Discrete Population Models Inspired by Applications in Cancer Research

Modèles discrètes de populations inspirés par des applications dans la recherche sur le cancer

Marek Kimmel

Rice University, Houston, TX, USA

Outline

- Mutations, genetic drift, and selection in proliferating cancer cells lead to populations with highly diverse genomes and complex dynamics.

- Mathematical tools to describe the resulting stochastic phenomena are classical population genetics models such as Wright-Fisher and Moran models, as well as more general branching processes (bp) and Markov processes.

- The diversity of cells and genomes frequently leads to models with infinite collections of types.

Plan d’action

- Les mutations, la dérive génétique et la sélection dans les cellules cancéreuses proliférantes conduisent à des populations aux génomes très diversifiés et à une dynamique complexe.
- Les outils mathématiques permettant de décrire les phénomènes stochastiques qui en résultent sont les modèles classiques de génétique des populations tels que les modèles de Wright-Fisher et de Moran, ainsi que les processus de ramification (bp) et les processus de Markov plus généraux.
- La diversité des cellules et des génomes conduit fréquemment à des modèles avec des collections infinies de types.

Outline

- This mini-course includes a review of several of these models, preceded by a brief introduction to cancer dynamics, including some statistical issues such as early detection paradoxes, followed by a brief overview of mathematical tools.
- The lecture will be mathematically elementary, with an emphasis on intuitions and model building, but examples of mathematical proofs and references to the body of published literature will be provided.
- The overall goal is to generate interest in the field, which is currently expanding and has many interesting connections to basic processes in living cells.
- Although these are biologically and mathematically diverse models, they share some “exotic” properties, which arise from a non-trivial interplay between cell growth and cell-like transitions.

Plan d’action

- Ce mini-cours comprend une revue de plusieurs de ces modèles, précédée d'une brève introduction à la dynamique du cancer, y compris certaines questions statistiques telles que les paradoxes de détection précoce, suivie d'un bref aperçu des outils mathématiques.
- L'exposé sera mathématiquement élémentaire, en mettant l'accent sur les intuitions et la construction de modèles, mais des exemples de preuves mathématiques et des références au corpus de la littérature publiée seront fournis.
- L'objectif général est de susciter l'intérêt pour le domaine qui est actuellement en pleine expansion et qui présente de nombreux liens intéressants avec les processus de base des cellules vivantes.
- Bien qu'il s'agisse de modèles biologiquement et mathématiquement divers, ils partagent certaines propriétés « exotiques », qui résultent d'une interaction non triviale entre la croissance cellulaire et les transitions de type

List of Meetings

- **Meeting 1:** Biological and mathematical background
- **Meeting 2:** Gene amplification. Role of quasi-stationarity
- **Meeting 3:** Telomere dynamics, from branching random walk to Greider’s model
- **Meeting 4:** Heavy-tail distributions in cell proliferation models
- **Meeting 5:** Tug-of-War model of competition between advantageous and deleterious mutations

Meeting 1: Biological and mathematical context

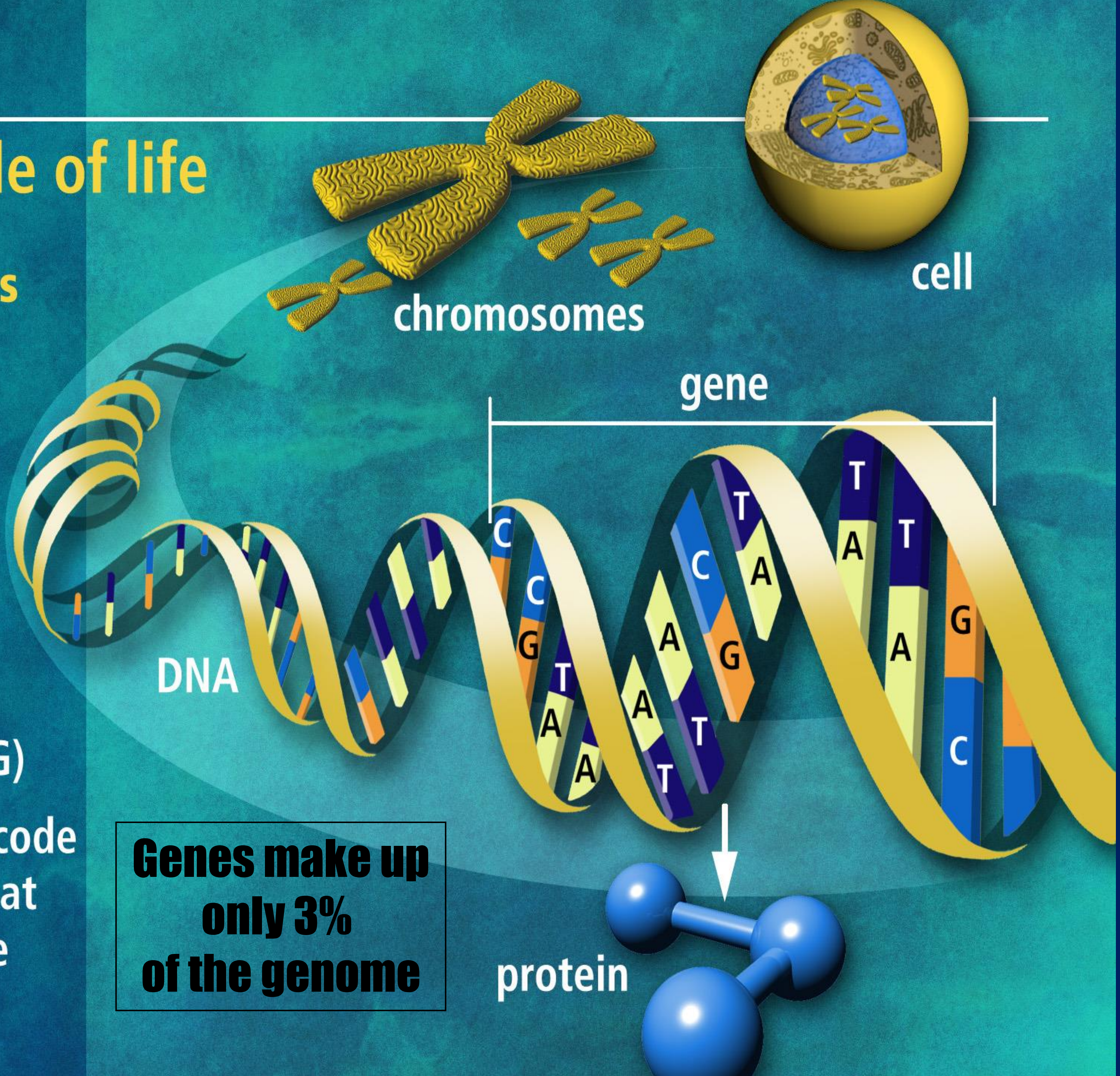
- DNA, RNA and proteins, the dogma of molecular biology, structure of human genomes
- quantitative theories of cancer
- population genetic models - genetic drift: continuous-time Moran model versus discrete-time Wright-Fisher model - mutations: infinite allele and infinite site models - neutral evolution and the coalescent
- branching processes - Galton-Watson (GW) processes: basic equations of the generating function and criticality, Yaglom's theorem - continuous-time processes and the Goldie-Coldman two-bp model of chemotherapy resistance

Réunion 1: Contexte biologique et mathématique

- ADN, ARN et protéines, le dogme de la biologie moléculaire, structure des génomes humains
- théories quantitatives du cancer
- modèles de génétique des populations - dérive génétique : modèle Moran continu dans le temps contre modèle Wright-Fisher discret en temps – mutations : modèles d'allèles infinis et de sites infinis - évolution neutre et le coalescent
- processus de ramification - processus de Galton-Watson (GW) : équations de base de la fonction génératrice et criticité, théorème de Yaglom - processus continus en temps et modèle Goldie-Coldman à deux types de bp de la résistance aux chimiothérapie

DNA

the molecule of life



Trillions of cells

Each cell:

- 46 human chromosomes
- 2 m of DNA
- 3 billion DNA subunits (the bases: A, T, C, G)
- 30,000 ~~80,000~~ genes code for proteins that perform all life functions

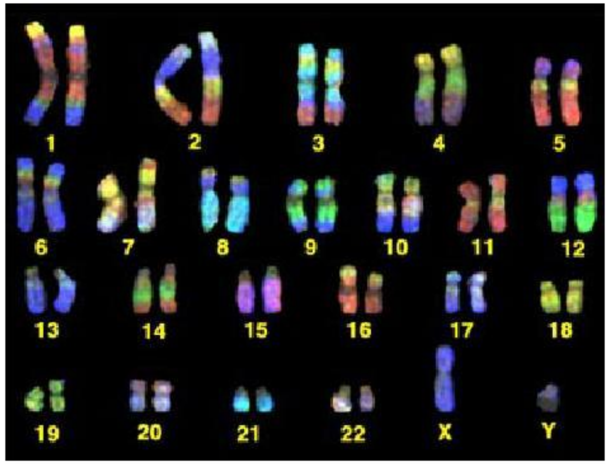
Genes make up only 3% of the genome

Genome Sizes

Human	3.0×10^9 base pairs
Mouse	3.0×10^9
Drosophila	1.1×10^8
Worm	1.0×10^8
Dictyostelium	3.4×10^7
Yeast	1.2×10^7
Bacteria	$1.0 - 5.0 \times 10^6$

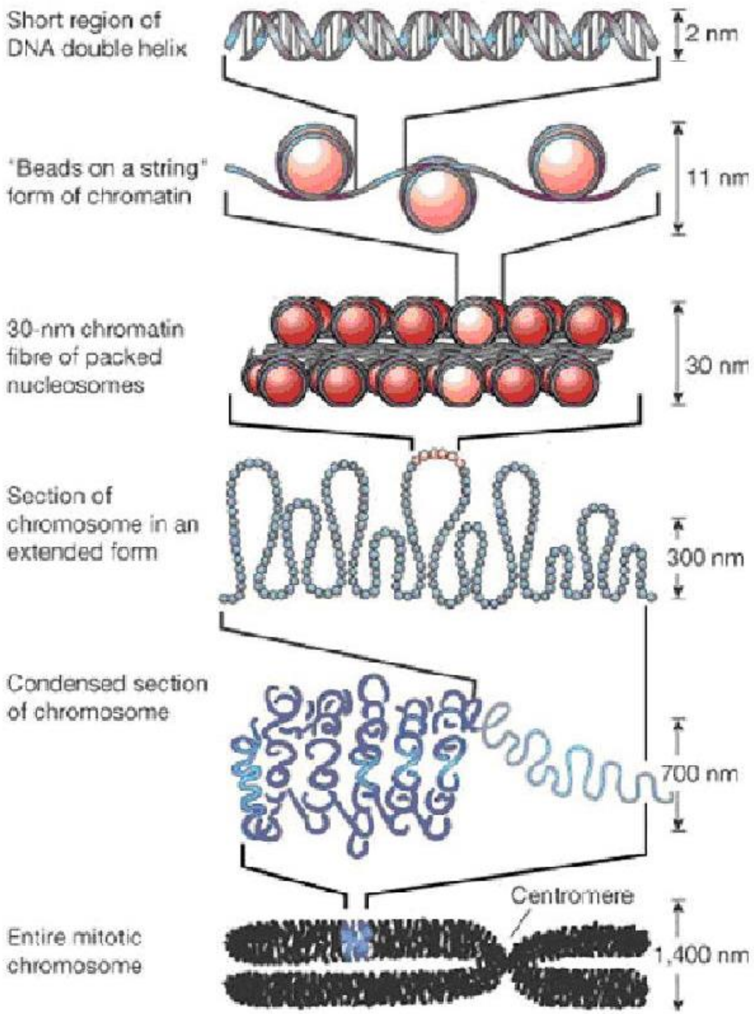
Deoxyribonucleic Acid (DNA) within Chromosomes

- Each somatic (non-gamete) cell in the human body contains two copies of 23 chromosomes within the cellular nucleus.
- One copy is inherited from the mother, one from the father.
- Humans are thus diploid organisms, with haploid gametes which fuse during reproduction.



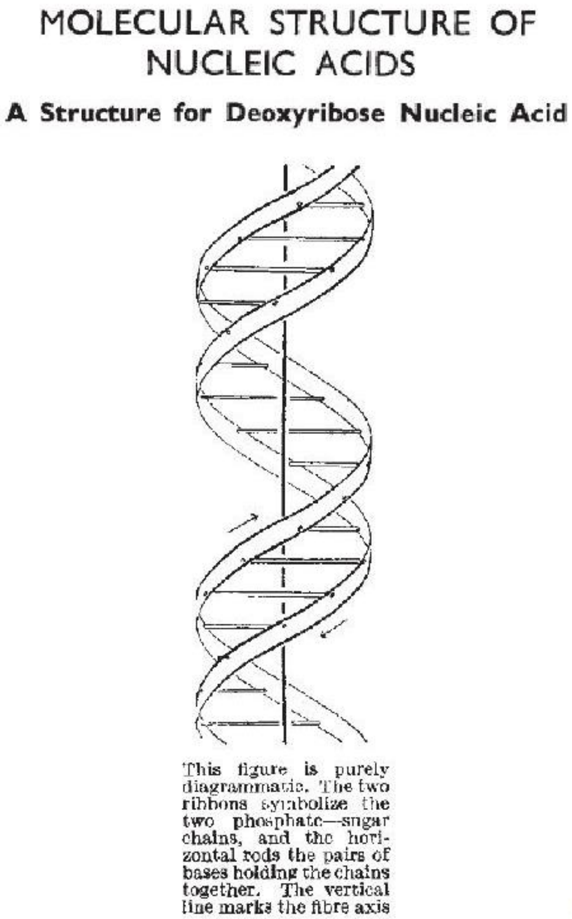
Chromosome Coiling

- Chromosomes are structures comprised of super-coiled DNA.


Short region of DNA double helix
2 nm
"Beads on a string" form of chromatin
11 nm
30-nm chromatin fibre of packed nucleosomes
30 nm
Section of chromosome in an extended form
300 nm
Condensed section of chromosome
700 nm
Centromere
Entire mitotic chromosome
1,400 nm

DNA Helix Structure

- Watson and Crick won the Nobel prize in 1962 for their discovery, initially in 1953, of the double-helical structure of DNA.



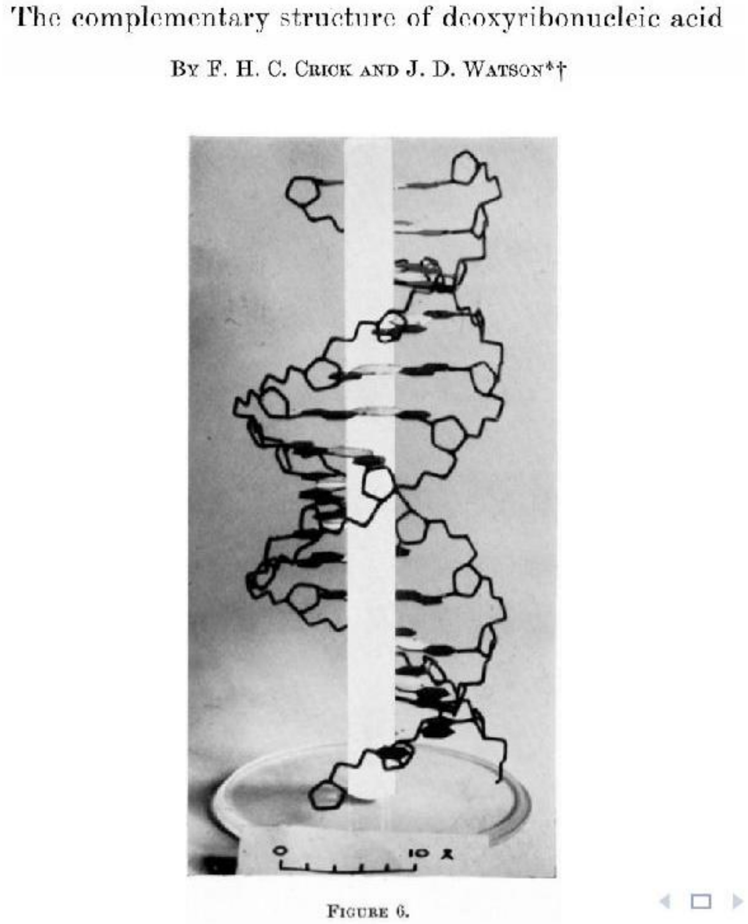
MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

This figure is purely diagrammatic. The two ribbons symbolize the two phosphate—sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

DNA Helix Structure

- Their 1953 *Nature* paper was more of an announcement with a detailed follow-up manuscript published by *The Royal Society* in 1954.



The complementary structure of deoxyribonucleic acid

By F. H. C. Crick and J. D. Watson*†

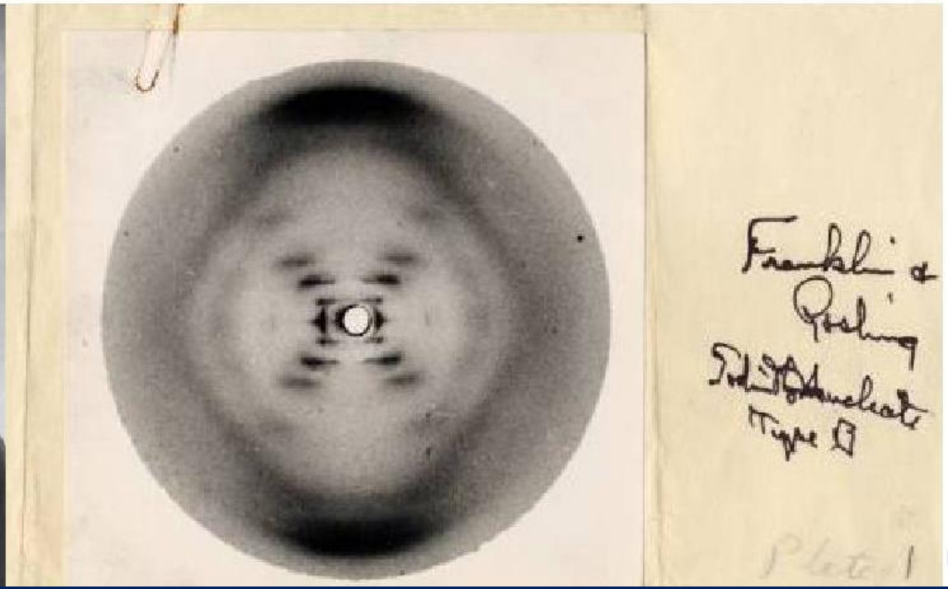
Figure 6.

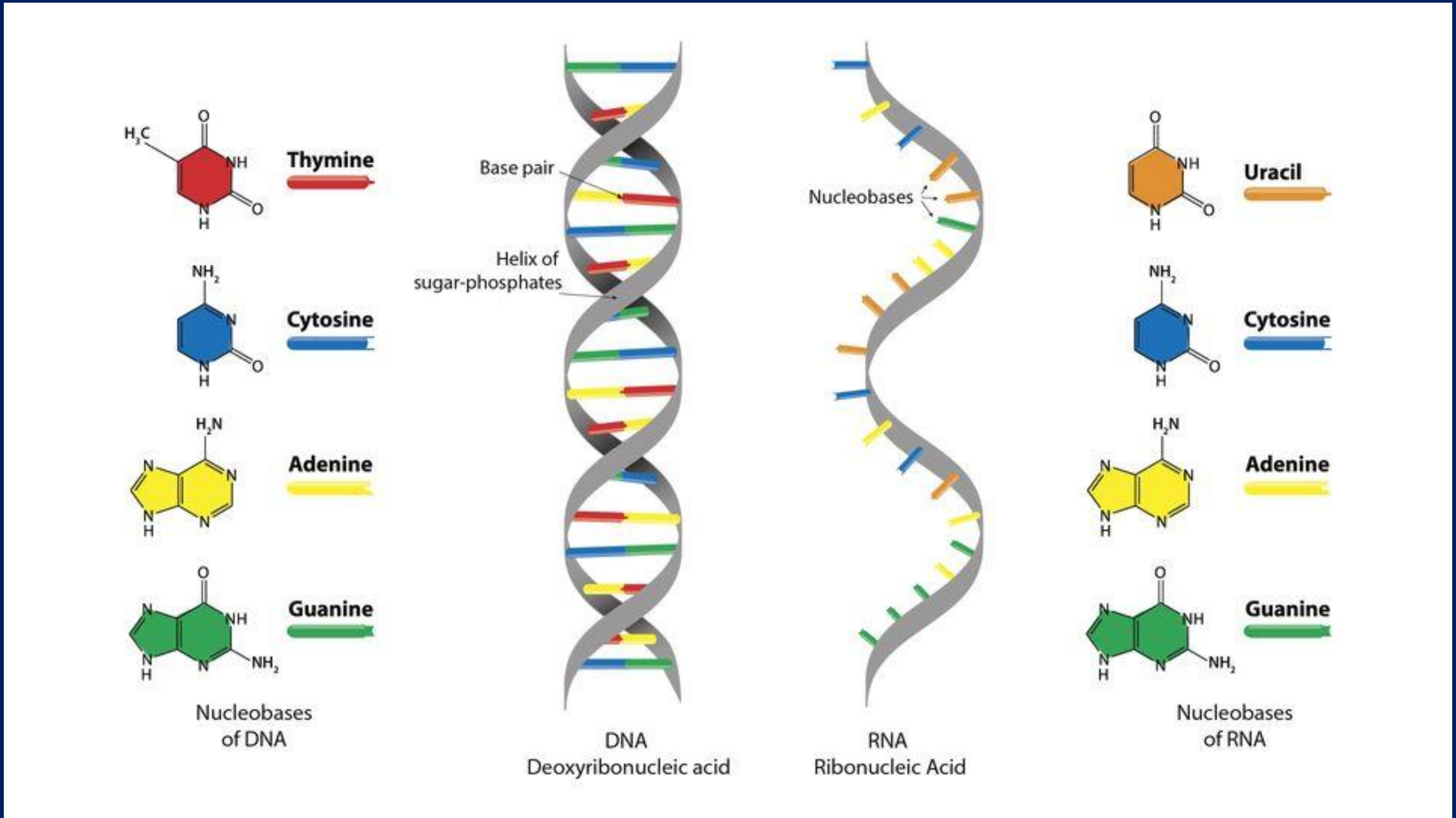
The Railroading of Rosalind Franklin

- The quest to find discover the structure of DNA was indeed a race.
- Linus Pauling (CalTech), Maurice Wilkins and Rosalind Franklin (King's College London), and James Watson and Francis Crick (Cavendish Laboratory at Cambridge University) all sought to be the first to publish the elucidated structure.
- Pauling beat Watson and Crick in finding that certain proteins are helical in shape.
- He went on to erroneously propose that DNA was a three-stranded structure with the bases facing outwards.

The Railroading of Rosalind Franklin

- James Watson was shown the famous X-ray crystallography "Photograph 51" without Franklin's permission or knowledge by Maurice Wilkins.
- The photograph revealed the antiparallel nature of the double helical strands and showed the bases to be in the center of the helix.
- Calculations made using the photograph helped Watson and Crick calculate size and structural parameters of the molecule.

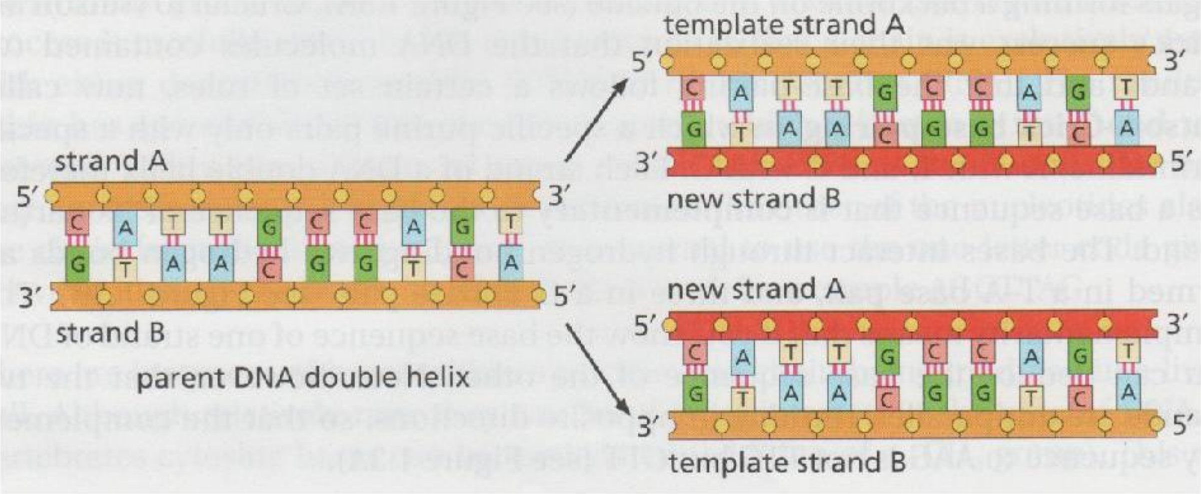




https://www.technologynetworks.com/genomics/articles/what-are-the-key-differences-between-dna-and-rna-296719

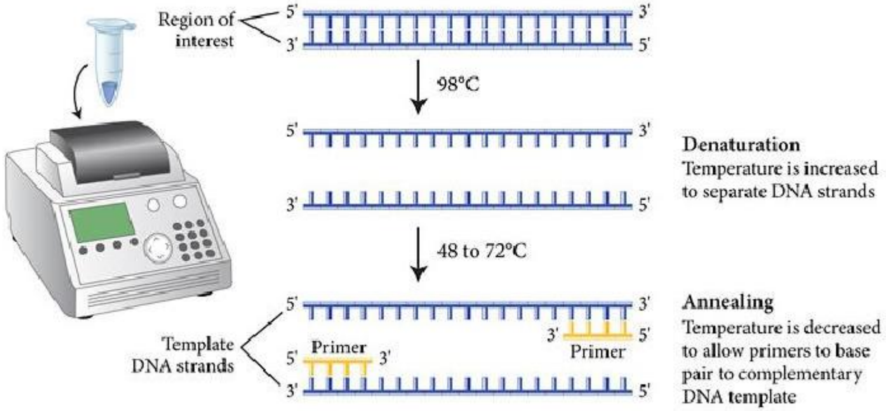
DNA Replication

- The hydrogen bonds between the purines and pyrimidines pairs are somewhat weak noncovalent bonds.
- Thus, their separation, for instance, during DNA replication prior to cellular division, is relatively easily accomplished.
- The DNA polymerase enzyme operates along both templates matching complementary nucleotides to their strands.



Polymerase Chain Reaction

- Absolutely indispensable method for DNA cloning/sequencing, gene analysis, etc.
- Creates an exponentially increasing number of DNA copies, through triggering and controlling DNA replication.
- Majority of methods are based on thermal cycling.

