Genetics and molecular biology

Bioinformatics · Universitat Politècnica de Catalunya

Alexandre Perera-Lluna

Challenges in Bioinformatics



Some challenges in Bio-informatics

- Bioinfo is a truly multidisciplinary science.
- Different researchers are trained with different backgrounds.
- They all speak different languages.
- They all have different goals.
- They need to understand a baseline of the research goals of the different areas.

Early cell theory

- In 1665, Robert Hooke described the compartments he saw in cork as “cells”.
- In 1838–1839, Matthias Schleiden and Theodor Schwann developed the idea that plants and animals consist of cells.
- Shared cellular structures provide a basis for comparing organisms.

Gregor Mendel

- Gregor Mendel (1822–1884) was an Augustinian friar working in Brno, then in the Austrian Empire and now in Czechia.
- His pea-plant experiments examined how selected traits pass between generations.
- He presented the results in 1865 and published them in 1866.



Inheritance before the gene

- Mendel studied observable traits and inferred discrete inherited factors.
- His chosen traits supported clear segregation patterns under controlled crosses.
- Many biological traits depend on several genes and environmental influences.

Mendel’s paper



The language of genetics

- **1866:** Mendel's paper on plant hybridisation was published.
- **1905:** William Bateson used the term *genetics* in a letter.
- **1909:** Wilhelm Johannsen introduced the term *gene*.

These terms preceded the identification of DNA as the principal hereditary material.

DNA as hereditary material

- Avery, MacLeod, and McCarty identified DNA as the transforming principle in 1944.
- Hershey and Chase provided further evidence using bacteriophages in 1952.
- Watson and Crick proposed the DNA double helix in 1953, informed by key experimental work including that of Franklin, Gosling, Wilkins, and colleagues.
- Chromosomes package long DNA molecules with associated proteins.

Familial hypercholesterolemia (FH)

…1972

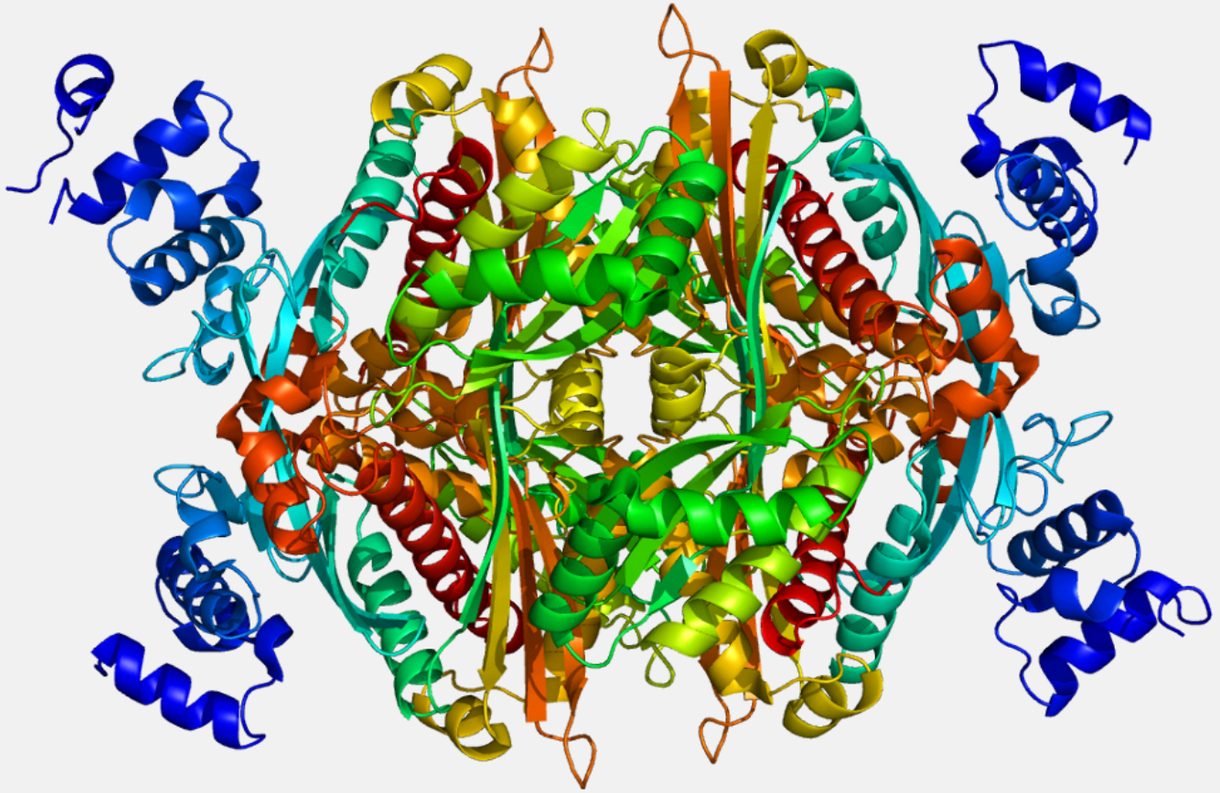
- M. S. Brown & J. L. Goldstein. Studied FH
- They applied the concept of *Inborn error of metabolism*.
- Postulated the existence of a genetic disease affecting the repression of the cholesterol synthesis.
- Patients with FH showed large indexes of cholesterol in blood -> early cardiovascular troubles.

from Genetics of Cardiovascular Disease, Circulation, 102:20 (2000)

Cholesterol regulation in FH (1/2)

HMG-CoA **reductase** catalyses a rate-limiting step in cholesterol synthesis. HMG-CoA itself is a metabolite, not the enzyme.

Brown and Goldstein’s studies connected LDL uptake to regulation of cholesterol synthesis. Severe receptor defects disrupt this feedback.



Cholesterol regulation in FH (2/2)



The LDL receptor hypothesis

- Brown and Goldstein linked defective LDL-receptor function to impaired cellular LDL uptake.
- Their work explained how LDL uptake participates in feedback control of cholesterol synthesis.
- Pathogenic variants can affect receptor production, transport, binding, internalisation, or recycling.

A genetic defect need not imply complete absence of the receptor.

Familial hypercholesterolemia

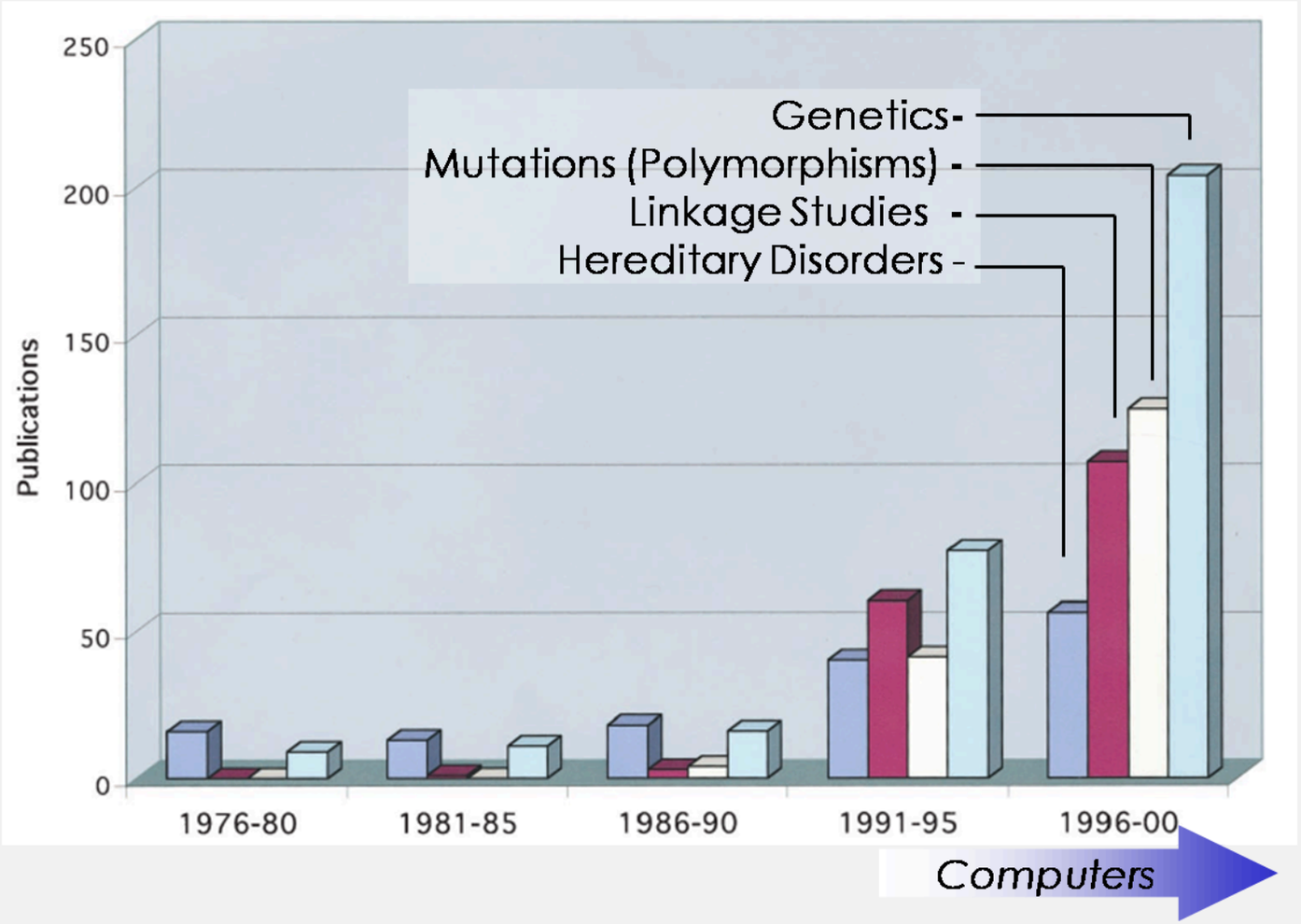
- FH causes elevated LDL cholesterol and increases the risk of premature atherosclerotic cardiovascular disease.
- Brown and Goldstein received the 1985 Nobel Prize for discoveries concerning cholesterol metabolism.
- Modern studies place heterozygous FH prevalence around 1 in 250–310, with variation by population and diagnostic criteria.
- Individual risk depends on genotype, exposure over time, treatment, and other risk factors.



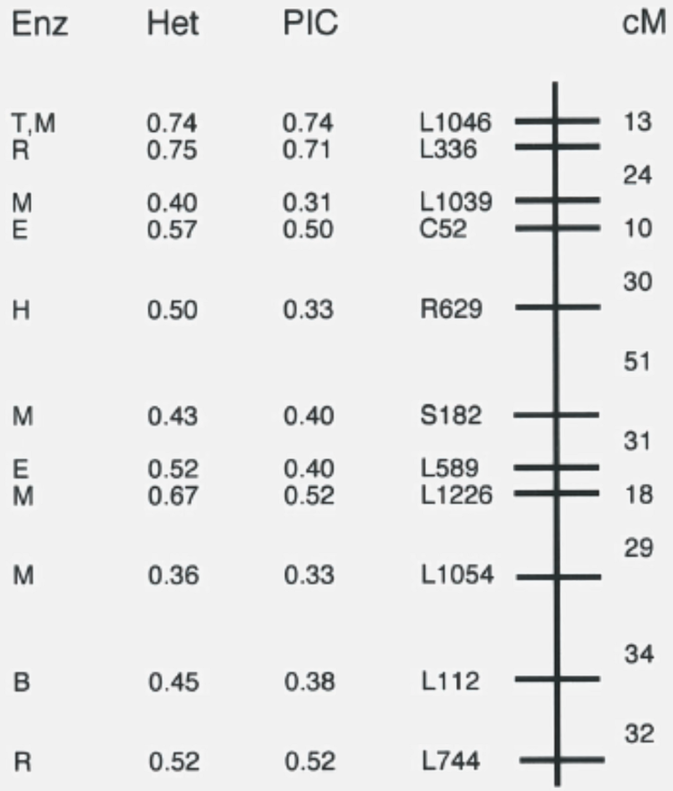
Genetic mechanisms and annotation

- FH is a classic example of a disorder that can result from a pathogenic variant in a single gene.
- Different genes can produce an FH phenotype, including **LDLR**, **APOB**, and **PCSK9**.
- Connecting a variant to a phenotype requires evidence about the sequence, gene, molecular mechanism, and clinical context.
- Recording these connections is a central purpose of biological annotation.

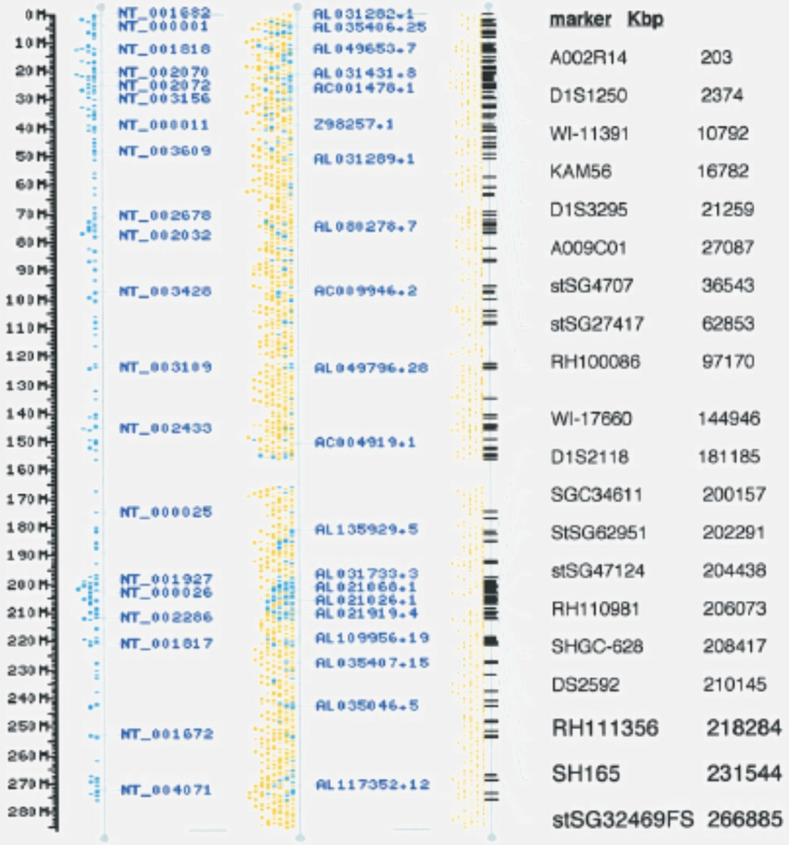
Evolution of publications in 25 years



Chromosome 1 annotation: 1988 and 2000

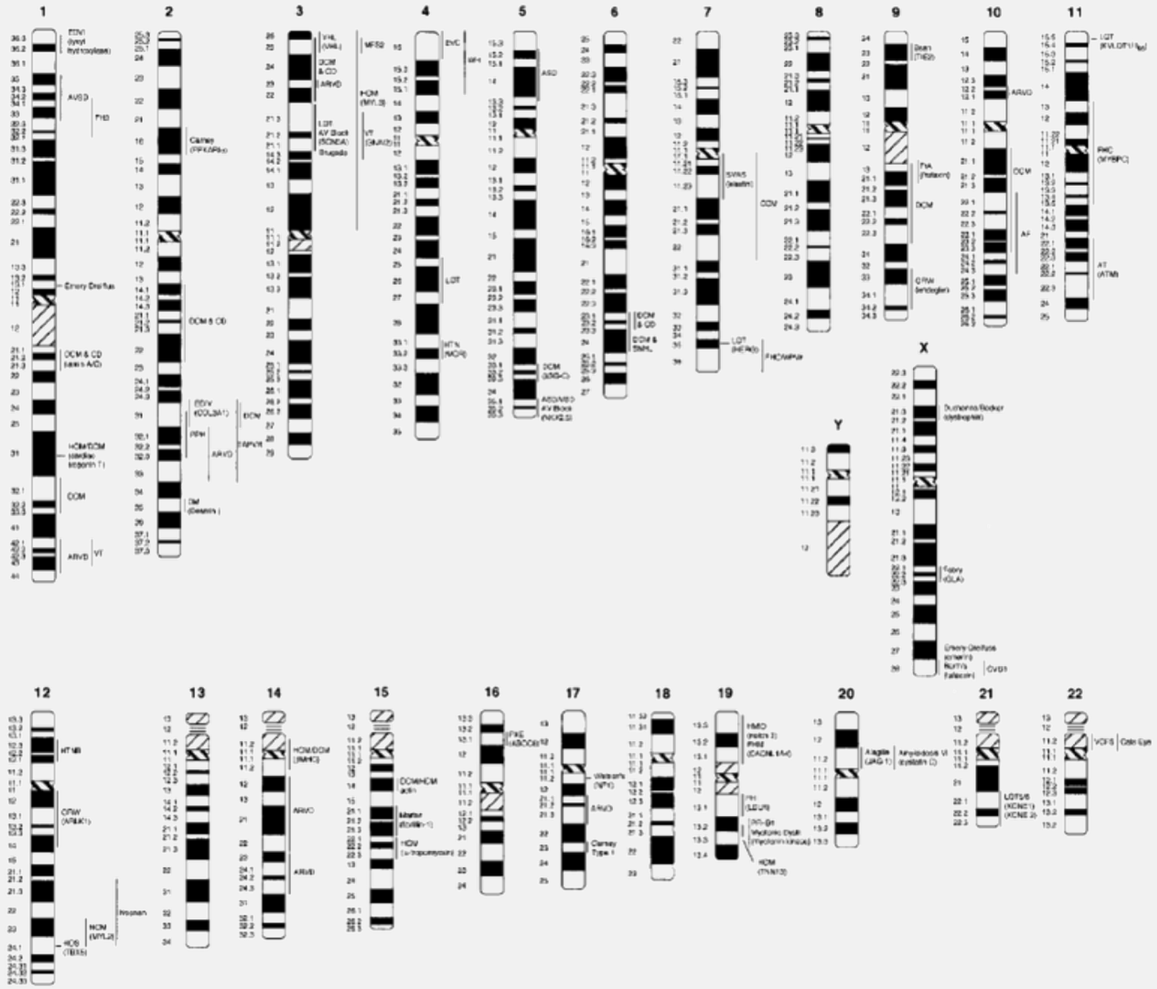


1988

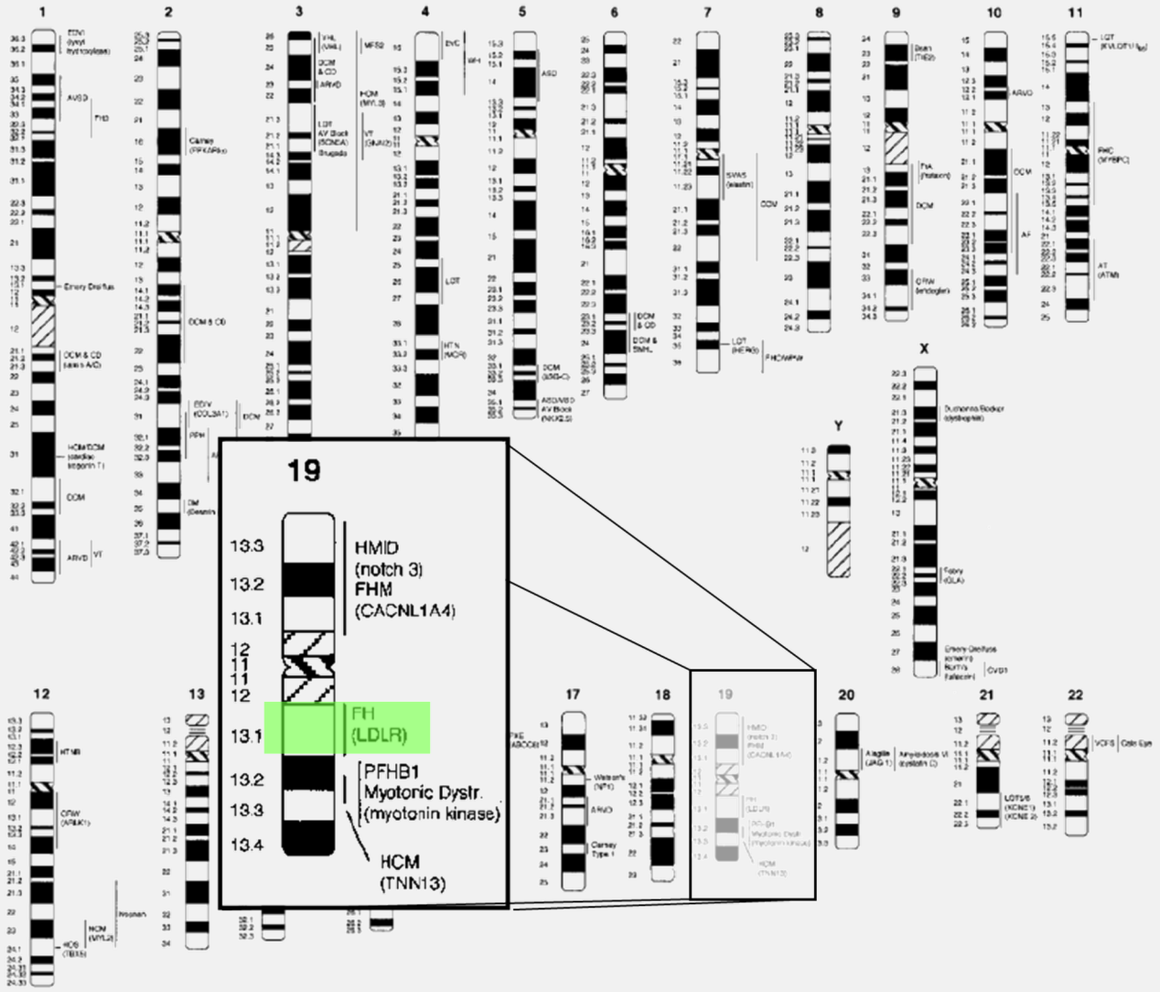


2000

Evolution on (cardiovascular) annotations on Genome



Evolution on (cardiovascular) annotations on Genome



Human Genome Project?

