still..... Proteins were thought to be the primary genetic materials... but..

# Avery's Experiment

A gene is made of DNA.

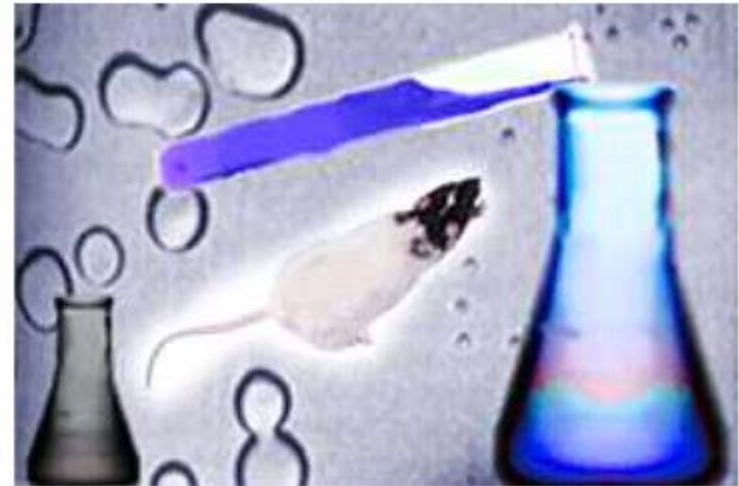
17

DNA FROM THE  
BEGINNING



In the 1920s, experiments showed that a harmless strain of bacteria can become infectious when mixed with a virulent strain of bacteria that had been killed. The dead bacteria apparently provide some chemical that "transforms" the harmless bacteria to infectious ones. This so-called "transforming principle" appeared to be a gene.

A team of scientists led by Oswald Avery at the Rockefeller Institute, rigorously followed up on these experiments in the 1940's. They found that a pure extract of the "transforming principle" was unaffected by treatment with protein-digesting enzymes but was destroyed by a DNA-digesting enzyme. This showed that the transforming principle is DNA – and, by extension, a gene is made of DNA. Still, many scientists were slow to accept this clear proof that DNA, not protein, is the genetic molecule.



# What does a gene produce?

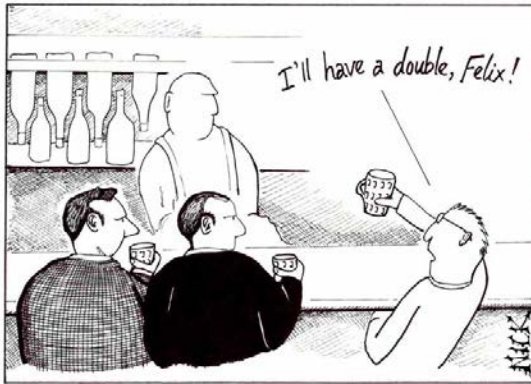
Gene → ... ?... → Protein

In 1902, Archibald Garrod described the inherited disorder alkaptonuria as an "inborn error of metabolism." He proposed that a gene mutation causes a specific defect in the biochemical pathway for eliminating liquid wastes. The phenotype of the disease – dark urine – is a reflection of this error.

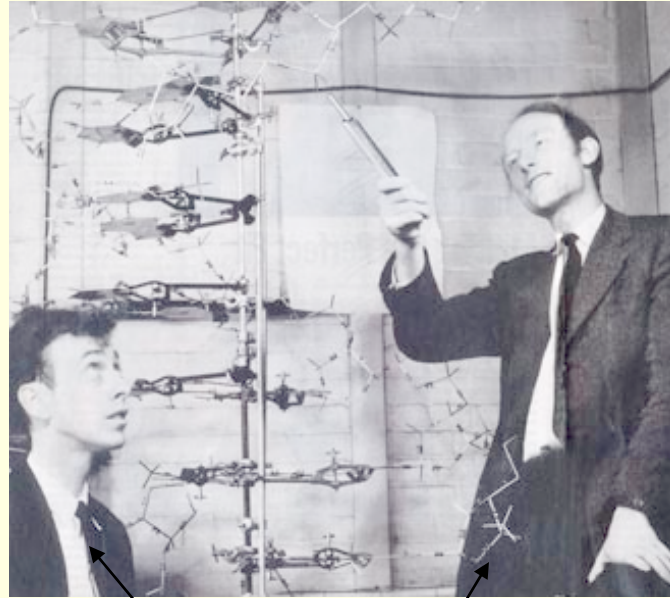
This hypothesis was rigorously proven in 1941 by George Beadle and Edward Tatum, using the simple bread mold *Neurospora*. First, they found that molds exposed to radiation lose the ability to produce essential nutrients, and this slowed, even stopped the growth of the mold. Then, they found that growth can be restored by providing the mutated mold with a specific supplement. They reasoned that each mutation must inactivate the enzyme (protein) needed to synthesize the nutrient. Thus, one gene carries the directions for making one protein.



# DNA: Birth of Molecular Biology (1953)



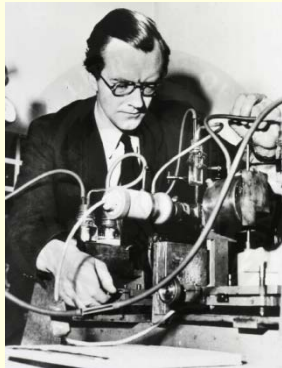
Cambridge, 1953. Shortly before discovering the structure of DNA, Watson and Crick, depressed by their lack of progress, visit the local pub.