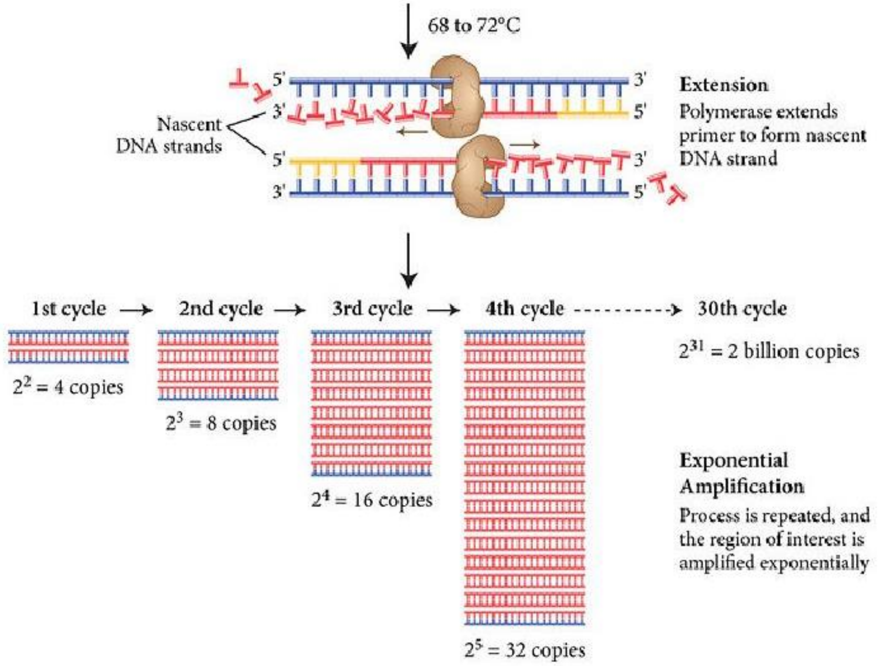


Denaturation
Temperature is increased to separate DNA strands

Annealing
Temperature is decreased to allow primers to base pair to complementary DNA template

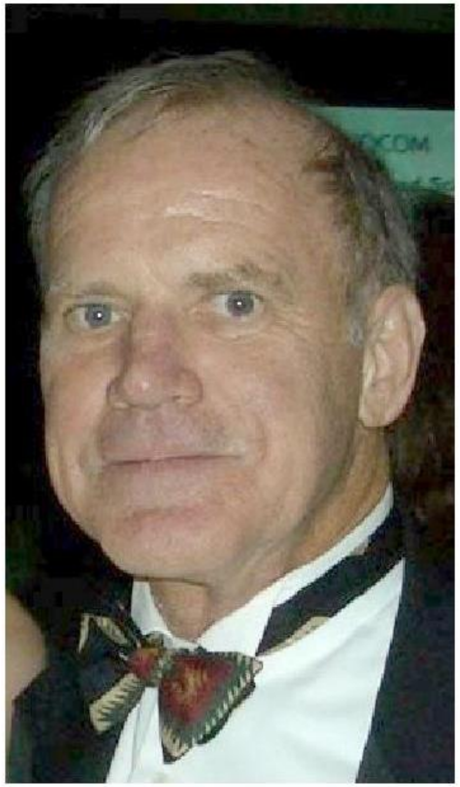
Polymerase Chain Reaction

- Introduced oglionucleotides or DNA primers are introduced and expanded on by heat-stable polymerases (Taq).
- The cycle is repeated numerous times for exponential yields of cloned DNA.



Almost More Fascinating than PCR: Kary Mullis

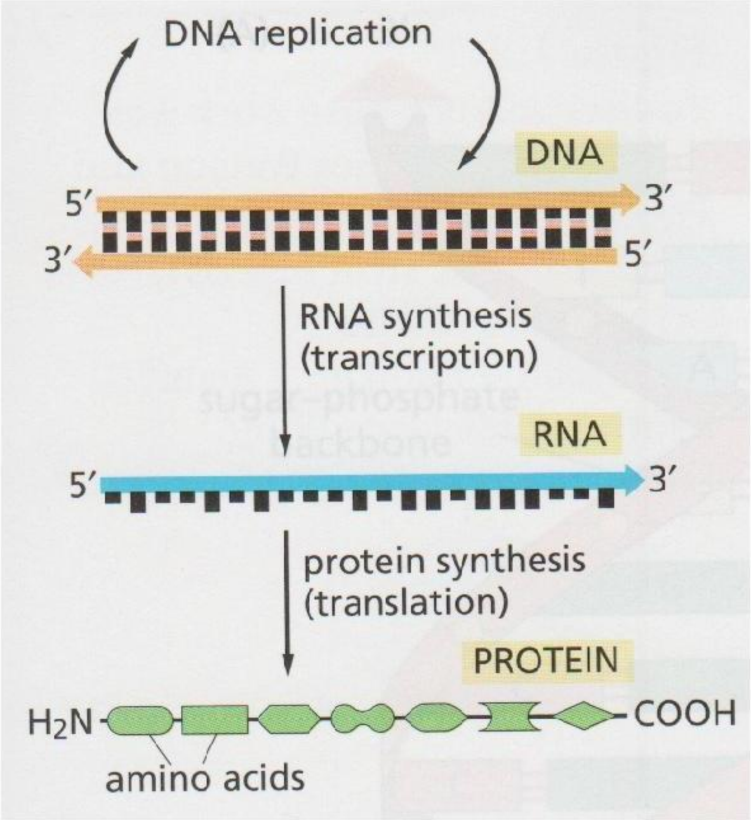
- PCR has vastly influenced the fields of biochemistry, molecular biology, genetics and genomics.
- Mullis won the Nobel Prize in 1993, ten years after first demonstrating PCR.
- Incredibly gifted and talented scientist...



The "Central Dogma of Molecular Biology"

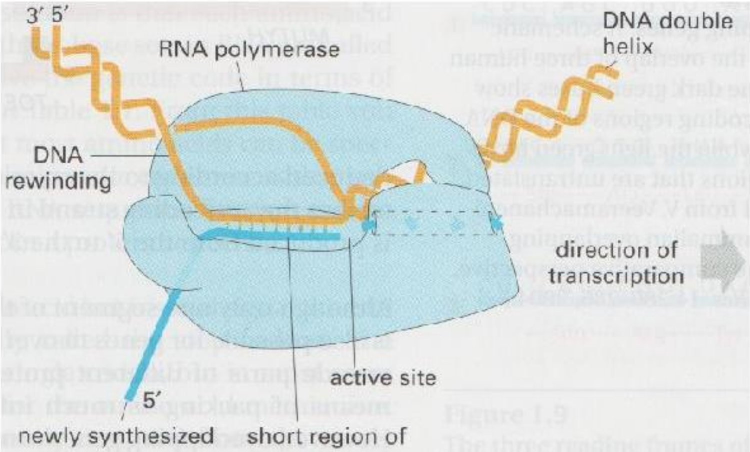
- There is a flow of genetic information:

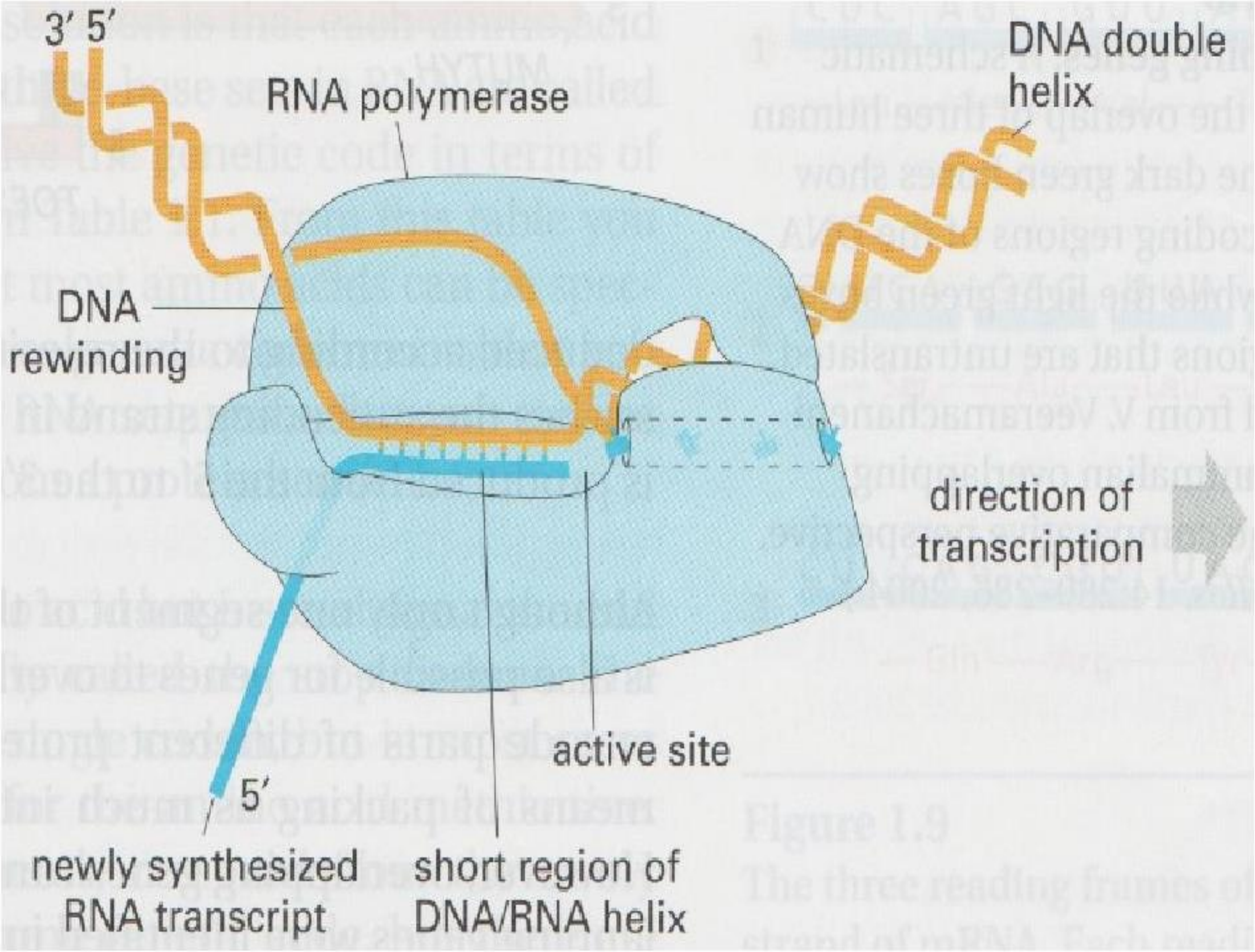
  DNA ⇒ RNA (mRNA) ⇒ Proteins
- DNA acts as the store of information, RNA as the translator of information into proteins, which are the actuators of life.



Transcription

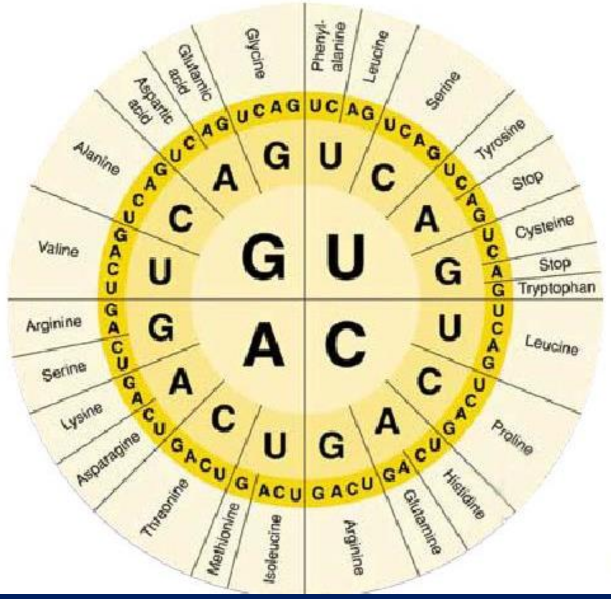
- Transcription is the process by which the information store (DNA) is expressed in the form of the translator (messenger RNA).
- In a fashion roughly akin to DNA polymerase, RNA polymerase unwinds the DNA.
- By operating on the noncoding/anticoding/antisense strand (3' to 5'), it transcribes an mRNA strand identical to the coding/sense DNA strand (5' to 3').
- As mentioned previously, uracil (U) replaces thymine (T) in RNA.




3′ 5′
RNA polymerase
DNA double helix
DNA rewinding
direction of transcription
active site
5′
newly synthesized RNA transcript
short region of DNA/RNA helix

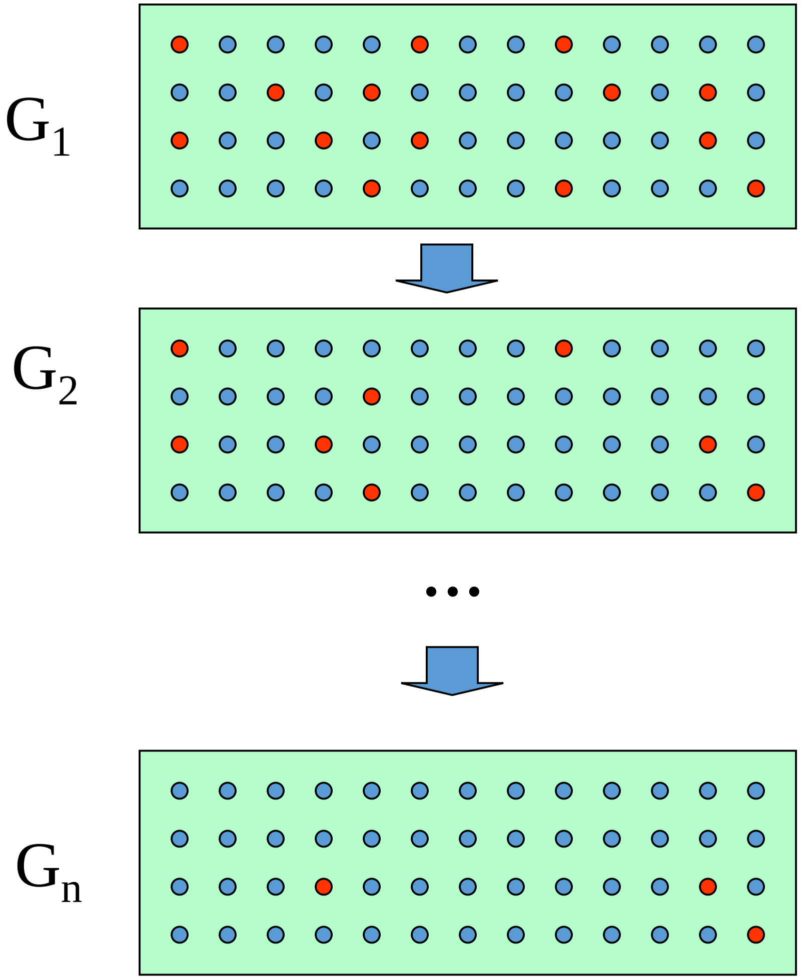
Translation

- Messenger RNA (mRNA) is then translated into a sequence of amino acids by ribosomes with the help of ribosomal RNA and tranfer RNA, rRNA and tRNA, respectively.
- Three-base sets of RNA make up amino acid forming codons.
- Translation is often signaled by the AUG codon, coding for Met (methionine).
- UAA, UAG and UGA are stop codons, signalling the end of translation.



Wright-Fisher Model of Genetic drift (ca. 1920)

Loss of variants due to random sampling of progeny from parental gene pool



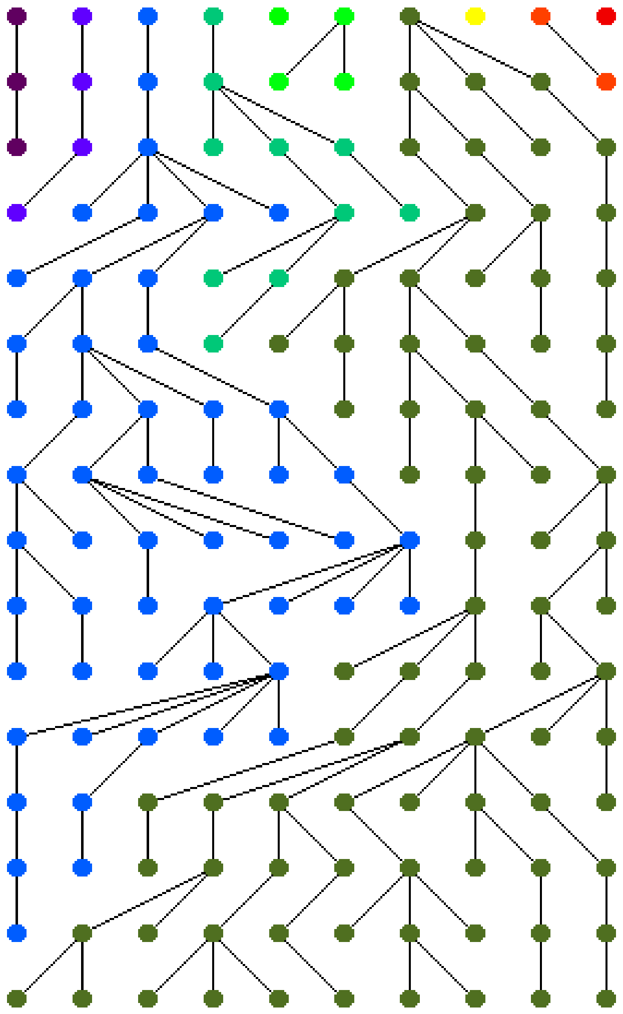
Alleles: A_1: ● A_2: ●

Replication = sampling with replacement

A_1 – becomes fixed

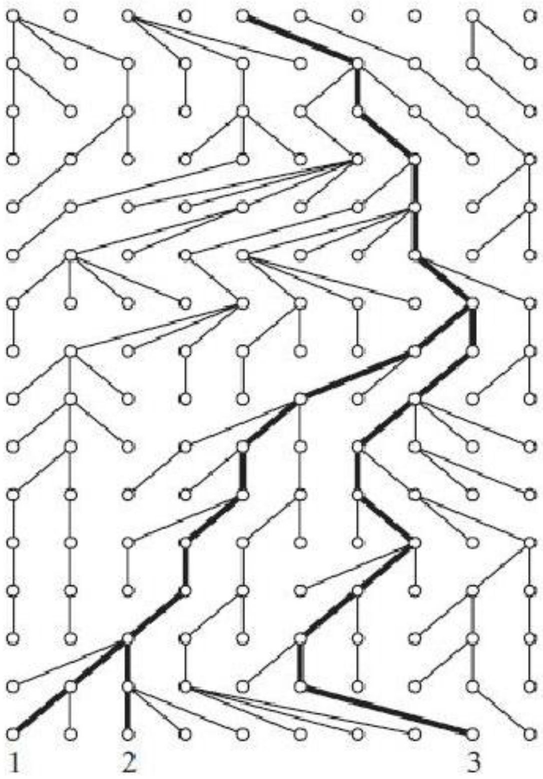
A_2 – becomes lost

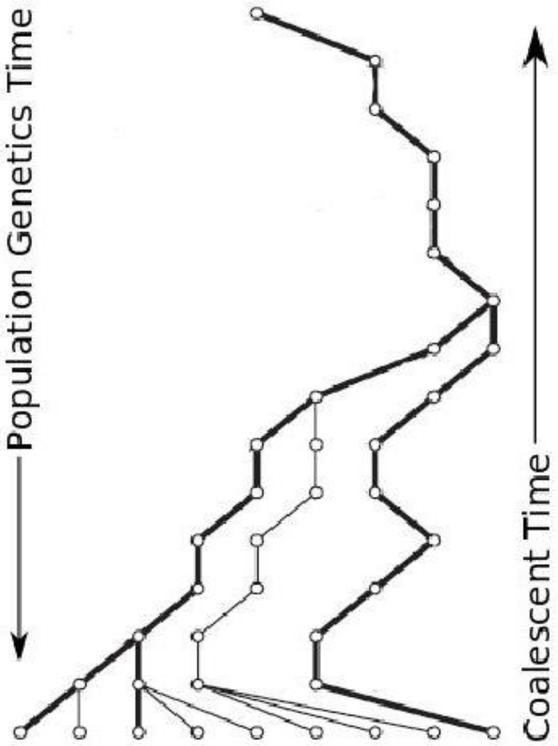
Coalescent: Drift seen in reverse time



Lineage mergers in reverse time correspond to

Loss of variants in forward time





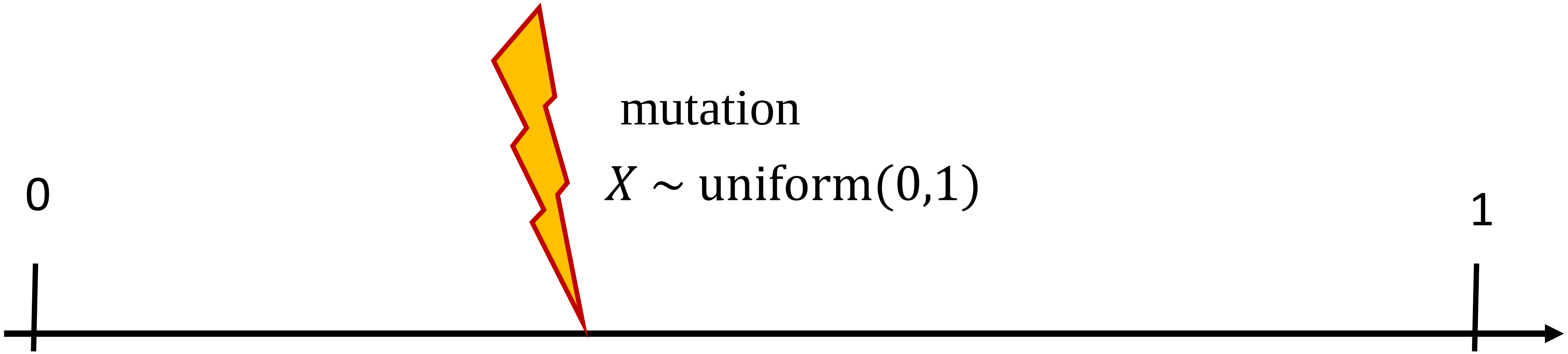
Mutation

Mutation times follow a **Poisson process** with intensity μ measured per locus (per site) per generation

Model used for rare mutations in long genomes:

Infinite Sites Model (ISM), where it is assumed that each mutation takes place at a DNA site that never mutated before

- Mutation sites can be represented as **iid uniform(0, 1) rv's**

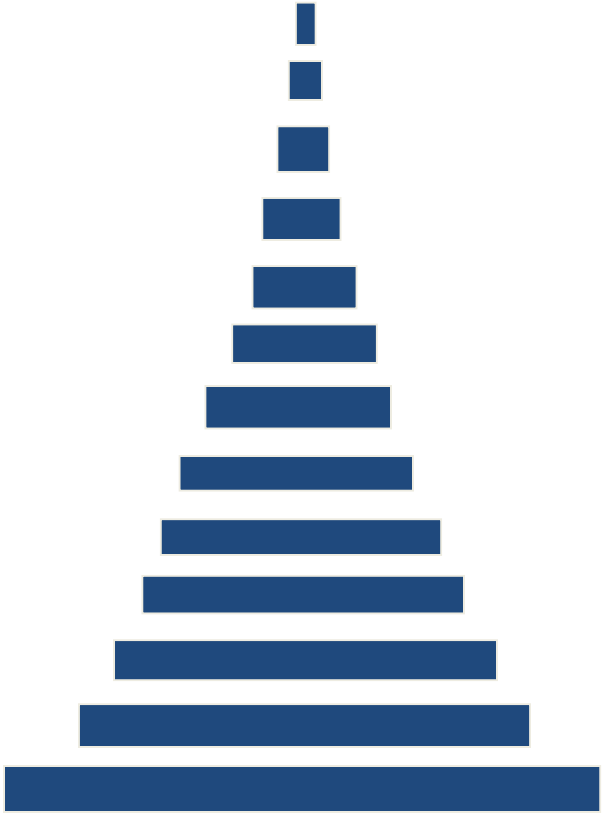


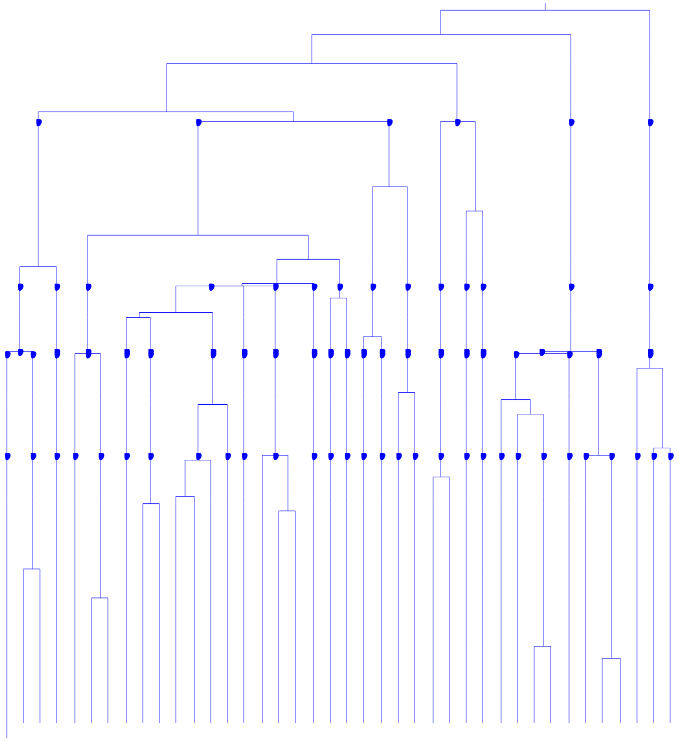
Genome = [0, 1]

Growing population scenario

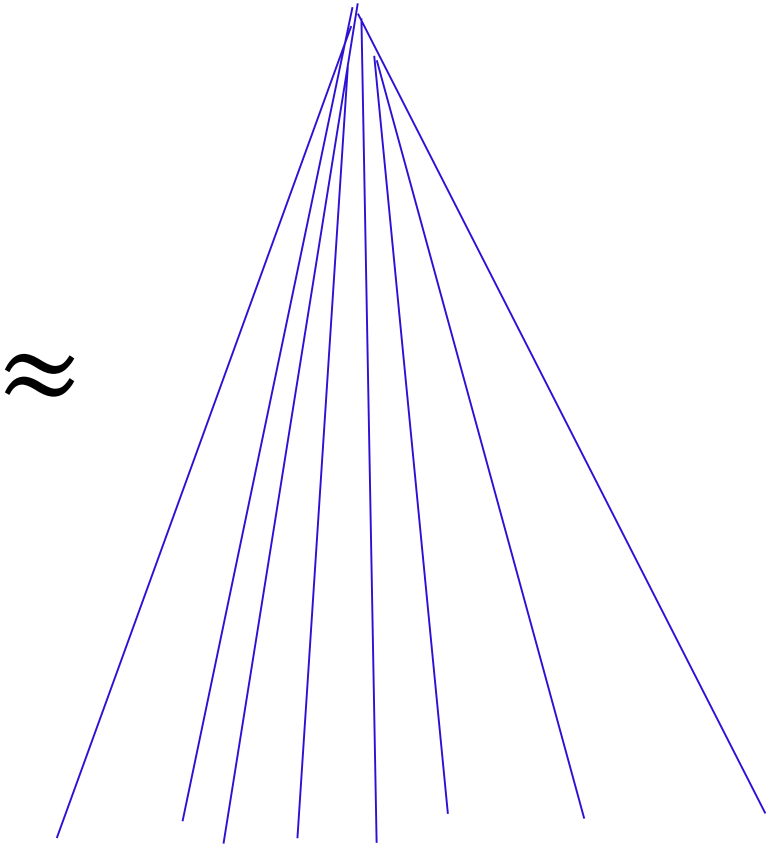
≈ tumor growth case

Coalescent tree “star-like”





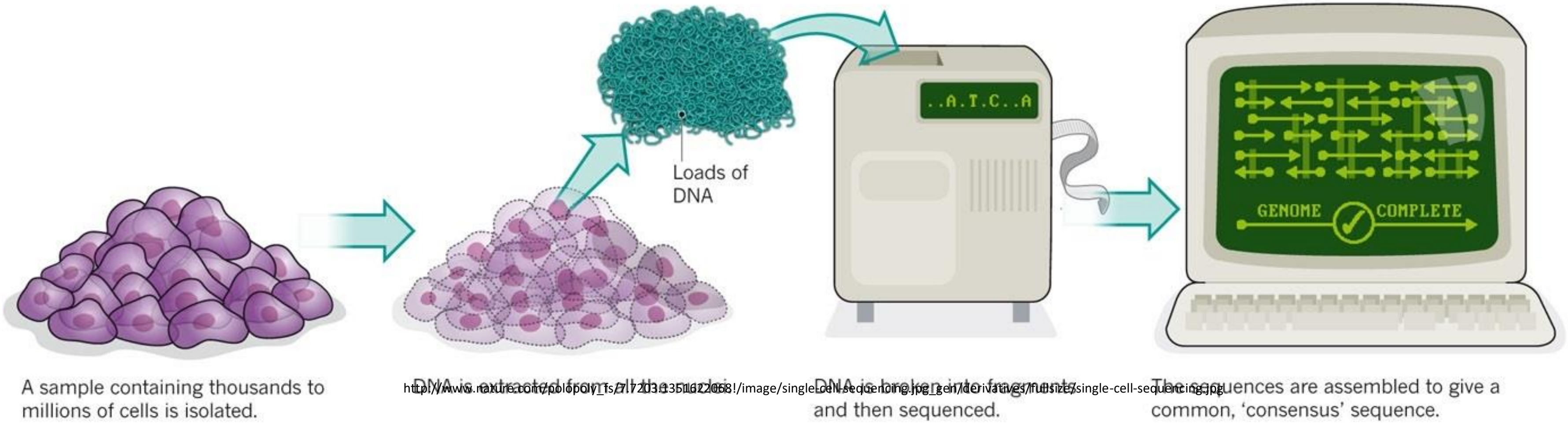
≈



Coalescence method

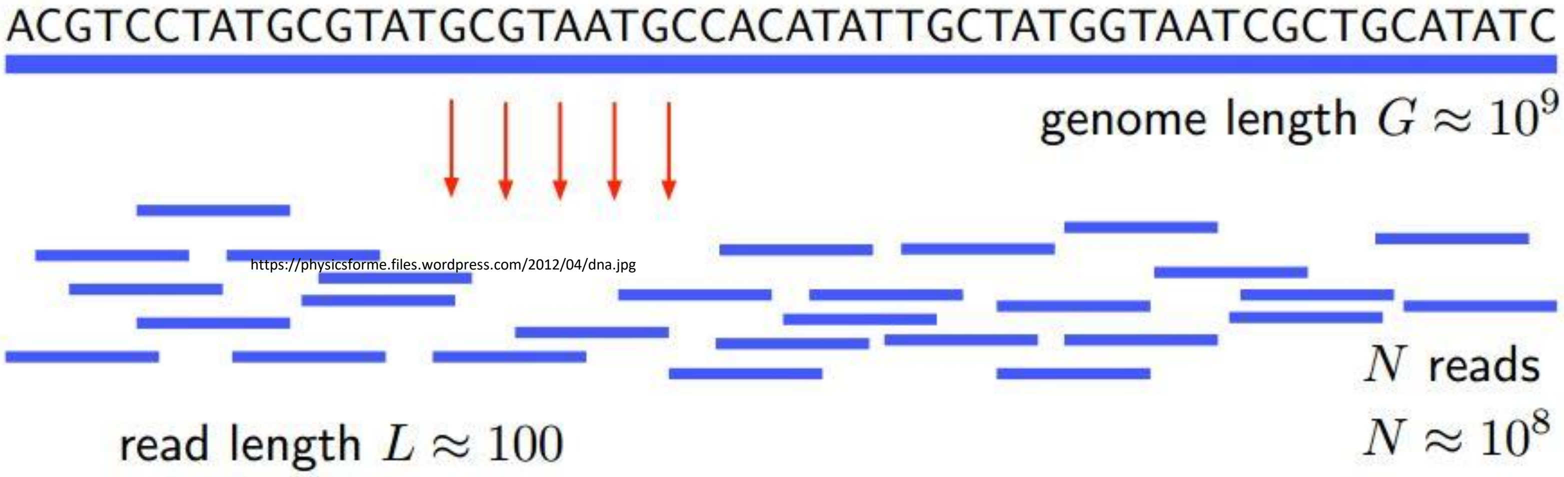
- Genetic drift viewed in reverse time
- Estimating the past of an *n* - sample of sequences taken at present.
- Possible events that happen in the past are
  - **coalescences** (lineage merges) leading to common ancestors of sequences, and
  - **mutations** along branches of ancestral tree each at a new site (“Poisson rain”).