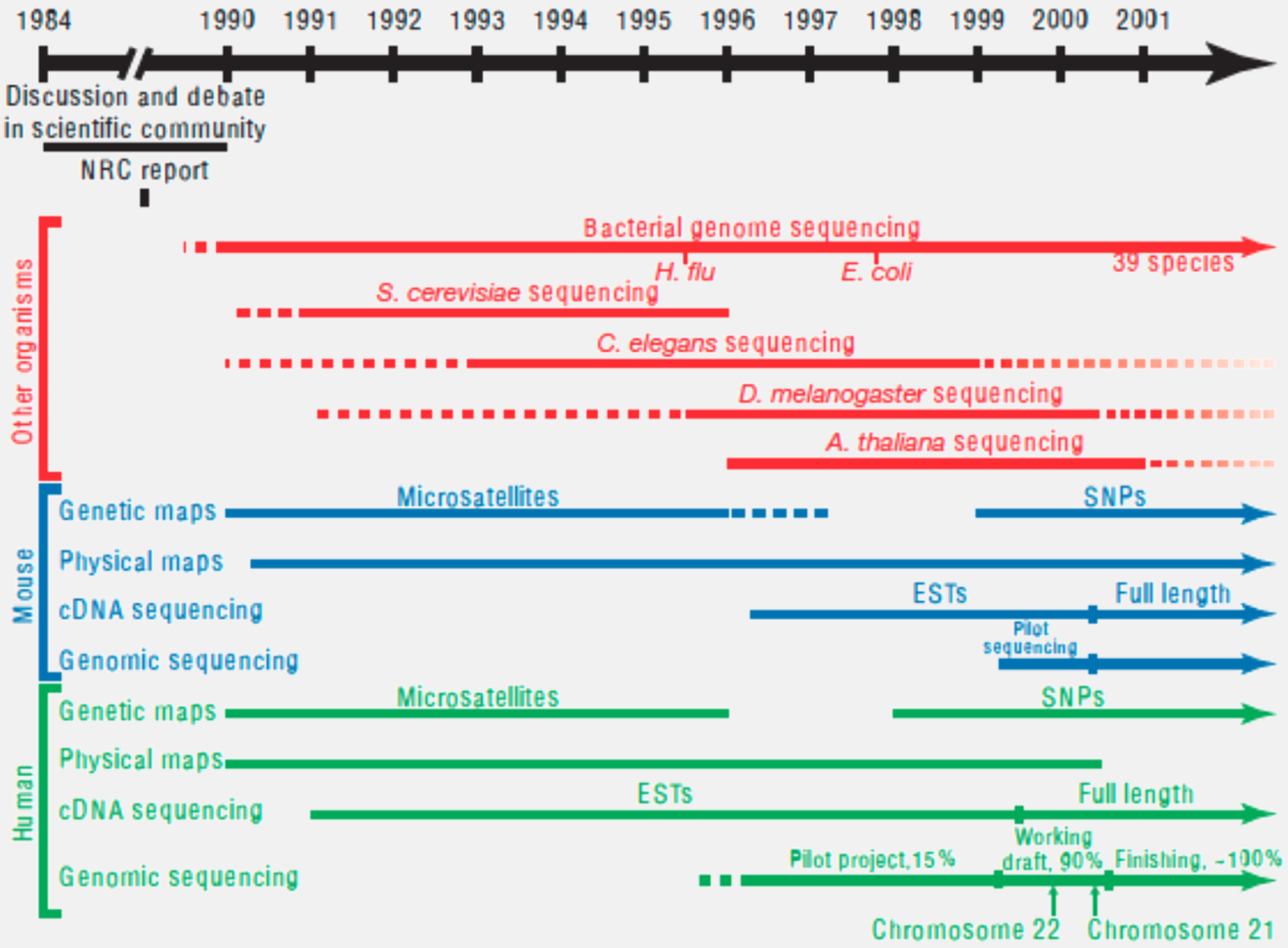
MAPPING AND SEQUENCING THE HUMAN GENOME

Committee on
Mapping and Sequencing the
Human Genome
Board on Basic Biology
Commission on Life Sciences
National Research Council

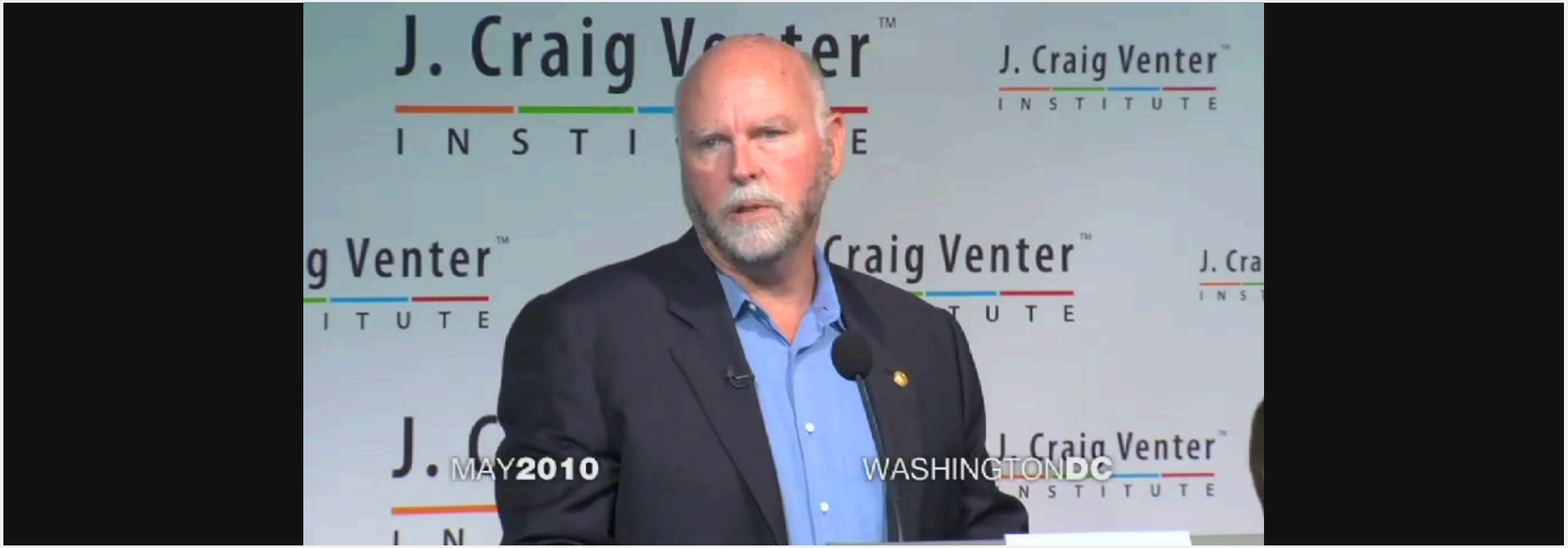
Click for pag

NATIONAL ACADEMY PRESS
Washington, D.C. 1988

Human Genome Project?



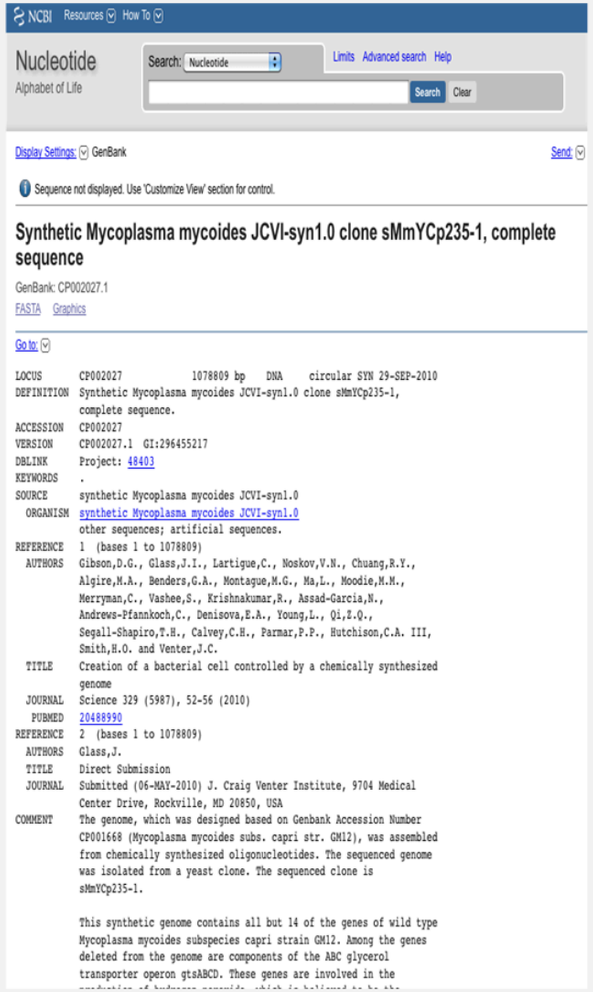
Craig Venter unveils “synthetic life”



A synthetic genome was transplanted into an existing recipient cell; the entire cell was not built from scratch.

TED · 18:18 · TED, 2010 · English subtitles from the published caption track.

Where are we now?



NCBI Resources How To

Nucleotide
Alphabet of Life

Search: Nucleotide Limits Advanced search Help
Search Clear

Display Settings: GenBank Send:

Sequence not displayed. Use 'Customize View' section for control.

Synthetic Mycoplasma mycoides JCVI-syn1.0 clone sMmYCp235-1, complete sequence

GenBank: CP002027.1
FASTA Graphics

Go to:

```
LOCUS       CP002027            1078809 bp    DNA     circular SYN 29-SEP-2010
DEFINITION  Synthetic Mycoplasma mycoides JCVI-syn1.0 clone sMmYCp235-1,
            complete sequence.
ACCESSION   CP002027
VERSION     CP002027.1  GI:296455217
DBLINK      Project: 48403
KEYWORDS    .
SOURCE      synthetic Mycoplasma mycoides JCVI-syn1.0
  ORGANISM  synthetic Mycoplasma mycoides JCVI-syn1.0
            other sequences; artificial sequences.
REFERENCE   1  (bases 1 to 1078809)
  AUTHORS   Gibson,D.G., Glass,J.I., Lartigue,C., Noskov,V.N., Chuang,R.Y.,
            Algire,M.A., Benders,G.A., Montague,M.G., Ma,L., Moodie,M.M.,
            Merryman,C., Vashee,S., Krishnakumar,R., Assad-Garcia,N.,
            Andrews-Pfannkoch,C., Denisova,E.A., Young,L., Qi,Z.Q.,
            Segall-Shapiro,T.H., Calvey,C.H., Parmar,P.P., Hutchison,C.A. III,
            Smith,H.O. and Venter,J.C.
  TITLE     Creation of a bacterial cell controlled by a chemically synthesized
            genome
  JOURNAL   Science 329 (5987), 52-56 (2010)
   PUBMED   20488990
REFERENCE   2  (bases 1 to 1078809)
  AUTHORS   Glass,J.
  TITLE     Direct Submission
  JOURNAL   Submitted (06-MAY-2010) J. Craig Venter Institute, 9704 Medical
            Center Drive, Rockville, MD 20850, USA
COMMENT     The genome, which was designed based on Genbank Accession Number
            CP001668 (Mycoplasma mycoides subs. capri str. GM12), was assembled
            from chemically synthesized oligonucleotides. The sequenced genome
            was isolated from a yeast clone. The sequenced clone is
            sMmYCp235-1.

            This synthetic genome contains all but 14 of the genes of wild type
            Mycoplasma mycoides subspecies capri strain GM12. Among the genes
            deleted from the genome are components of the ABC glycerol
            transporter operon gtsABCD. These genes are involved in the
```

nature biotechnology

ishing Group http://www.nature.com/naturebiotechnology

What would you do if you could sequence everything?

Avak Kahvejian[1], John Quackenbush[2] & John F Thompson[1]

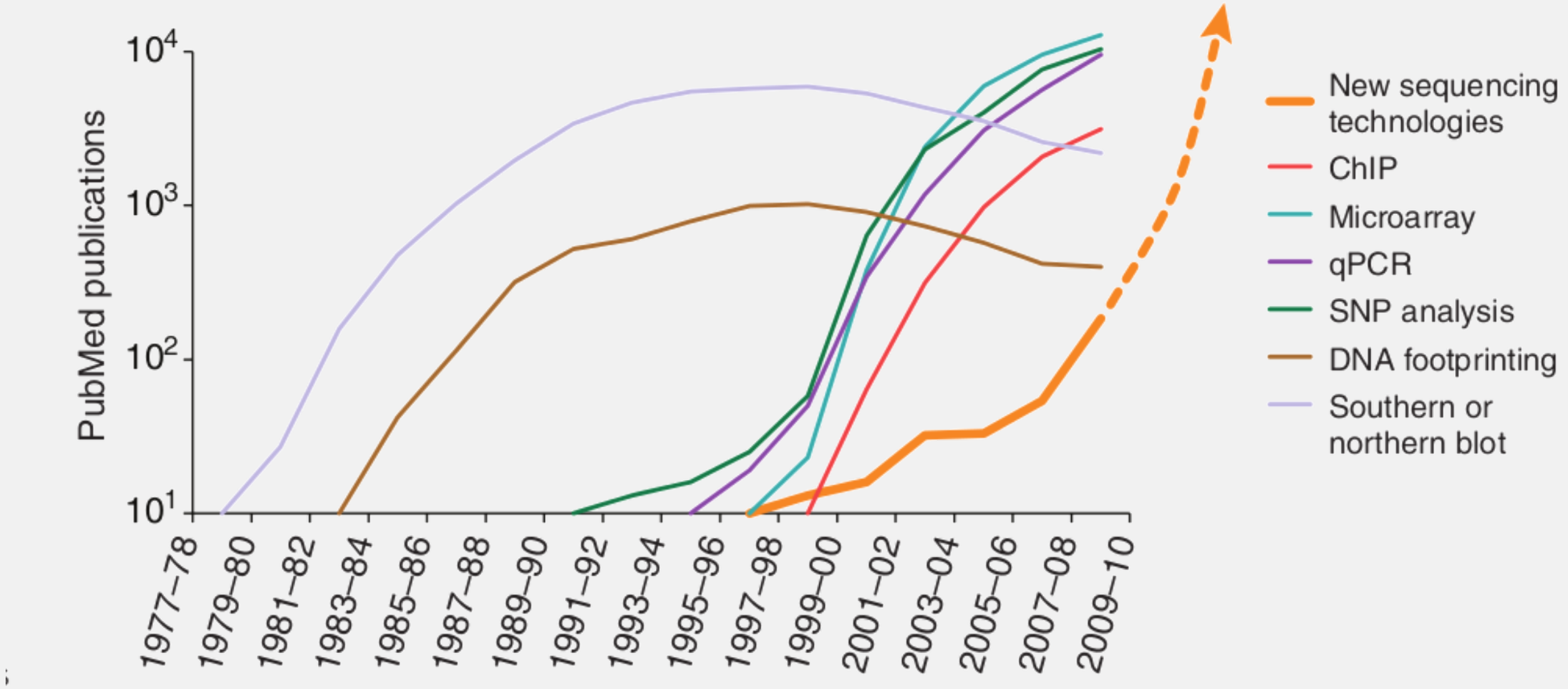
It could be argued that the greatest transformative aspect of the Human Genome Project has been not the sequencing of the genome itself, but the resultant development of new technologies. A host of new approaches has fundamentally changed the way we approach problems in basic and translational research. Now, a new generation of high-throughput sequencing technologies promises to again transform the scientific enterprise, potentially supplanting array-based technologies and opening up many new possibilities. By allowing DNA/RNA to be assayed more rapidly than previously possible, these next-generation platforms promise a deeper understanding of genome regulation and biology. Significantly enhancing sequencing throughput will allow us to follow the evolution of viral and bacterial resistance in real time, to uncover the huge diversity of novel genes that are currently inaccessible,

ing technologies, be leveraged with insightful approaches to biology and medicine to maximize the benefits to all? DNA sequence is no longer just an end in itself, but it is rapidly becoming the digital substrate replacing analog chip-based hybridization signals; it is the barcode tracking of enormous numbers of samples; and it is the readout indicating a host of chemical modifications and intermolecular interactions. Sequence data allow one to count mRNAs or other species of nucleic acids precisely, to determine sharp boundaries for interactions with proteins or positions of translocations and to identify novel variants and splice sites, all in one experiment with digital accuracy.

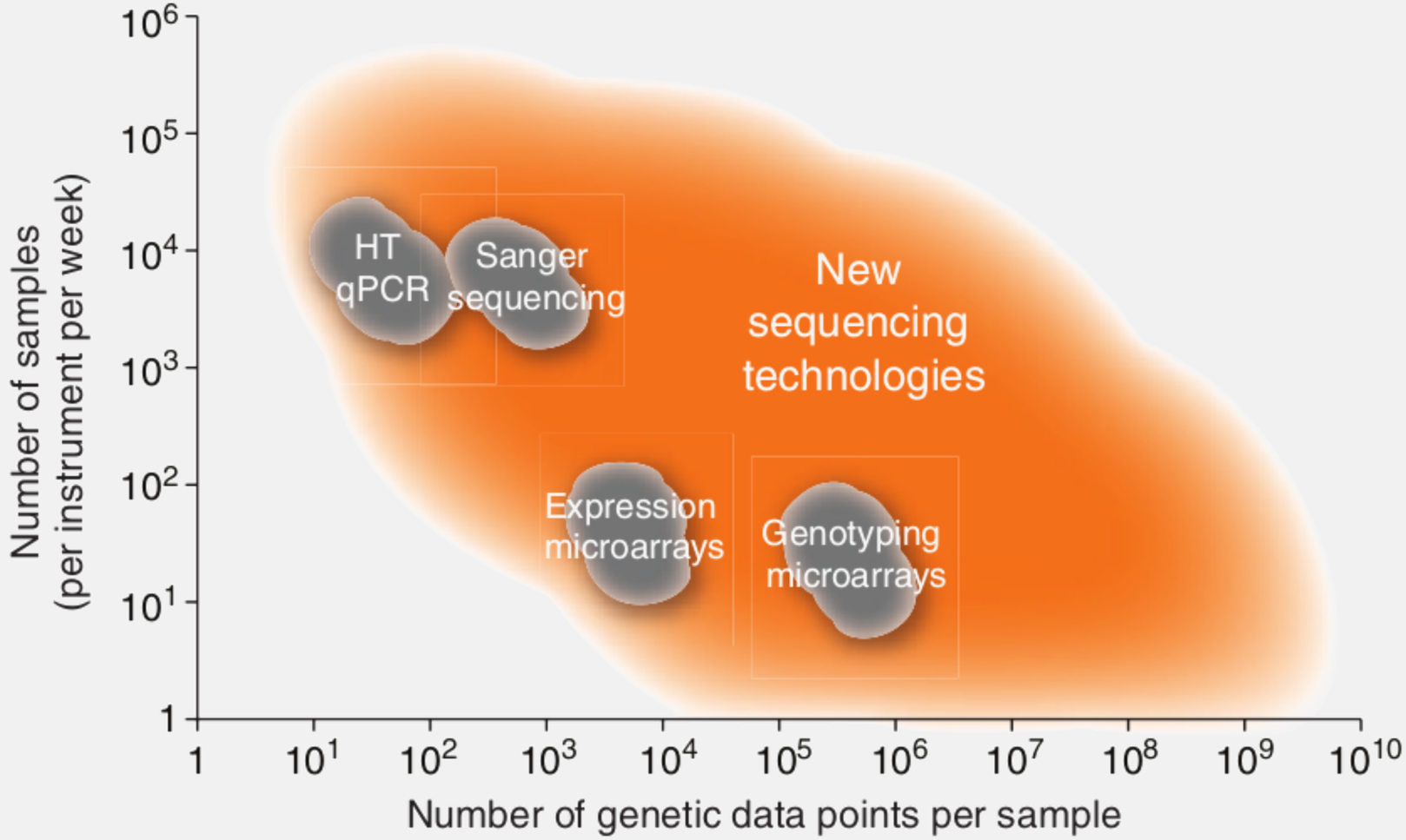
The brief history of molecular biology and genomic technologies has been marked by the introduction of new technologies, their rapid uptake and then a steady state or slow decline in use as newer techniques are developed that supersede them. For example, as measured by publications listed in PubMed, gel-based hybridization techniques for analyzing

Kahvejian, A., Quackenbush, J., & Thompson, J. F. (2008). What would you do if you could sequence everything?. Nature biotechnology, 26(10), 1125-1133.

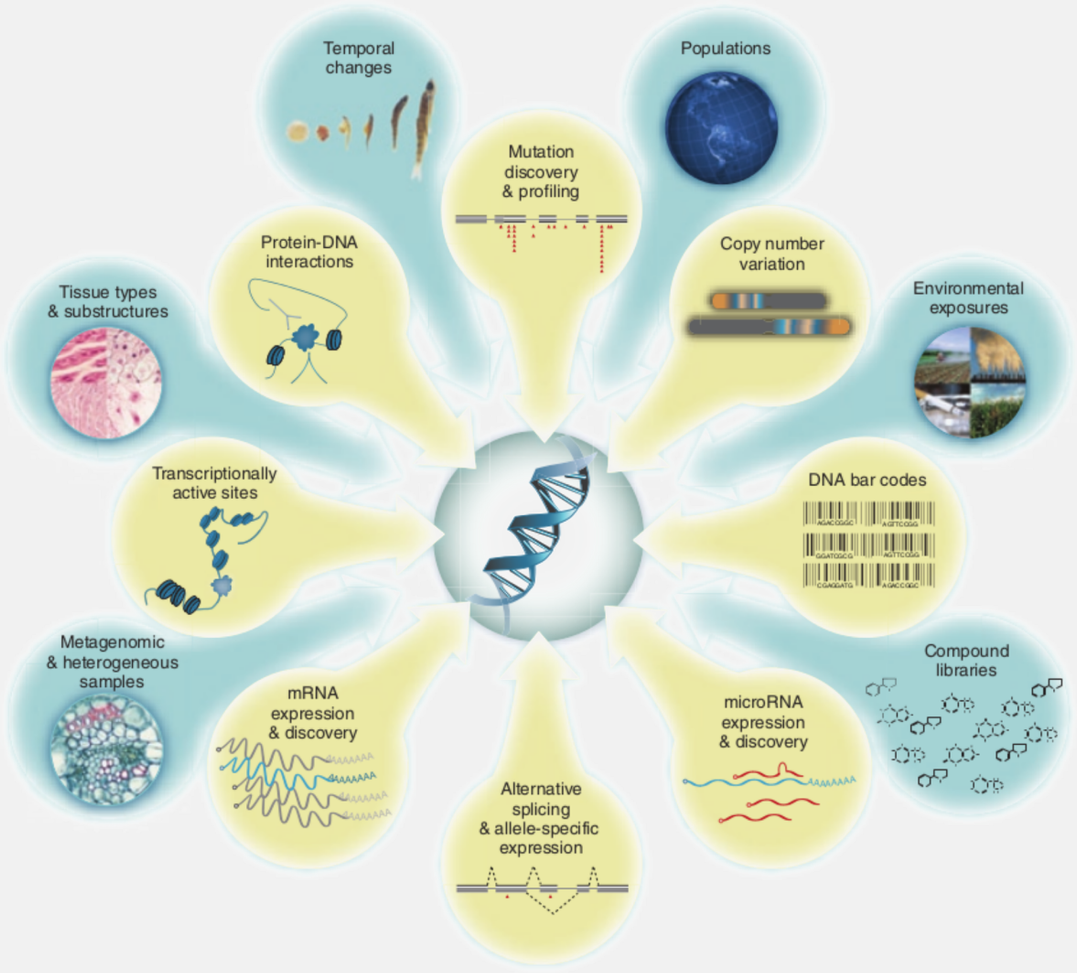
Evolution in genotypic technologies



Evolution in genotypic technologies



Evolution in genotypic technologies



DNA, RNA, and proteins

DNA

Stores hereditary information in a nucleotide sequence.