J.D. Watson

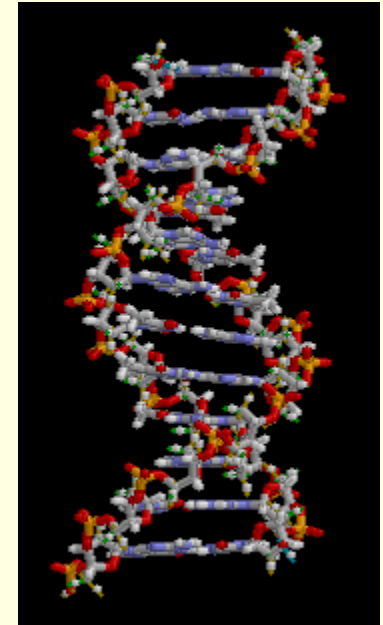
F. Crick

@Cavendish Lab, Cambridge

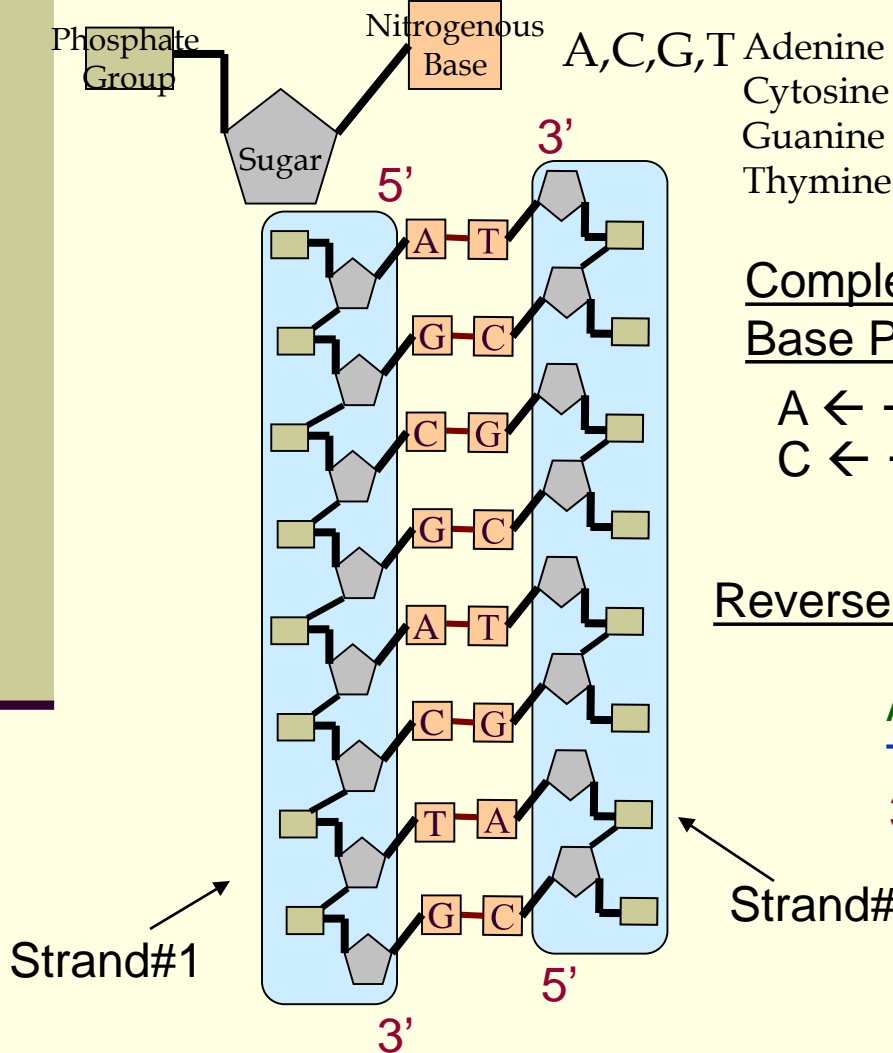


M.H.F. Wilkins R. Franklin

@King's College, London



# DNA: A Double Helix



A,C,G,T Adenine  
Cytosine  
Guanine  
Thymine

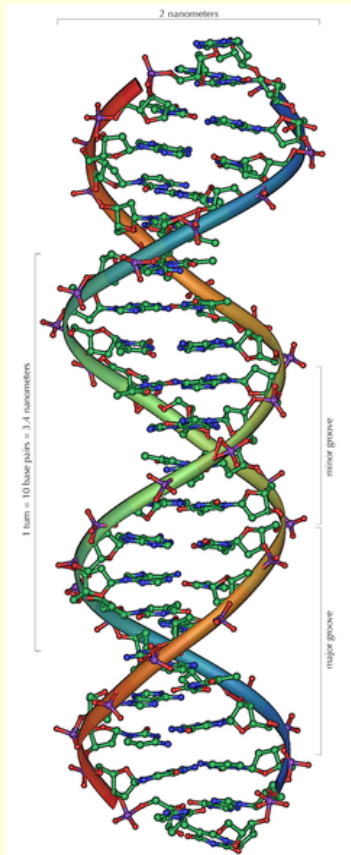
Complementary  
Base Pairing Rule:

A  $\leftarrow \rightarrow$  T  
C  $\leftarrow \rightarrow$  G

Reverse Complement:

5'  $\longrightarrow$  3'  
AGCGACTG  
TCGCTGAC  
3'  $\longleftarrow$  5'

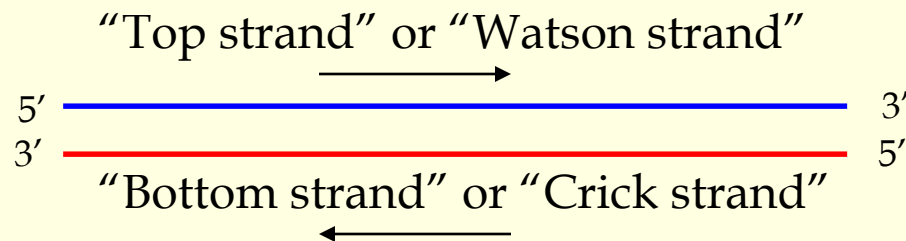
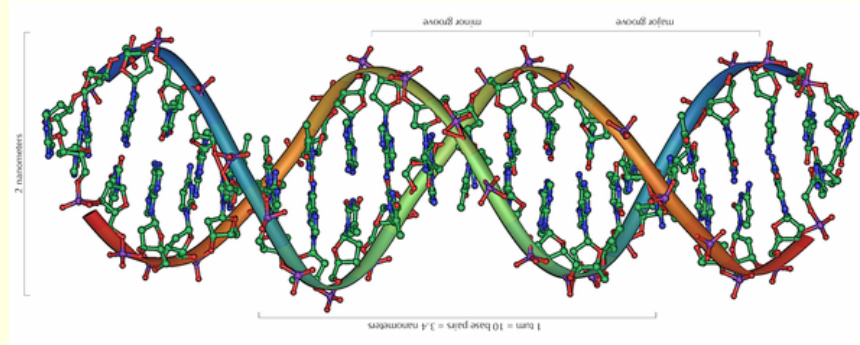
Double Helix in 3D



Each position is a "base pair"

# A little convention for convenience

- Let us use a straight line from now on to represent a DNA strand (or equivalently, its sequence)



# Genome

- The collection of all DNA in a cell
- Every organism has its own genome

## The human genome:

- Humans have 23 pairs of chromosomes
- Each chromosome is one long DNA molecule (hence, also a DNA sequence)
- The "human genome"
  - = 23 x 2 DNA sequences
  - = approximately 3 billion base pairs (haploid copy)