Sequencing cancer DNA



Loads of DNA
..A.T.C..A
GENOME COMPLETE
A sample containing thousands to millions of cells is isolated.
DNA is extracted from all the nuclei
DNA is broken into fragments and then sequenced.
The sequences are assembled to give a common, ‘consensus’ sequence.

http://www.nature.com/polopoly_fs/7.7203.1351622068!/image/single-cell-sequencing.jpg_gen/derivatives/fullsize/single-cell-sequencing.jpg



ACGTCCTATGCGTATGCGTAATGCCACATATTGCTATGGTAATCGCTGCATATC
genome length $G \approx 10^9$
N reads
$N \approx 10^8$
read length $L \approx 100$

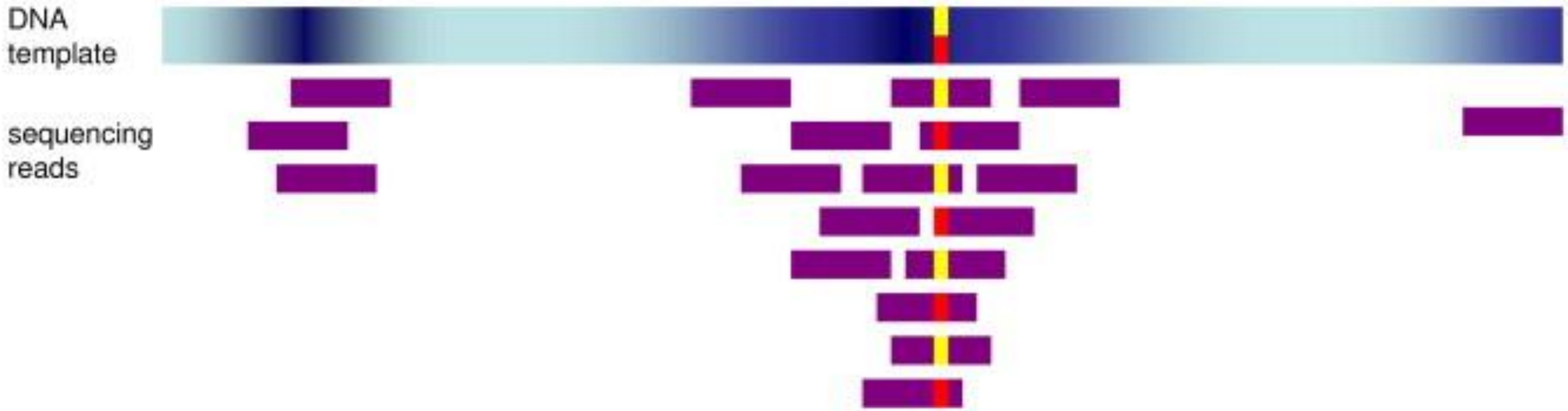
https://physicsforme.files.wordpress.com/2012/04/dna.jpg

Mutations, or rather “called somatic variants” at sites (DNA nucleotides) where reads differ

…TATATGCTAGCTAGCTACGGCGCGCTG…

↓

…TATATGCTAGCTAGATACGGCGCGCTG…



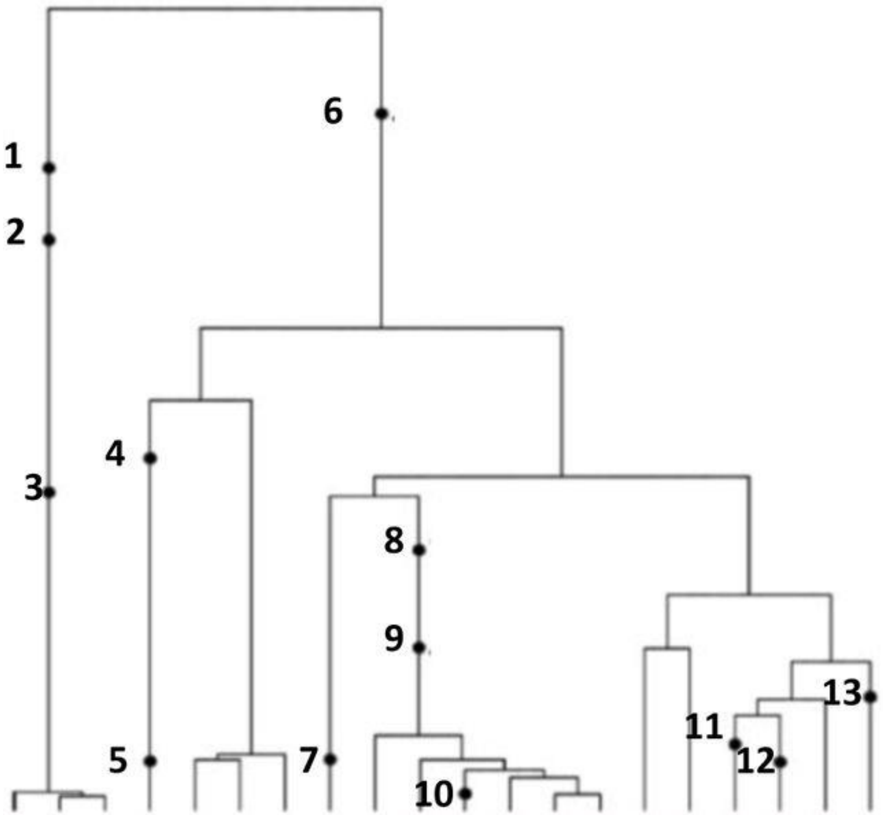
O. Morozova, M. A. Marra, Genomics, 2008

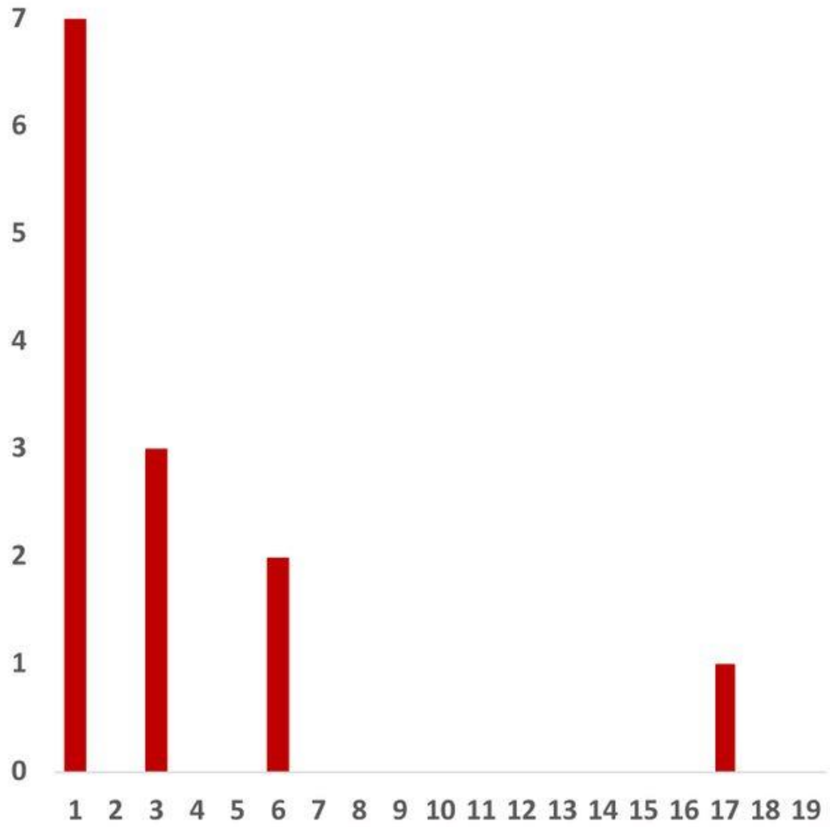
Site Frequency Spectrum (SFS) = Variant Allele Frequency (VAF)

Statistic for neutral mutation distribution (neutral mutations are ticks of molecular clock)

SFS is bar chart of $\boldsymbol{\eta_i}$ = # mutations represented in $\boldsymbol{i}$ out of $\boldsymbol{n}$ cells

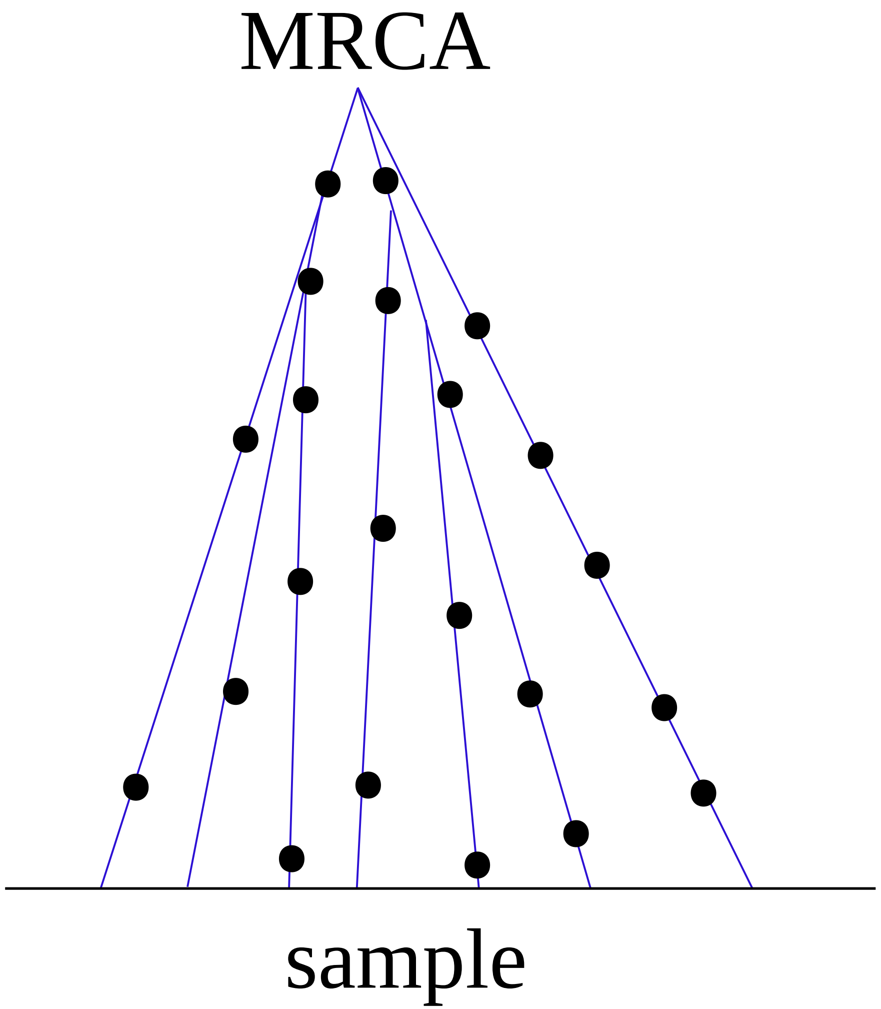
$\eta = \{\eta_1, \eta_2, \ldots, \eta_{20}\} = \{7, 0, 3, 0, 0, 2, 0, 0, 0, 1, \ldots, 0\}$, $\Sigma_{i=1}^{n-1}\eta_i = s = \#$ segregating sites





Consequences of star-shaped coalescence:

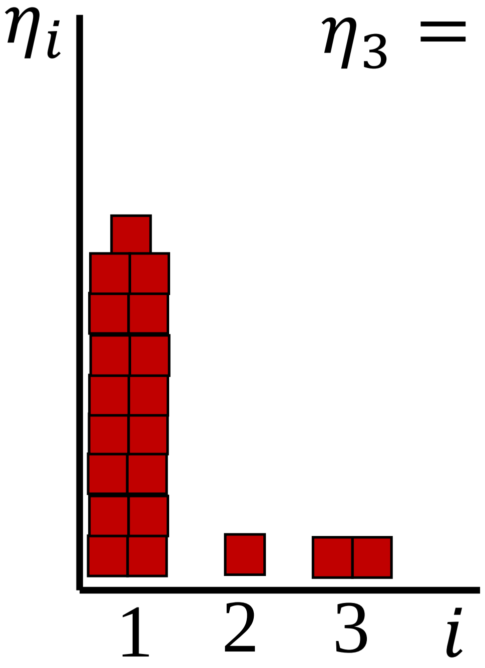
Exaggerated singletons (● = mutation)



$\eta_1 = 17$, singletons

$\eta_2 = 1$, dublets

$\eta_3 = 2$, triplets



Expected SFS based on GT-coalescent

q_b = probability that a mutation is present in

b $= \; 1, 2, \dots, n$ sequences out of the sample of n sequences

Griffiths and Tavare, 1998

$$q_b = \frac{\sum_{k=2}^{n} p_k^n(b) k E(S_k)}{\sum_{k=2}^{n} k E(S_k)}$$

Depend on metrics of the tree:
Expectations of inter-merger times

Probability that a mutation at the level where there are k ancestors will grow to b copies at the bottom of the tree

$$p_k^n(b) = \frac{\binom{n-b-1}{k-2}}{\binom{n-1}{k-1}}$$

Polanski and Kimmel M (2003) New Explicit Expressions… Genetics

How to reconstruct past dynamics ?

- Chief problem: how to measure **time** ?

- But are mutations **not** accumulating in/with **time** ?

- The difficulty is clear from the **asymptotic** Griffiths-Tavaré/Durrett formula

$$q_m \begin{cases} \sim \frac{n\nu}{\gamma} \ln(N\gamma)\,, & m = 1 \text{ (singletons)} \\ \to \frac{n\nu}{\gamma} \frac{1}{m(m-1)}, & m = 2, \dots, n-1 \end{cases} \quad \text{as } N \to \infty$$

where $\boldsymbol{\frac{\nu}{\gamma} \ln(N\gamma) = \nu t}$, since $N = N(t) = \gamma \exp(\gamma t)$

- Notice q_m decays almost exactly **quadratically** (so, log-log slope $= \boldsymbol{-2}$)

- **Time is associated with singleton count**, but in genome data singletons are considered indistinguishable from **sequencing errors** and discarded

Lambert - Stadler - MK SFS

SFS based on a sample of size n from a b-d process grown to size N, with $p = \frac{n}{N}$

$$\mathbb{E}S_n(k) = \theta \int_0^{\infty} \left(1 - W(t)^{-1}\right)^{k-1} \left((n-k-1)W(t)^{-2} + 2W(t)^{-1}\right) dt$$

where $W(x) = 1 + \frac{b}{r}(e^{rx} - 1), \quad x \in [0, \infty)$

$$\mathbb{E}S_n(k) = \frac{\theta}{r}\left(\frac{n-k-1}{k(k+1)}F([1,2];k+2,\alpha) + \frac{2}{k}F([1,1];k+1,\alpha)\right) \quad \alpha = 1 - pb/r$$

$$F([1,2];i,\alpha) = (i-1)(i-2)\alpha^{1-i}\int_0^{\alpha}(\alpha - t)^{i-3}\frac{t}{1-t}dt$$

e.g. $F([1,2];3,\alpha) = -2\alpha^{-2}(\ln(1-\alpha) + \alpha)$ etc.

Three curves (G-T, Durrett’s, Lambert-MK)

Lambert-MK based on b-d process sampling

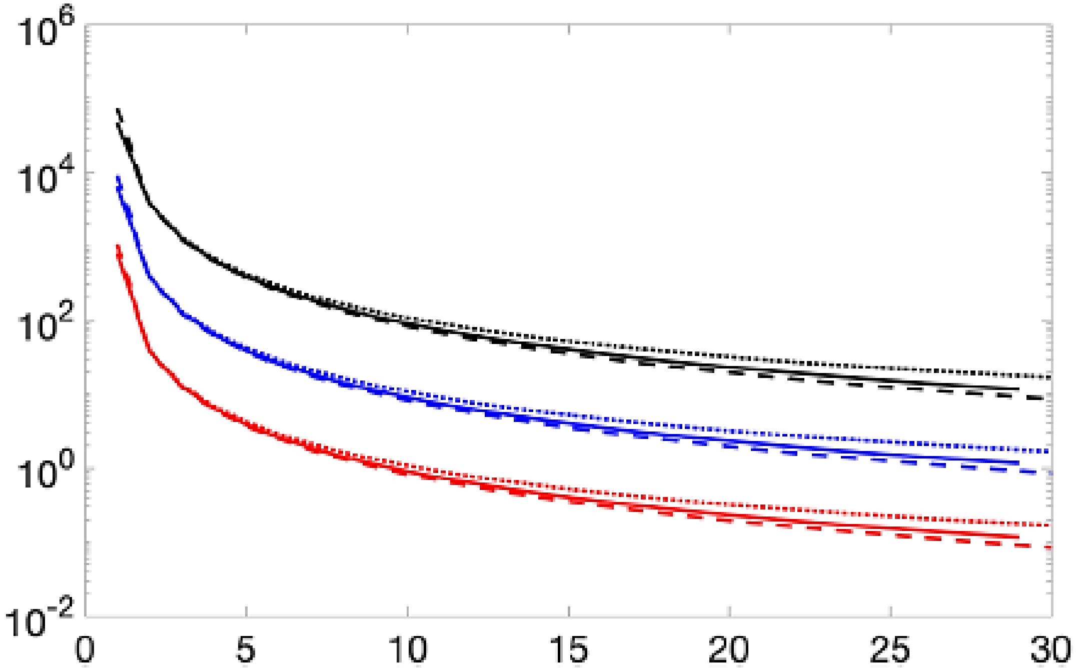
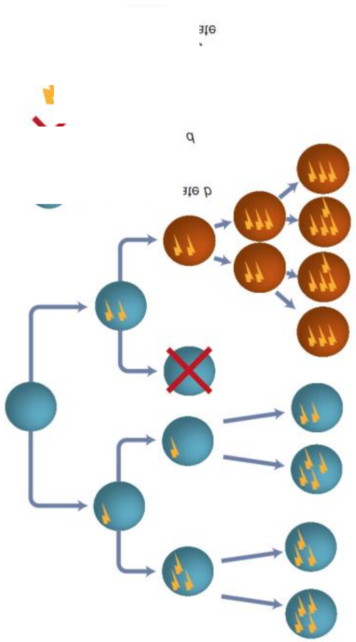
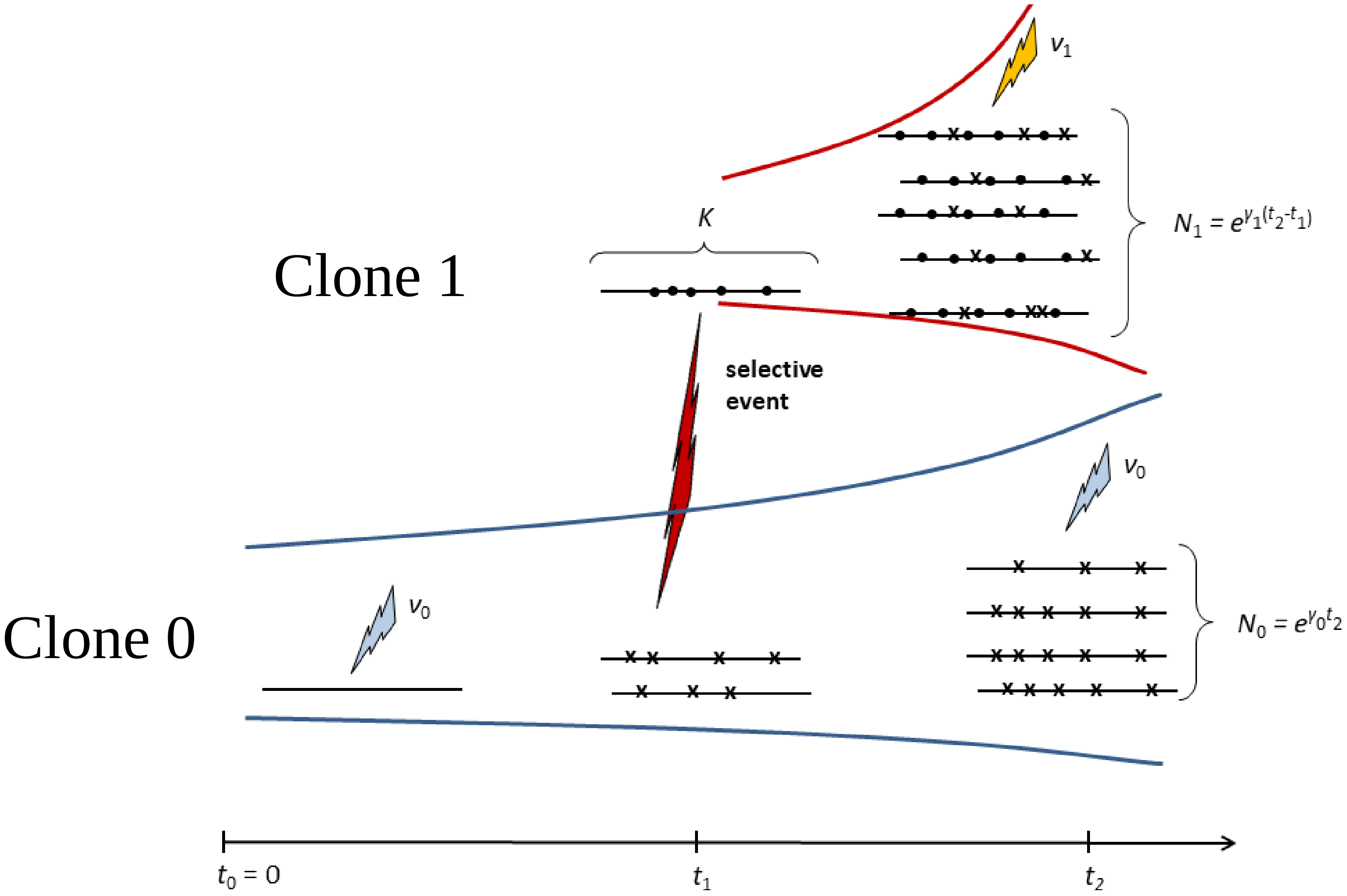


FIG. 3. *Comparison of expected SFS based on the hypergeometric formula (8) with parameters as in Table 1 (dotted lines), Griffiths–Tavaré theory (continuous lines), and Durrett’s approximation (dashed lines). Three cases as in Table 1, fast-growing tumors (red), moderate–growing (blue), and slow growing ones (black) are considered. $\theta = 1$ has been assumed. Unscaled parameters listed in Table 1, can be converted to scaled ones, using Table 2.*

Neutral evolution with a selective sweep in a tumor

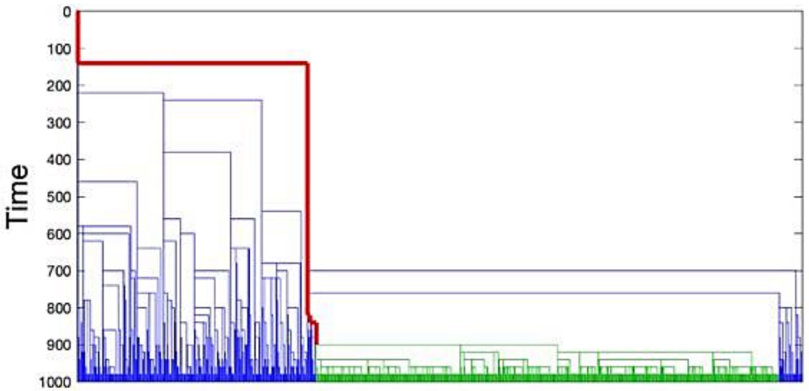
(Dinh et al Stat Sci 2020)



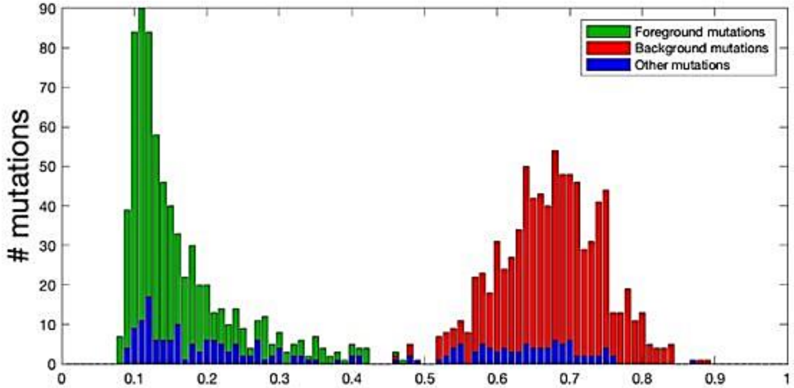


How to interpret the SFS? Suppose that ….