RNA

Participates in information transfer, regulation, catalysis, and cellular structure.

Proteins

Perform catalytic, structural, transport, and signalling functions. Protein sequences use 20 standard amino acids, with specialised exceptions.

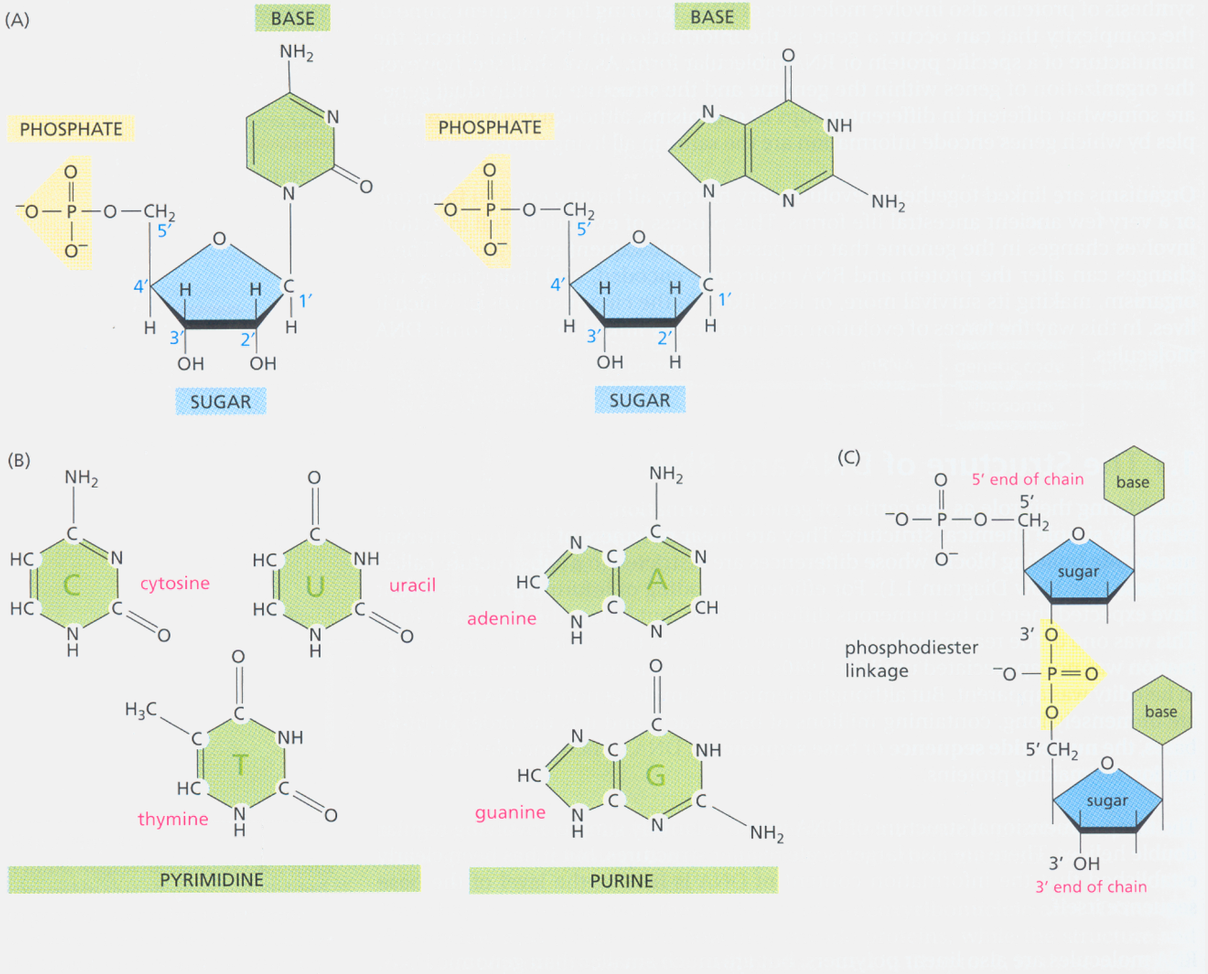
DNA nucleotides

A DNA nucleotide contains a nitrogenous base, deoxyribose sugar, and phosphate.

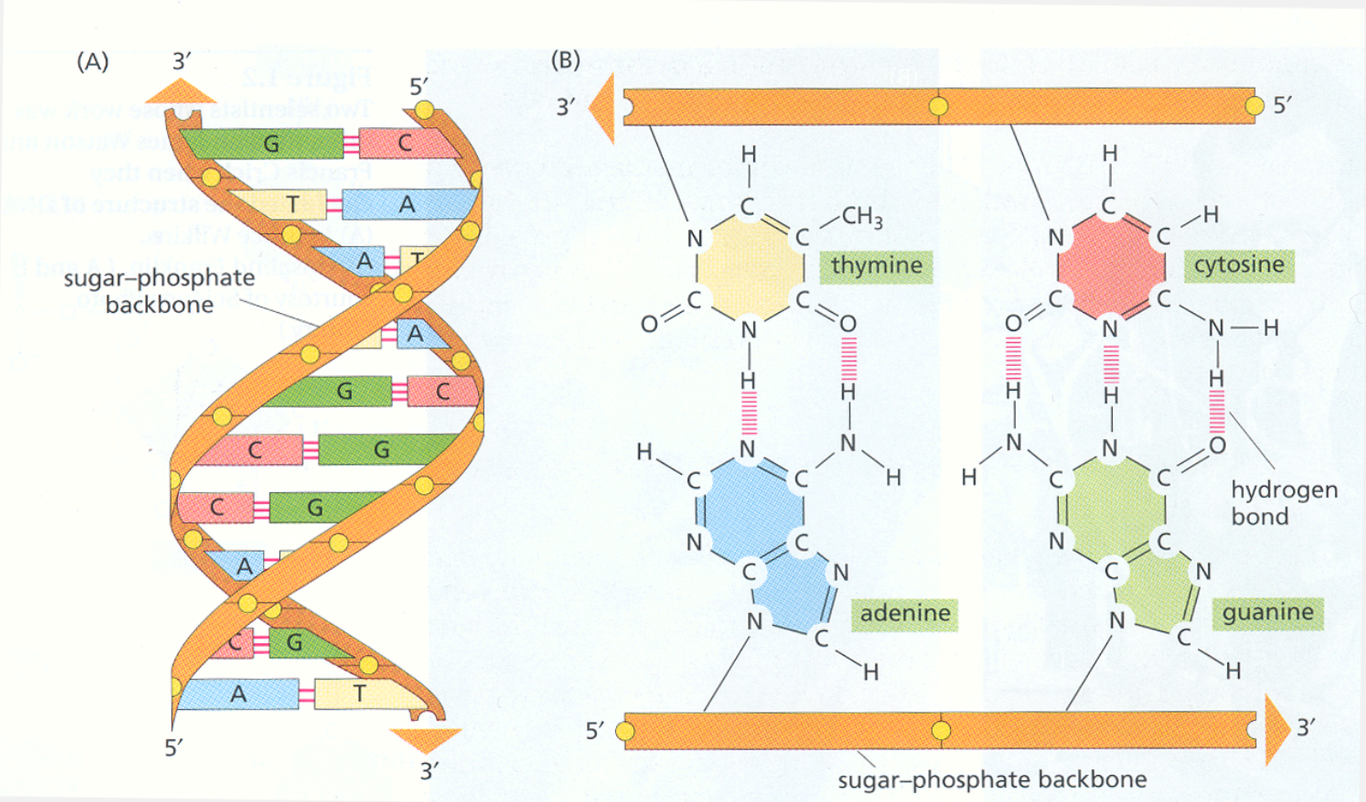
Base class	DNA bases
Purines	Adenine (A), guanine (G)
Pyrimidines	Cytosine (C), thymine (T)

RNA contains uracil (U) in place of thymine. A haploid human nuclear genome contains roughly 3 billion base pairs.

DNA structure

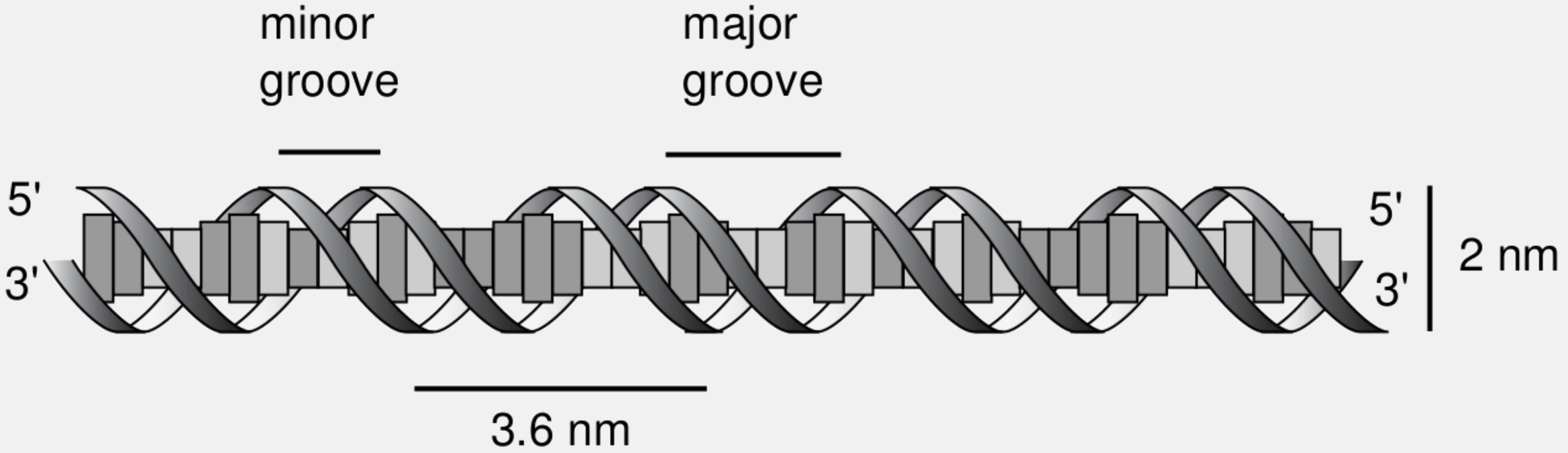


DNA structure



The resulting chain has a 5’ end (free phosphate group) and a 3’ end (free hydroxyl group). The base sequence is defined as the order of the nucleotides along the chain from the 5’ to the 3’ end.

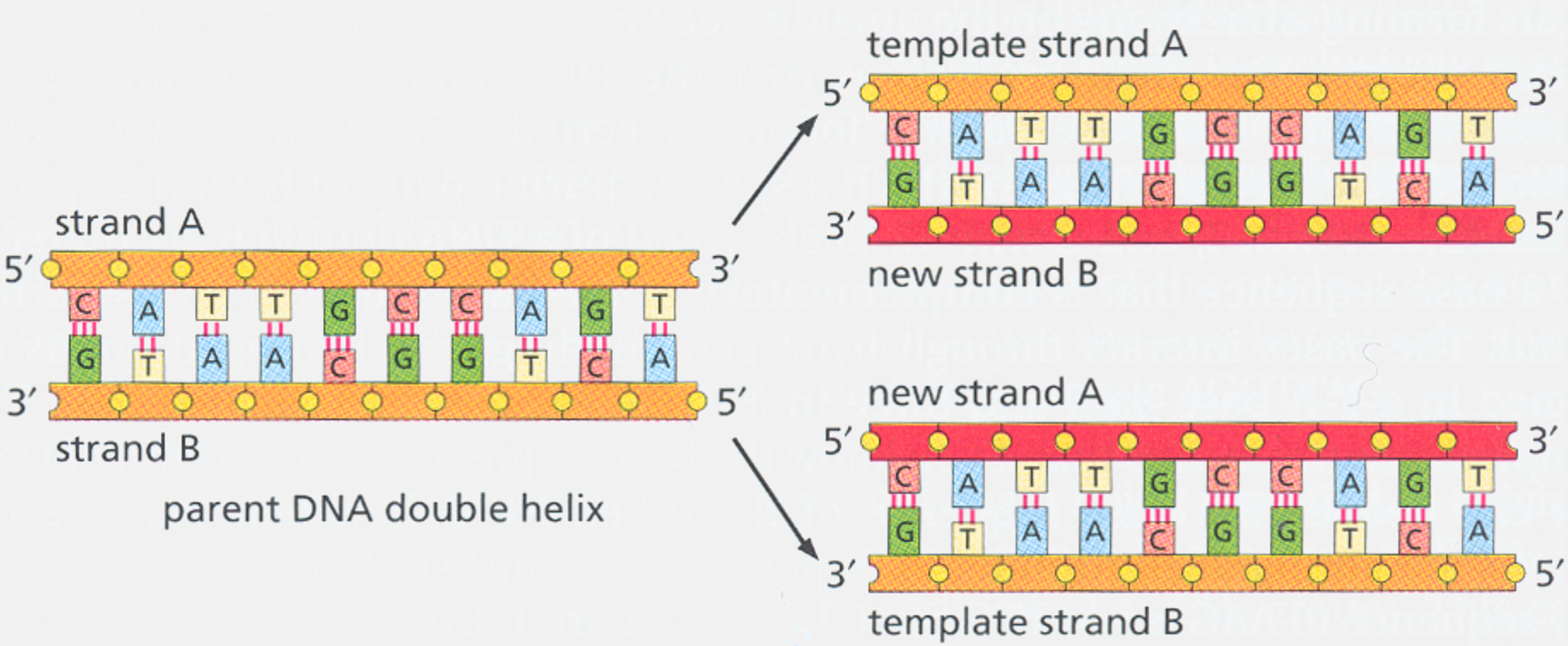
DNA structure



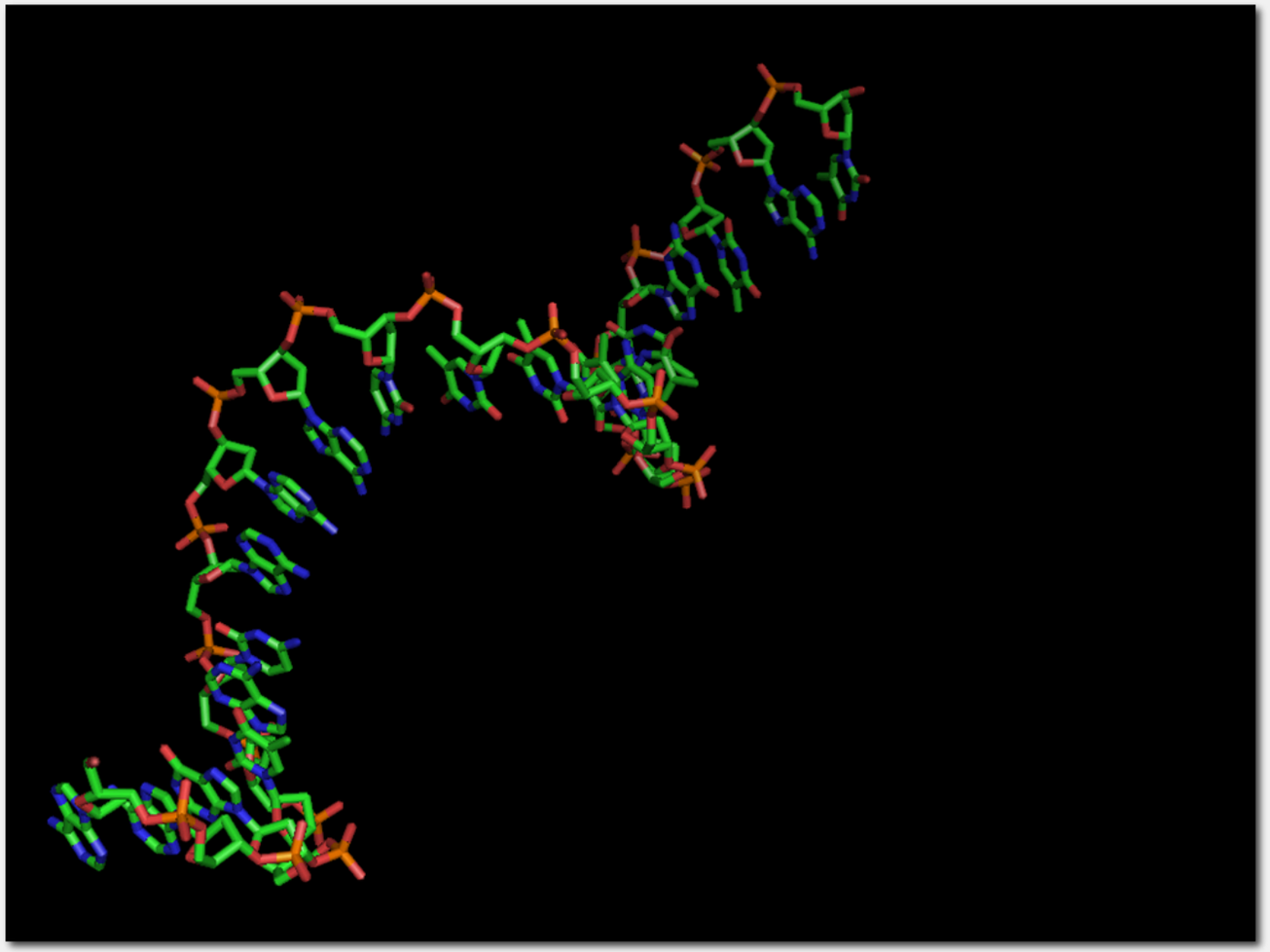
DNA replication

- Complementary strands pair through hydrogen bonds; base stacking also stabilises the double helix.
- The sugar–phosphate backbones lie on the outside of the helix.
- Polymerases synthesise DNA in the **5′ to 3′ direction** using a template strand.
- Proofreading and repair improve fidelity. Error rates depend on the organism and on which stage of replication or repair is measured.
- Complementary DNA and RNA sequences can hybridise.