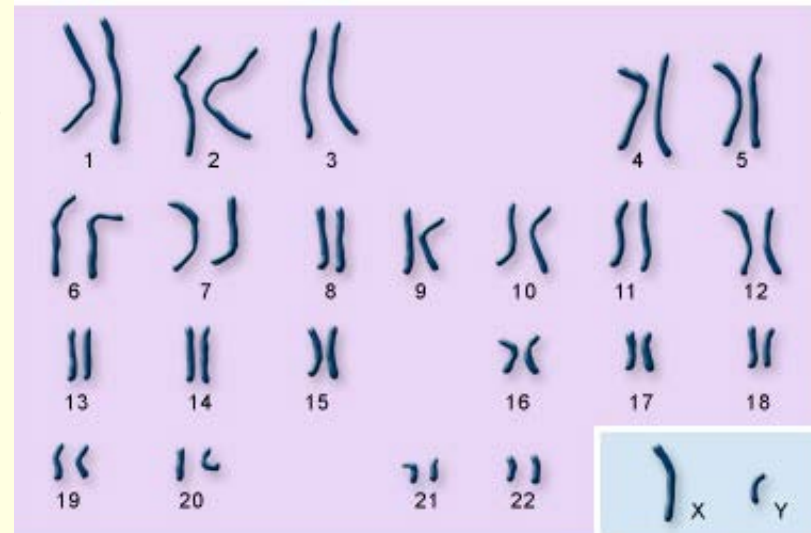


Fig source: <https://www.edinformatics.com/>

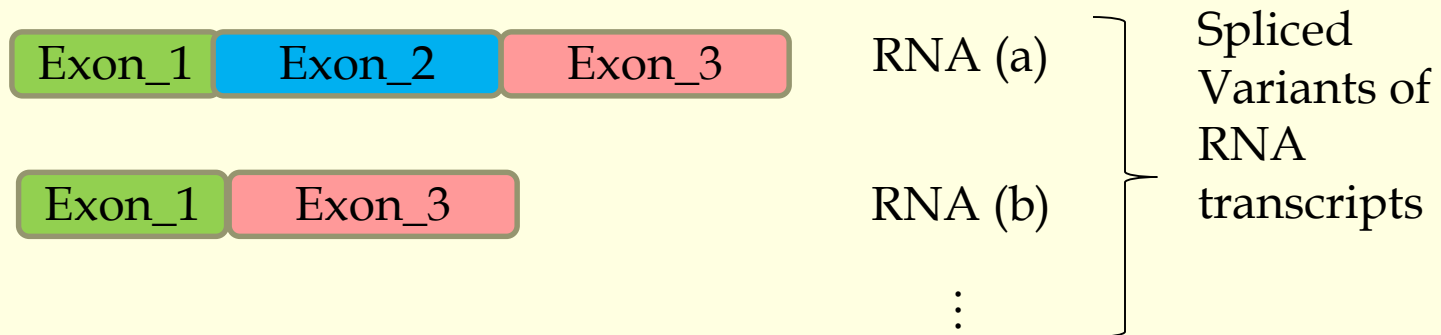
Genomes are of varied size and complexity

Then, what are "genes"?



# Genes are coding parts of a genome

- Genes are internally made of exons (coding segments) and introns (non-coding segments)



- Each gene can code for one or more RNA product (via alternative splicing)

# Central Dogma:

DNA → RNA → Protein

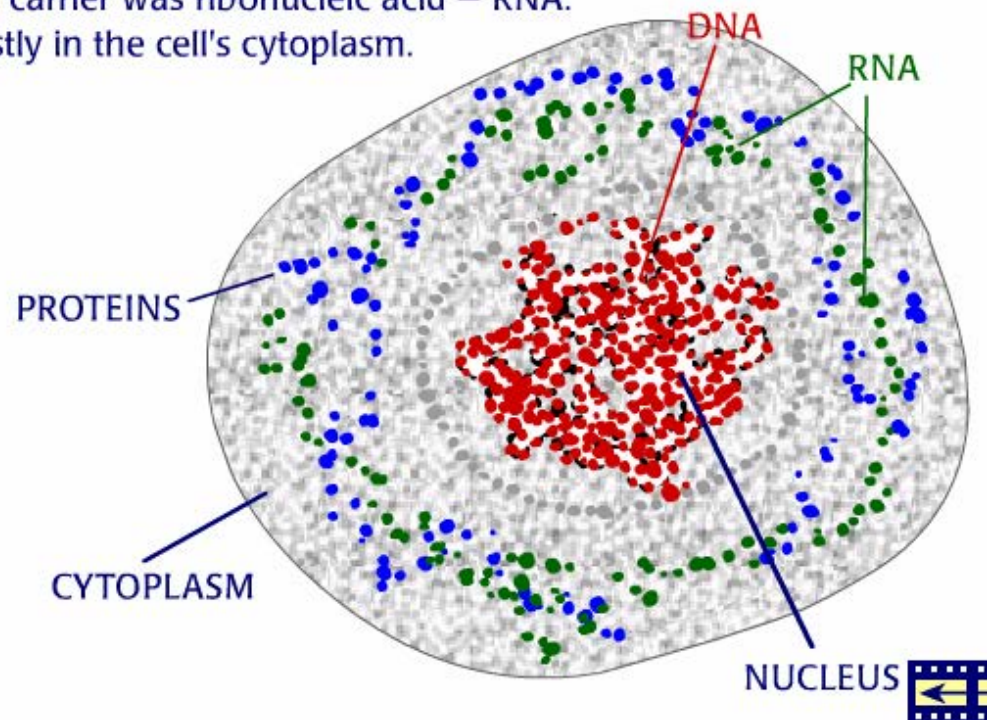
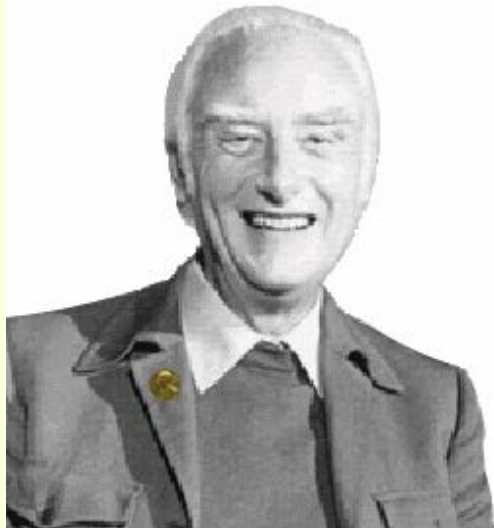
RNA is an intermediary  
between DNA and protein.