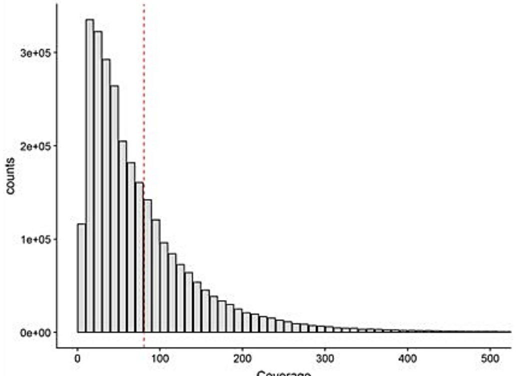
- Blue cells grow and mutate as a birth-death process
- At time around 900, a green subclone emerges and proliferates **faster**
- Mutation accumulate neutrally, except that
    - A (red) subset of mutations leads to the ancestral cell of the green clone
- When we sample reads from cells, the SFS will show **red hump**



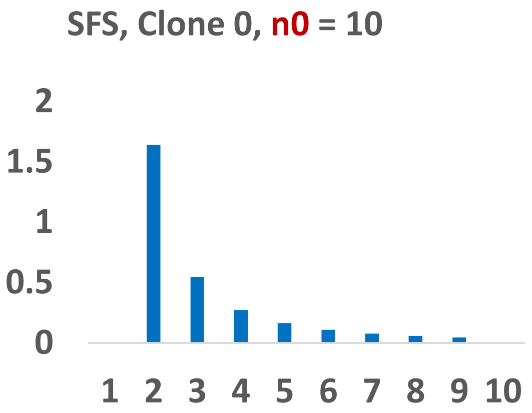
At T_N: selective=63%, total=821 cells

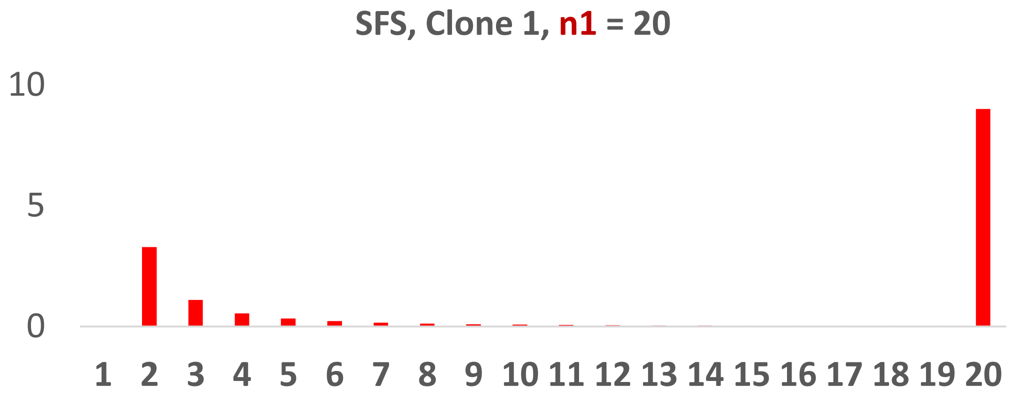


SFS (binomial, mean=50); population at T_N=821; selective=63%

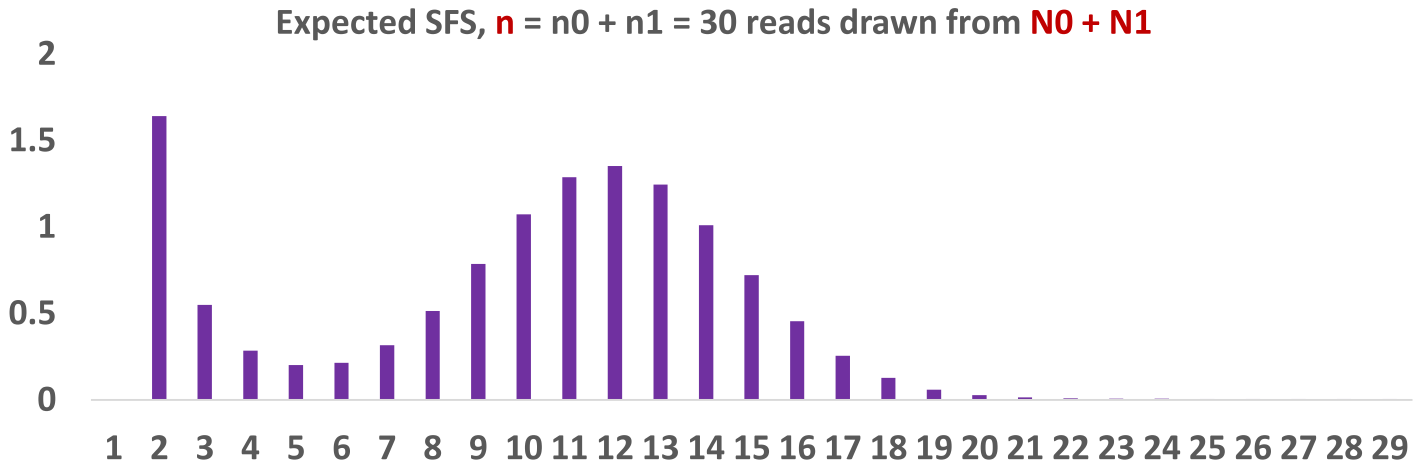


- The skew component correspond to neutral accumulation of mutations
- More subclones → more humps?


SFS, Clone 0, n0 = 10


SFS, Clone 1, n1 = 20

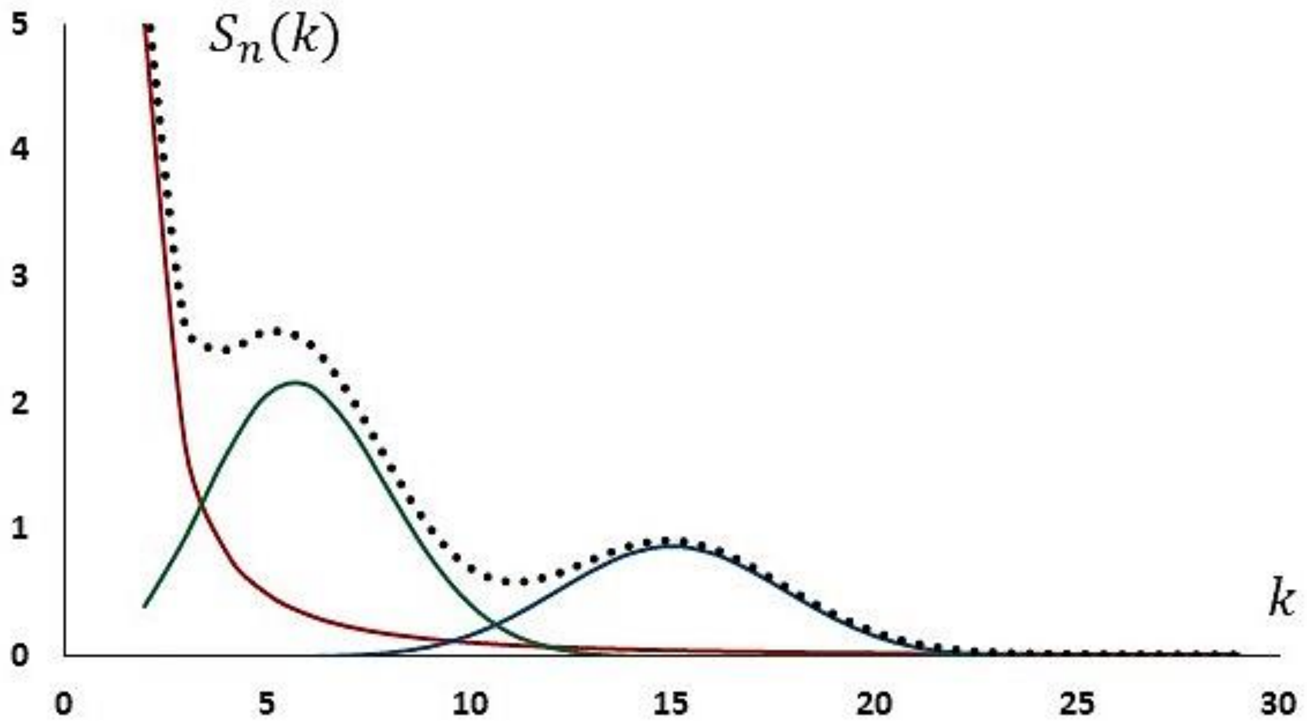
$$q^0_m = \frac{p_0 n_0 v_0}{\gamma_0}\frac{1}{m(m-1)}, m = 2, \dots, n_0 - 1 \qquad q^1_m = \frac{p_1 n_1 v_1}{\gamma_1}\frac{1(m < n_1)}{m(m-1)} + K\delta_{m,n_1}, m = 2, \dots, n_1$$


Expected SFS, n = n0 + n1 = 30 reads drawn from N0 + N1

$$Q_m = A\frac{1}{m(m-1)} = n\left(\frac{p_0 v_0}{\gamma_0} + \frac{p_1 v_1}{\gamma_1}\right)\frac{1}{m(m-1)} + K\binom{n}{m}{p_0}^{n-m}{p_1}^{m}, \qquad m = 2, \dots, n-1$$

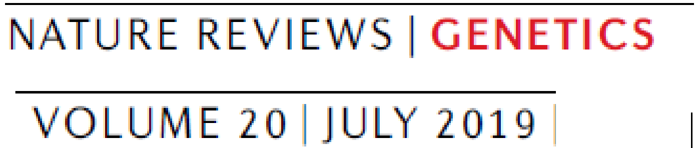
Where A is the total of neutral mutations, np_1 is the mode of the hump, and K is the area under the hump.

3 clones?



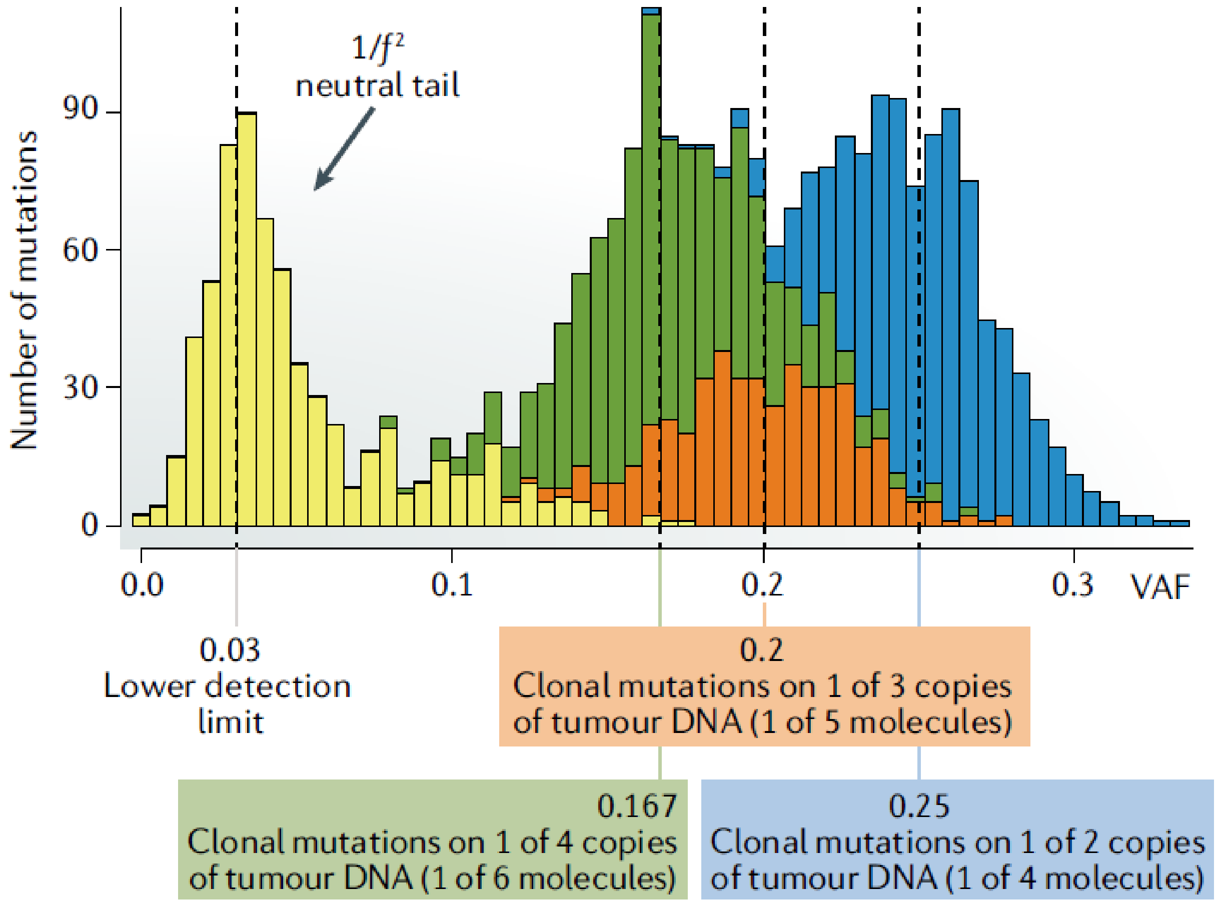
Resolving genetic heterogeneity in cancer

Samra Turajlic[1,2,7], *Andrea Sottoriva*[3,7], *Trevor Graham*[4]* *and Charles Swanton*[1,5,6]*



NATURE REVIEWS | GENETICS

VOLUME 20 | JULY 2019 |



0.03
Lower detection limit

0.2
Clonal mutations on 1 of 3 copies of tumour DNA (1 of 5 molecules)

0.167
Clonal mutations on 1 of 4 copies of tumour DNA (1 of 6 molecules)

0.25
Clonal mutations on 1 of 2 copies of tumour DNA (1 of 4 molecules)

Expanded Biological Background

Introduction to DNA (and other) Sequences

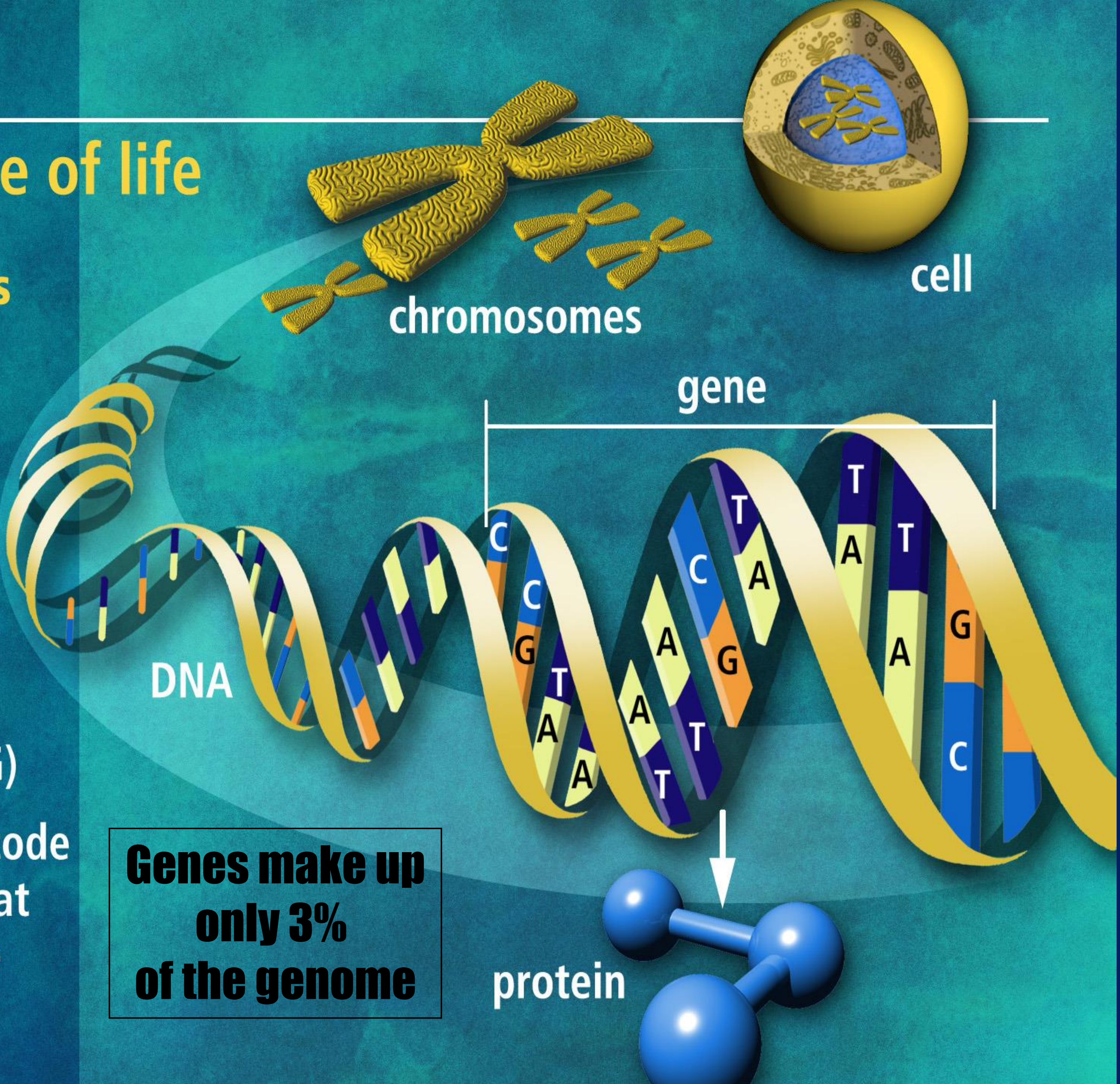
DNA

the molecule of life

Trillions of cells

Each cell:

- 46 human chromosomes
- 2 m of DNA
- 3 billion DNA subunits (the bases: A, T, C, G)
- 30,000 ~~80,000~~ genes code for proteins that perform all life functions



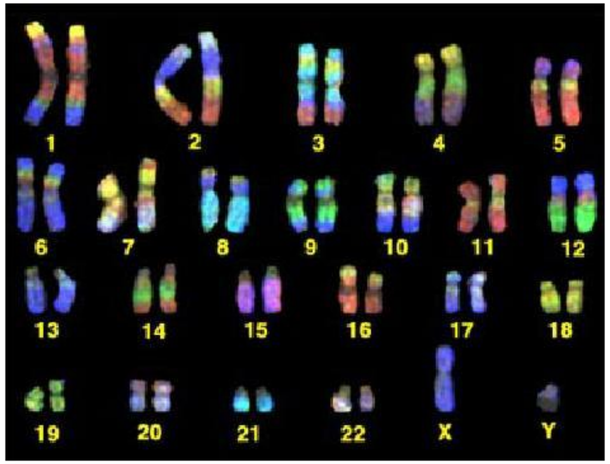
Genes make up only 3% of the genome

Genome Sizes

Human	3.0 x 10^{9} base pairs
Mouse	3.0 x 10^{9}
Drosophila	1.1 x 10^{8}
Worm	1.0 x 10^{8}
Dictyostelium	3.4 x 10^{7}
Yeast	1.2 x 10^{7}
Bacteria	1.0 - 5.0 x 10^{6}

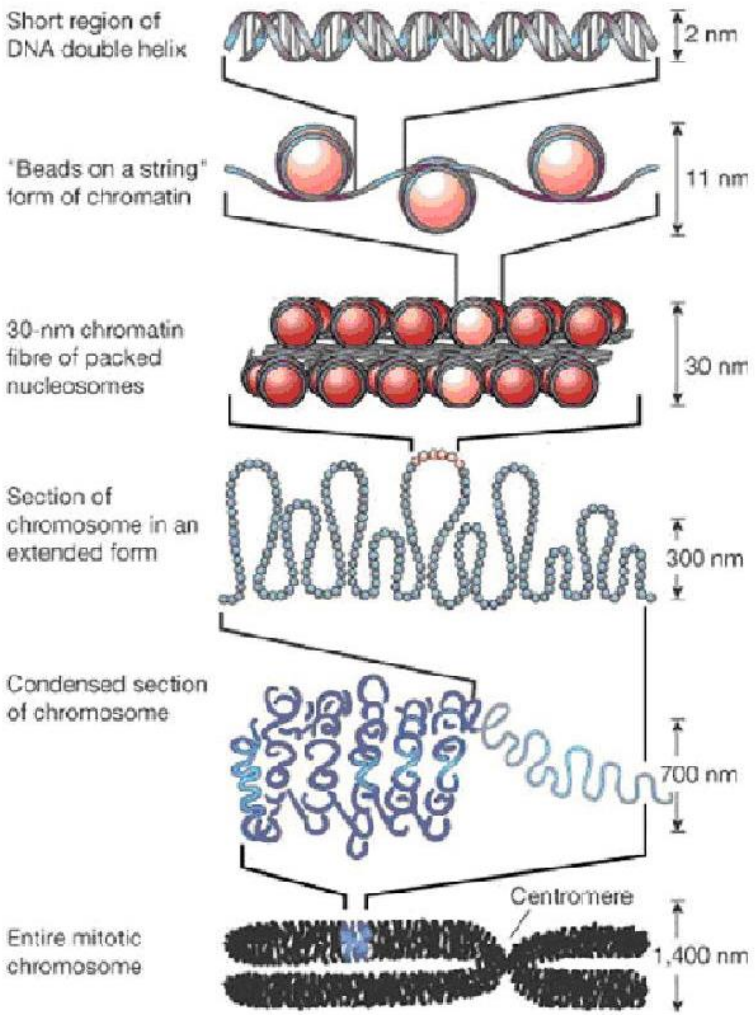
Deoxyribonucleic Acid (DNA) within Chromosomes

- Each somatic (non-gamete) cell in the human body contains two copies of 23 chromosomes within the cellular nucleus.
- One copy is inherited from the mother, one from the father.
- Humans are thus diploid organisms, with haploid gametes which fuse during reproduction.

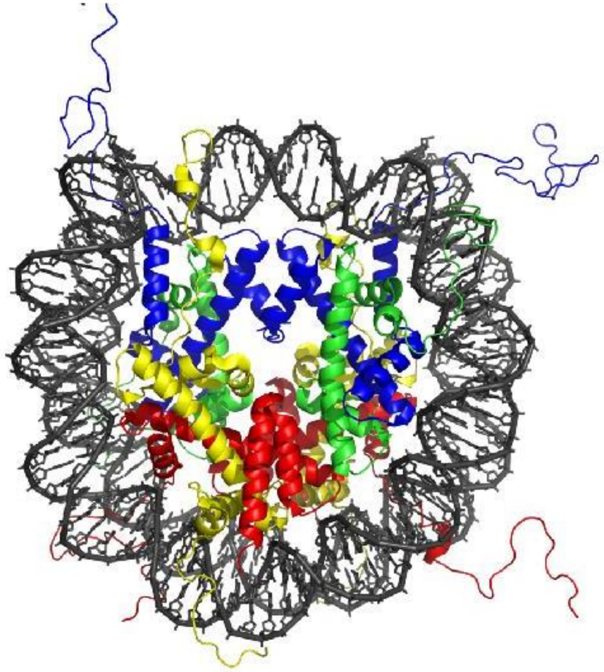


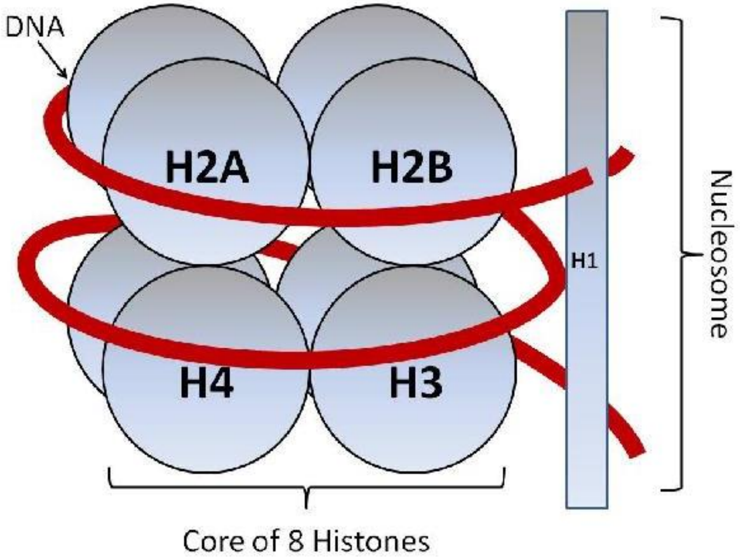
Chromosome Coiling

- Chromosomes are structures comprised of super-coiled DNA.


Short region of DNA double helix
2 nm
"Beads on a string" form of chromatin
11 nm
30-nm chromatin fibre of packed nucleosomes
30 nm
Section of chromosome in an extended form
300 nm
Condensed section of chromosome
700 nm
Centromere
Entire mitotic chromosome
1,400 nm

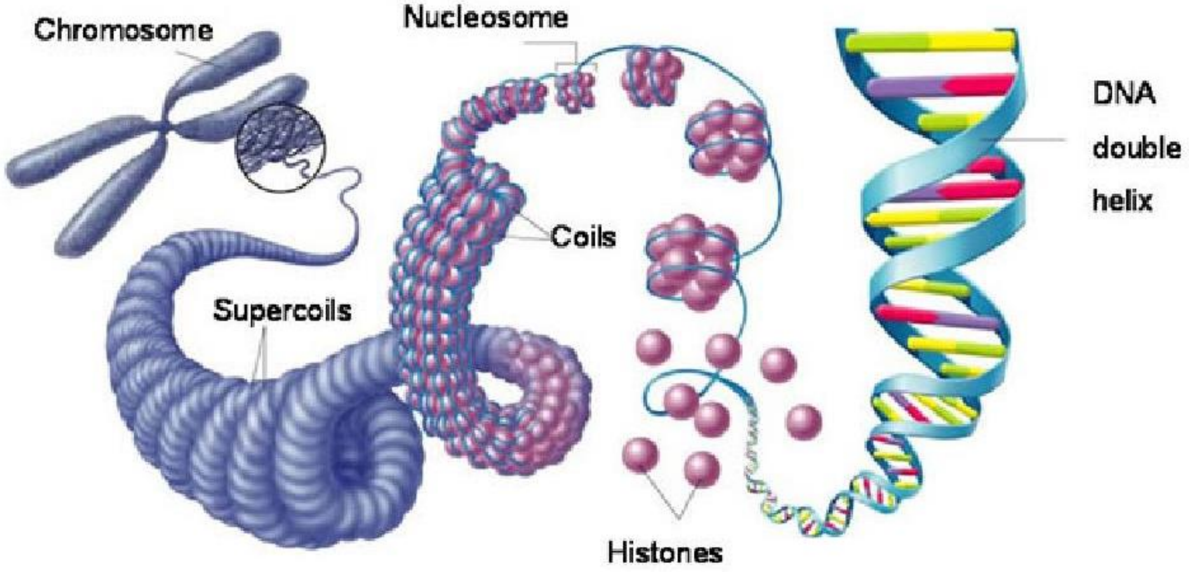
Histone “Beads on a String”





Chromosome Coiling

- Chromosomes are structures comprised of super-coiled DNA.

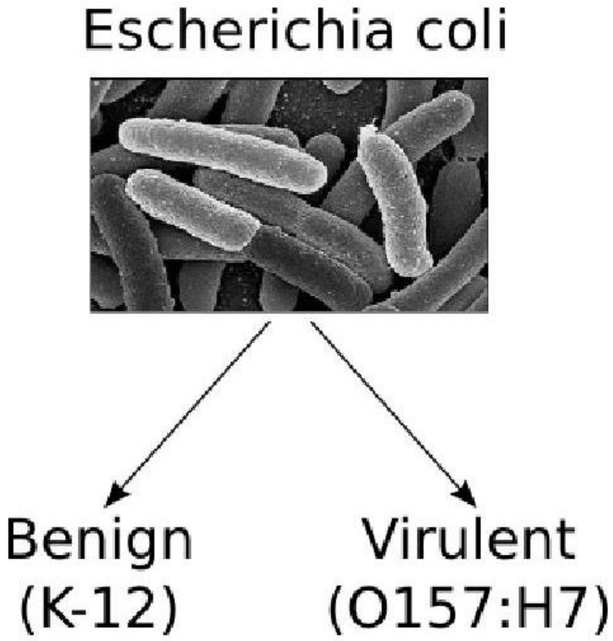


DNA Fun Facts

- Each somatic cell contains 2m of (uncoiled) DNA.
- Chromosomes range in length from 85cm (Chromosome 1) to 16cm (Chromosome 21).
- There are 3 billion nucleotide base pairs in the human genome.
- $\sim$21,000 genes code for the proteins which perform all our life functions.
- Protein-coding genes only make up 1-2% of our DNA.

Relationship between Organism Complexity and Genome Length

- ... is often tenuous.



Escherichia coli

Benign (K-12)

Virulent (O157:H7)

- The virulent strain contains $5,416$ genes in 5.44×10^6 DNA base pairs.
- $1,387$ of these genes are not found in the harmess laboratory strain.