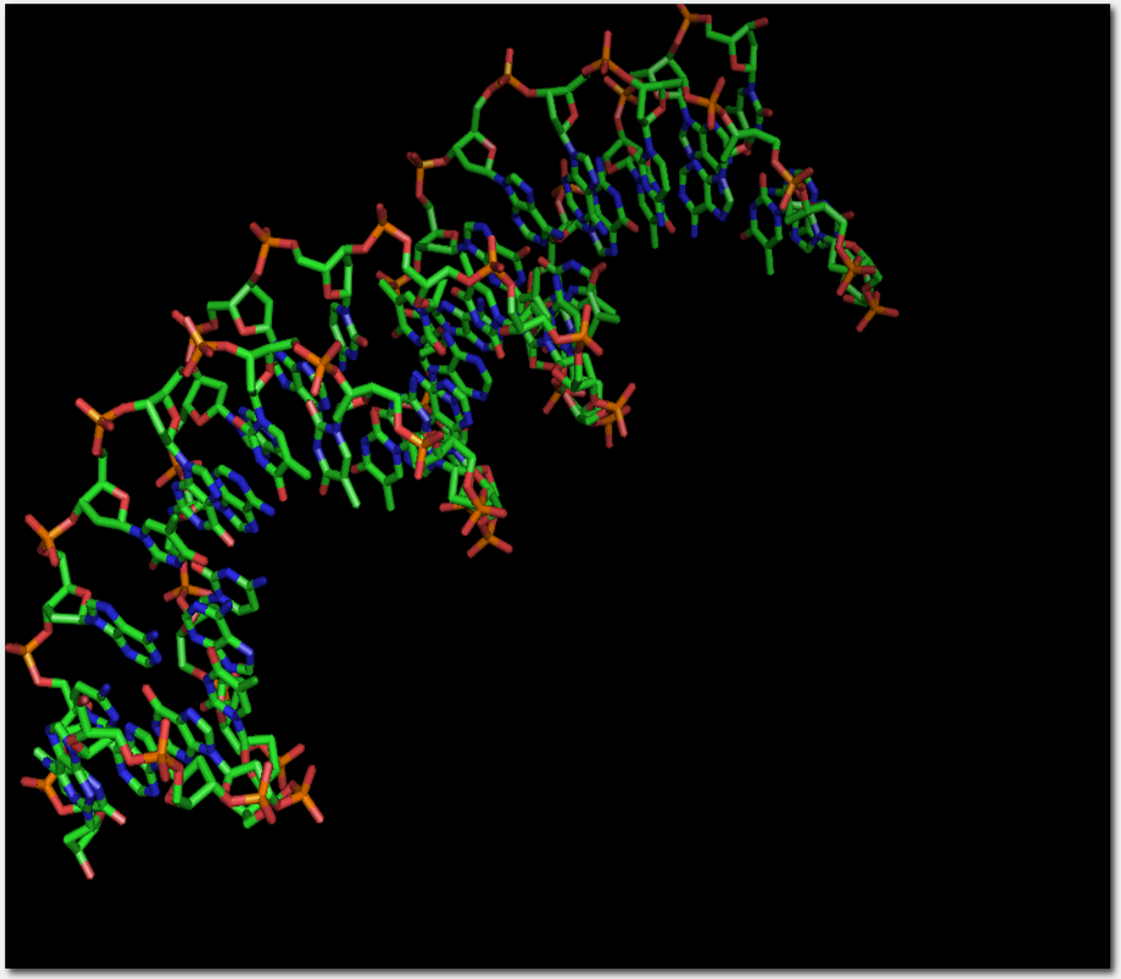
DNA double helix means DNA replication



DNA (1/15)



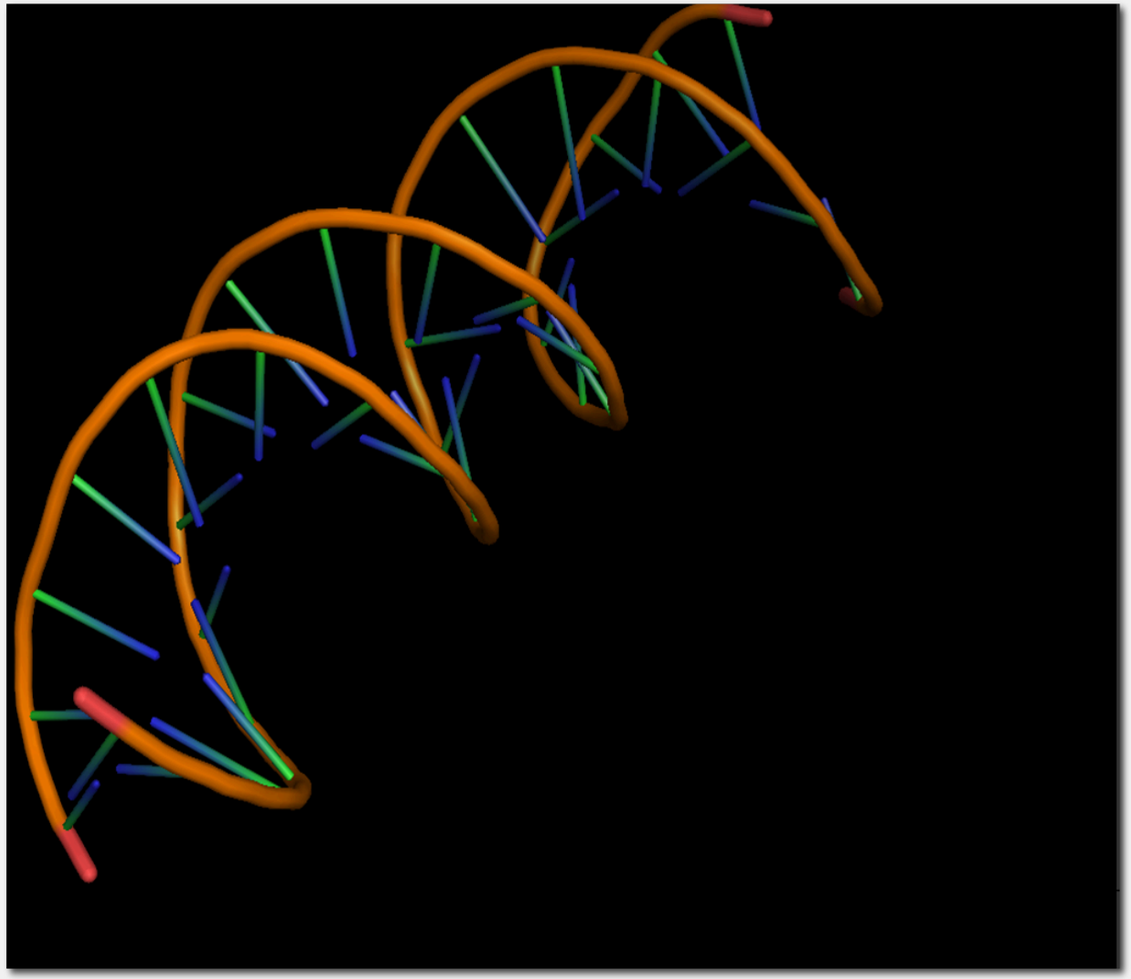
DNA (2/15)



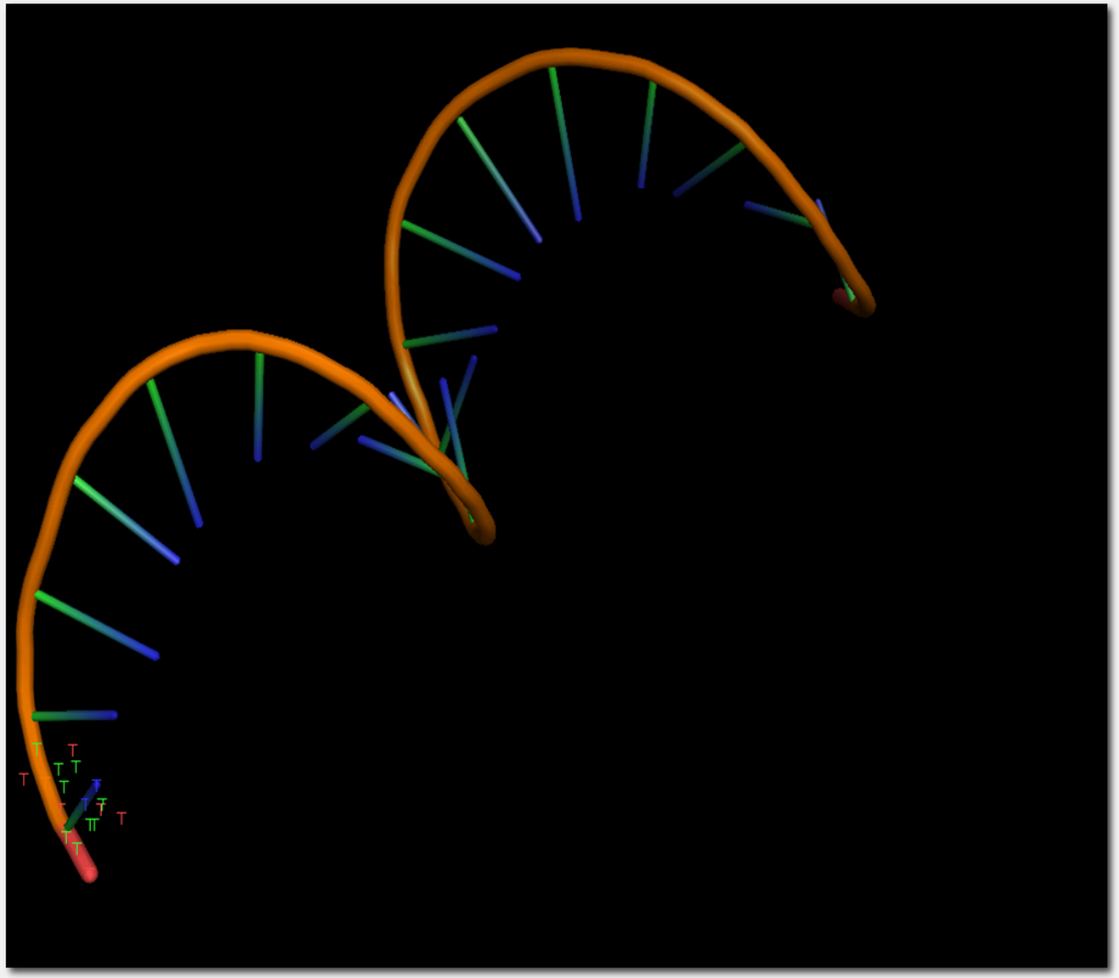
DNA (3/15)



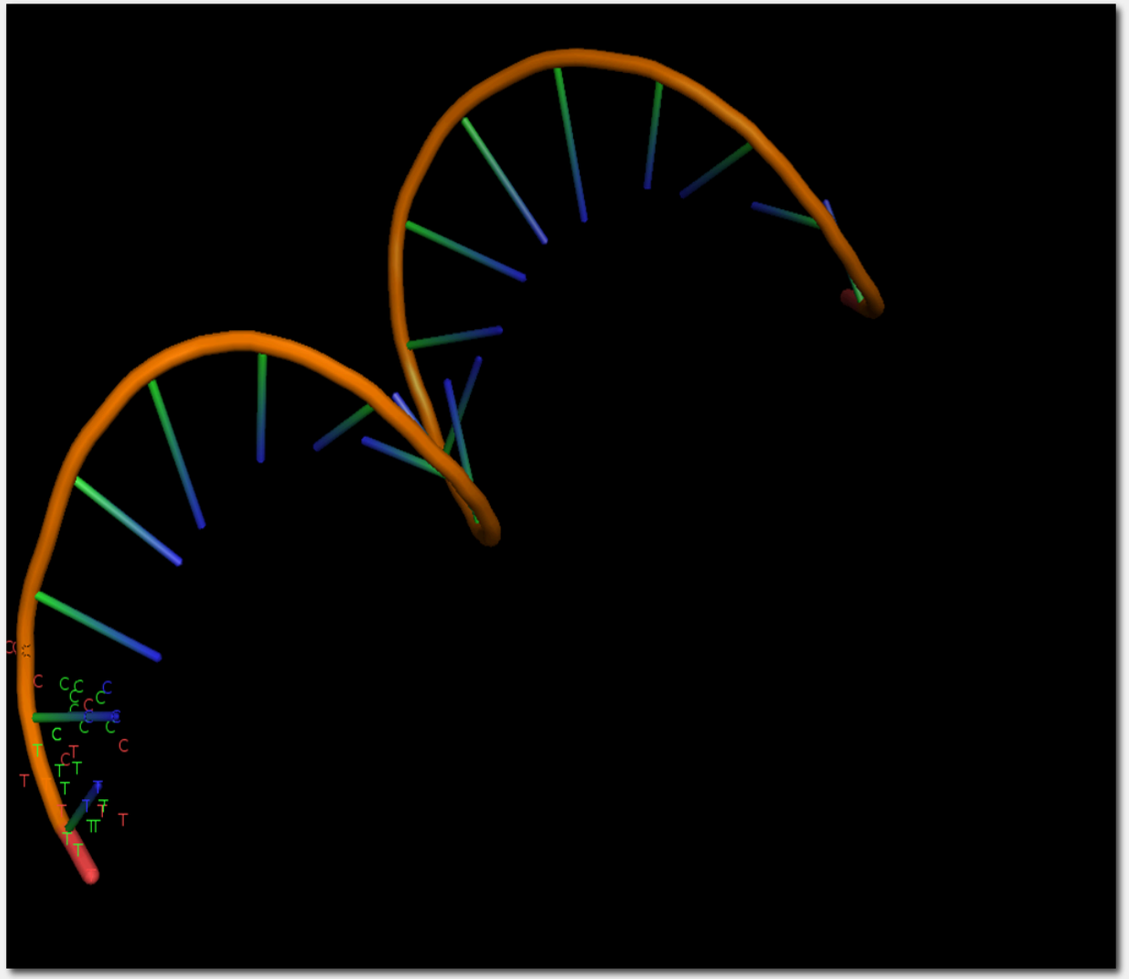
DNA (4/15)



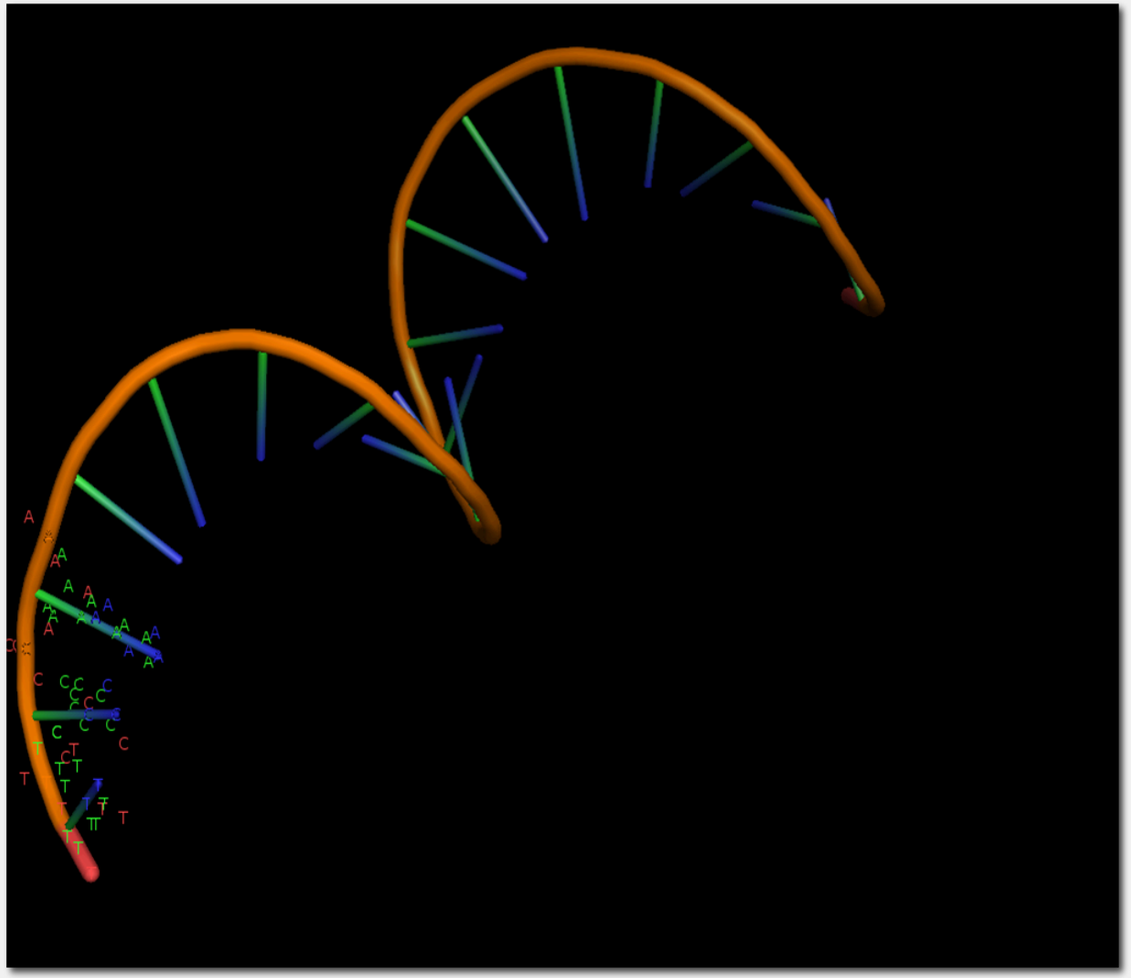
DNA (5/15)



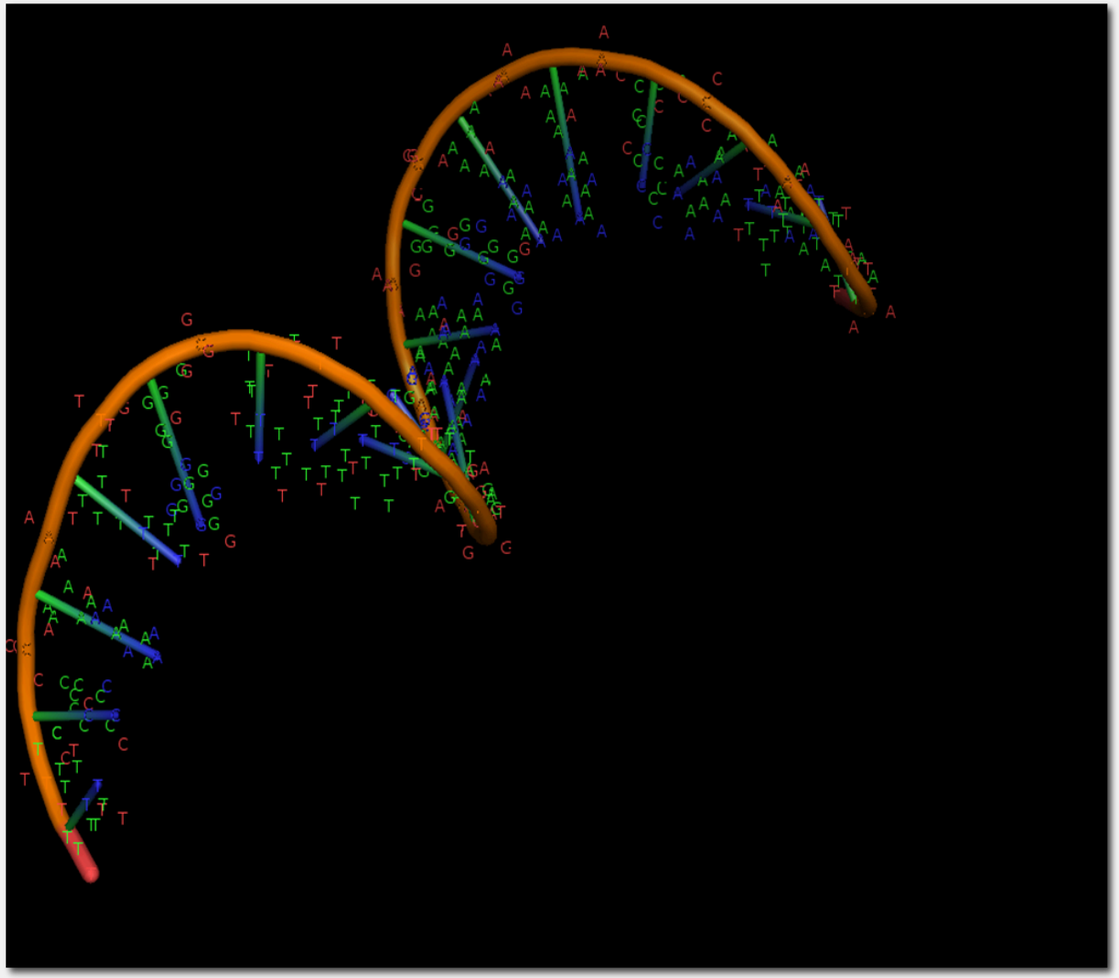
DNA (6/15)



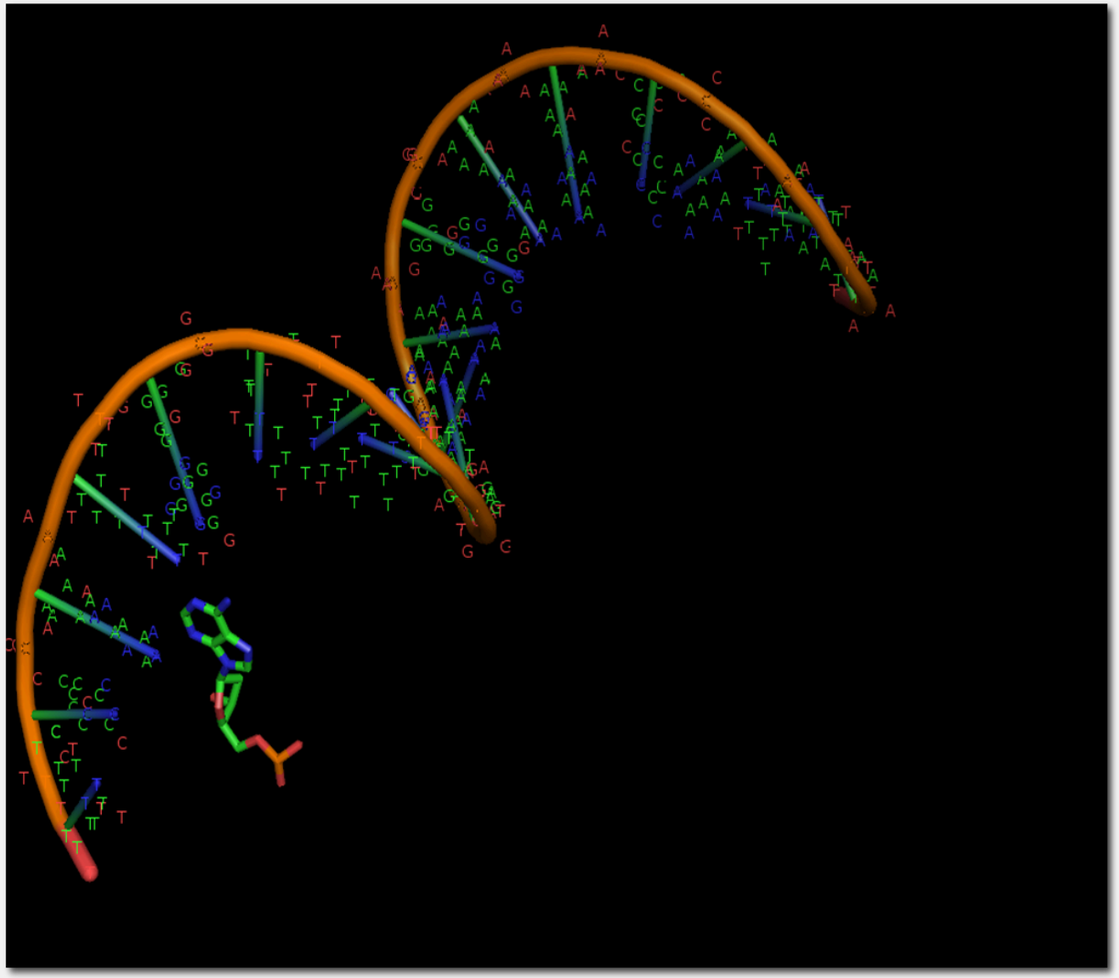
DNA (7/15)