21

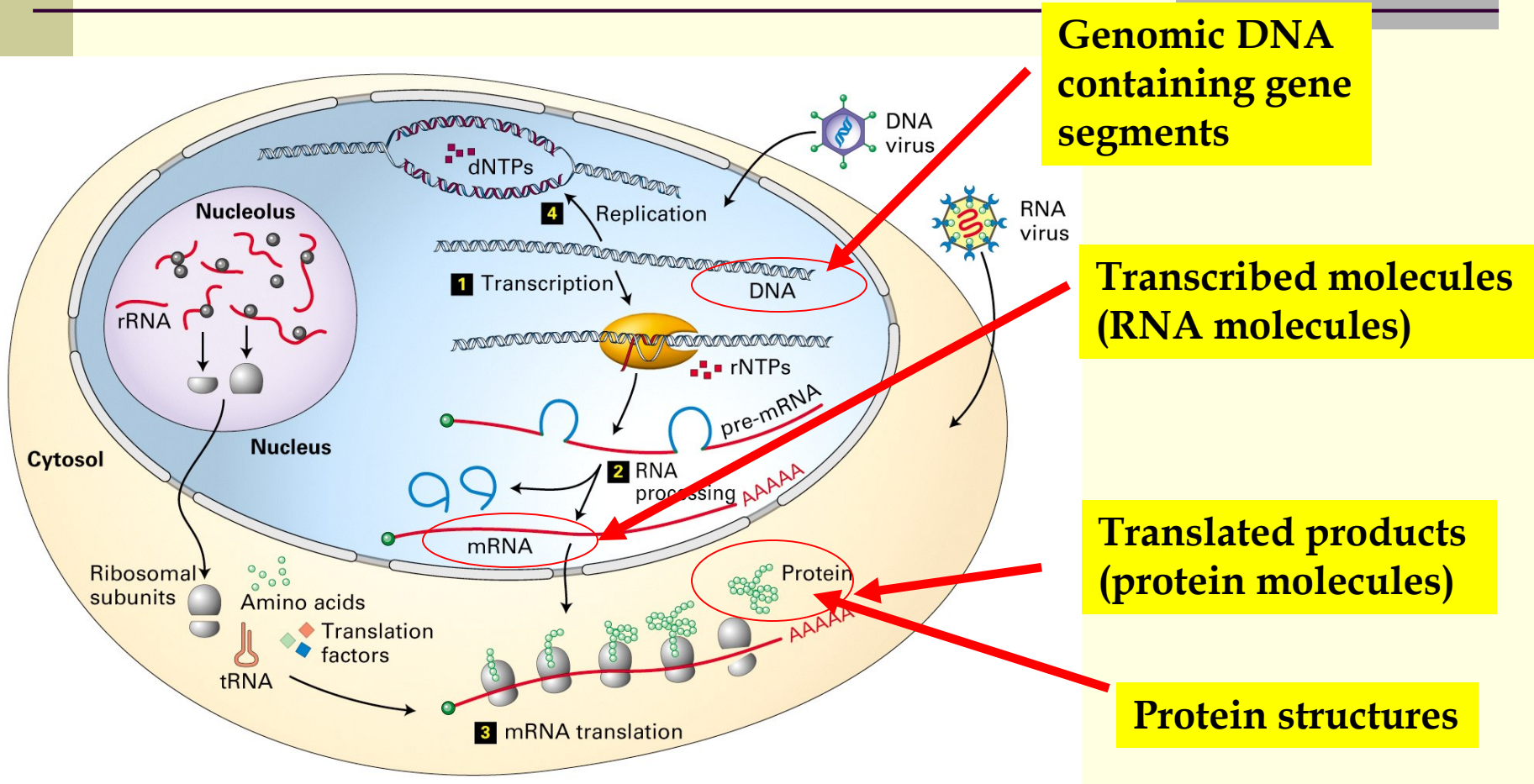
DNA FROM THE  
BEGINNING



A candidate for this information carrier was ribonucleic acid – RNA.  
RNA is a nucleic acid found mostly in the cell's cytoplasm.



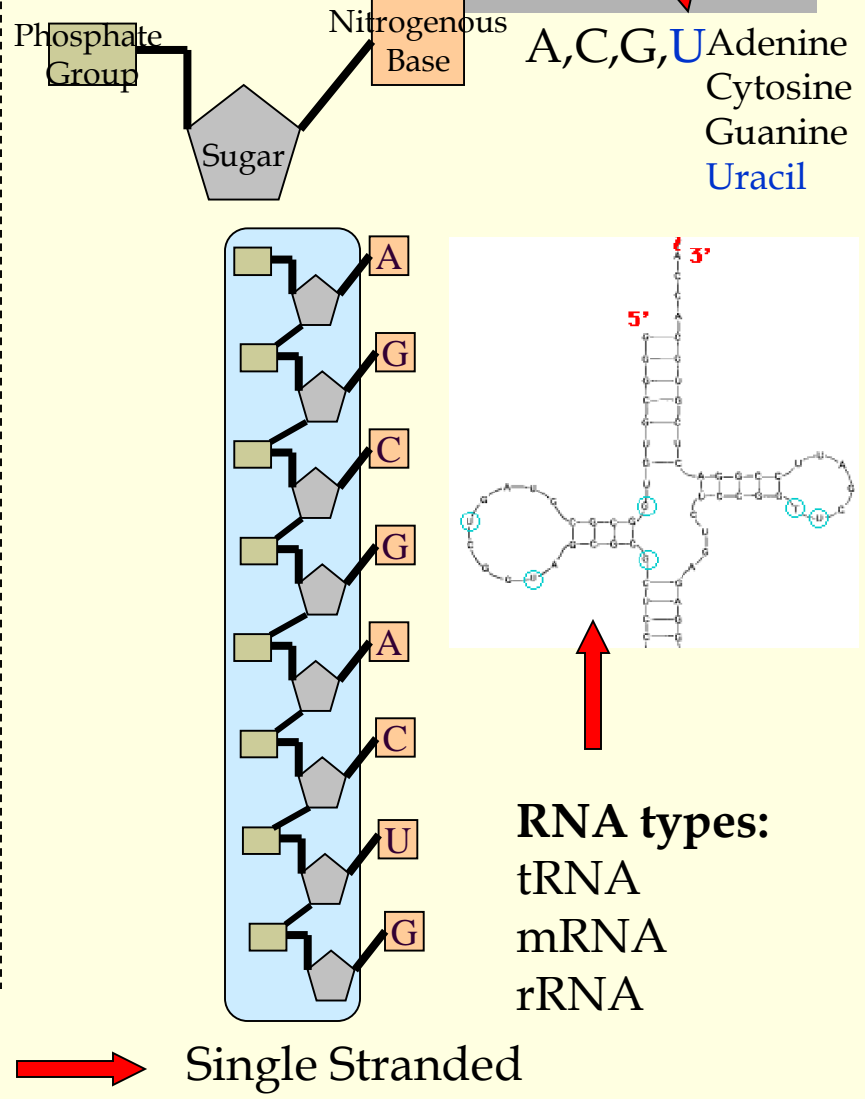
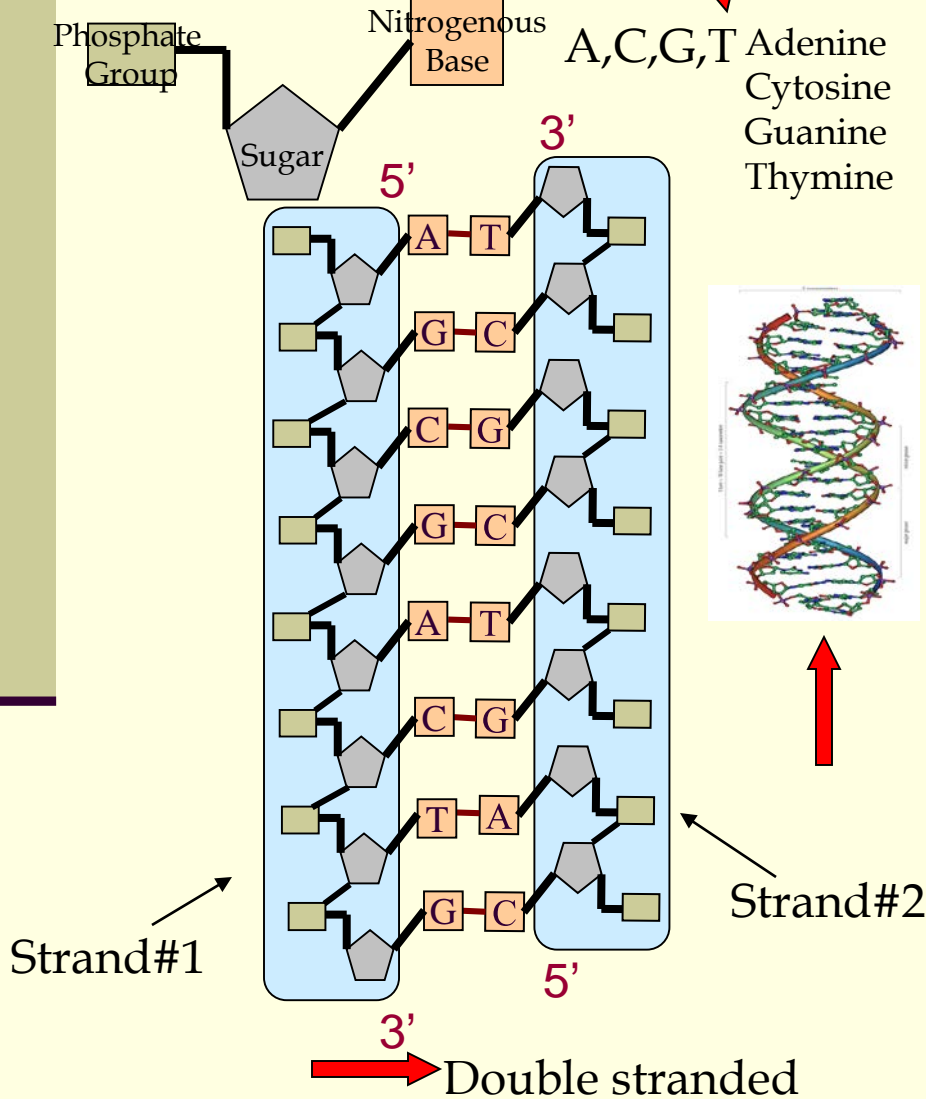
# The Central Dogma & Biological Data



Spot the difference!