Arabidopsis thaliana and Psilotum nudum

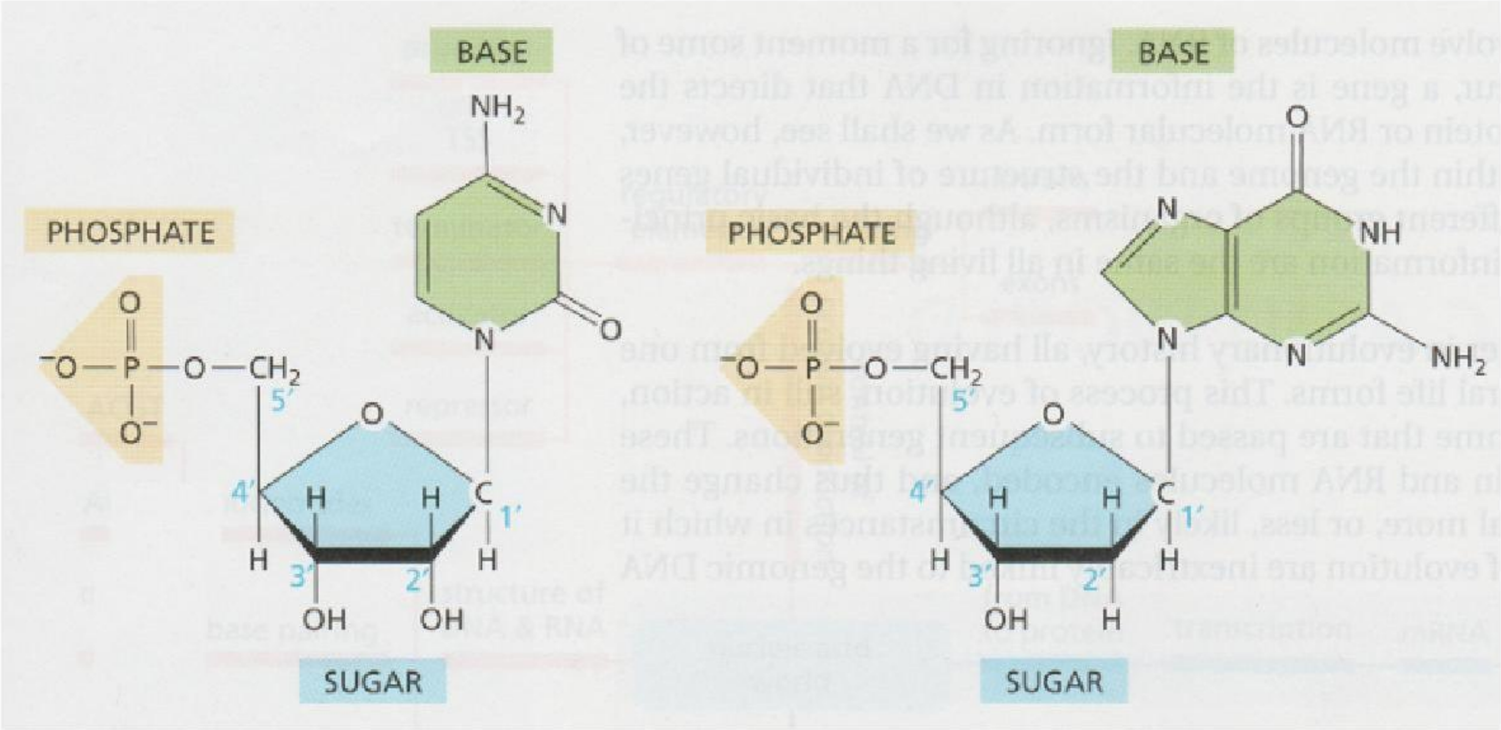
Arabidopsis thaliana

Psilotum nudum

- Psilotum nudum (the whisk fern) is far less complex an organism than Arabidopsis (no true leaves, flowers or fruit).
- However, its genome is made up of 2.5×10^{11} base pairs, **3000** times more than Arabidopsis!
- Recall, the human genome has “only” 3×10^{9} base pairs.
- As in many amphibians with far more DNA than us, the difference is in repetitive DNA.

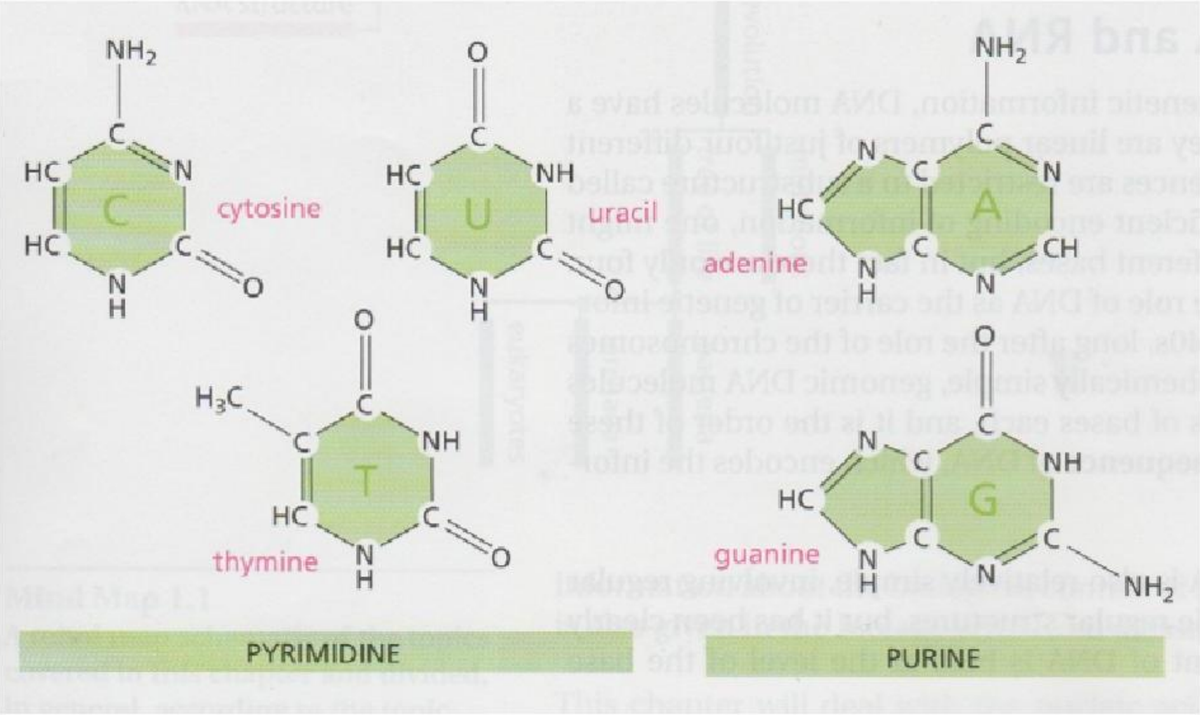
DNA and RNA Structure

- Nucleic acid consists of a pentose sugar with a phosphate group and base attached.
- DNA and RNA differ only in their sugar groups, deoxyribose and ribose, respectively.



DNA and RNA Structure

- DNA and RNA bases are either pyrimidines or purines, both planar ring structures.
- Cytosine (C), thymine (T) in DNA and uracil (U) in RNA are all pyrimidines.
- Adenine (A) and guanine (G) make up the purines.



DNA and RNA Chaining

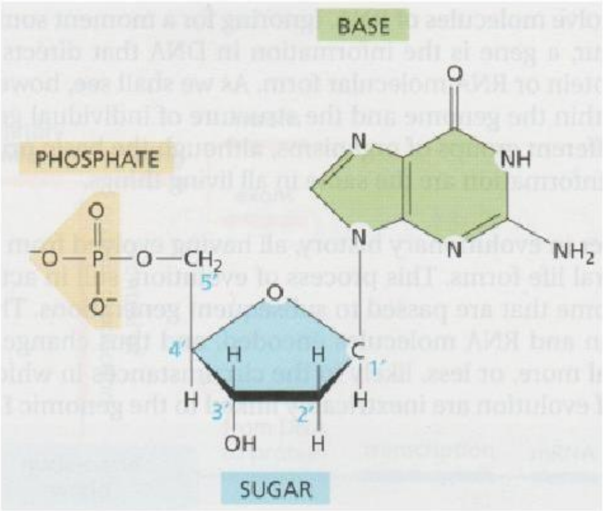
- DNA and RNA are chained together through phosphodiester linkages between the phosphate group of one nucleotide linked to the 3' carbon of the sugar of the other nucleotide.
- Thus, in any chain, one end has a free phosphate group on the **5'** carbon.
- The other end has a free hydroxyl group on the **3'** carbon.
- Nucleotide sequences are written by convention from the **5' end** to the **3' end**.

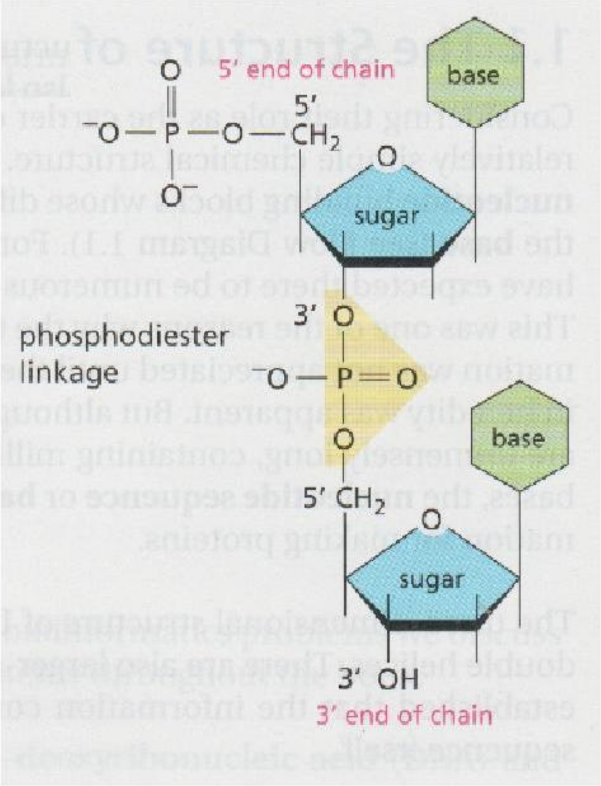
5' ACGTAGCTTATTAGA **3'**

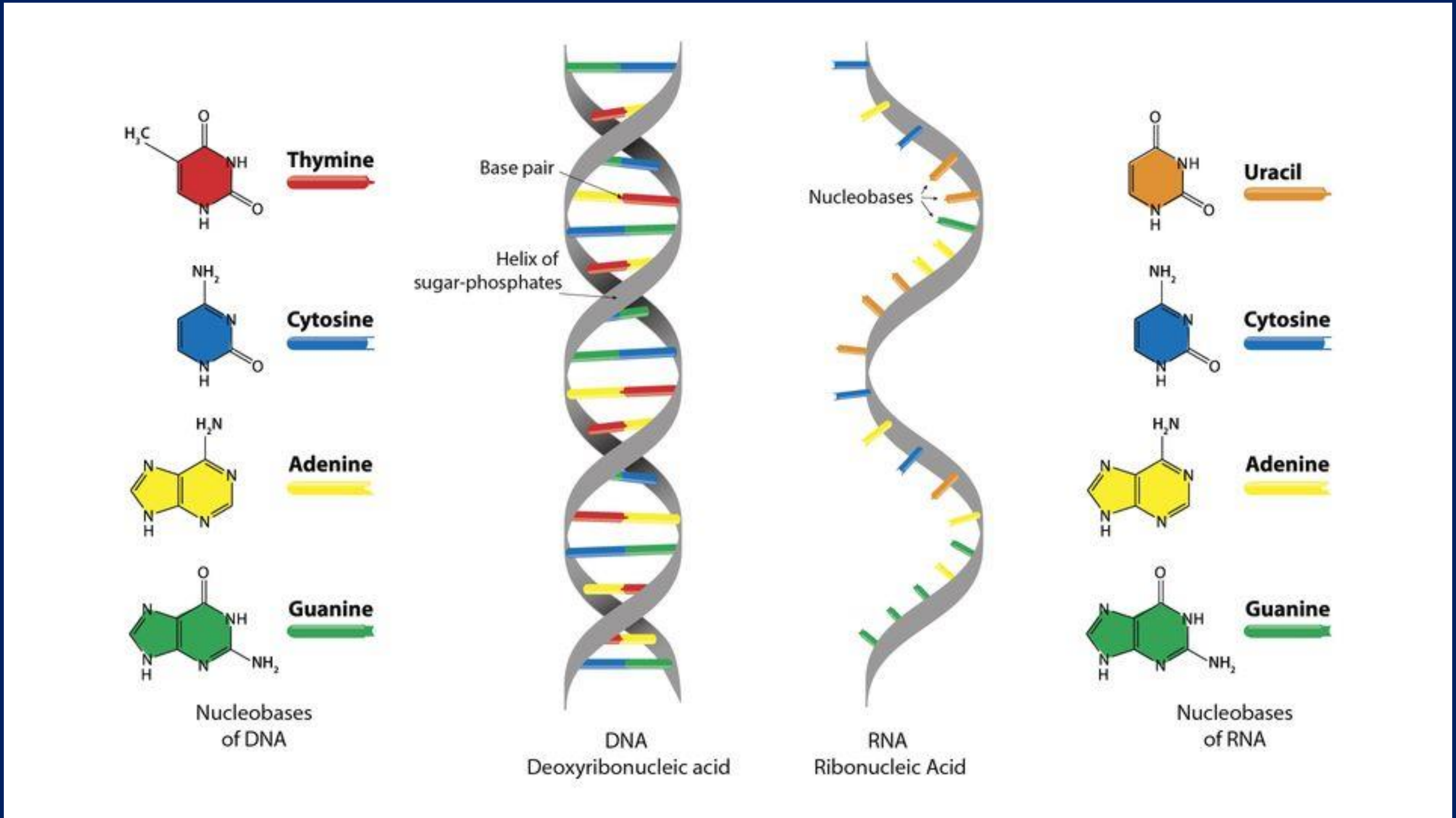
DNA and RNA Chaining

- Nucleotide sequences are written by convention from the **5' end** to the **3' end**.

5' ACGTAGCTTATTAGA **3'**







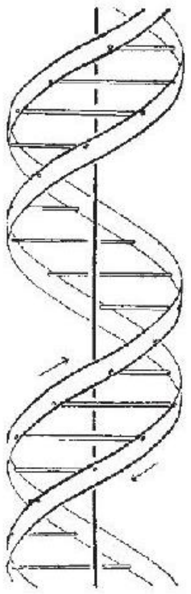
https://www.technologynetworks.com/genomics/articles/what-are-the-key-differences-between-dna-and-rna-296719

DNA Helix Structure

- Watson and Crick won the Nobel prize in 1962 for their discovery, initially in 1953, of the double-helical structure of DNA.

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid



This figure is purely diagrammatic. The two ribbons symbolize the two phosphate—sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

DNA Helix Structure

- Their 1953 *Nature* paper was more of an announcement with a detailed follow-up manuscript published by *The Royal Society* in 1954.

The complementary structure of deoxyribonucleic acid

By F. H. C. Crick and J. D. Watson*†

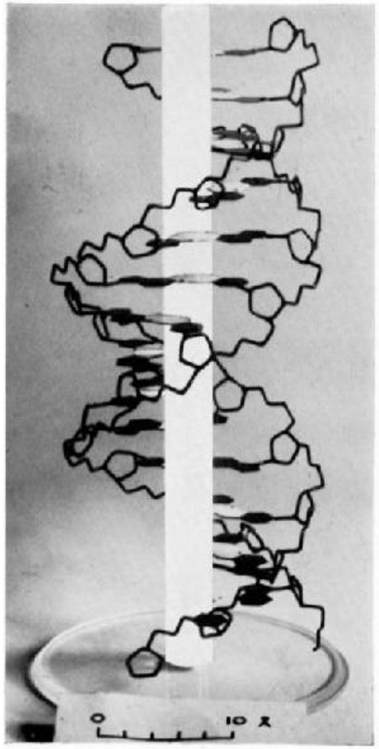


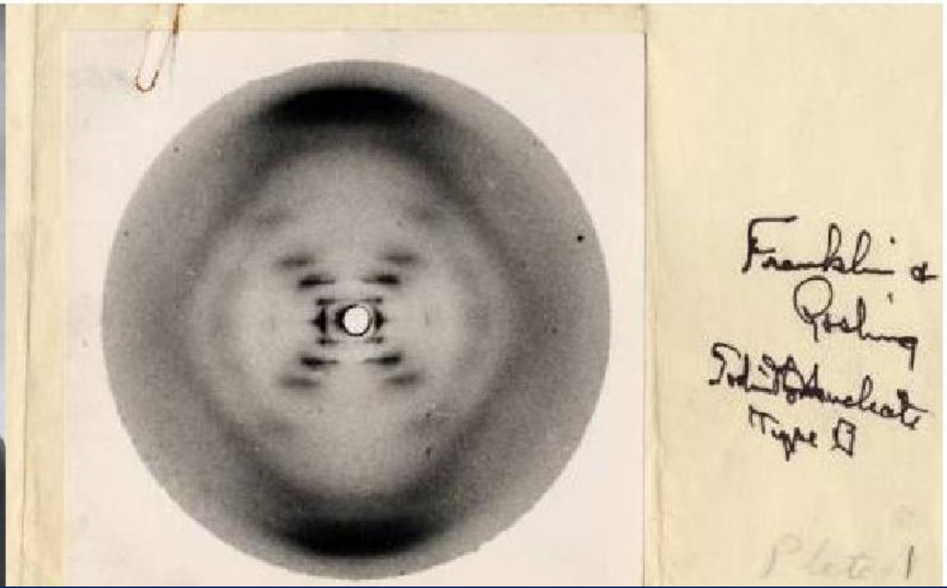
Figure 6.

The Railroading of Rosalind Franklin

- The quest to find discover the structure of DNA was indeed a race.
- Linus Pauling (CalTech), Maurice Wilkins and Rosalind Franklin (King's College London), and James Watson and Francis Crick (Cavendish Laboratory at Cambridge University) all sought to be the first to publish the elucidated structure.
- Pauling beat Watson and Crick in finding that certain proteins are helical in shape.
- He went on to erroneously propose that DNA was a three-stranded structure with the bases facing outwards.

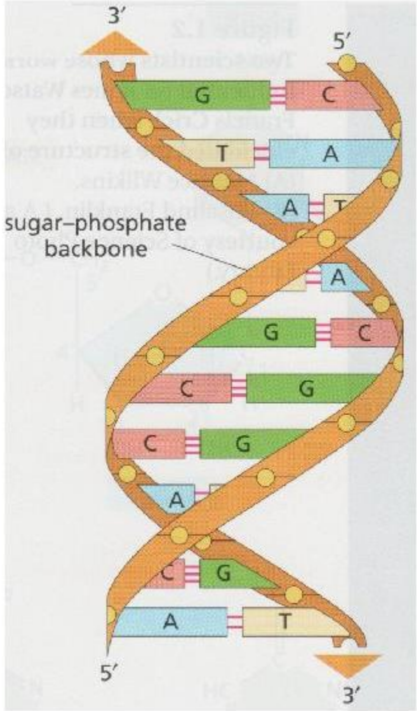
The Railroading of Rosalind Franklin

- James Watson was shown the famous X-ray crystallography "Photograph 51" without Franklin's permission or knowledge by Maurice Wilkins.
- The photograph revealed the antiparallel nature of the double helical strands and showed the bases to be in the center of the helix.
- Calculations made using the photograph helped Watson and Crick calculate size and structural parameters of the molecule.



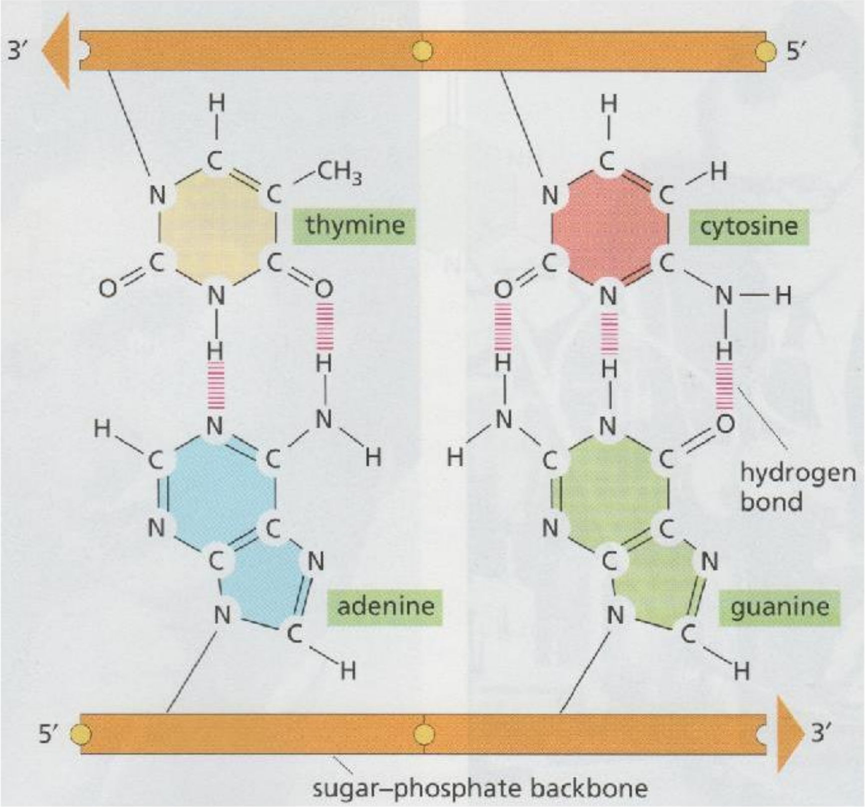
DNA Helix Structure

- Watson-Crick base-pairing describes how specific purines only pair with specific pyrimidines.
- Specifically, A pairs with T (double hydrogen bond) and C pairs with G (triple hydrogen bond).



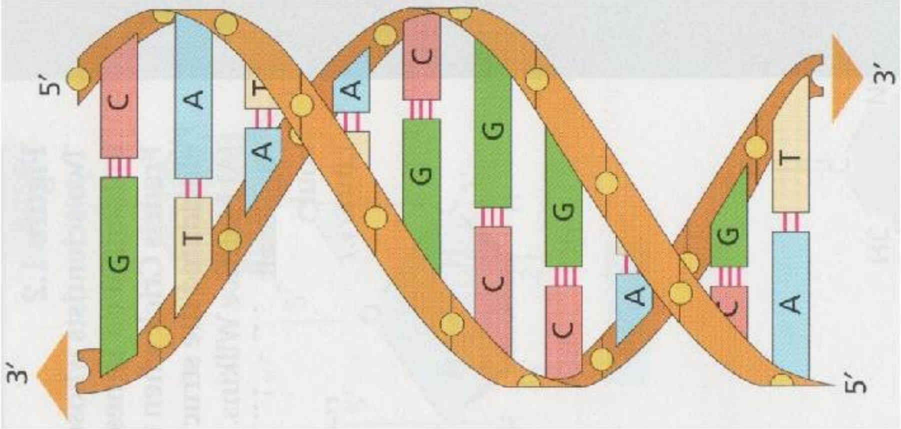
DNA Helix Structure

- Watson-Crick base-pairing describes how specific purines only pair with specific pyrimidines.
- Specifically, A pairs with T (double hydrogen bond) and C pairs with G (triple hydrogen bond).



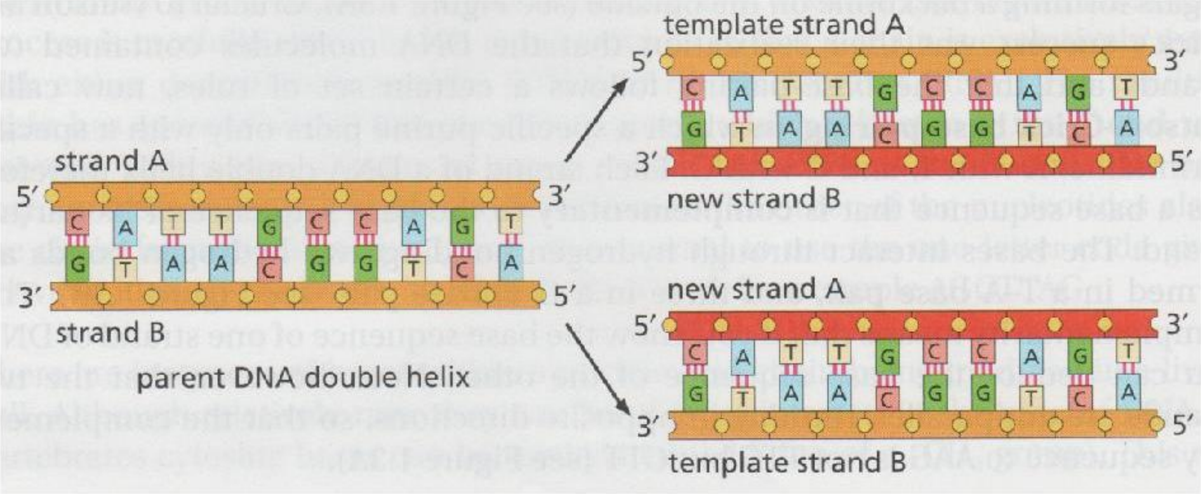
DNA Helix Structure

- Hence, if one knows the sequence of one strand of DNA, one can deduce the other strand.
- Complementary DNA strands are antiparallel, running in opposite directions.
- Hence, take care in that the complement of, for example, **TGA** is not **ACT**, but **TCA**.



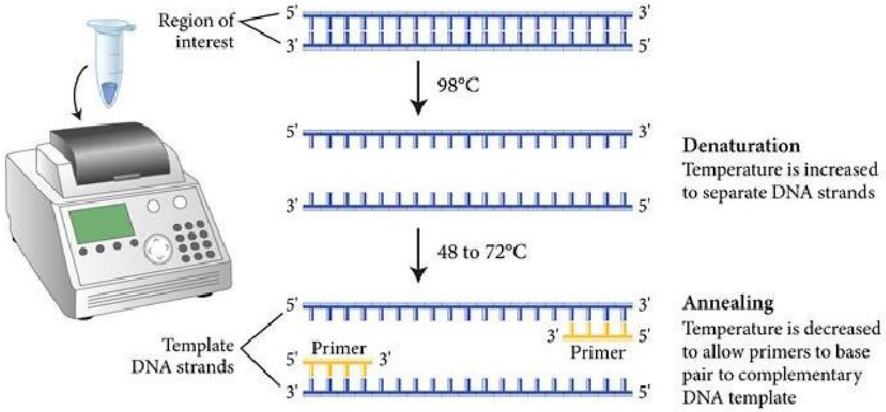
DNA Replication

- The hydrogen bonds between the purines and pyrimidines pairs are somewhat weak noncovalent bonds.
- Thus, their separation, for instance, during DNA replication prior to cellular division, is relatively easily accomplished.
- The DNA polymerase enzyme operates along both templates matching complementary nucleotides to their strands.



Polymerase Chain Reaction

- Absolutely indispensable method for DNA cloning/sequencing, gene analysis, etc.
- Creates an exponentially increasing number of DNA copies, through triggering and controlling DNA replication.
- Majority of methods are based on thermal cycling.