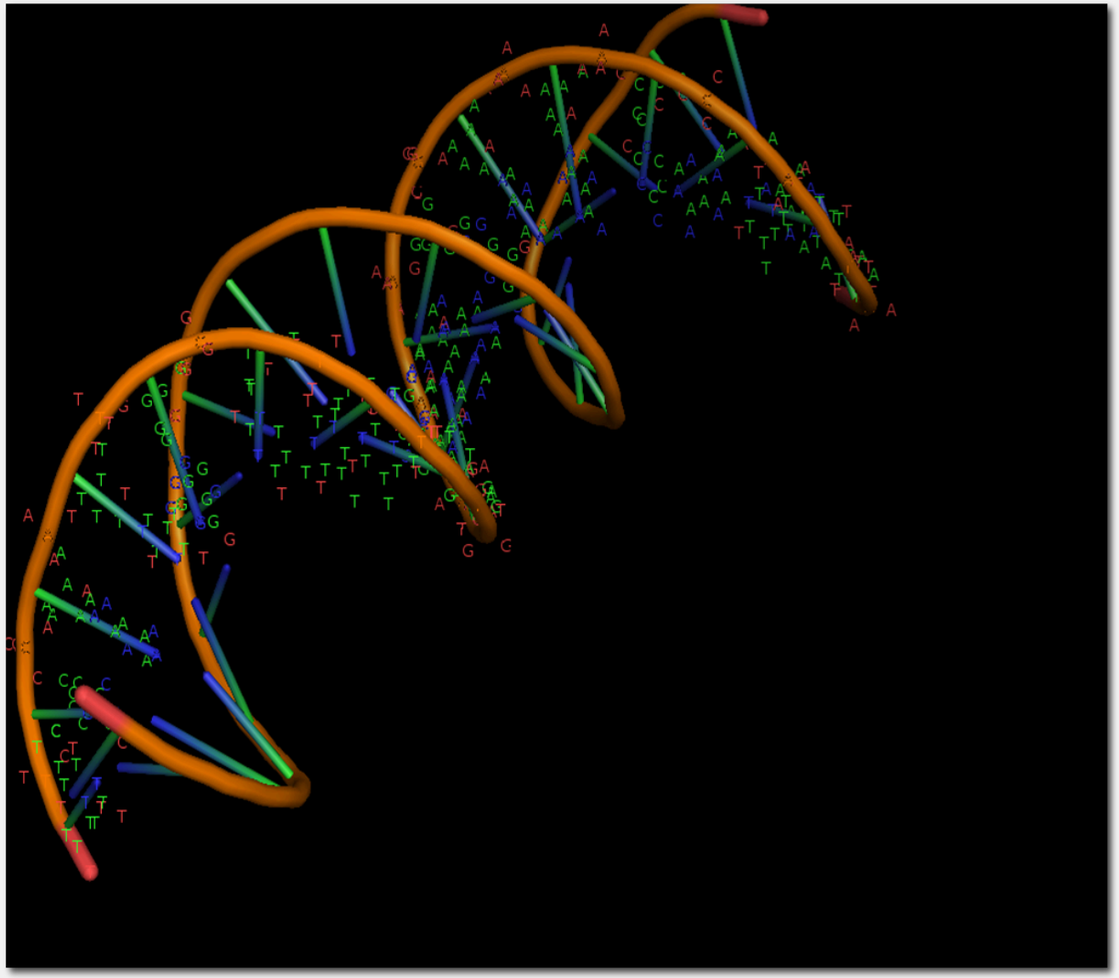
DNA (8/15)



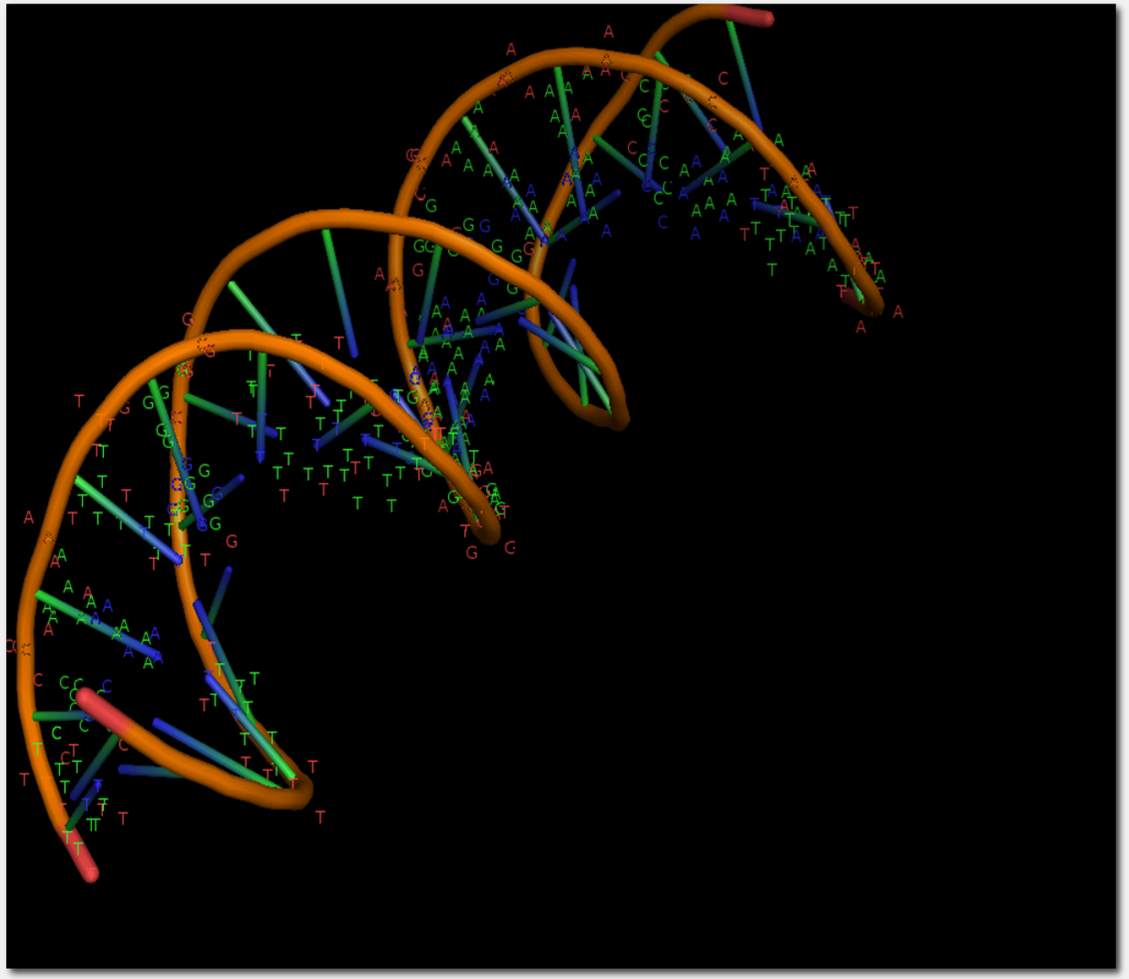
DNA (9/15)



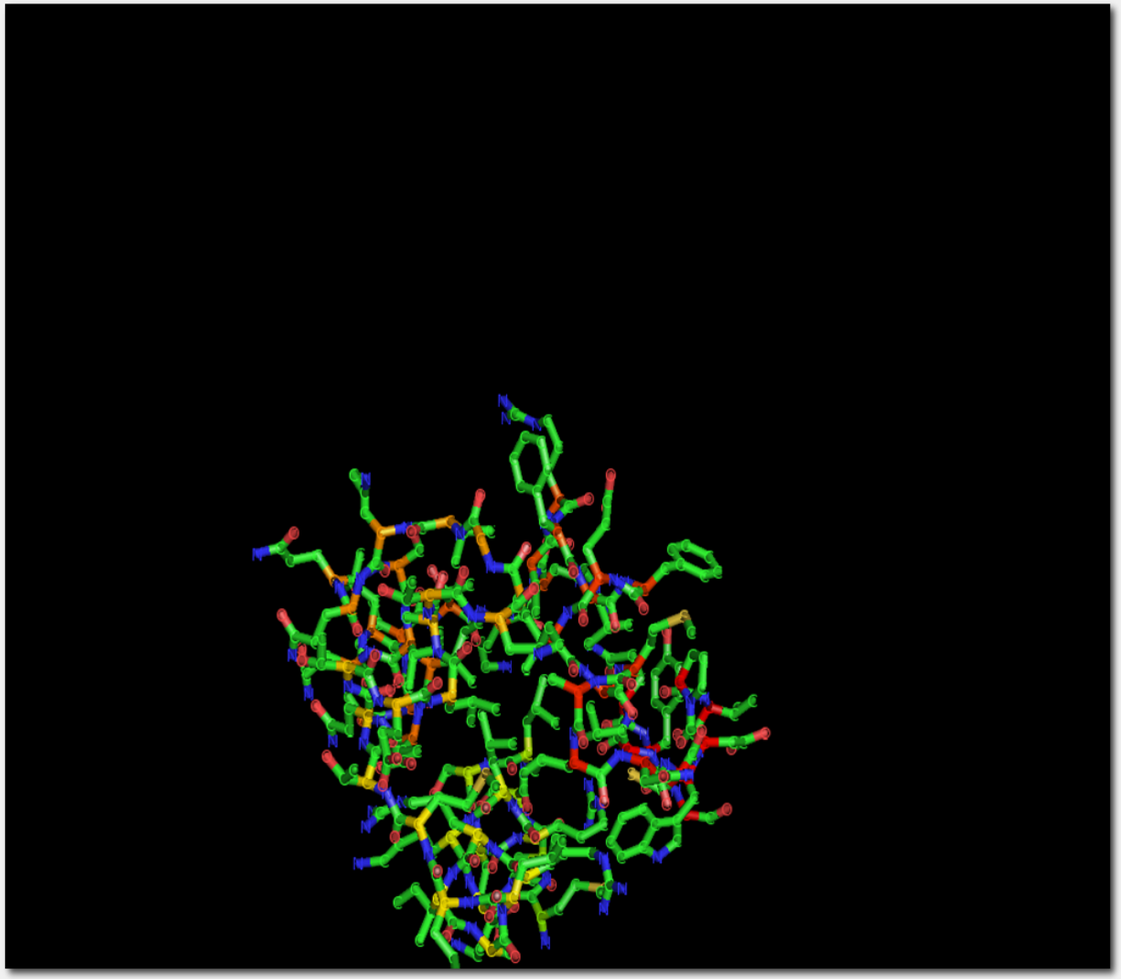
DNA (10/15)



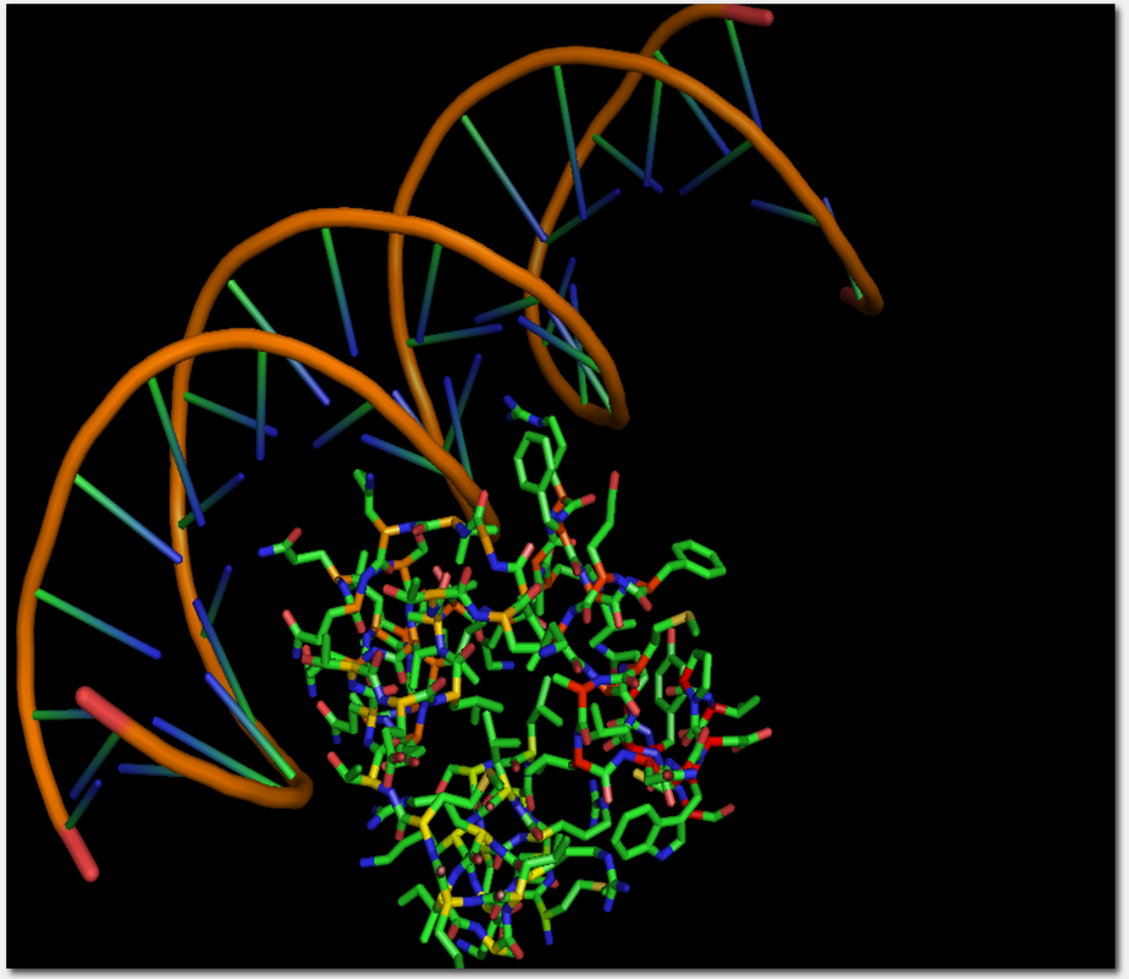
DNA (11/15)



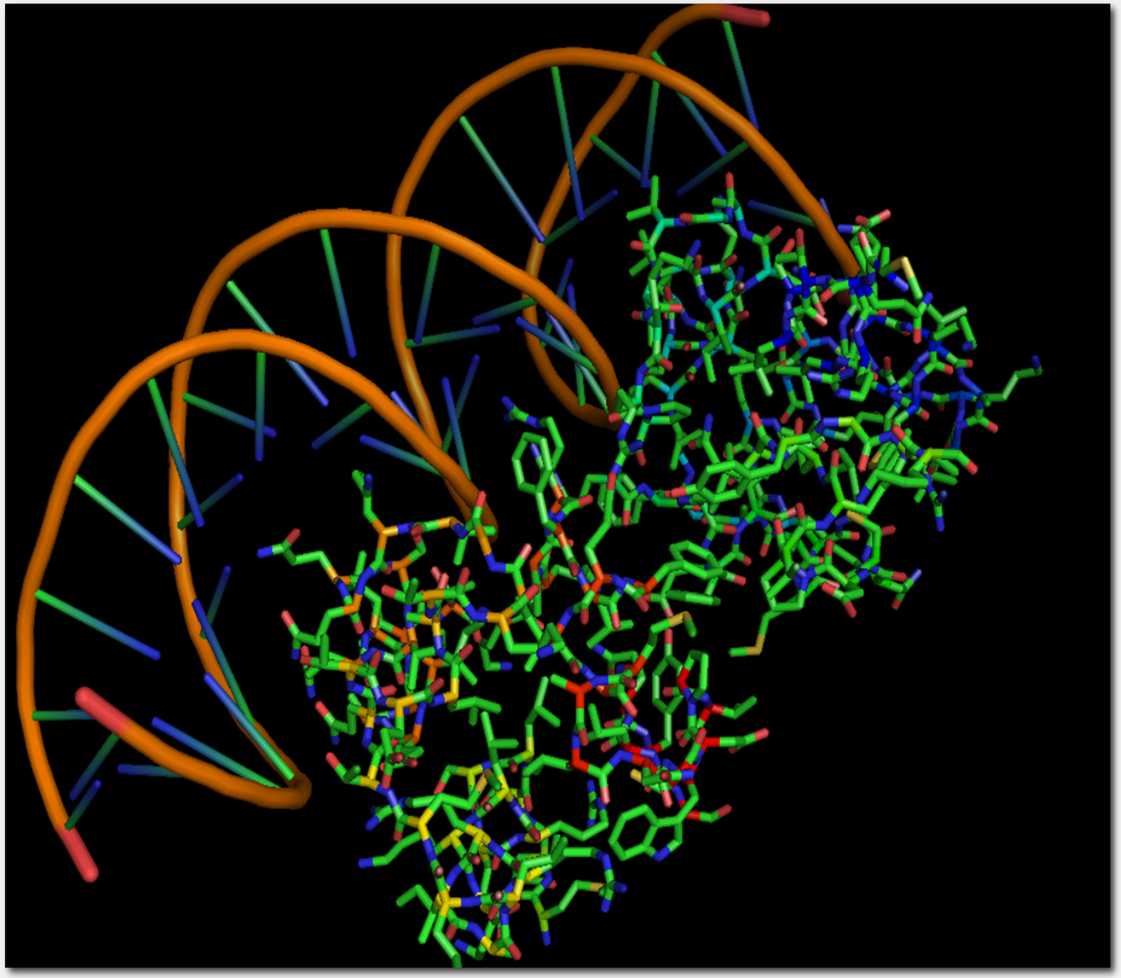
DNA (12/15)



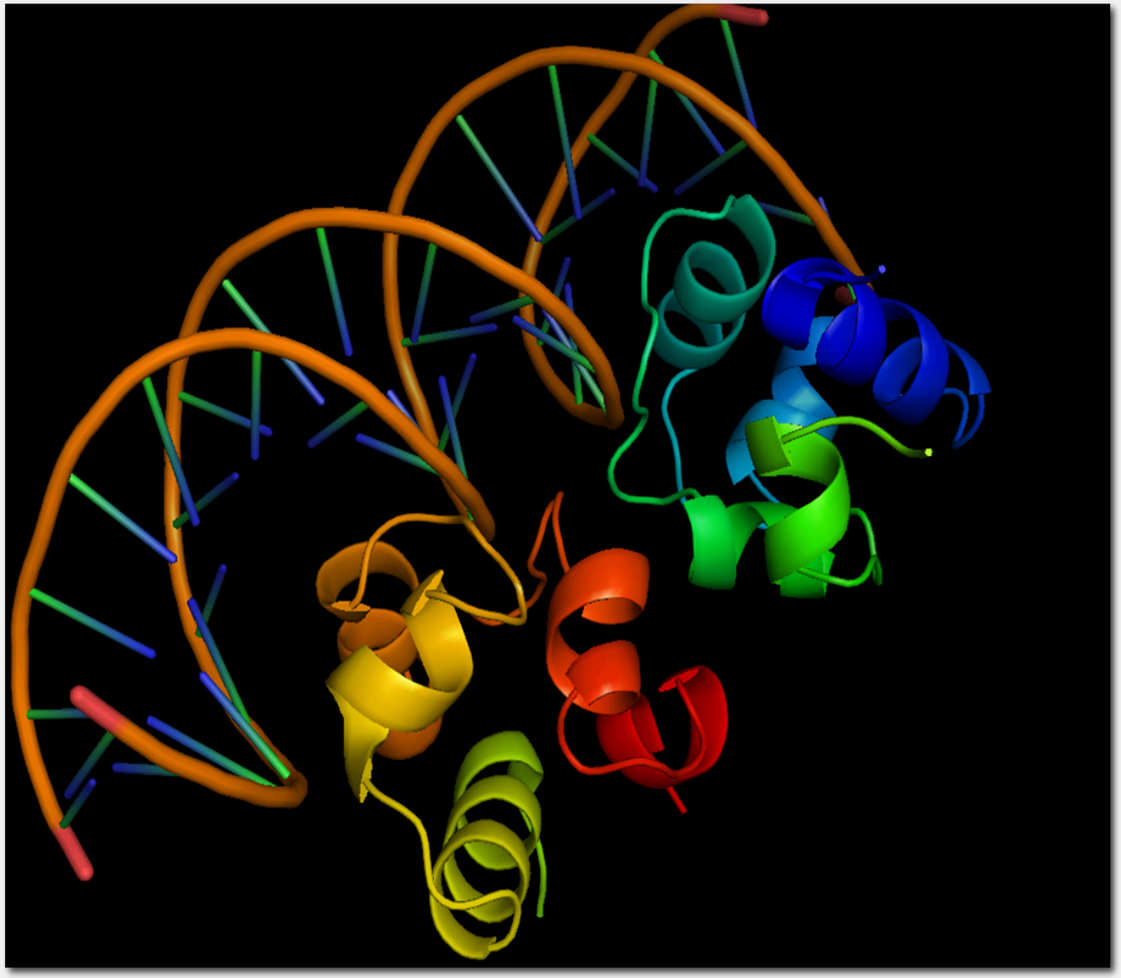
DNA (13/15)



DNA (14/15)



DNA (15/15)



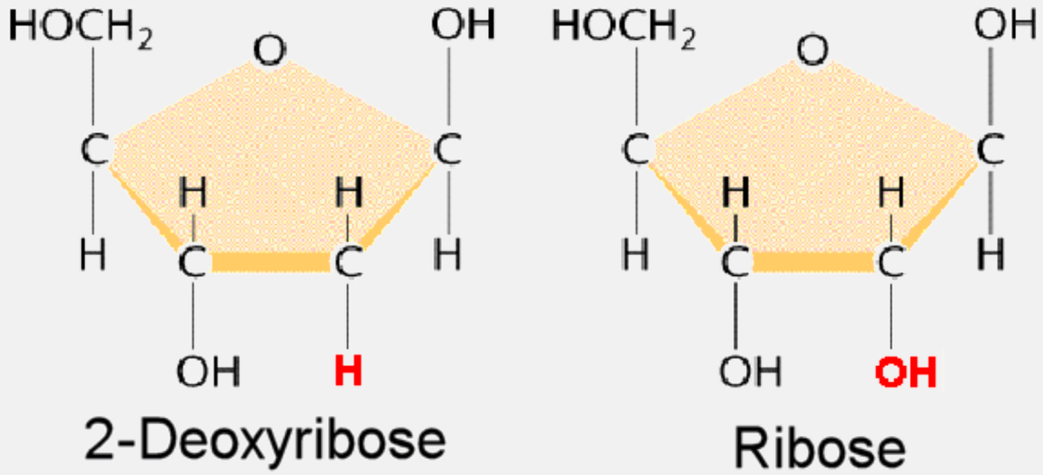
RNA structure and function

- RNA contains ribose and normally uses uracil in place of thymine.
- Many cellular RNAs are single strands that fold through internal base pairing.
- Double-stranded RNA also exists.
- Folded RNA can have structural, regulatory, or catalytic functions.