# DNA

# RNA





# Genetic Code: Khorana, Holley and Nirenberg, 1968



		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

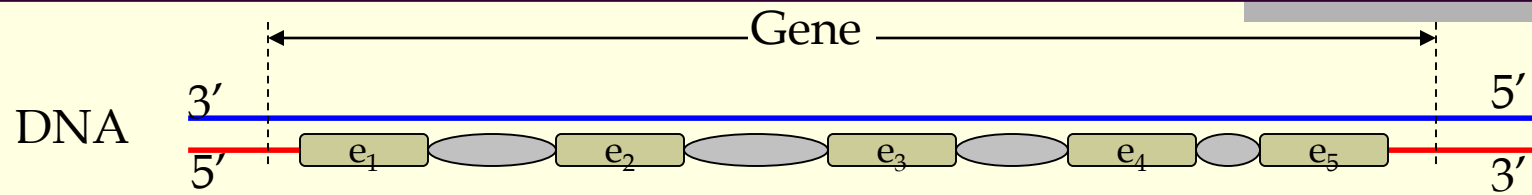
Third letter

## Combinatorial Logic:

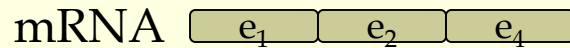
$$4^2 < 20 < 4^3$$

→ Hence 3 nucleotides in a codon

# Information Flow During Protein Synthesis



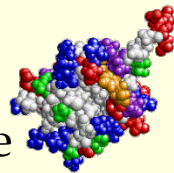
↓ *Transcription*



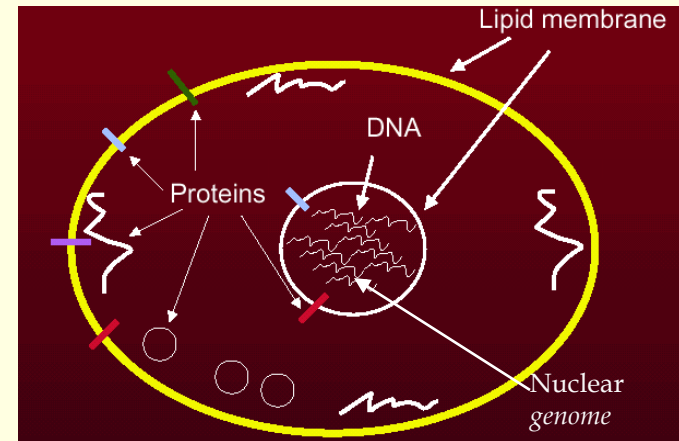
+ tRNA ↓ *Translation*





↓ *Folding*



One gene can code for many proteins! (*alternative splicing in eukaryotes*)



 Coding (*exons*)  
 Non-Coding (*introns*)

# Genetic imperfections

---

- Mutations are changes (edits) on the genome
- Point mutations (single character edits) are referred to as “single nucleotide polymorphisms” (SNPs)
- Point mutations can possibly change the protein product



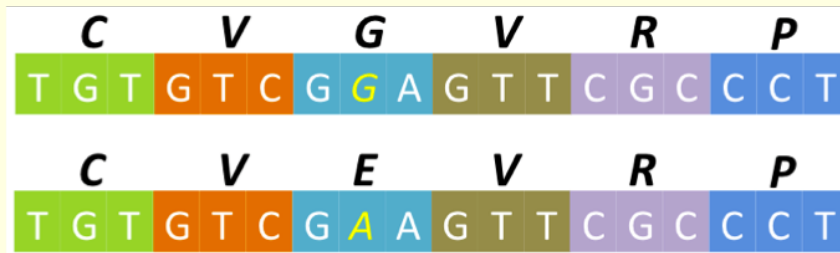
# Genetic imperfections

---

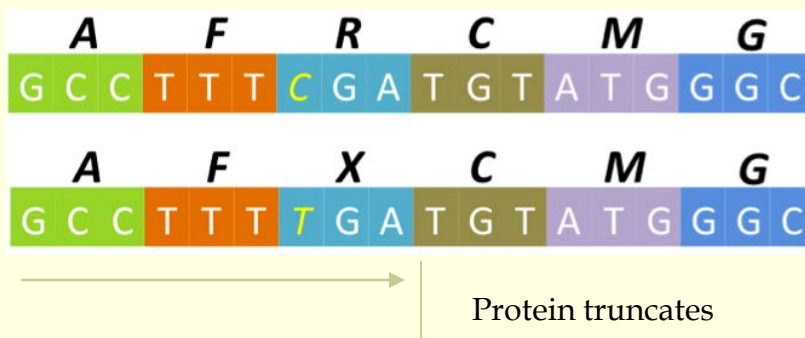
- If a point mutation is in the coding part of a gene, it is one of the three kinds:
  - Synonymous: doesn't change the amino acid product  
  
e.g., a codon changes from  
... CCA... → ... CCC...  
both yield Proline (as the amino acid product)