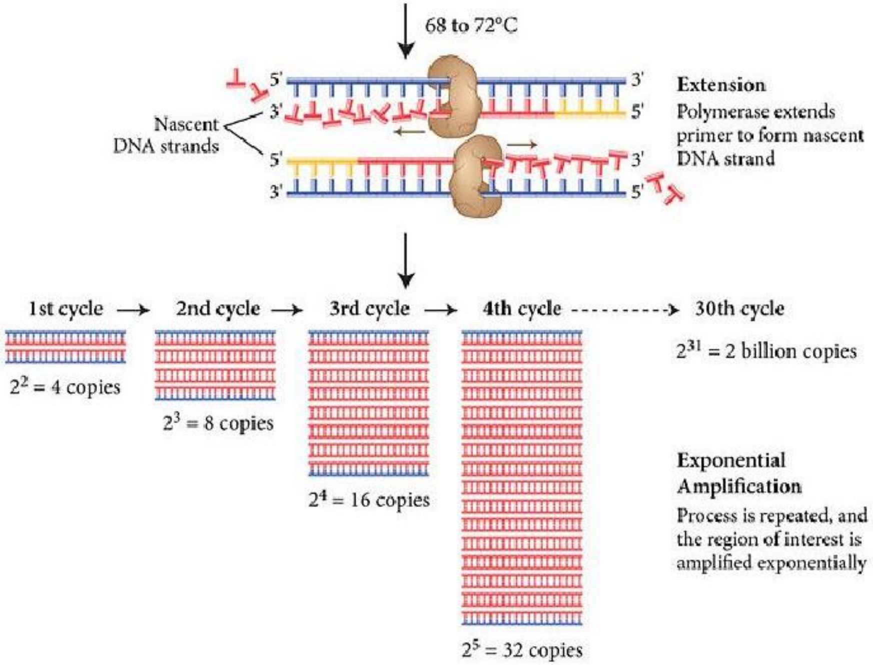


Denaturation
Temperature is increased to separate DNA strands

Annealing
Temperature is decreased to allow primers to base pair to complementary DNA template

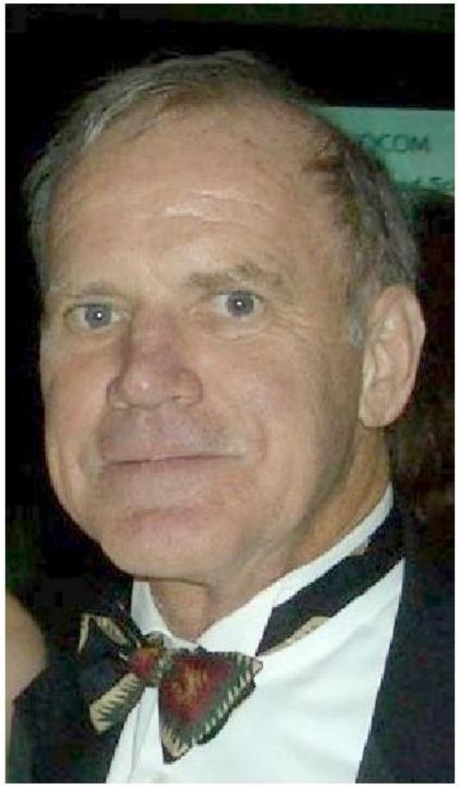
Polymerase Chain Reaction

- Introduced oglionucleotides or DNA primers are introduced and expanded on by heat-stable polymerases (Taq).
- The cycle is repeated numerous times for exponential yields of cloned DNA.


68 to 72°C
5'
3'
5'
3'
3'
5'
3'
5'
Nascent DNA strands
Extension
Polymerase extends primer to form nascent DNA strand
1st cycle
2nd cycle
3rd cycle
4th cycle
30th cycle
$2^2 = 4$ copies
$2^3 = 8$ copies
$2^4 = 16$ copies
$2^5 = 32$ copies
$2^{31} = 2$ billion copies
Exponential Amplification
Process is repeated, and the region of interest is amplified exponentially

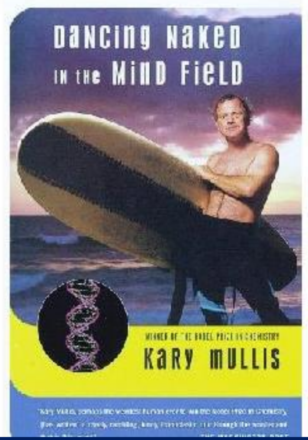
Almost More Fascinating than PCR: Kary Mullis

- PCR has vastly influenced the fields of biochemistry, molecular biology, genetics and genomics.
- Mullis won the Nobel Prize in 1993, ten years after first demonstrating PCR.
- Incredibly gifted and talented scientist...



Almost More Fascinating than PCR: Kary Mullis

- ...Almost completely nuts IRL.
- Denies link between HIV and AIDS.
- Seriously doubts if he would have invented PCR if not having had taken copious quantites of LSD while at Berkley.
- Fervent believer in Astrology.
- Once recounted coming in contact with an extraterrestrial irradiated racoon at his cabin in the woods of Northern California.
- Take-home: don't ascribe authority to realms outside of others' specialization!

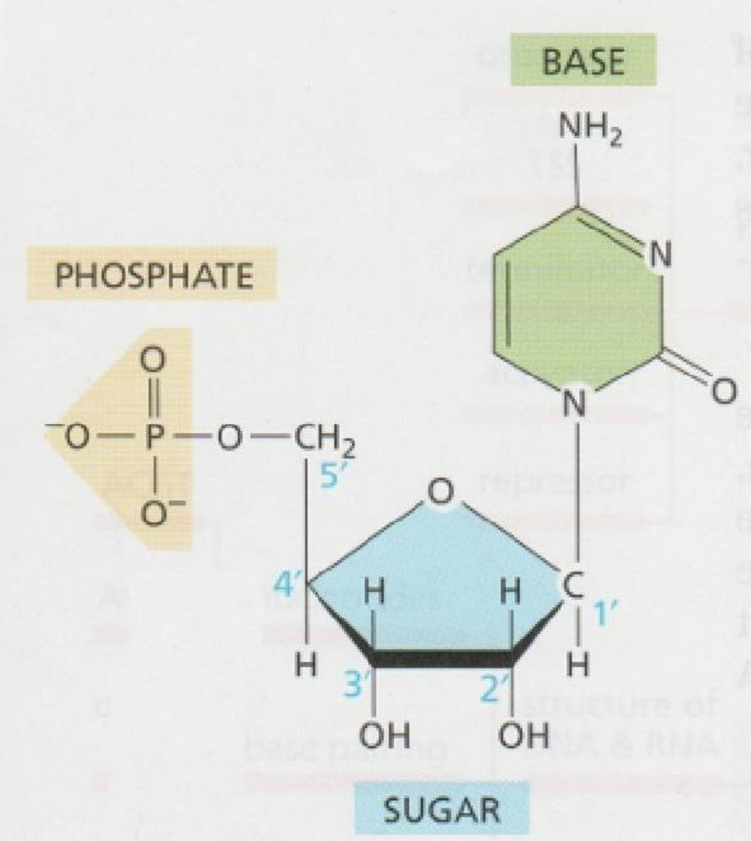


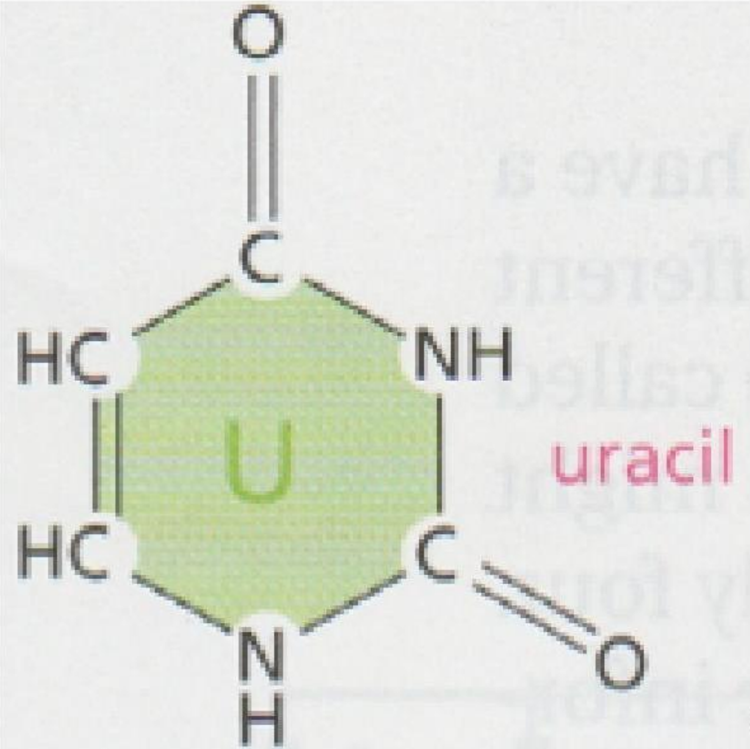
RNA Structure

- Unlike DNA, most RNA molecules are single stranded.
- Due to this property, RNA molecules are far more structurally flexible than DNA.
- RNA plays roles in the coding, decoding, regulation, and expression of genes.
- Its structural flexibility also allows it to act as a chemical catalyst in some cases, essentially forming a non-protein enzyme.

RNA Structure

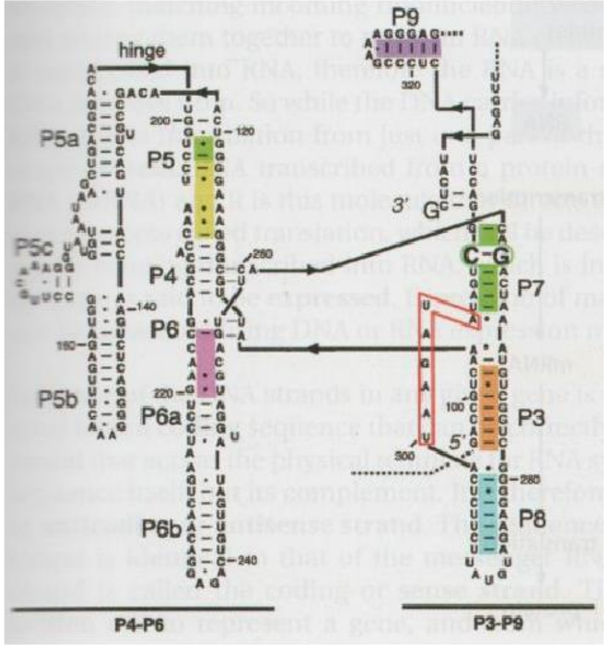
- A reminder: RNA differs structurally from DNA by:
  1. having a backbone made of ribonucleic acid and not deoxyribonucleic acid, as in DNA.
  2. uracil being the bonding complement of adenine and not thymine, as in DNA.





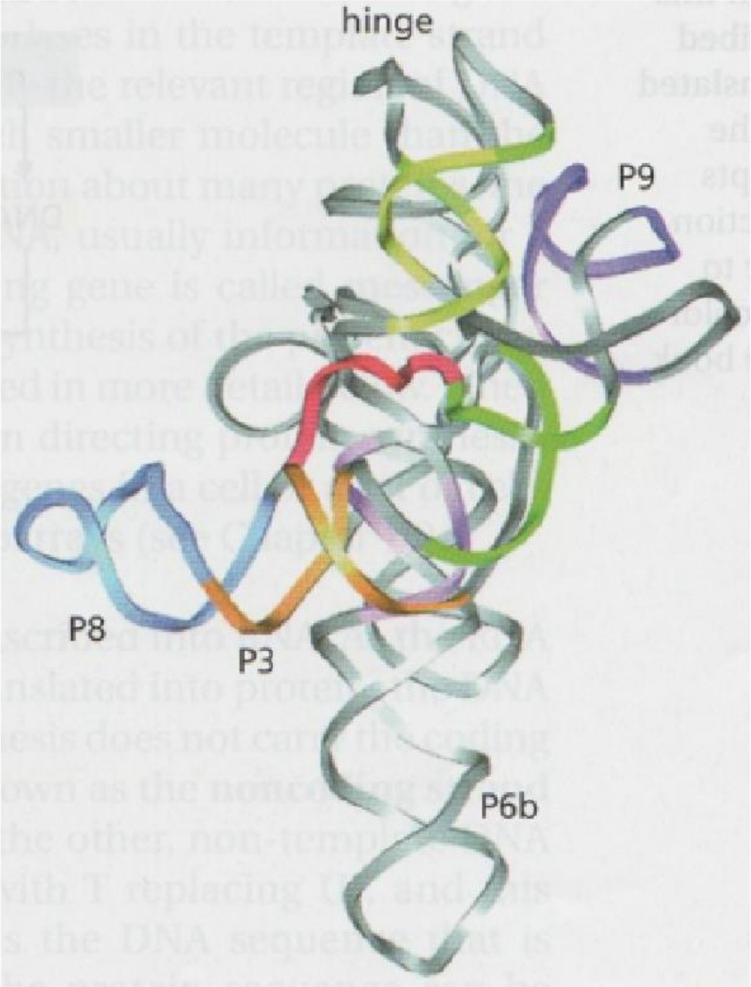
RNA Structure

- Being made up of free-floating bases allows for the same energetically favorable hydrogen bonding as occurs in DNA.
- Hence, complementary strands of RNA will “double back” on each other, forming double-helical loops.
- Often times, the pairings are not standard Watson-Crick pairings.

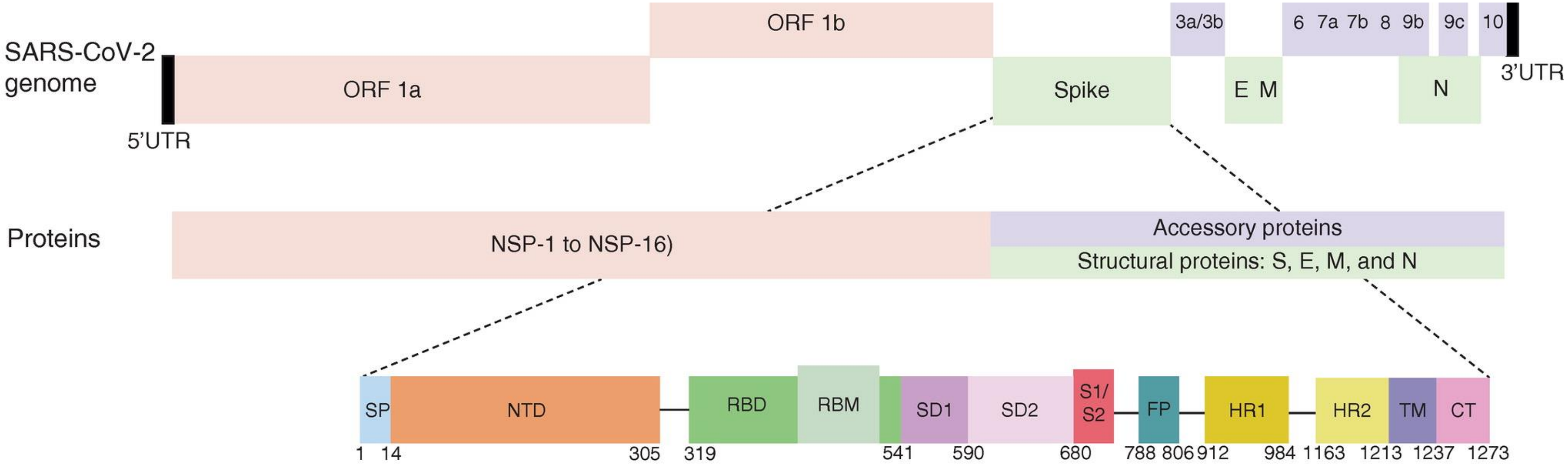


RNA Structure

- Often, not unlike DNA in the chromosome, RNA will fold even further on itself, forming tightly packed, rigid molecules.



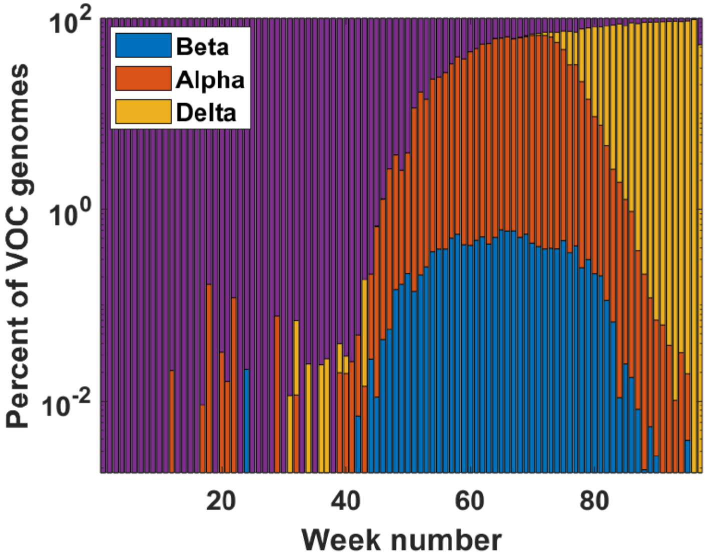
Structure of the SARS-Cov-2 genome
Single-stranded RNA virus (ca 30 kb in size)

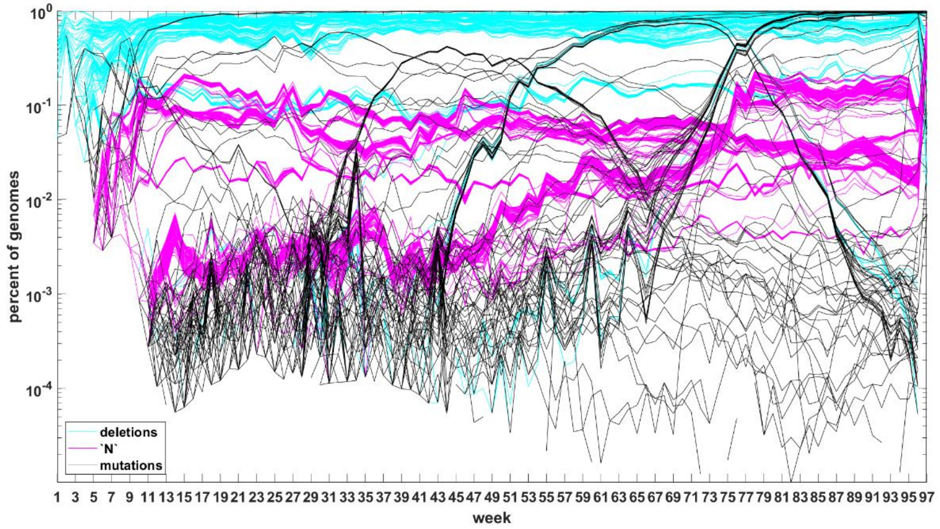


Interesting:

- Spike (S) protein
- Untranslated (possibly regulatory) regions 5 'UTR and 3' UTR

SARS-Cov-2 Genomic Evolution in 1 Slide



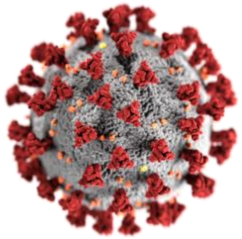


Can we get a glimpse of the evolutionary processes based on mutation trajectories at a given count of sites?

Recorded diversity of viral genomes (tip of the iceberg?)

Hidden diversity of viral genomes

New selectively advantageous variants emerge



Kurpas et al. (2022) Viruses 14(11): 2375

Underlying processes

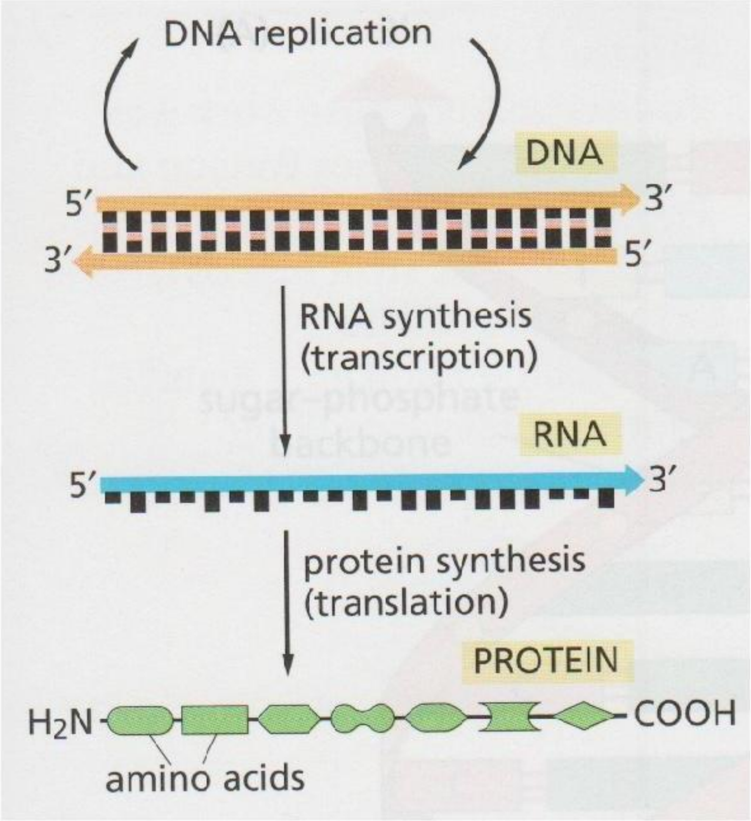
- Mutation
- Drift
- Selection
- Recombination (?)

The "Central Dogma of Molecular Biology"

- There is a flow of genetic information:

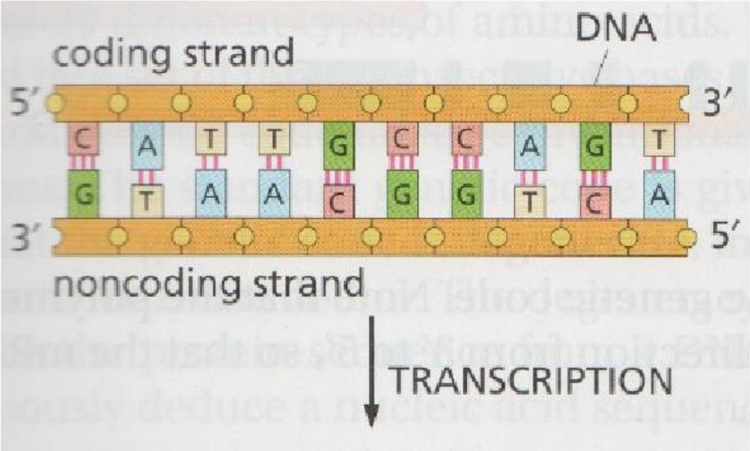
$$\text{DNA} \Rightarrow \text{RNA (mRNA)} \Rightarrow \text{Proteins}$$

- DNA acts as the store of information, RNA as the translator of information into proteins, which are the actuators of life.



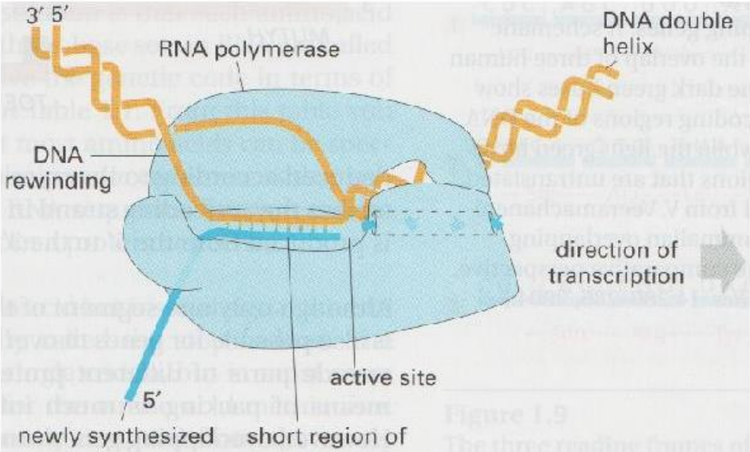
Transcription

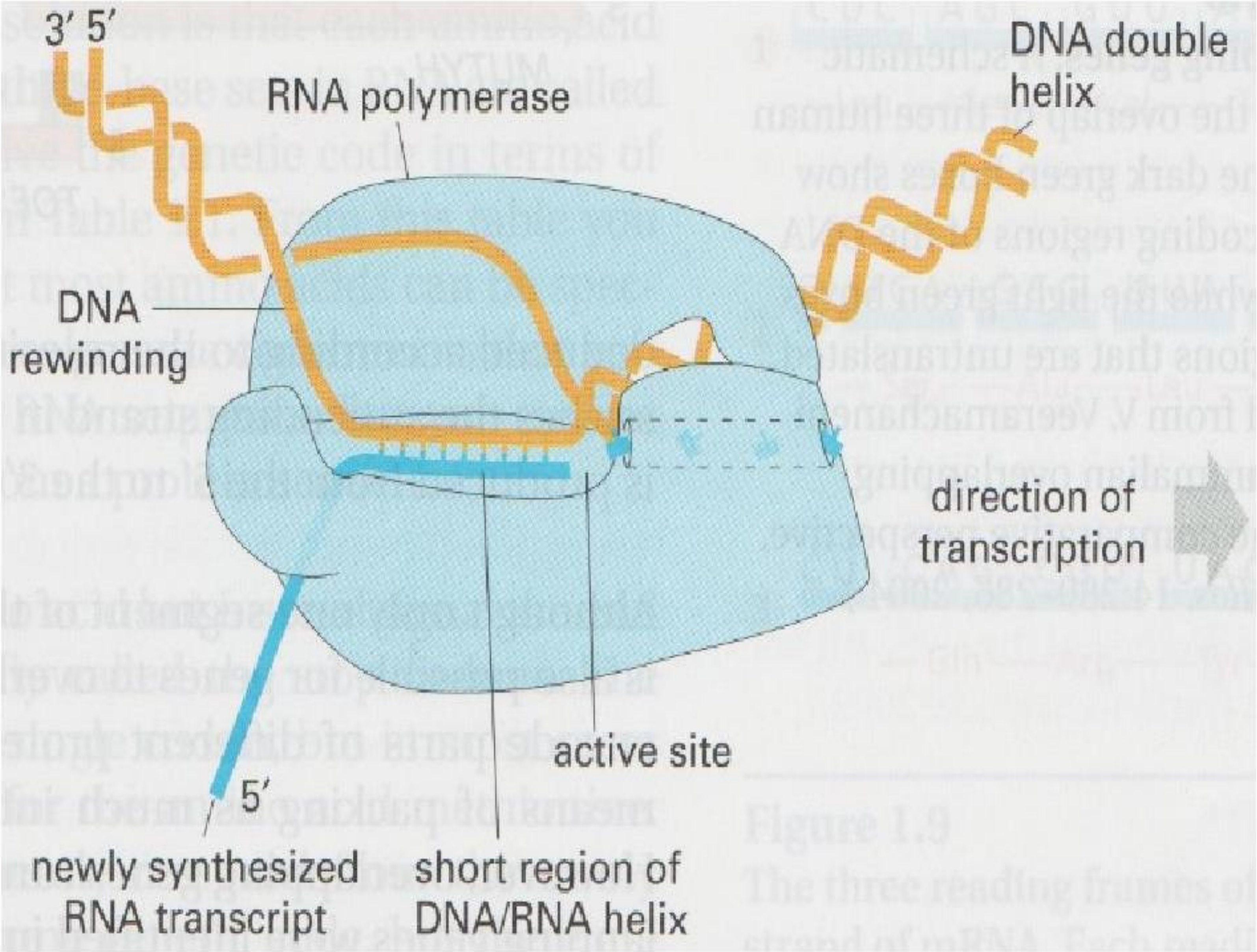
- Transcription is the process by which the information store (DNA) is expressed in the form of the translator (messenger RNA).
- In a fashion roughly akin to DNA polymerase, RNA polymerase unwinds the DNA.
- By operating on the noncoding/anticoding/antisense strand (3' to 5'), it transcribes an mRNA strand identical to the coding/sense DNA strand (5' to 3').
- As mentioned previously, uracil (U) replaces thymine (T) in RNA.



Transcription

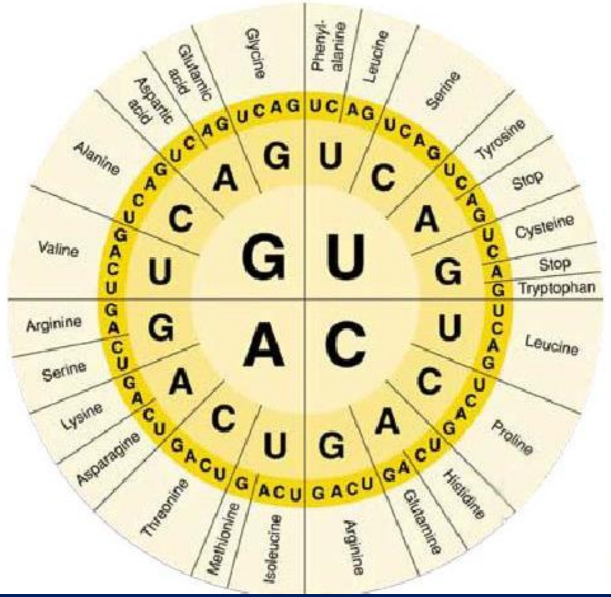
- Transcription is the process by which the information store (DNA) is expressed in the form of the translator (messenger RNA).
- In a fashion roughly akin to DNA polymerase, RNA polymerase unwinds the DNA.
- By operating on the noncoding/anticoding/antisense strand (3' to 5'), it transcribes an mRNA strand identical to the coding/sense DNA strand (5' to 3').
- As mentioned previously, uracil (U) replaces thymine (T) in RNA.