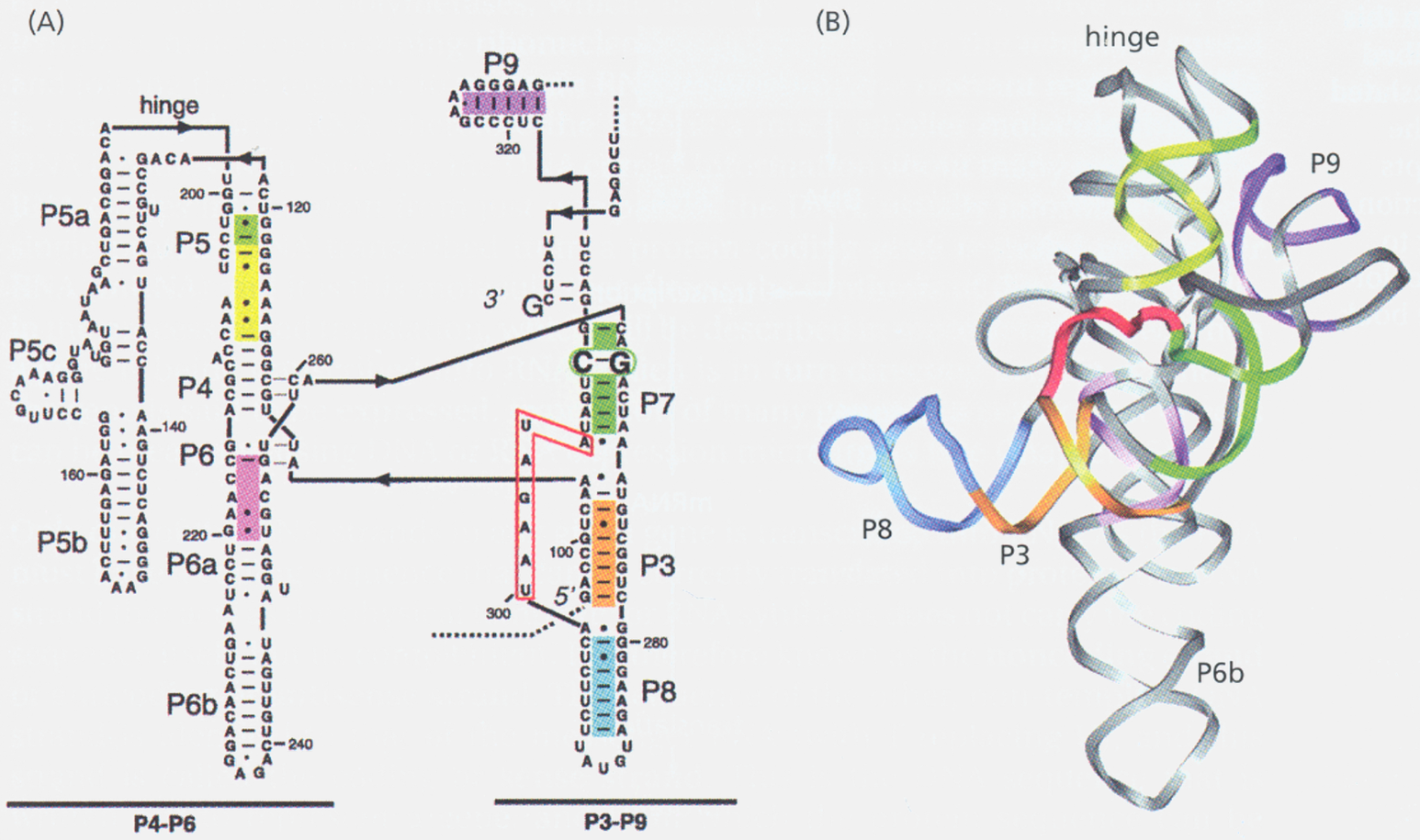
RNA secondary structure

Ribose has a **2′ hydroxyl group**; deoxyribose has hydrogen at that position.

RNA can form stable double helices, usually with A-form geometry. The 2′ hydroxyl affects geometry and chemical stability; it does not prevent double-stranded regions.



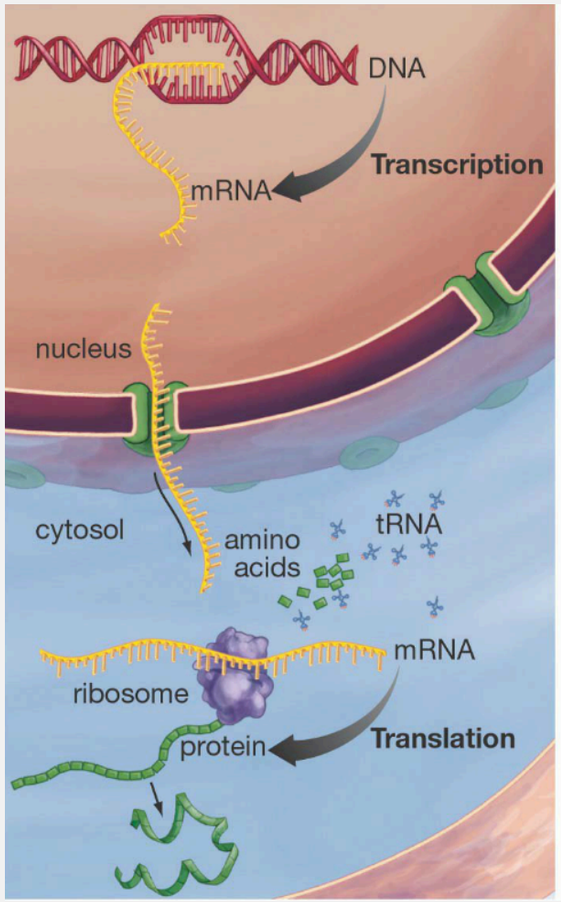
RNA Secondary structure



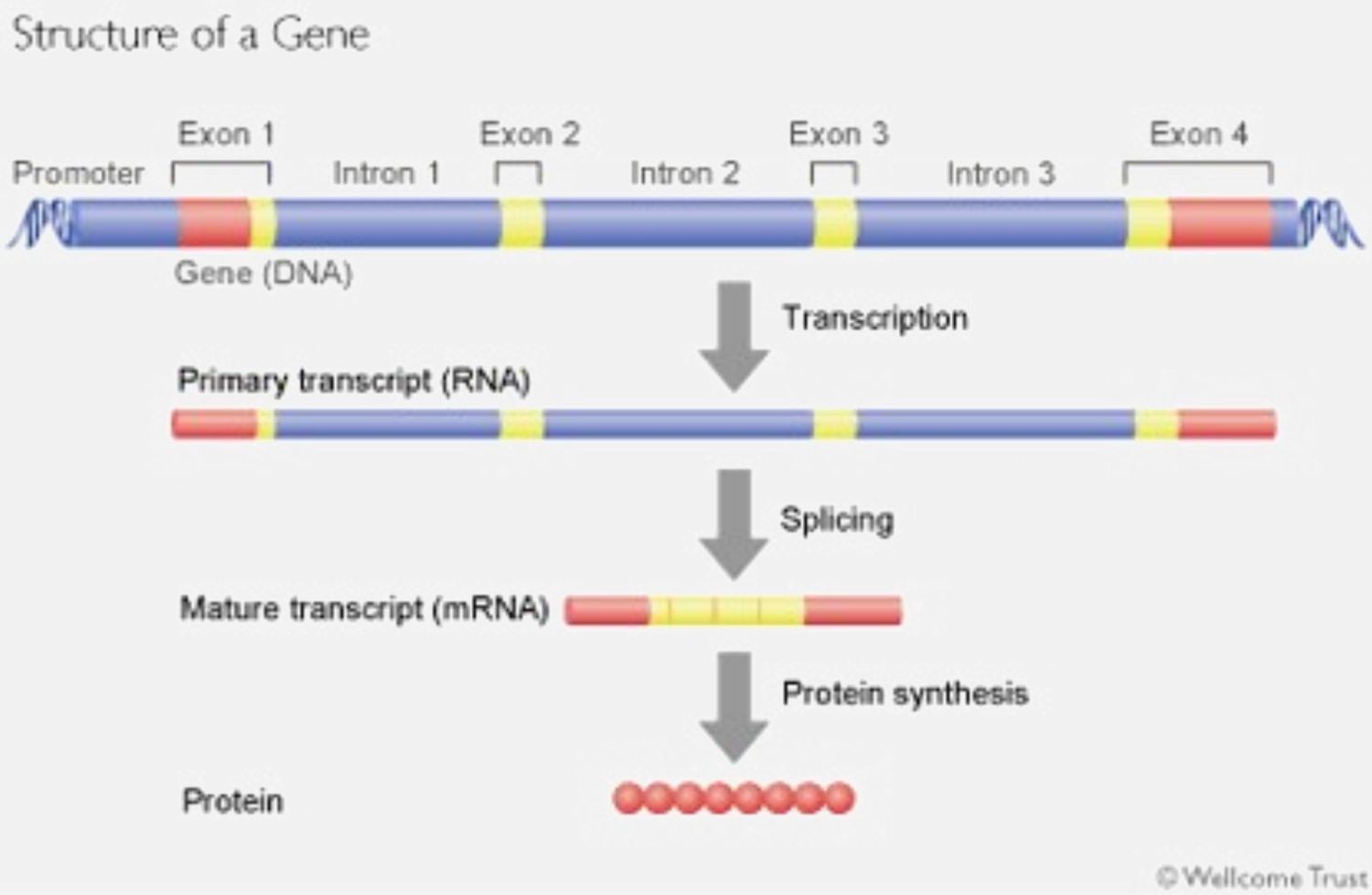
Example of three dimensional structure of RNA (Tetrahymena ribozyme)

- **Transcription:** a DNA sequence serves as a template for RNA synthesis.
- **Translation:** ribosomes decode an mRNA sequence to synthesise a polypeptide.
- Many RNAs function without being translated.

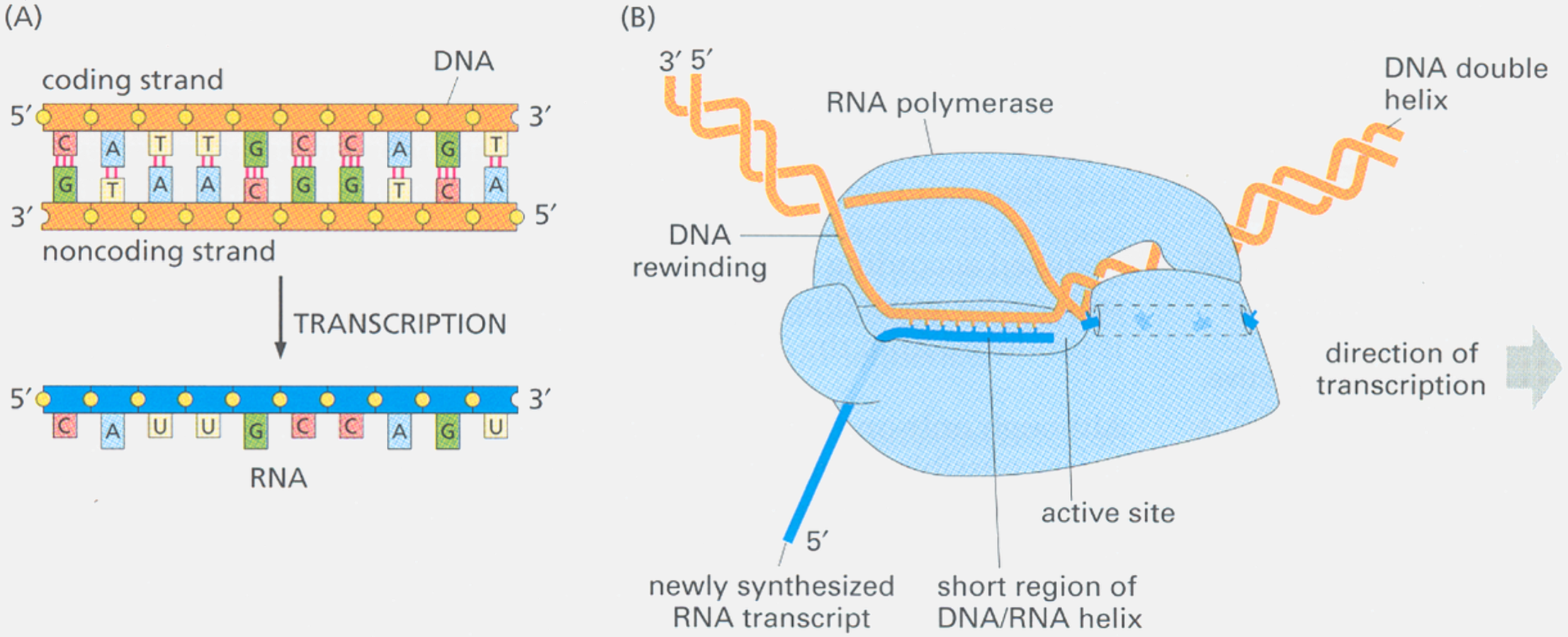
Reverse transcription transfers information from RNA to DNA. It does not imply sequence-information transfer from protein back to nucleic acid.



The Central Dogma



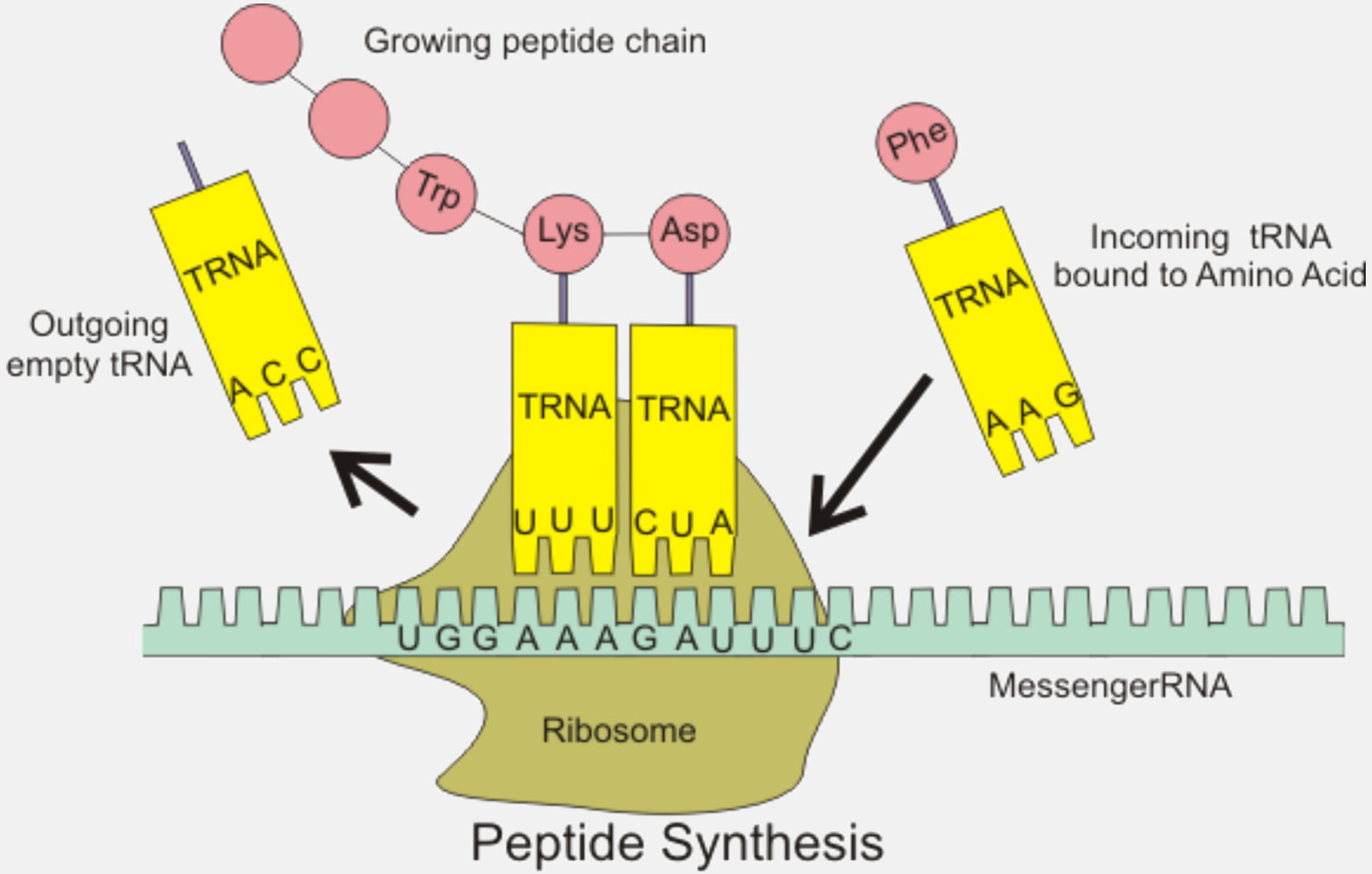
The Central Dogma



Translation

- The basic process of protein production is addition of one amino acid at a time to the end of a protein.
- This operation is performed by a ribosome.
- The ribosome molecules translate a RNA code to a specific sequence of amino acids.
- tRNAs are small noncoding RNA chains (74-93 nucleotides) that transport amino acids to the ribosome.

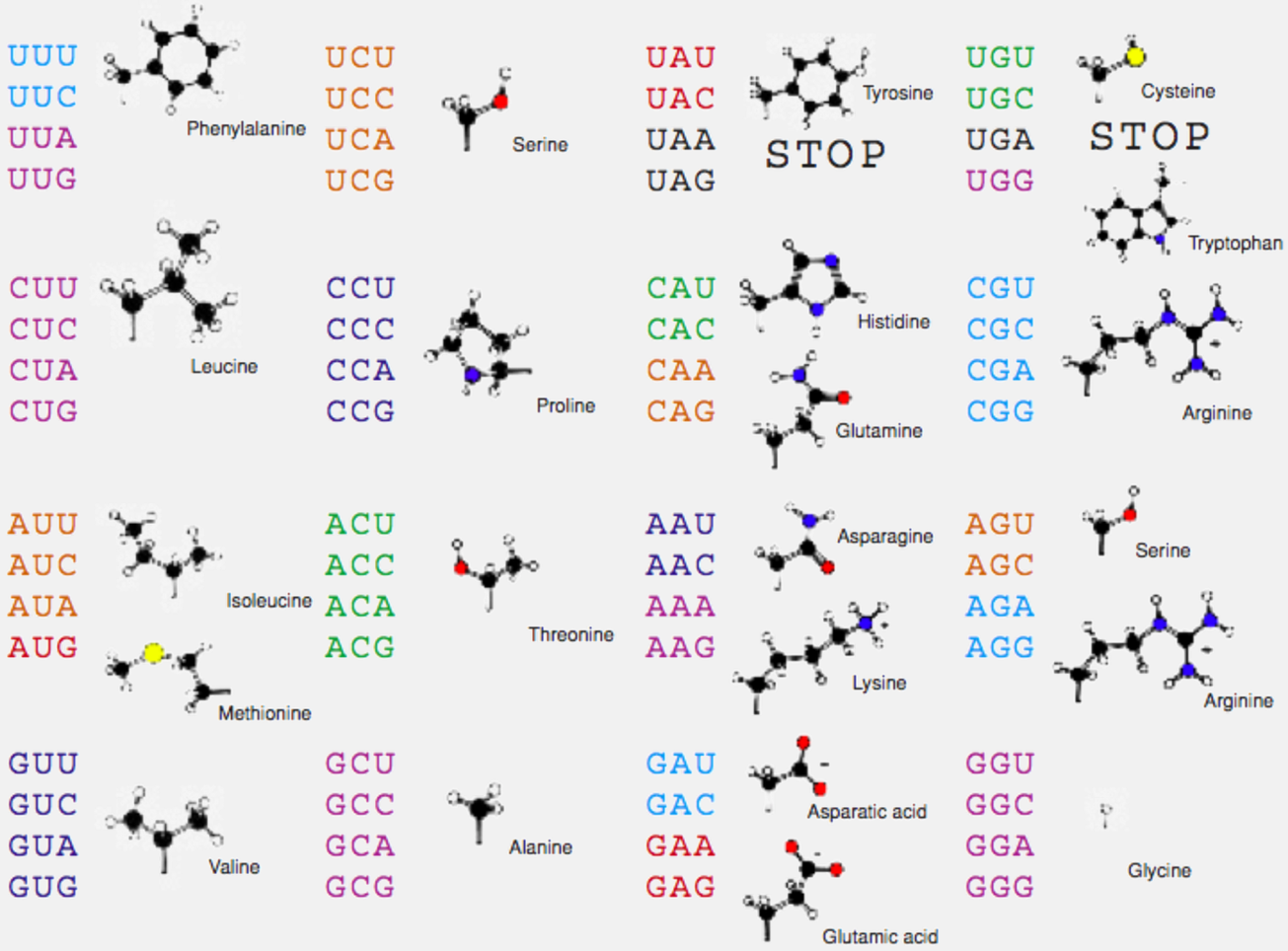
The Central Dogma



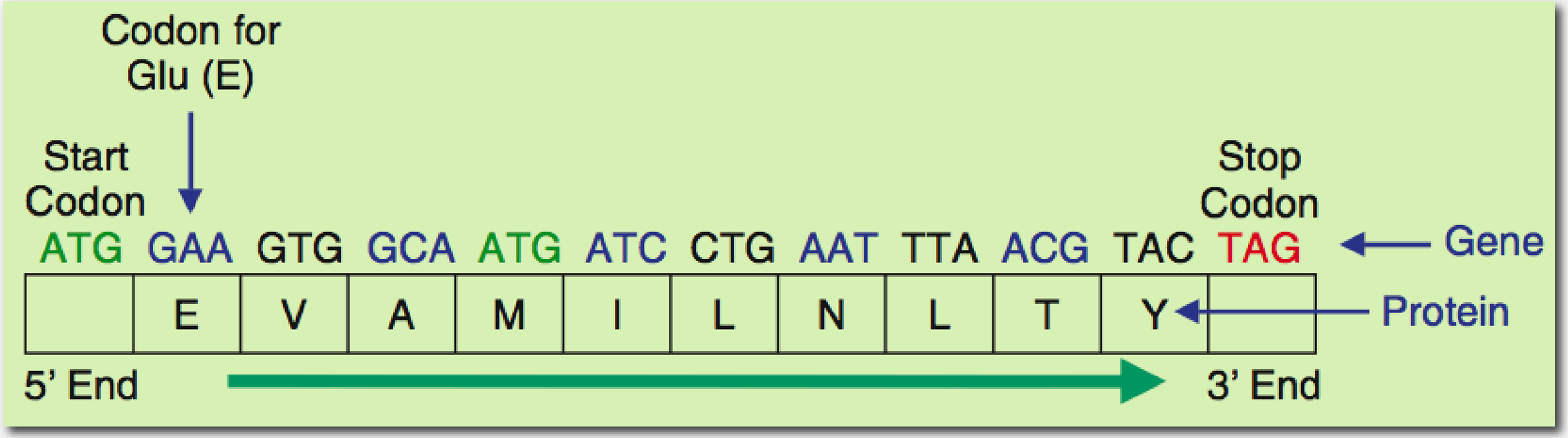
The Central Dogma

AAA: K (Lys)	*GAA:* E (Glu)	*TAA:* Stop	*CAA:* Q (Gln)
AAG: K (Lys)	*GAG:* E (Glu)	*TAG:* Stop	*CAG:* Q (Gln)
AAT: N (Asn)	*GAT:* D (Asp)	*TAT:* Y (Tyr)	*CAT:* H (His)
AAC: N (Asn)	*GAC:* D (Asp)	*TAC:* Y (Tyr)	*CAC:* H (His)
AGA: R (Arg)	*GGA:* G (Gly)	*TGA:* Stop	*CGA:* R (Arg)
AGG: R (Arg)	*GGG:* G (Gly)	*TGG:* W (Trp)	*CGG:* R (Arg)
AGT: S (Ser)	*GGT:* G (Gly)	*TGT:* C (Cys)	*CGT:* R (Arg)
AGC: S (Ser)	*GGC:* G (Gly)	*TGC:* C (Cys)	*CGC:* R (Arg)
ATA: I (Ile)	*GTA:* V (Val)	*TTA:* L (Leu)	*CTA:* L (Leu)
ATG: M (Met)	*GTG:* V (Val)	*TTG:* L (Leu)	*CTG:* L (Leu)
ATG = Start			
ATT: I (Ile)	*GTT:* V (Val)	*TTT:* F (Phe)	*CTT:* L (Leu)
ATC: I (Ile)	*GTC:* V (Val)	*TTC:* F (Phe)	*CTC:* L (Leu)
ACA: T (Thr)	*GCA:* A (Ala)	*TCA:* S (Ser)	*CCA:* P (Pro)
ACG: T (Thr)	*GCG:* A (Ala)	*TCG:* S (Ser)	*CCG:* P (Pro)
ACT: T (Thr)	*GCT:* A (Ala)	*TCT:* S (ser)	*CCT:* P (Pro)
ACC: T (Thr)	*GCC:* A (Ala)	*TCC:* S (Ser)	*CCC:* P (Pro)