# Genetic imperfections

- Missense: changes the amino acid product using a substituted amino acid

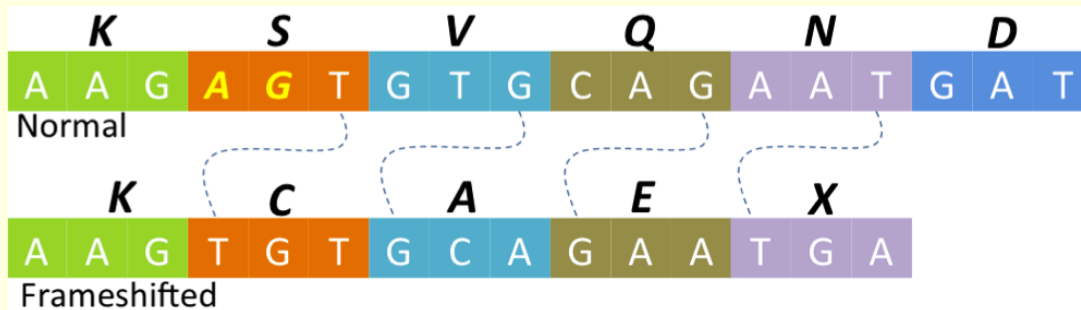


- Nonsense: truncates the protein product because of a premature stop codon



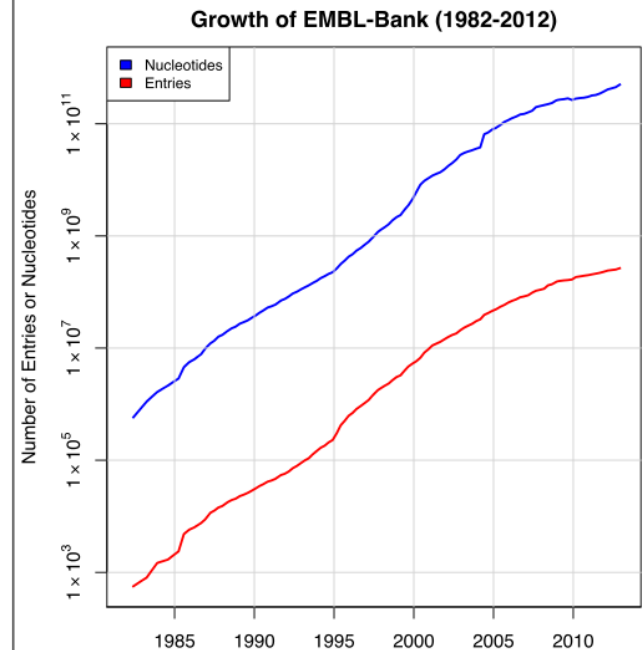
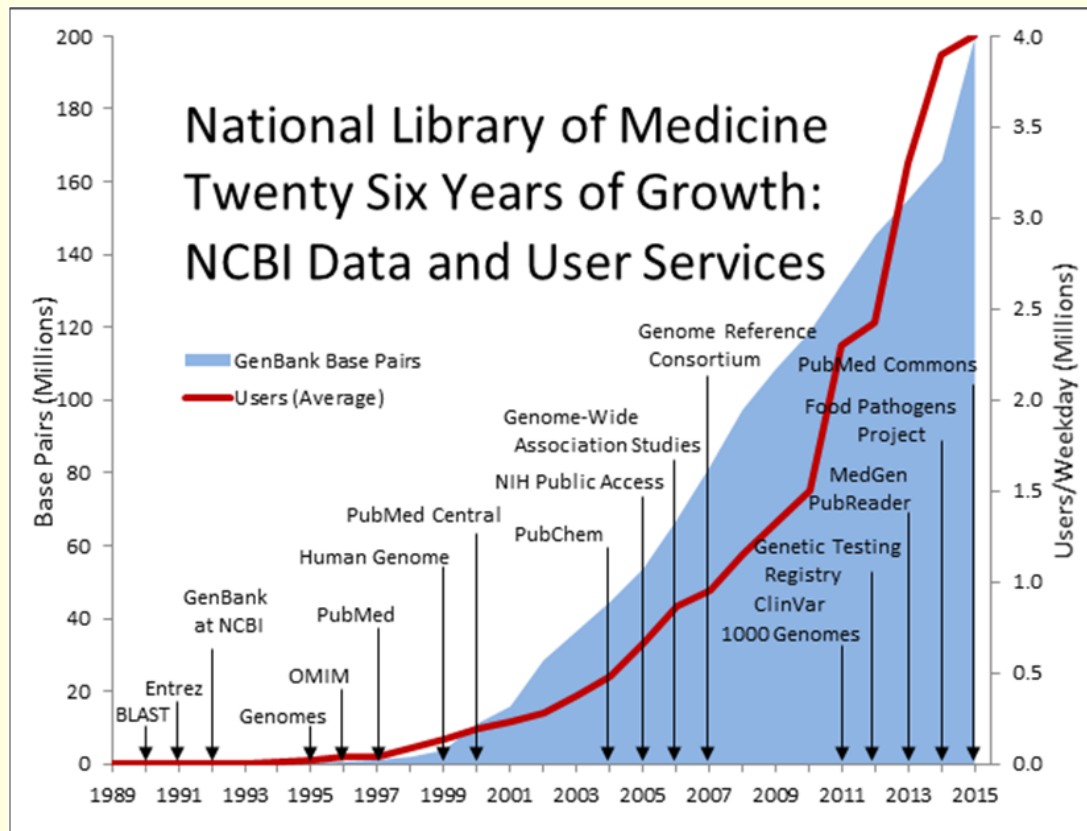
# Genetic imperfections

- Frameshift errors: happens when the *deletion* (or *insertion*) of a nucleotide could result in shifting of the open reading frame (used in transcription)



# Genomic databases

*“An annotated collection of all publicly available nucleotide and amino acid sequences.”*



Source: NCBI GenBank, EMBL websites

<https://www.nlm.nih.gov/about/2017CJ.html>

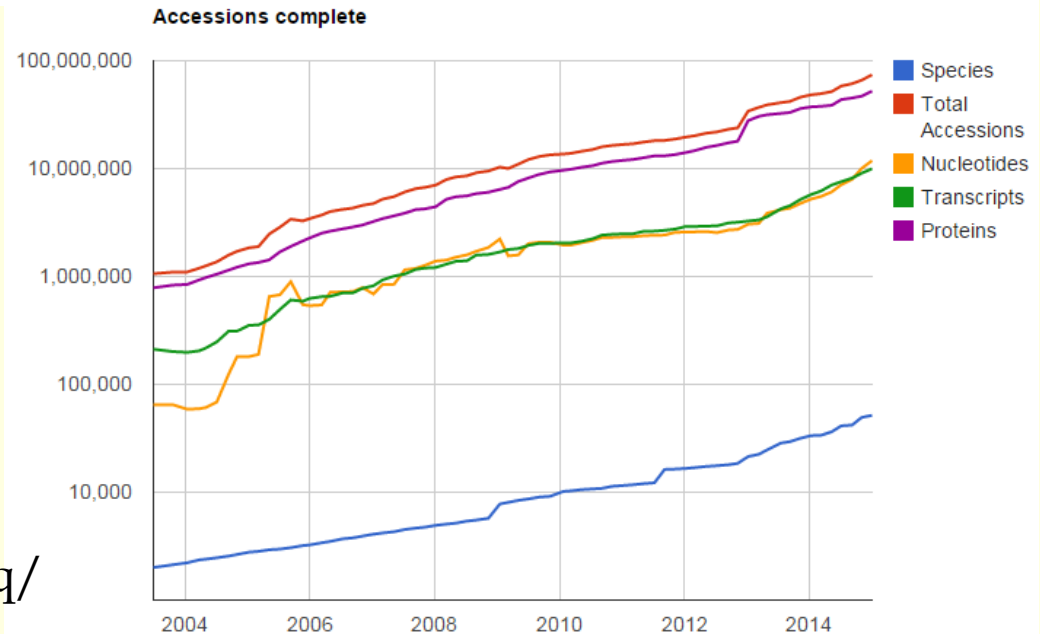
# NCBI RefSeq database

- “A comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein.”

Counts of accessions and basepairs/residues per molecule type:

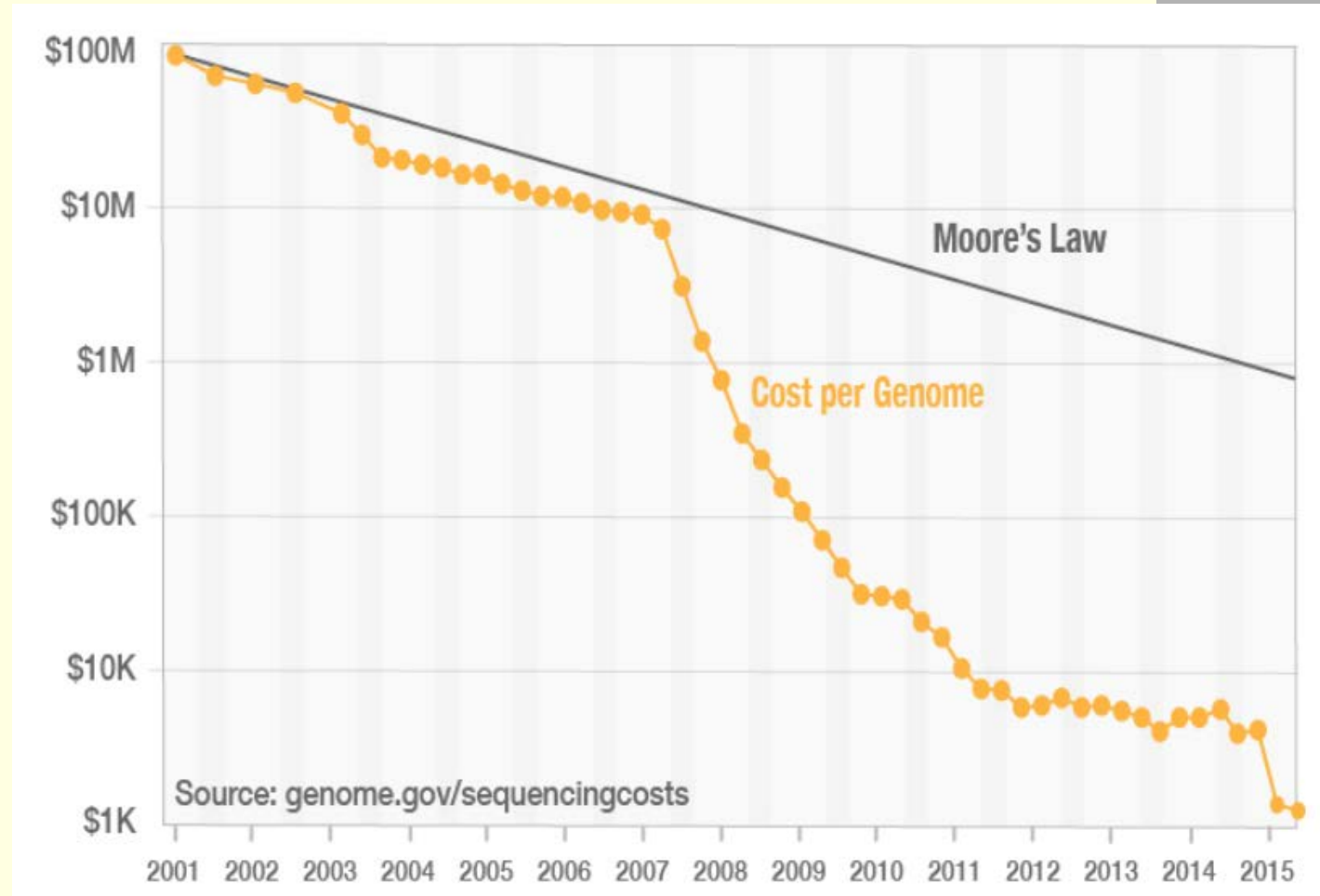
	Accessions	Basepairs/Residues
Genomic:	11852380	574044044515
RNA:	9973568	20408631127
Protein:	52276468	18690872100
Wgs master:	24603	0

~600GB of data

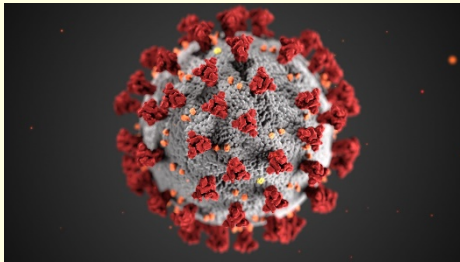


<http://www.ncbi.nlm.nih.gov/refseq/>

# Cost to sequence a genome



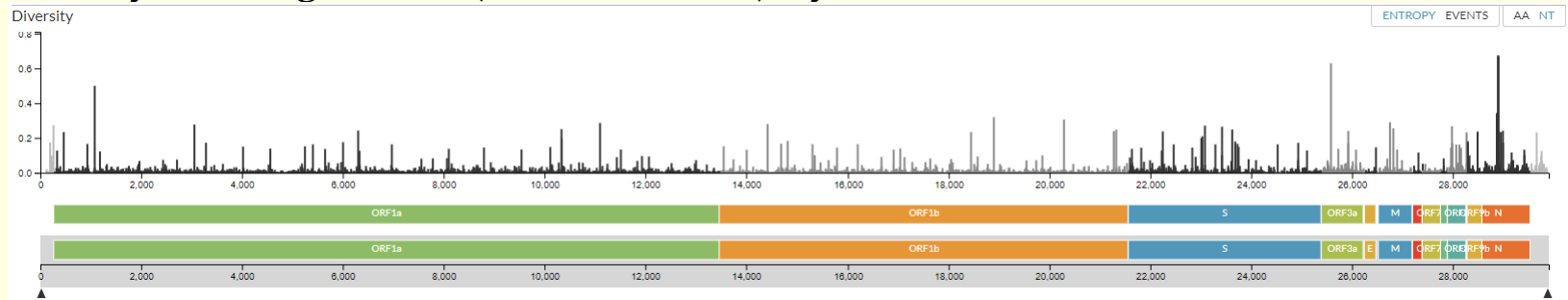
# COVID-19 genome strains



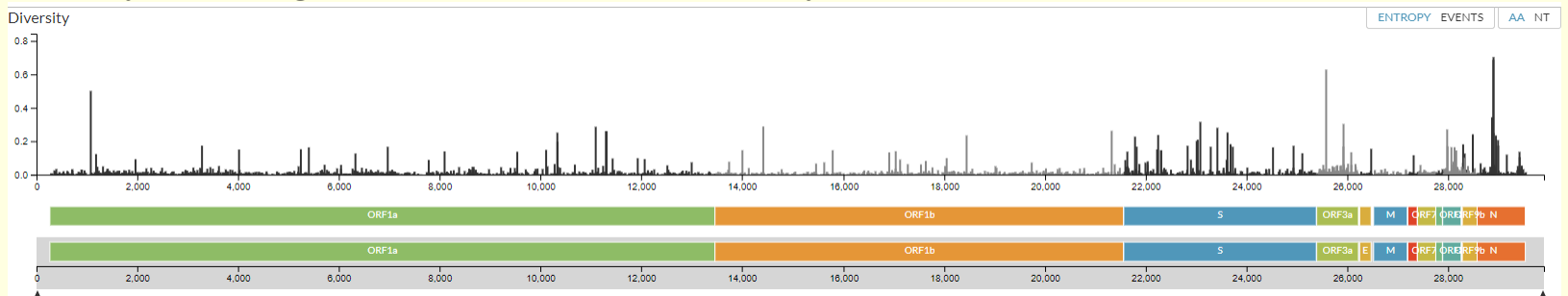
## COVID-19 Genome:

- RNA virus
- Approximately 30Kbp genome size
- No. strains recorded till date: 4046