3′ 5′
RNA polymerase
DNA double helix
DNA rewinding
direction of transcription
active site
5′
newly synthesized RNA transcript
short region of DNA/RNA helix

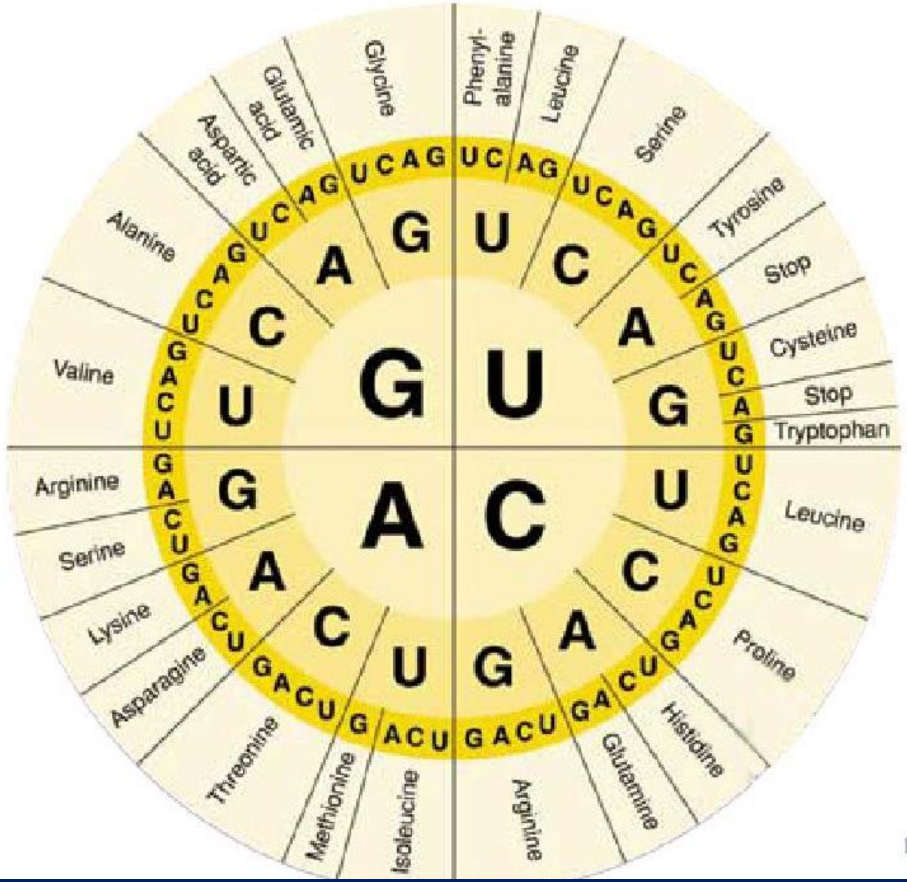
Translation

- Messenger RNA (mRNA) is then translated into a sequence of amino acids by ribosomes with the help of ribosomal RNA and tranfer RNA, rRNA and tRNA, respectively.
- Three-base sets of RNA make up amino acid forming codons.
- Translation is often signaled by the AUG codon, coding for Met (methionine).
- UAA, UAG and UGA are stop codons, signalling the end of translation.



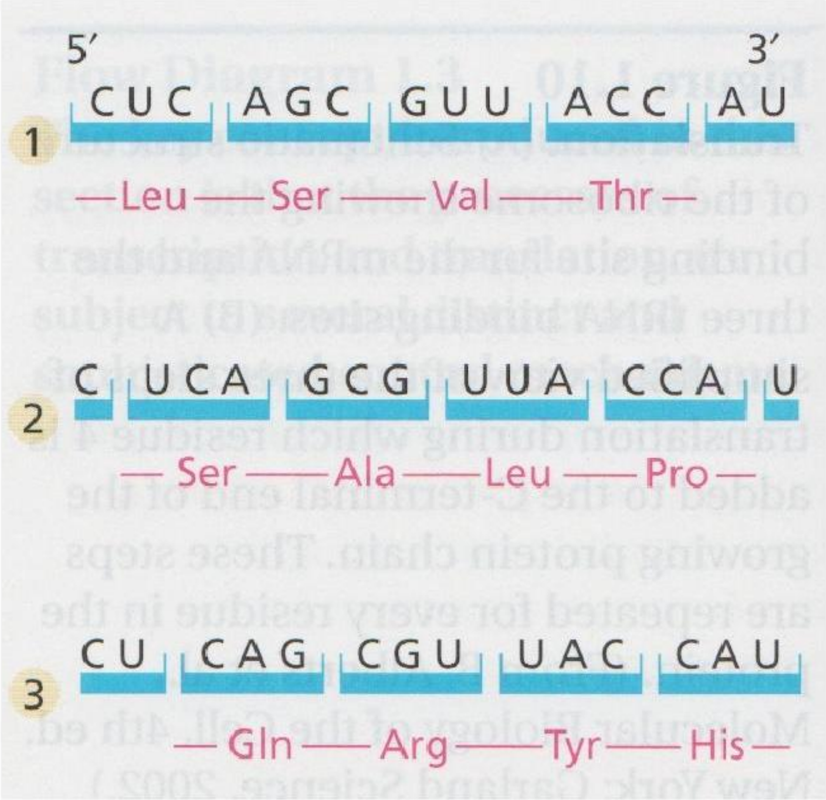
Translation

- Of importance: the genetic coding for animo acids is degenerate.
- That is, most amino acids can be coded for by more than one codon.



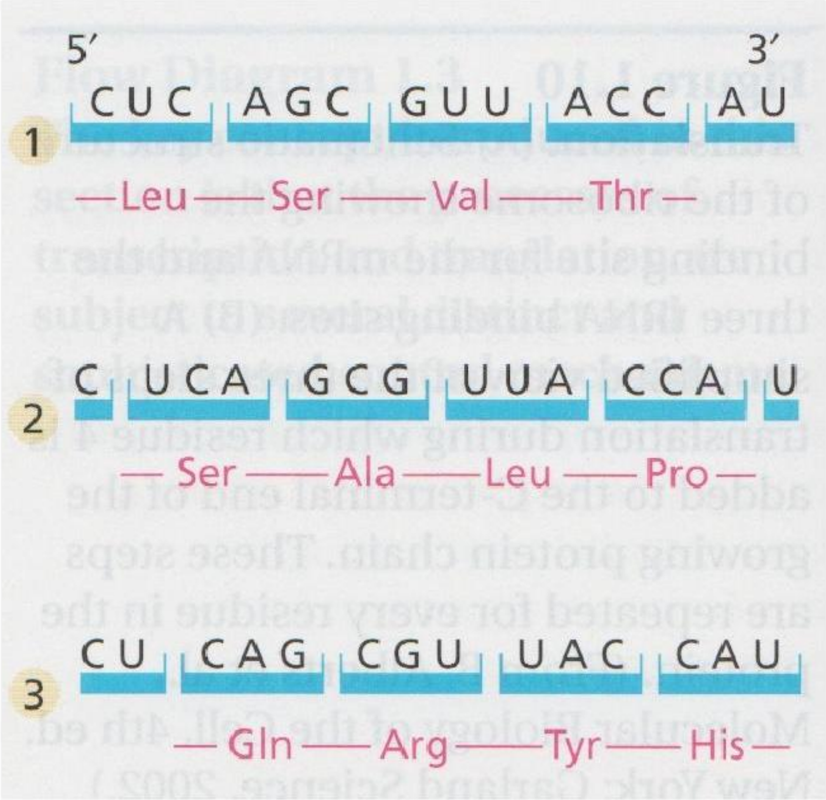
mRNA Reading Frames

- Upon assessing an mRNA sequence, where one begins is obviously of utmost importance.
- Due to codons being of length three, each sequence of mRNA has three possible reading frames.



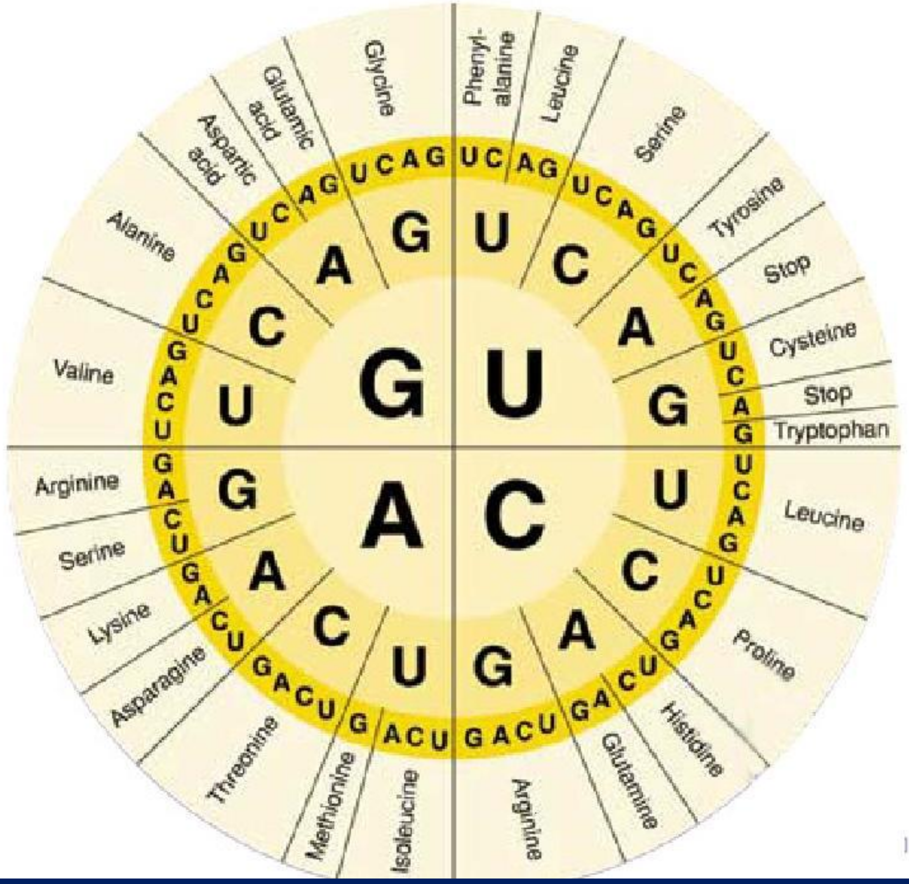
mRNA Reading Frames

- Scientists predicting protein-coding DNA sequences would use known control signals to determine the start of the frame.
- Also helpful: most proteins are at least 100 animo acids in length.



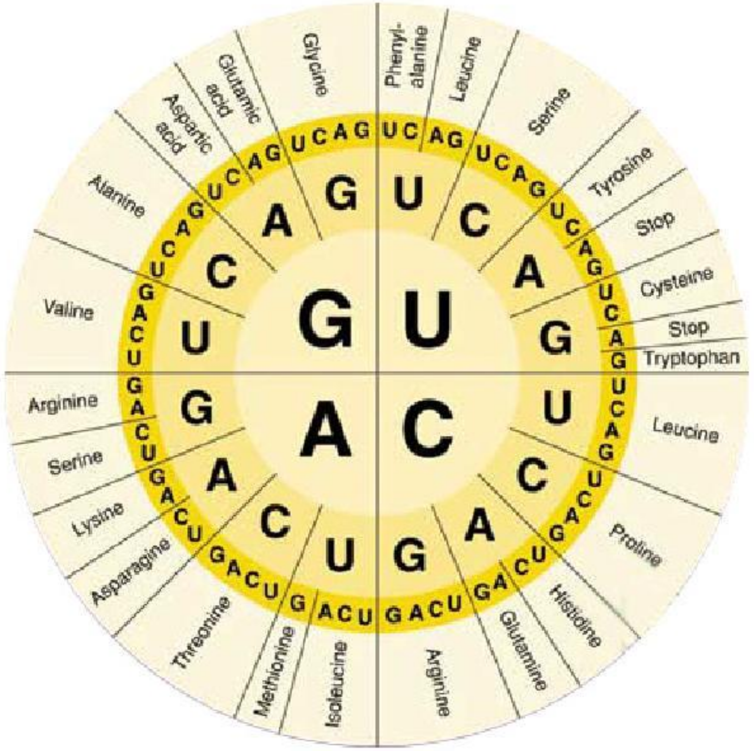
Mutation Effects on Translation

- Of importance: the genetic coding for animo acids is degenerate.
- That is, most amino acids can be coded for by more than one codon.



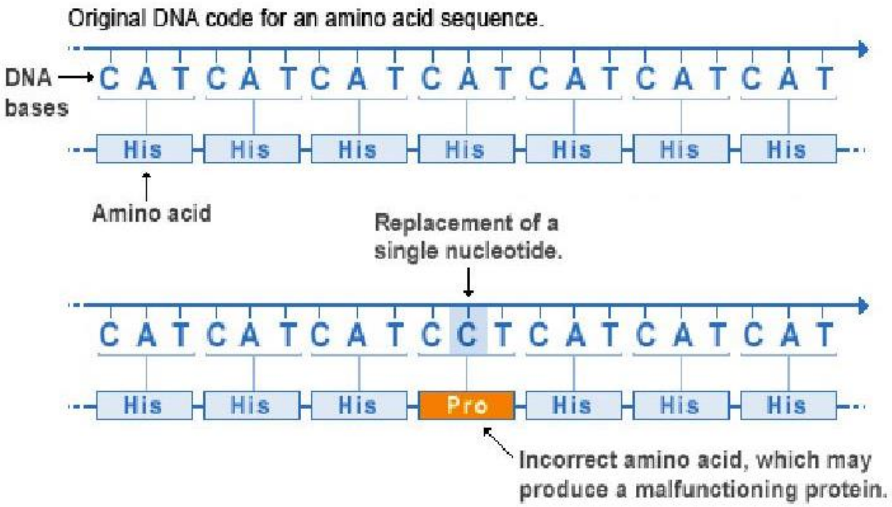
Mutation Effects on Translation

- Thus, it is possible for a DNA mutation will not result in the change of a translated protein.
- When occuring in a gene coding for a protein, this is known as a synonymous or silent substitution.
- For example, GCA or GCC will both code for alanine.



Missense Mutations

- Mutations which through one DNA base pair change causes the substitution of one animo acid for another are know as missense mutations.
- Also known as nonsynonymous mutation.
- Such amino acid change can be
  - conservative (similar in physiochemical properties),
  - semi-conservative (negative to positively charged amino acid, or vice versa), or
  - radical (vastly different amino acid).

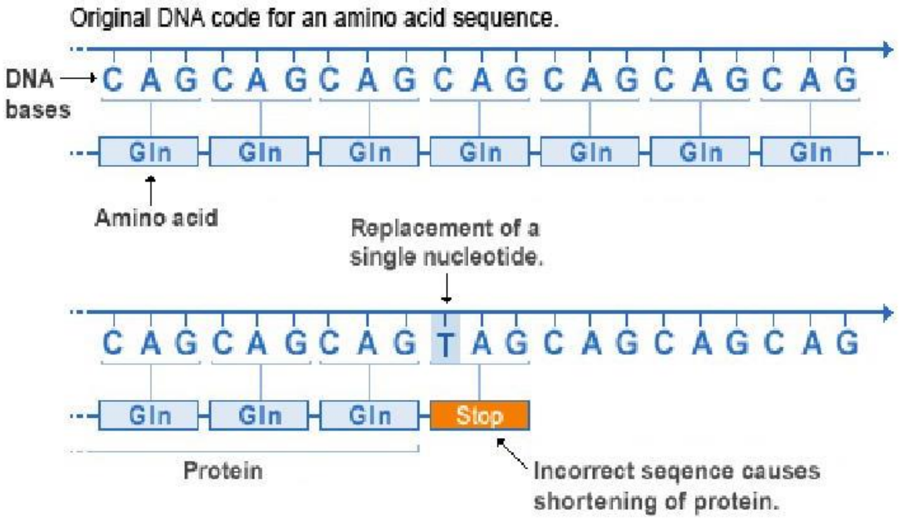


Original DNA code for an amino acid sequence.

DNA bases
C A T C A T C A T C A T C A T C A T C A T
His His His His His His His
Amino acid
Replacement of a single nucleotide.
C A T C A T C A T C C T C A T C A T C A T
His His His Pro His His His
Incorrect amino acid, which may produce a malfunctioning protein.

Nonsense Mutations

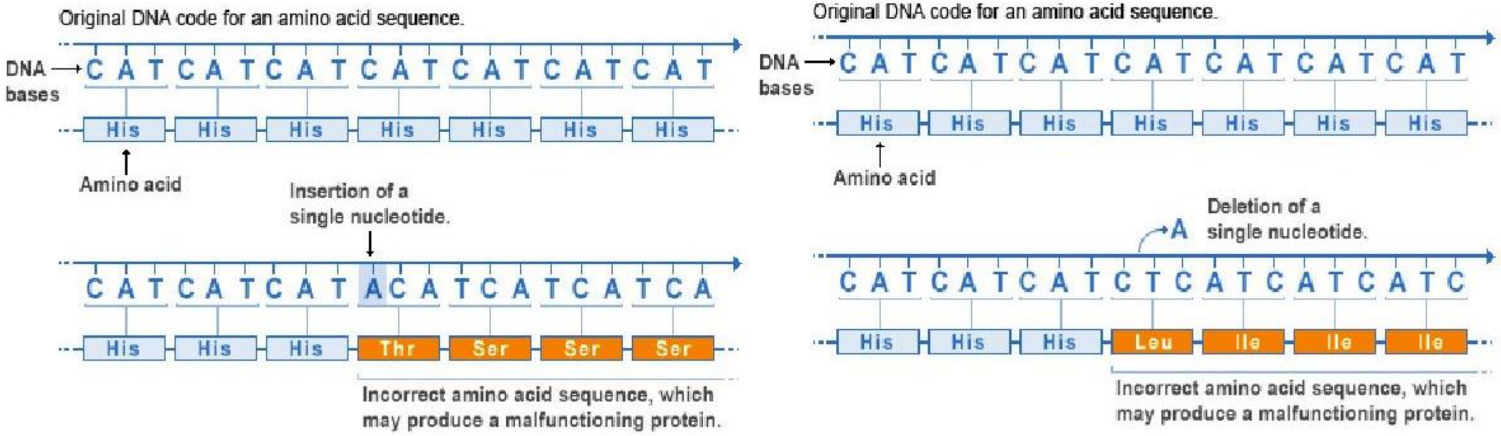
- Nonsense mutations are those which trigger the erroneous shortening of a protein.
- In the example, the **TAG** DNA would be transcribed into the **UAG** RNA stopping codon.



Original DNA code for an amino acid sequence.
DNA bases
C A G C A G C A G C A G C A G C A G C A G
Gln Gln Gln Gln Gln Gln Gln
Amino acid
Replacement of a single nucleotide.
C A G C A G C A G T A G C A G C A G C A G
Gln Gln Gln Stop
Protein
Incorrect seqence causes shortening of protein.

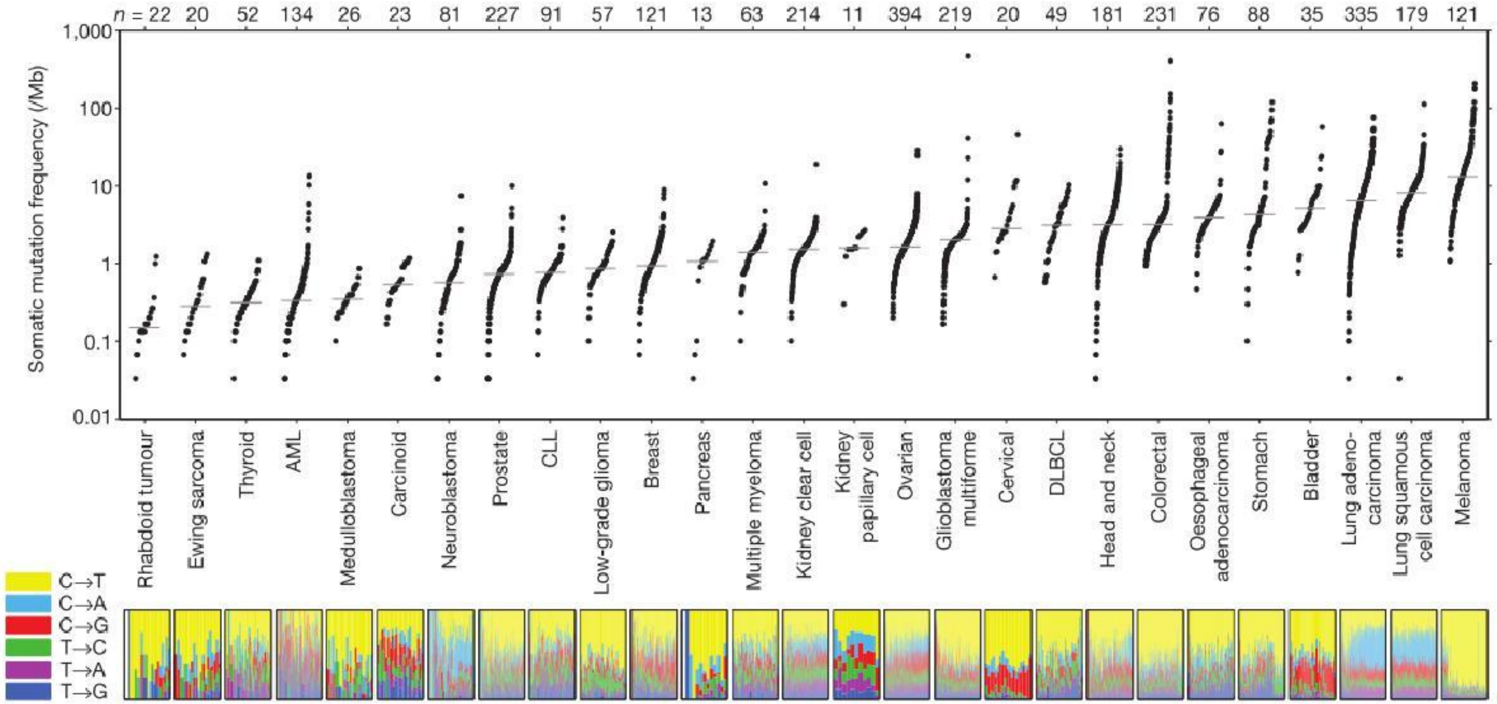
Insertion and Deletion Mutations

- Insertion (deletion) mutations are those which through insertion (deletion) of a base pair, cause the incorrect formation of a protein.
- Collectively shortened as InDels in bioinformatic nomenclature.
- NIH link for more types of gene mutations.


Original DNA code for an amino acid sequence.
DNA bases
C A T C A T C A T C A T C A T C A T C A T
His His His His His His His
Amino acid
Insertion of a single nucleotide.
C A T C A T C A T A C A T C A T C A T C A
His His His Thr Ser Ser Ser
Incorrect amino acid sequence, which may produce a malfunctioning protein.
Original DNA code for an amino acid sequence.
DNA bases
C A T C A T C A T C A T C A T C A T C A T
His His His His His His His
Amino acid
A
Deletion of a single nucleotide.
C A T C A T C A T C T C A T C A T C A T C
His His His Leu Ile Ile Ile
Incorrect amino acid sequence, which may produce a malfunctioning protein.

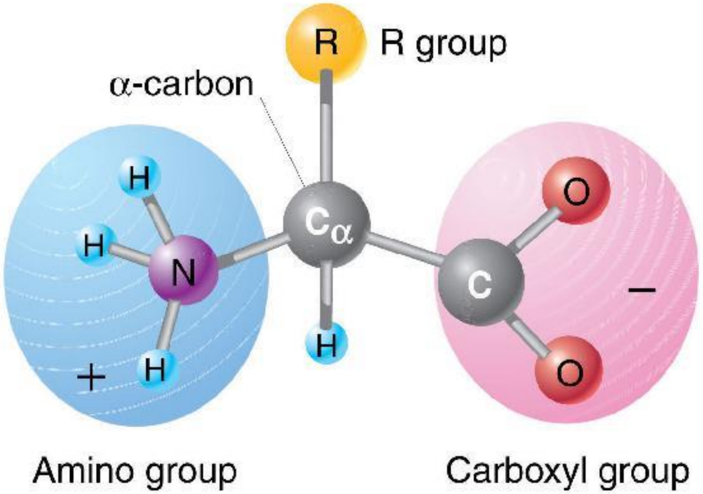
Different Pathologies Associated with Different Types and Numbers of Mutations

- There is much difference between the number and type of mutations associated with different cancer types!



Protein Structure - Primary

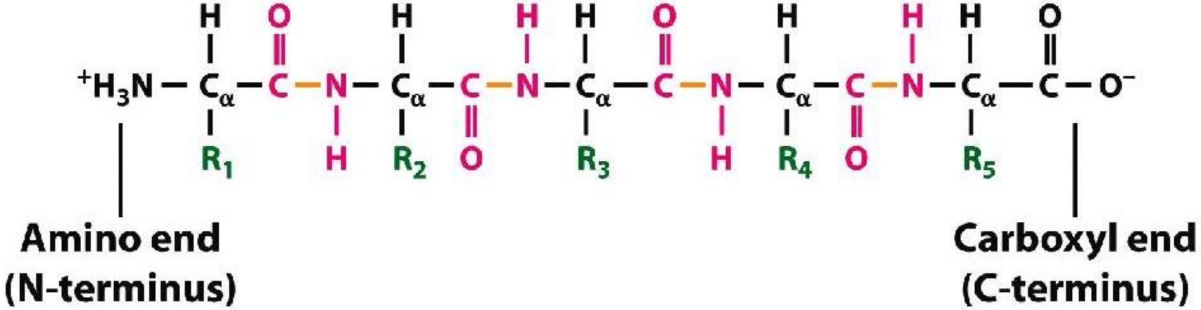
- As stressed before, 20 different types of amino acids form the building blocks of proteins.
- Amino acids are comprised solely of carbon, nitrogen, oxygen and hydrogen atoms (besides **Cys**tenine and **Met**hionine, which also contain sulphur).
- The amino acid structure is composed of a main chain (aminyl group and carboxyl group) and a side chain (varying R-group).
- The side chain is attached to the main chain by the α-carbon C_α.



Protein Structure - Primary

- Amino acids are chained together through peptide bonds, forming polypeptide chains.
- The polypeptide structure has a recative amino group at one end (N-terminus) and a reactive carboxyl group at the other end (C-terminus).
- Translation of mRNA synthesizes the N terminus first, hence the protein sequence writing convention.

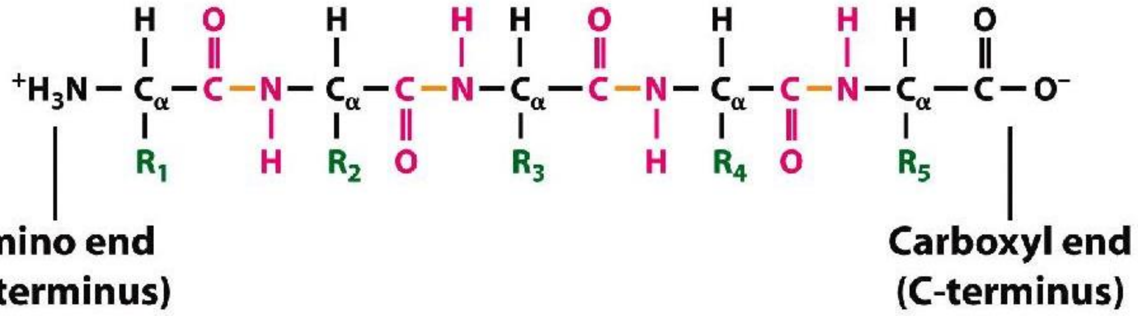
N- MYCATISEATINGFISHANDMEATANDWATER -**C**



Protein Structure - Primary

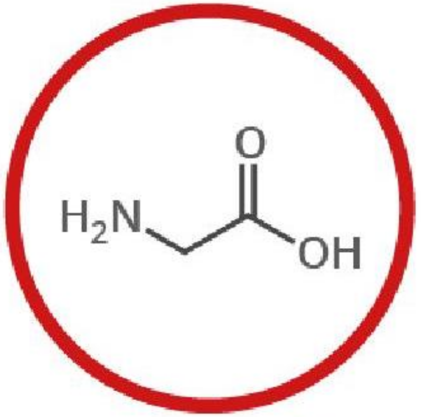
- Amino acids are chained together through peptide bonds, forming polypeptide chains.
- The polypeptide structure has a recative amino group at one end (N-terminus) and a reactive carboxyl group at the other end (C-terminus).
- Translation of mRNA synthesizes the N terminus first, hence the protein sequence writing convention.