The Central Dogma

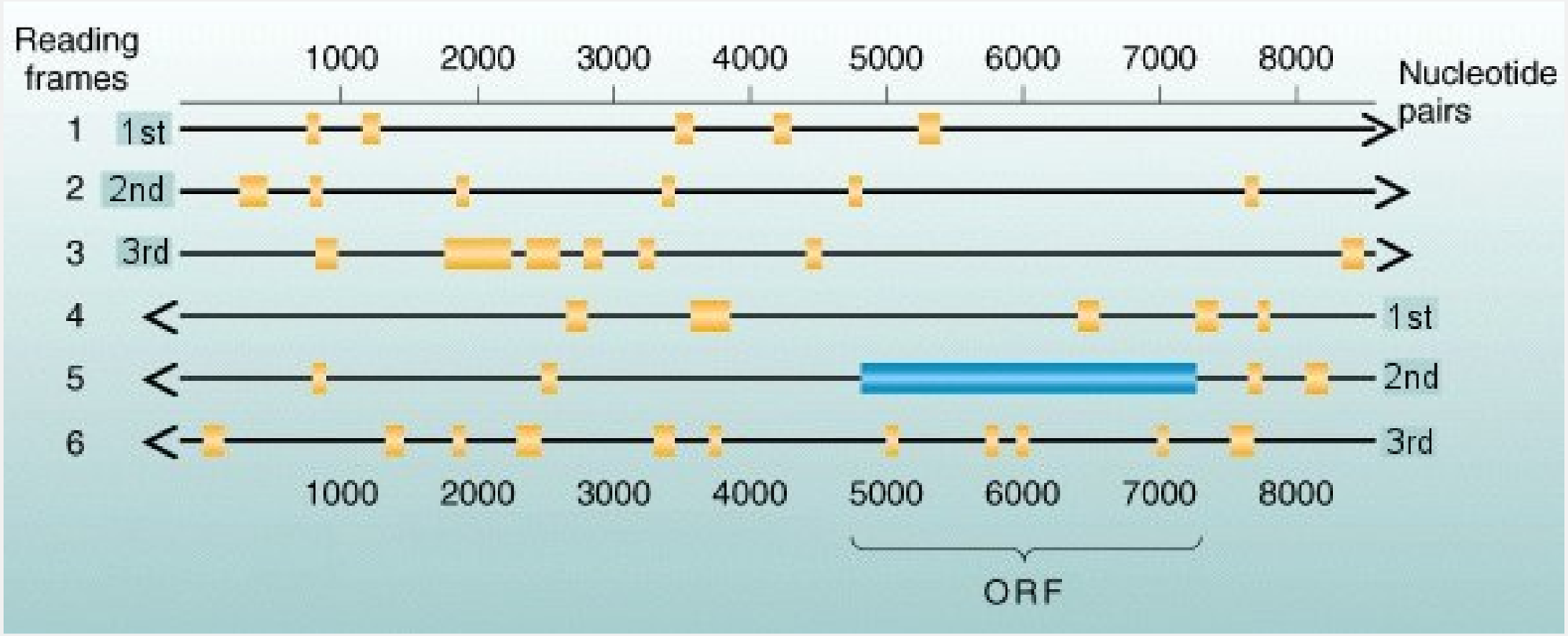


The Central Dogma



Open Reading Frames

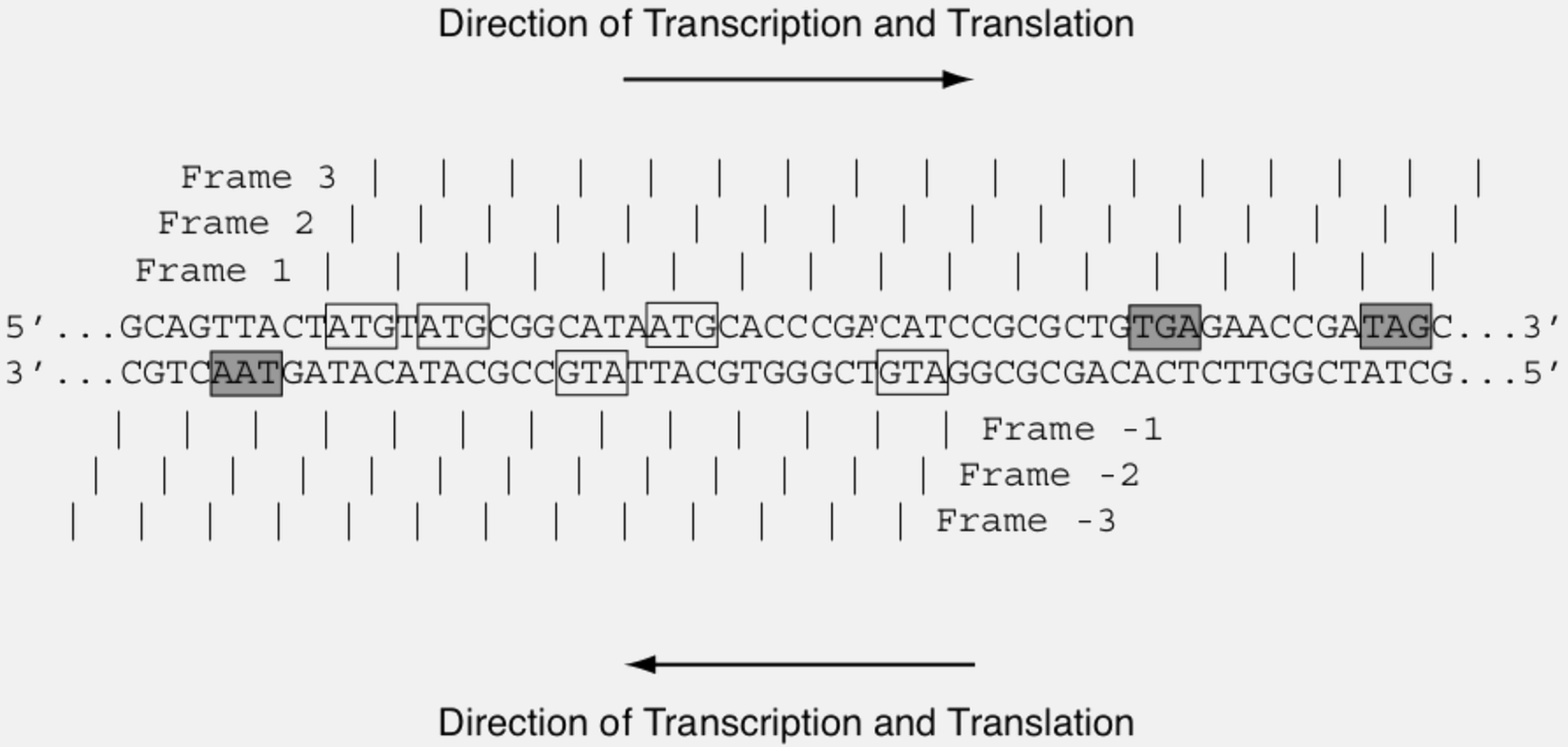
- Any piece of dsDNA can be read in six possible ways, or *reading frames*.
- For each of the two strands, either of which can be read in the *5' to the 3'* direction.
- Three possible starting points 1^{st},2^{nd},3^{rd} nuclotides.



Open Reading Frames

Start Codon: AUG (ATG in DNA)

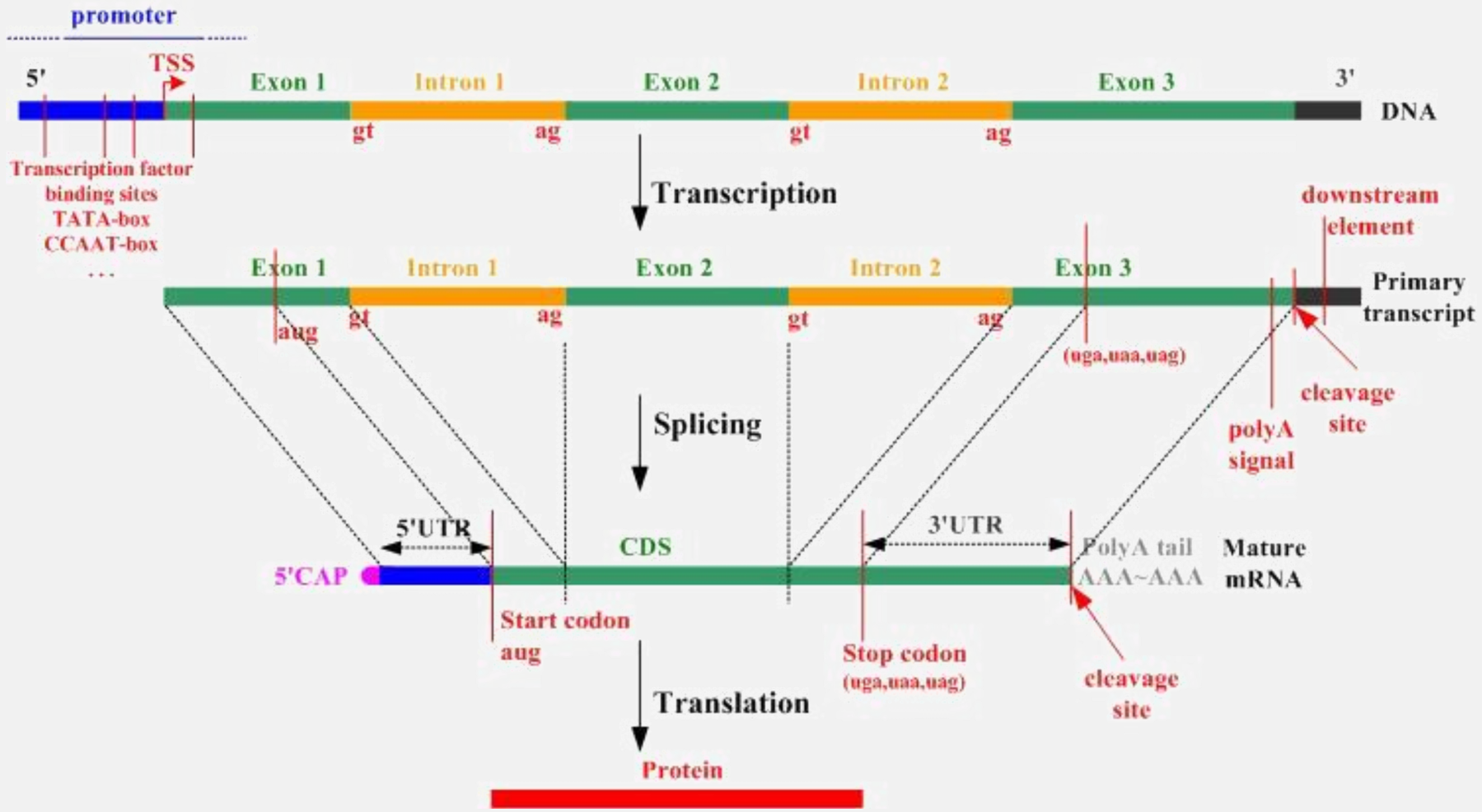
Stop Codons: UAA (TAA in DNA)
UAG (TAG in DNA)
UGA (TGA in DNA)



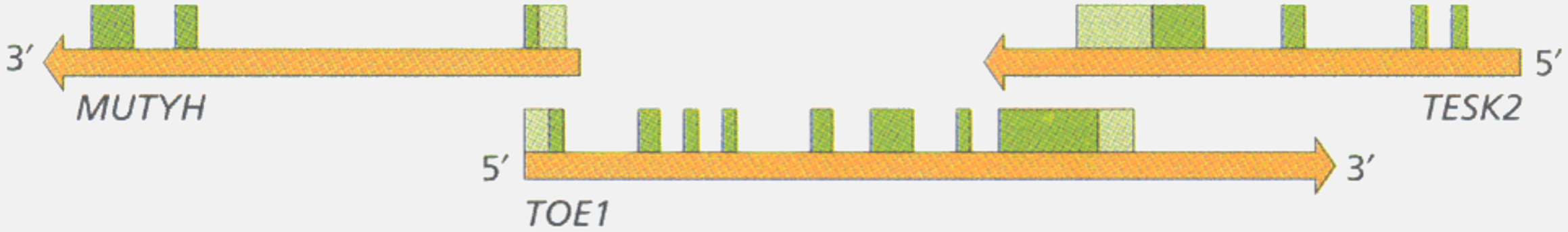
Open reading frames

- A reading frame defines how a sequence is divided into codons.
- An open reading frame has no internal in-frame stop codon between its start and end under the selected genetic code.
- Under the standard code, UAA, UAG, and UGA are stop codons.
- A long ORF is evidence for a possible coding region, but it does not prove expression or protein function.
- Short ORFs can also encode proteins.