## Diversity of the genome (across strains) by nucleotides:



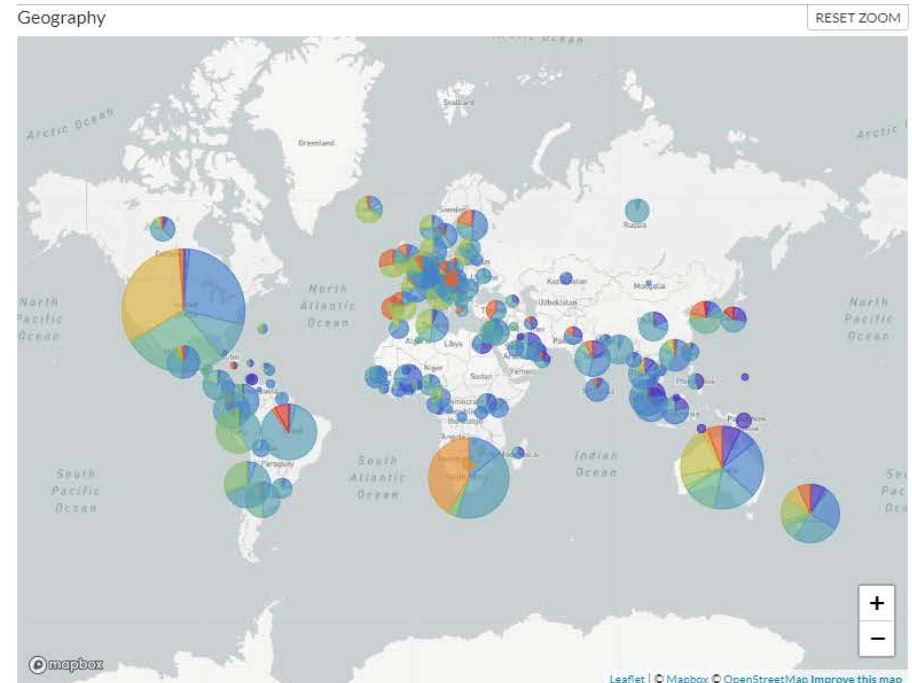
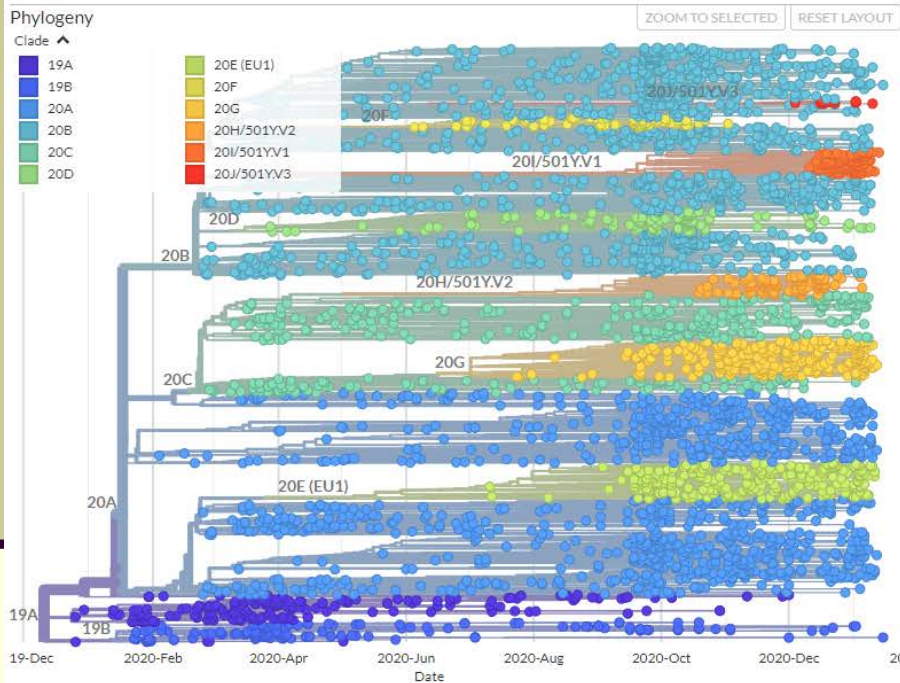
## Diversity of the genome (across strains) by Amino Acid product:



# COVID-19 genome strain evolution

## Genomic epidemiology of novel coronavirus - Global subsampling

Maintained by the Nextstrain team. Enabled by data from **GISAI**  
Showing 4046 of 4046 genomes sampled between Dec 2019 and Jan 2021.





# Several Questions Leading Up to Today's Computational Biology and Bioinformatics

---

- What are the nucleotides in a DNA molecule? (*problem of sequencing*)
- What DNAs make up the genome of a species? (*problem of genome sequencing, genome assembly*)
- What are the genes within a genome? (*gene identification/discovery*)
- What protein and RNA products does a gene produce? (*annotation*)
- What is the native 3D structure of a protein and how does it get there? (*protein folding, structure prediction*) Similar questions can be asked of RNAs too.

# Several Questions ....

---

- Are there non-protein coding genes? (*pseudo-genes*)
- Under what conditions does a gene express itself, and are there genes that are more active than others under experimental conditions? (*gene expression analysis, microarrays*)
- Are there a subset of genes that co-operate, and does a gene's activity get affected by others? (*gene regulatory networks*)
- How do genes look and behave in closely related species? What distinguishes them? (*gene and species evolution*)
- What is the ``TREE OF LIFE''? (*phylogenetic tree reconstruction*)
- How does a protein know where to go next within a cellular complex? (*localization, signal peptide prediction*)
- AND MANY MORE ....

# Computational Biology & Bioinformatics: Problem Areas

## Sequence Discovery

Genome  
Gene  
Regulatory elements  
RNA products  
Proteins

## Function

Gene to protein annotation  
Gene expression analysis  
Microarray experiments  
RNA interference  
Metabolic networks/pathway

## Structure

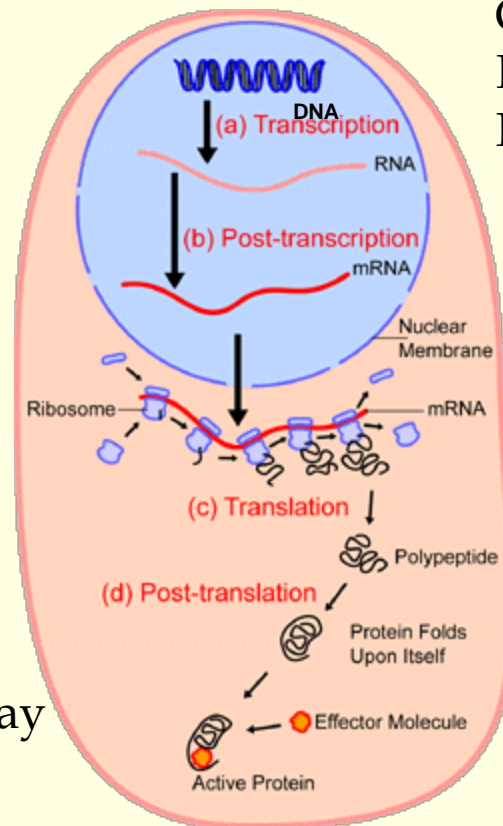
Gene structure prediction  
RNA structure prediction  
Protein structure prediction

## Evolutionary Studies

Tree of life  
Speciation

## Population Genetics

Haplotype analysis  
Nucleotide polymorphism



# Computational Biology and Bioinformatics

---

- A rapidly evolving field
  - Technology – biological and computational
  - Capabilities
  - Concepts
  - Knowledge and Science
- A plethora of grand challenge questions
- An Ante-disciplinary Science?
  - An interesting read:
    - “Antedisciplinary” Science, Sean R. Eddy, PLoS Computational Biology, 1(1):e6

# Referred Slide Materials, Acknowledgments, and Web Resources

---

- “DNA From the Beginning” (<http://www.dnaftb.org>), Dolan DNA Learning Center, Cold Spring Harbor Laboratory
- Stanford University, CS 262: Computational Genomics
- NCBI website
- Wikipedia
- J.D. Watson, *The Double Helix: A personal account of the discovery of the structure of DNA*