N- MYCATISEATINGFISHANDMEATANDWATER **-C**



Protein Structure - Primary

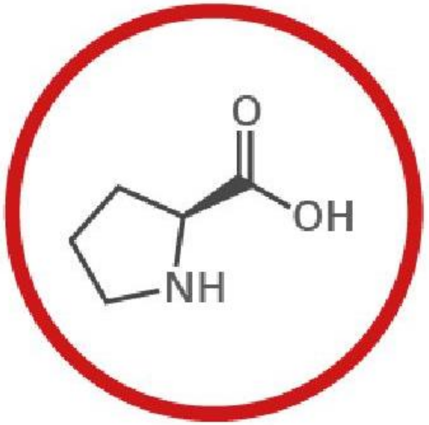
- Amino acids have very different physiochemical properties (acidic, basic, uncharged polar and nonpolar).



GLYCINE G

Gly

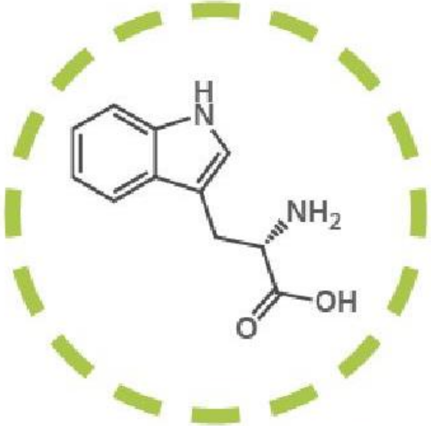
GGT, GGC, GGA, GGG



PROLINE P

Pro

CCT, CCC, CCA, CCG



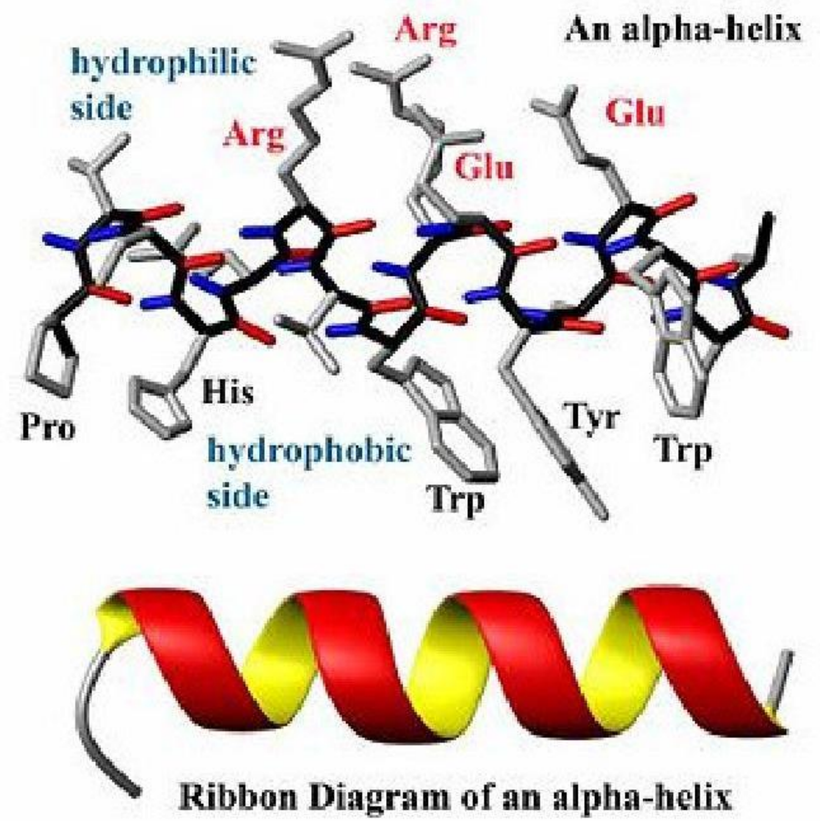
TRYPTOPHAN W

Trp

TGG

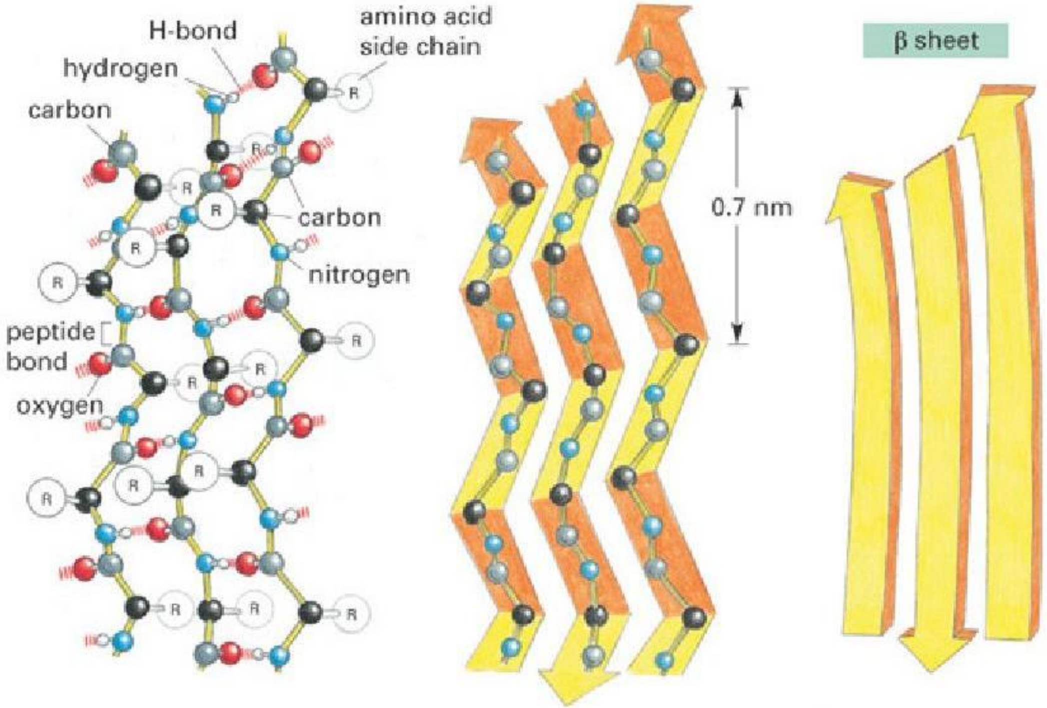
Protein Structure - Secondary

- Repetitive angles between the bonds flanking the peptide bonds (ϕ, ψ angles) often cause amino acid chains to form a right-handed helix.
- The α-helix structure repeats itself every 0.54 nm.



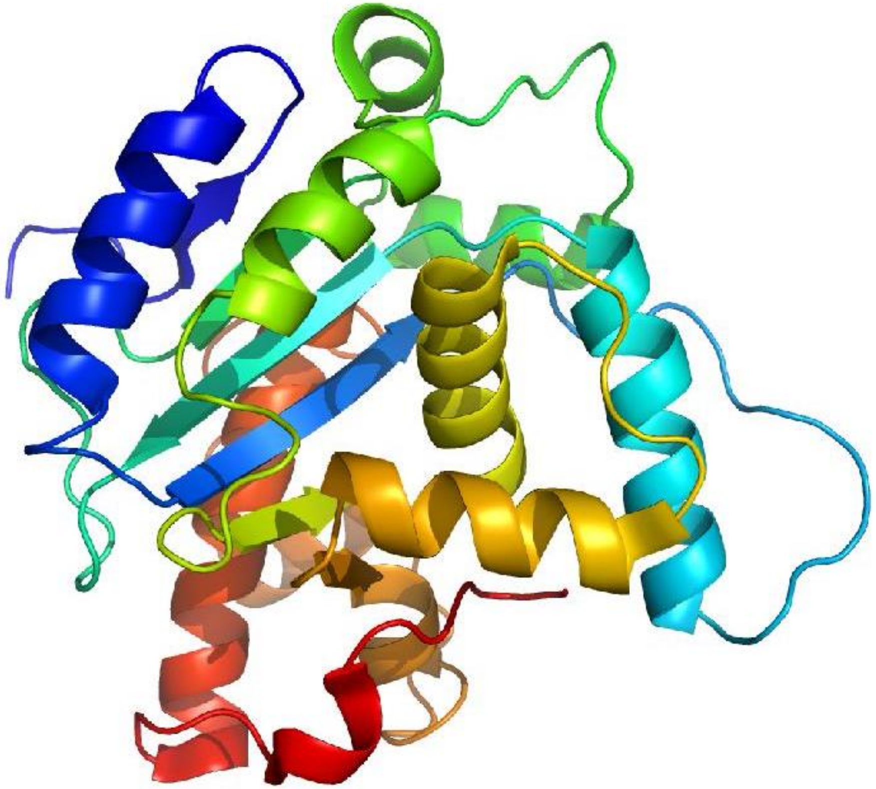
Protein Structure - Secondary

- Extended chains of amino acids favorable to hydrogen bonding form β-chains.
- The β-strand arrows denote the direction from the N- to the C-terminus.
- Sets of β-chains flanking one another bond and are called β-sheets.



Protein Structure - Tertiary

- The protein tertiary structure is concerned with the manner in which proteins are folded back upon themselves.
- Determing this structure is a significant challenge in Bioinformatics!



Holy Graal of Computational Biology

The protein folding problem

= deducing the tertiary structure of protein based on AA sequence

“Cracked” ca. 2020 – 2021 by Alpha Fold AI code

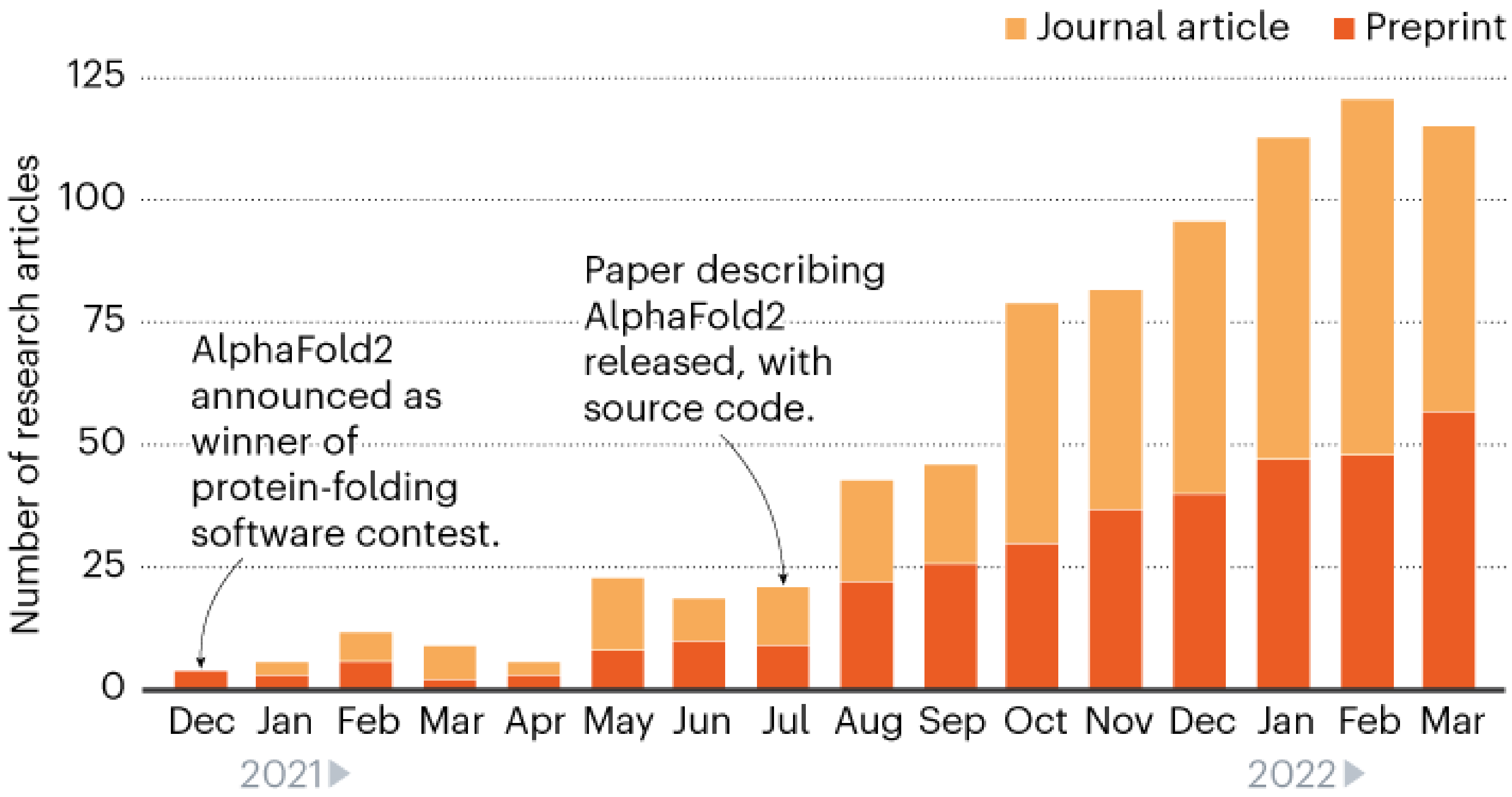
https://alphafold.ebi.ac.uk/

Based on “rules”, “learned” from PDB structures

See comments in https://doi.org/10.1073/pnas.22144231

ALPHAFOLD MANIA

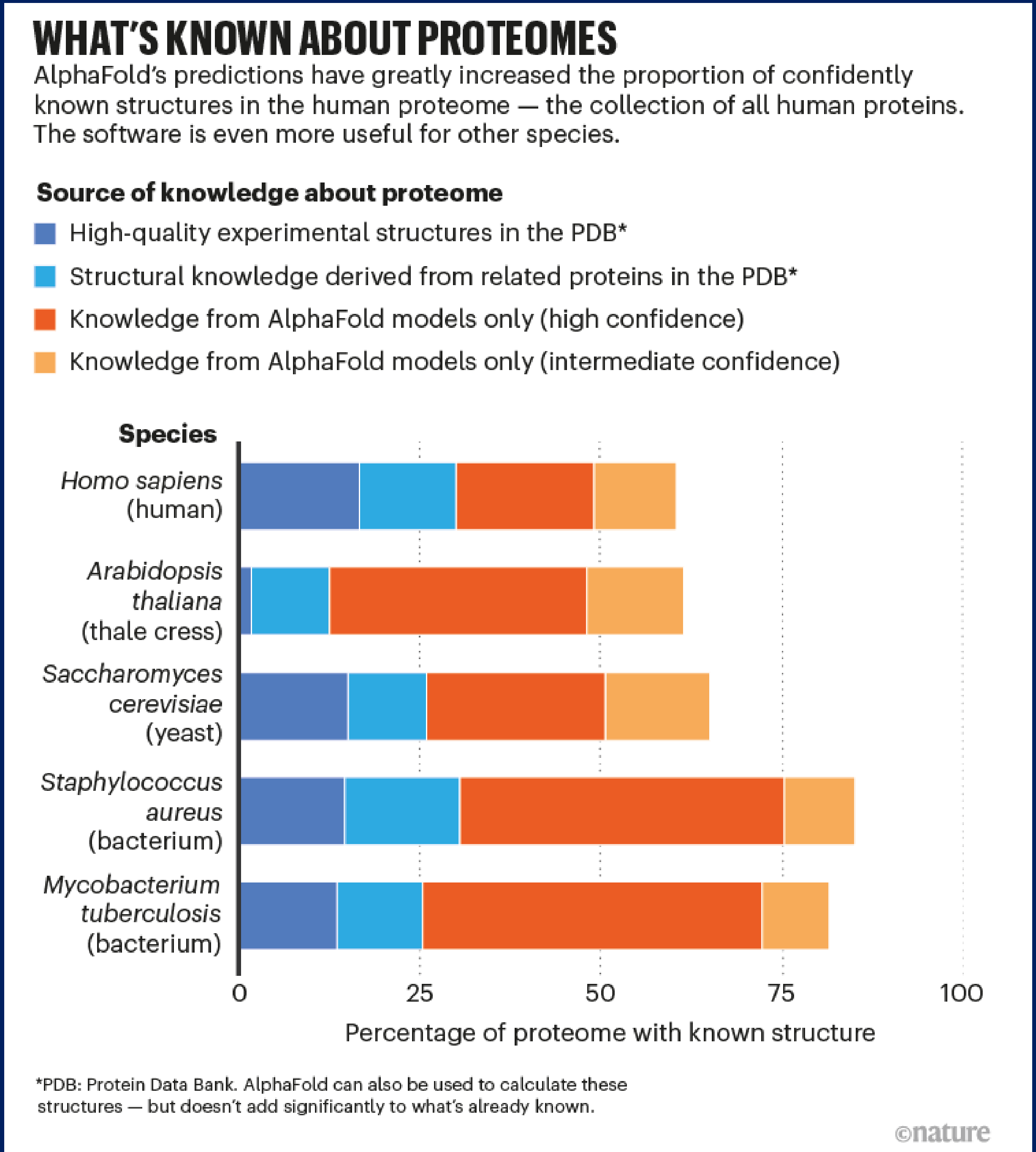
The number of research papers and preprints citing the AlphaFold2 AI software has shot up since its source code was released in July 2021*.



AlphaFold2 announced as winner of protein-folding software contest.

Paper describing AlphaFold2 released, with source code.

*_Nature_ analysis using Dimensions database; removing duplicate preprints and papers/R. Van Noorden, E. Callaway.

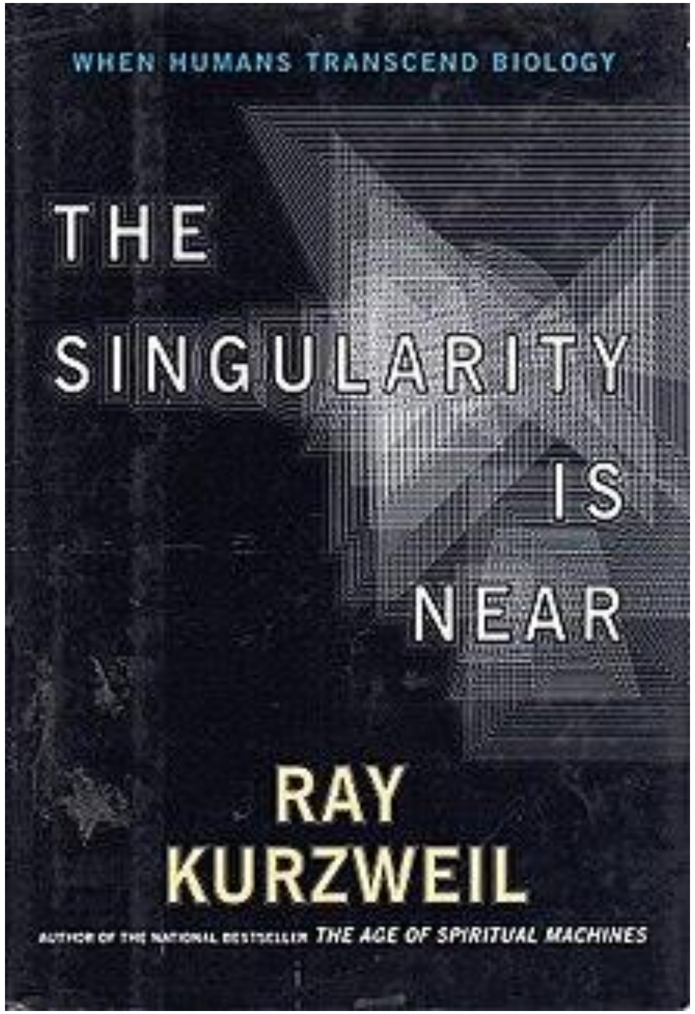


WHAT'S KNOWN ABOUT PROTEOMES

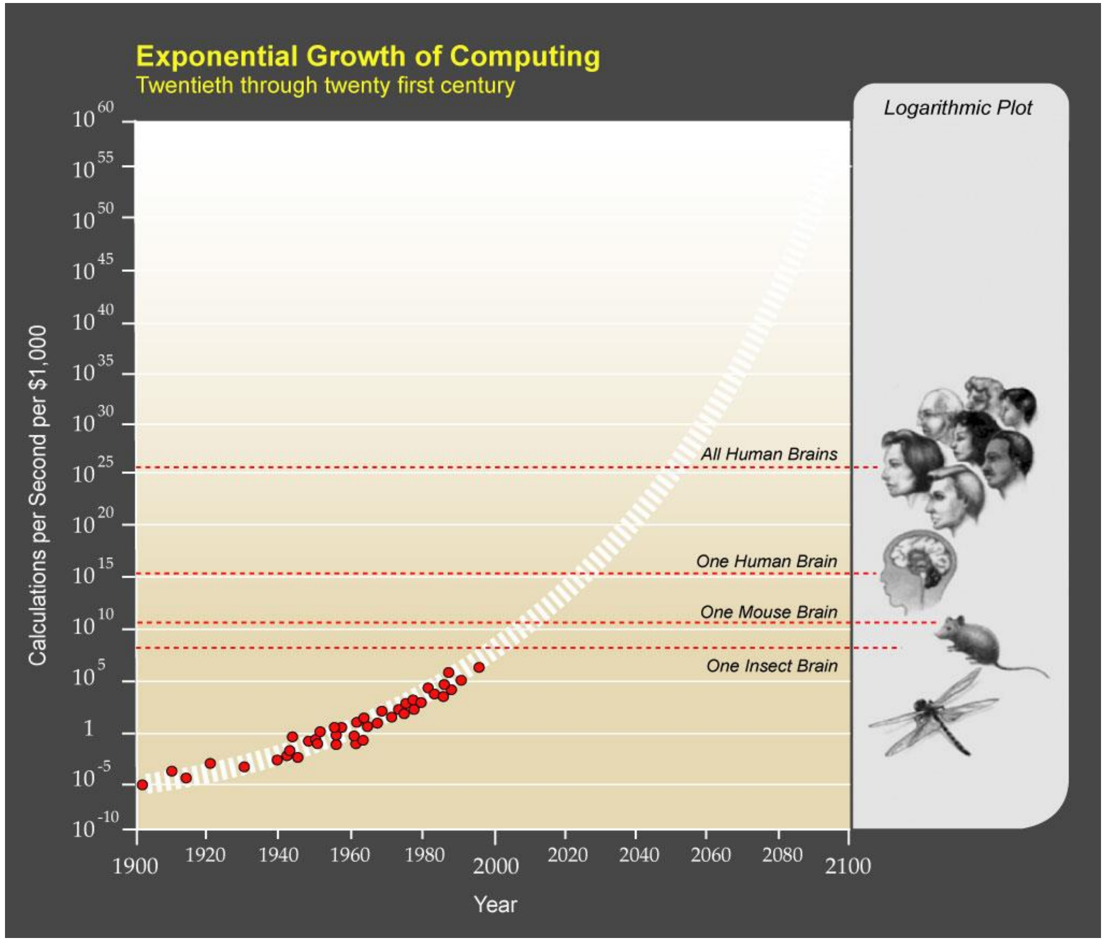
AlphaFold's predictions have greatly increased the proportion of confidently known structures in the human proteome — the collection of all human proteins. The software is even more useful for other species.

*PDB: Protein Data Bank. AlphaFold can also be used to calculate these structures — but doesn't add significantly to what's already known.

©nature

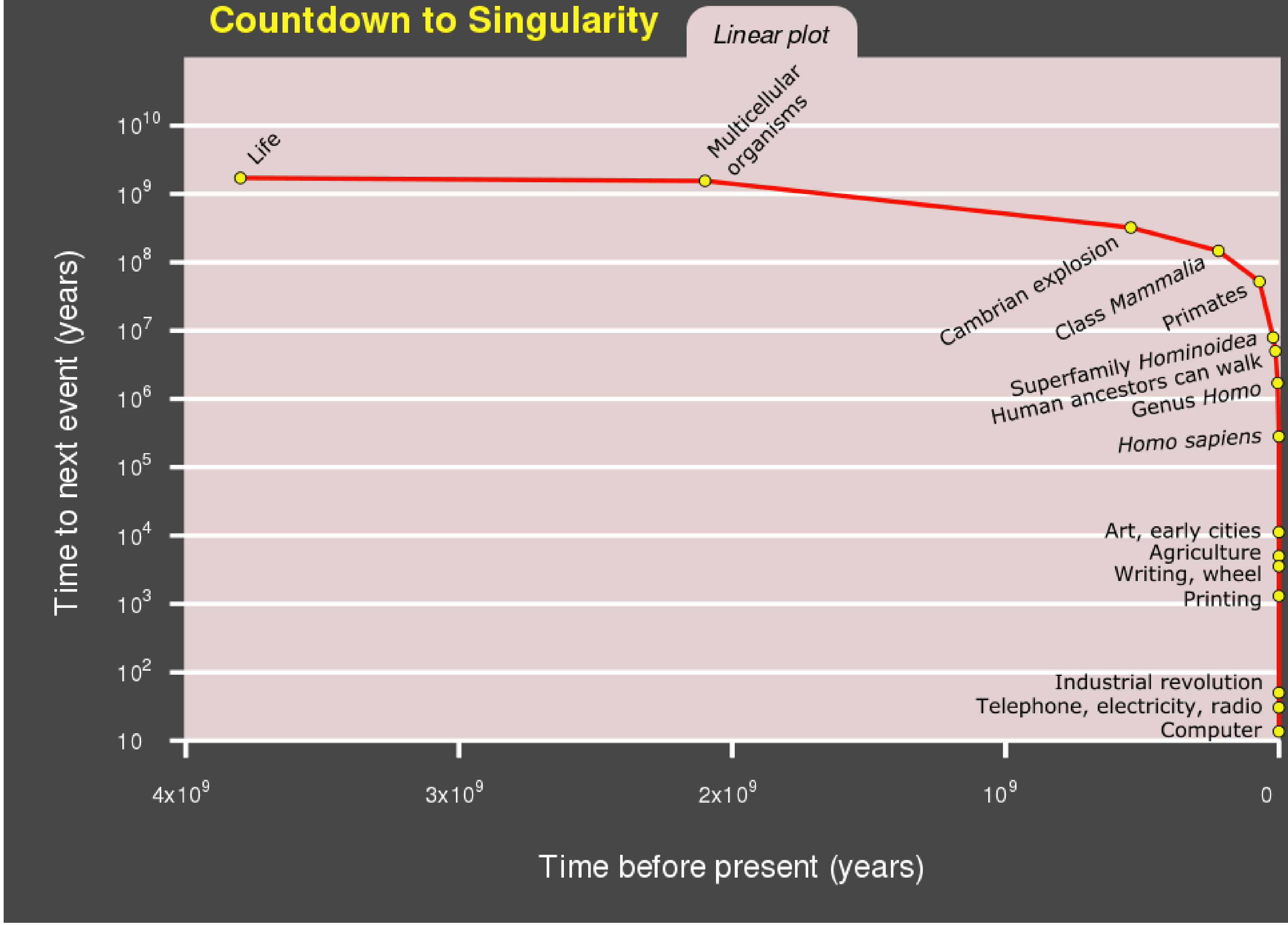


WHEN HUMANS TRANSCEND BIOLOGY
THE SINGULARITY IS NEAR
RAY KURZWEIL
AUTHOR OF THE NATIONAL BESTSELLER THE AGE OF SPIRITUAL MACHINES



Exponential Growth of Computing
Twentieth through twenty first century
Logarithmic Plot
Calculations per Second per $1,000
10^60
10^55
10^50
10^45
10^40
10^35
10^30
10^25
10^20
10^15
10^10
10^5
1
10^-5
10^-10
All Human Brains
One Human Brain
One Mouse Brain
One Insect Brain
1900
1920
1940
1960
1980
2000
2020
2040
2060
2080
2100
Year

Countdown to Singularity


Linear plot
Life
Multicellular organisms
Cambrian explosion
Class Mammalia
Primates
Superfamily Hominoidea
Human ancestors can walk
Genus Homo
Homo sapiens
Art, early cities
Agriculture
Writing, wheel
Printing
Industrial revolution
Telephone, electricity, radio
Computer
Time to next event (years)
10^10
10^9
10^8
10^7
10^6
10^5
10^4
10^3
10^2
4x10^9
3x10^9
2x10^9
10^9
Time before present (years)