Open Reading Frames

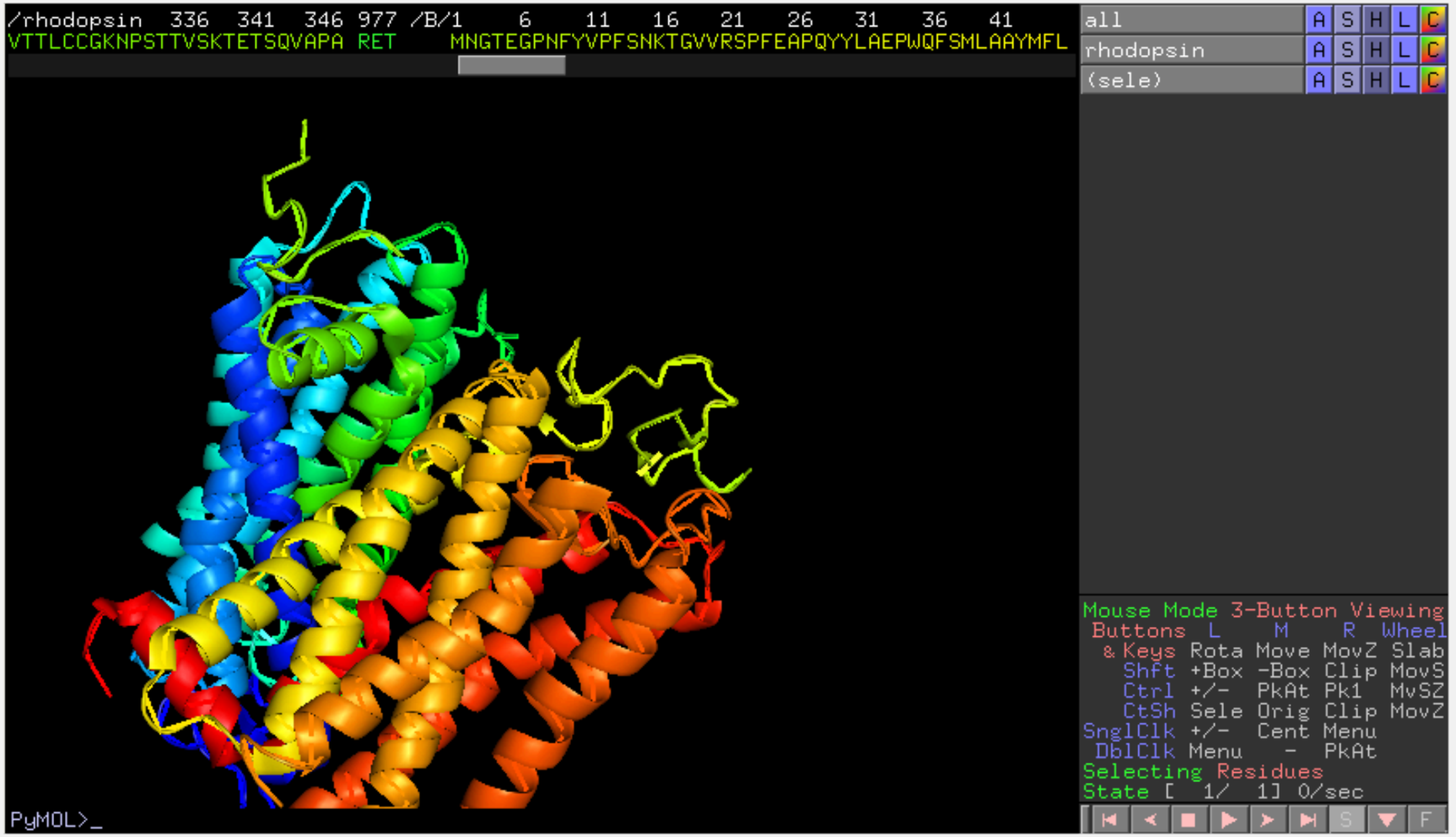


Overlapping genes

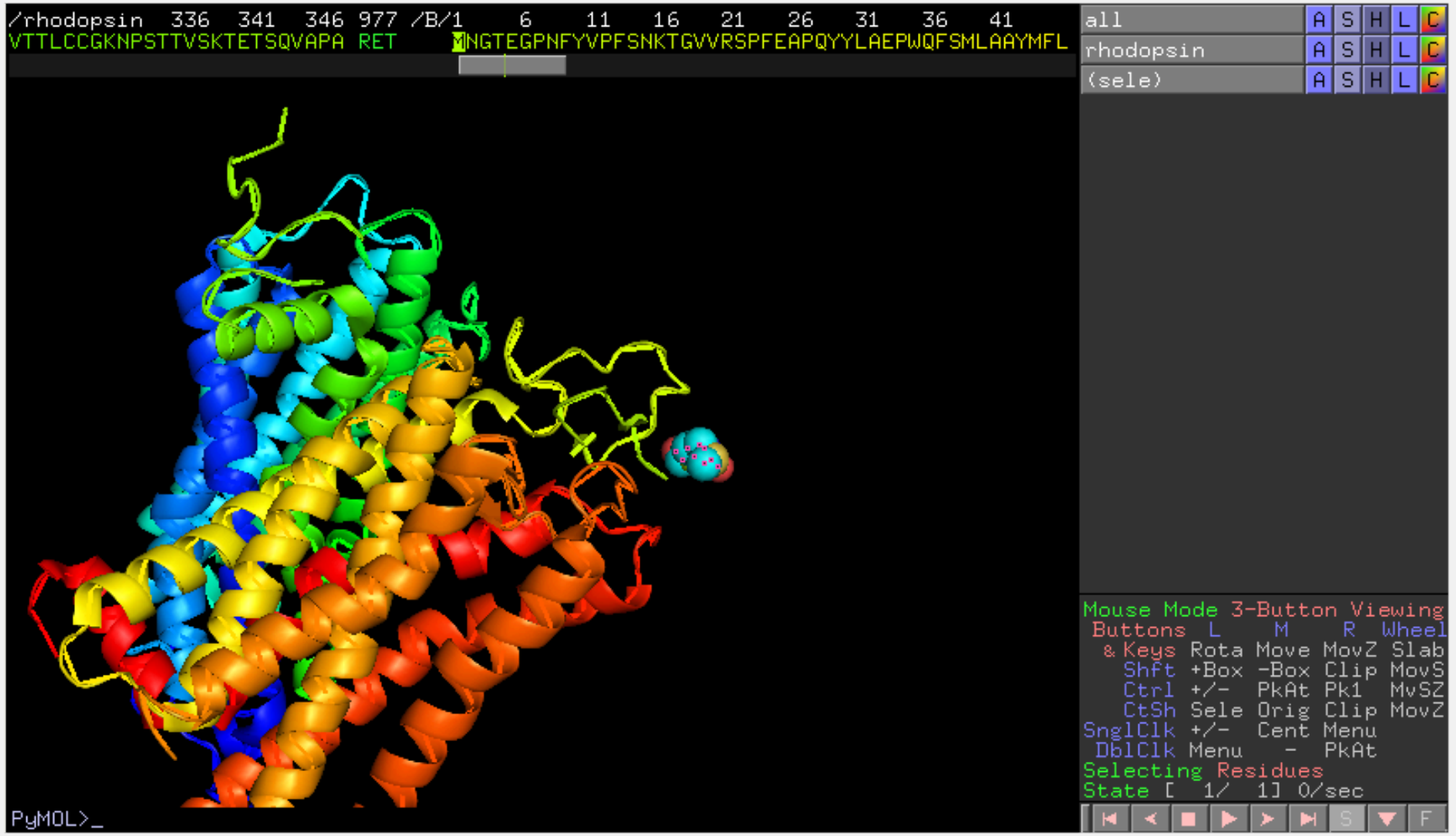


V. Veeramachenenu Mammalian overlapping genes: the comparative perspective. Genome Res. 14:280-286,2004

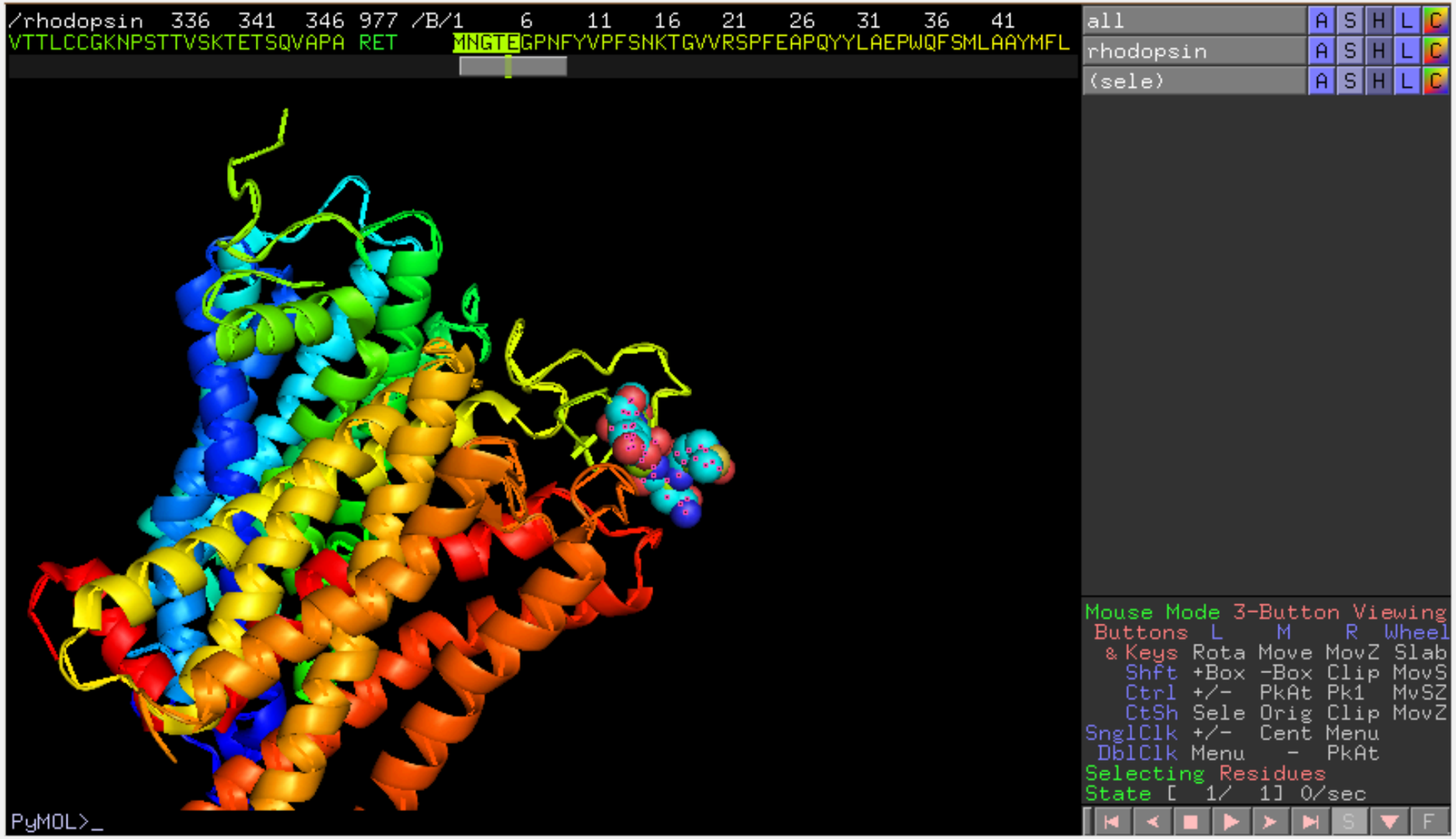
DNA (1/6)


/rhodopsin 336 341 346 977 /B/1 6 11 16 21 26 31 36 41
VTTLCCGKNPSTTVSKTETSQVAPA RET MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFL
all A S H L C
rhodopsin A S H L C
(sele) A S H L C
Mouse Mode 3-Button Viewing
Buttons L M R Wheel
& Keys Rota Move MovZ Slab
Shft +Box -Box Clip MovS
Ctrl +/- PkAt Pk1 MvSZ
CtSh Sele Orig Clip MovZ
SnglClk +/- Cent Menu
DblClk Menu - PkAt
Selecting Residues
State [1/ 1] 0/sec
PyMOL>_

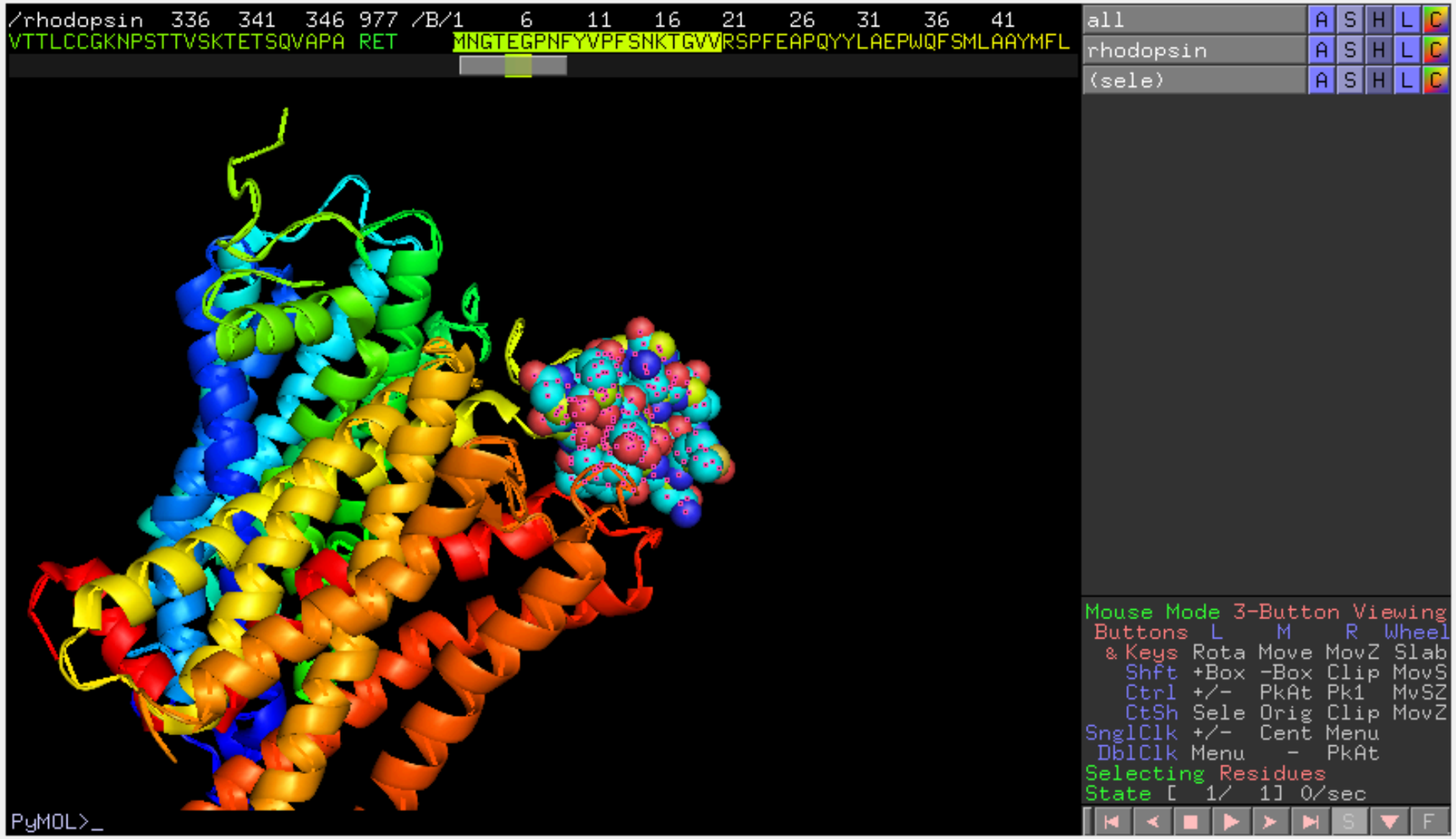
DNA (2/6)


/rhodopsin 336 341 346 977 /B/1 6 11 16 21 26 31 36 41
VTTLCCGKNPSTTVSKTETSQVAPA RET MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFL
all A S H L C
rhodopsin A S H L C
(sele) A S H L C
Mouse Mode 3-Button Viewing
Buttons L M R Wheel
& Keys Rota Move MovZ Slab
Shft +Box -Box Clip MovS
Ctrl +/- PkAt Pk1 MvSZ
CtSh Sele Orig Clip MovZ
SnglClk +/- Cent Menu
DblClk Menu - PkAt
Selecting Residues
State [1/ 1] 0/sec
PyMOL>_

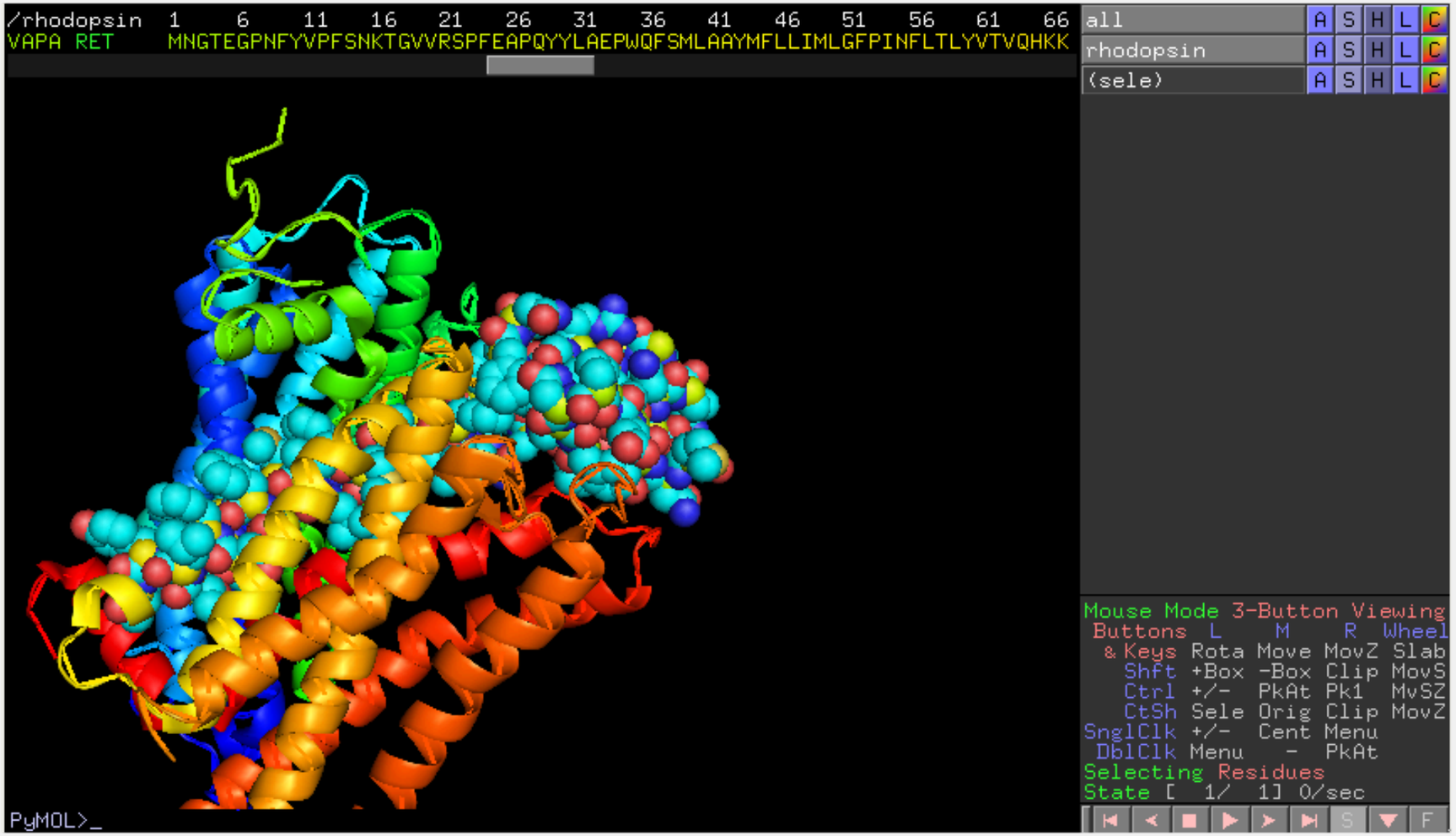
DNA (3/6)


/rhodopsin 336 341 346 977 /B/1 6 11 16 21 26 31 36 41
VTTLCCGKNPSTTVSKTETSQVAPA RET MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFL
all A S H L C
rhodopsin A S H L C
(sele) A S H L C
Mouse Mode 3-Button Viewing
Buttons L M R Wheel
& Keys Rota Move MovZ Slab
Shft +Box -Box Clip MovS
Ctrl +/- PkAt Pk1 MvSZ
CtSh Sele Orig Clip MovZ
SnglClk +/- Cent Menu
DblClk Menu - PkAt
Selecting Residues
State [1/ 1] 0/sec
PyMOL>_

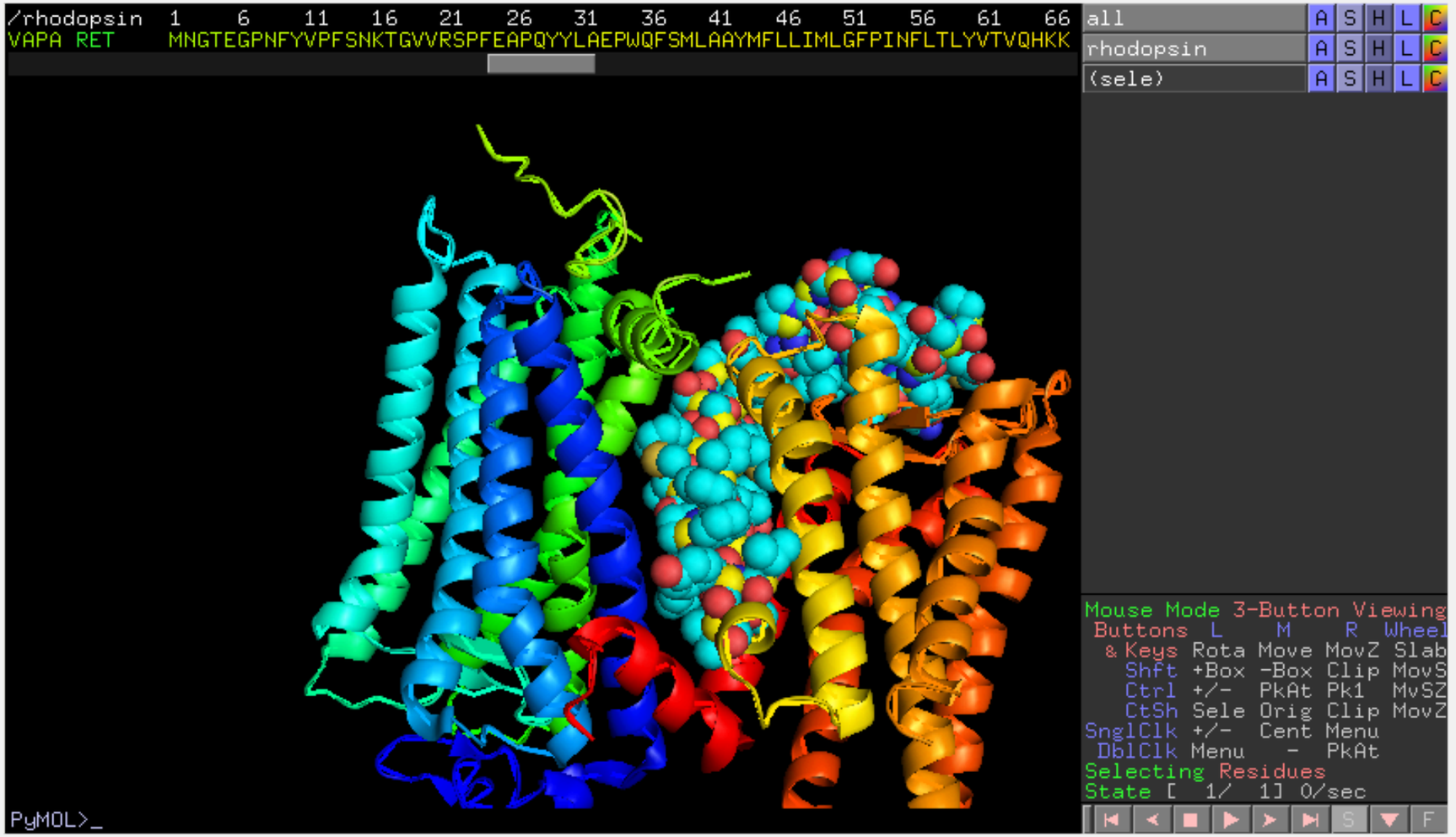
DNA (4/6)


/rhodopsin 336 341 346 977 /B/1 6 11 16 21 26 31 36 41
VTTLCCGKNPSTTVSKTETSQVAPA RET MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFL
all A S H L C
rhodopsin A S H L C
(sele) A S H L C
Mouse Mode 3-Button Viewing
Buttons L M R Wheel
& Keys Rota Move MovZ Slab
Shft +Box -Box Clip MovS
Ctrl +/- PkAt Pk1 MvSZ
CtSh Sele Orig Clip MovZ
SnglClk +/- Cent Menu
DblClk Menu - PkAt
Selecting Residues
State [1/ 1] 0/sec
PyMOL>_

DNA (5/6)


/rhodopsin 1 6 11 16 21 26 31 36 41 46 51 56 61 66
VAPA RET MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIMLGFPINFLTLYVTVQHKK
all A S H L C
rhodopsin A S H L C
(sele) A S H L C
Mouse Mode 3-Button Viewing
Buttons L M R Wheel
& Keys Rota Move MovZ Slab
Shft +Box -Box Clip MovS
Ctrl +/- PkAt Pk1 MvSZ
CtSh Sele Orig Clip MovZ
SnglClk +/- Cent Menu
DblClk Menu - PkAt
Selecting Residues
State [1/ 1] 0/sec
PyMOL>_

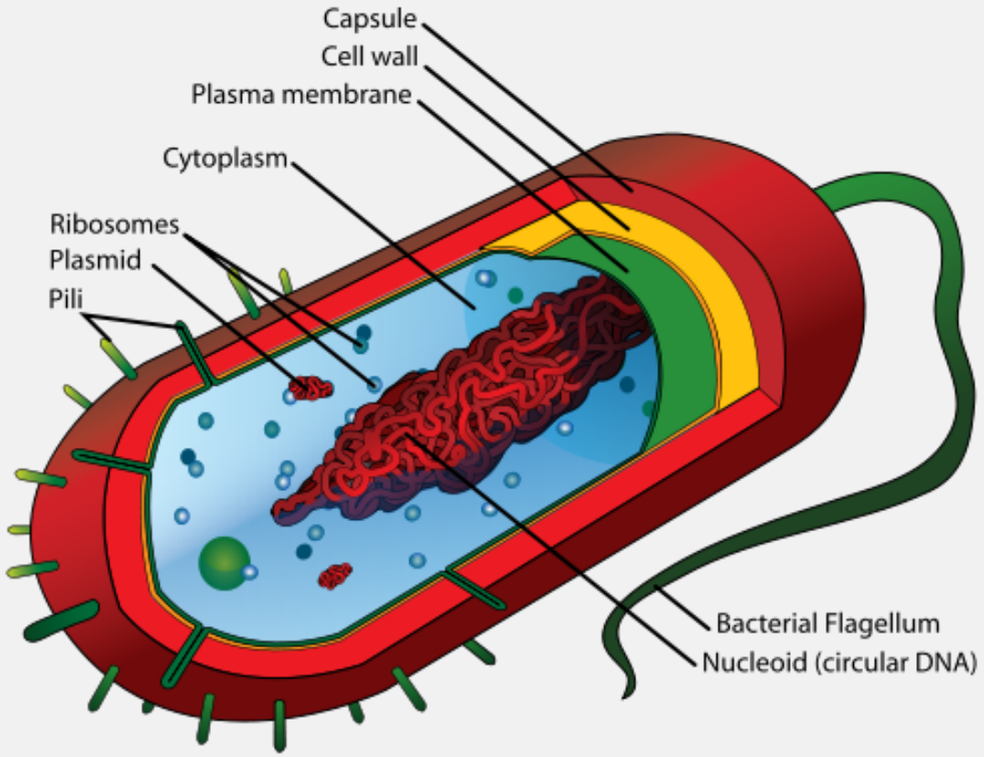
DNA (6/6)


/rhodopsin 1 6 11 16 21 26 31 36 41 46 51 56 61 66
VAPA RET MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIMLGFPINFLTLYVTVQHKK
all A S H L C
rhodopsin A S H L C
(sele) A S H L C
Mouse Mode 3-Button Viewing
Buttons L M R Wheel
& Keys Rota Move MovZ Slab
Shft +Box -Box Clip MovS
Ctrl +/- PkAt Pk1 MvSZ
CtSh Sele Orig Clip MovZ
SnglClk +/- Cent Menu
DblClk Menu - PkAt
Selecting Residues
State [1/ 1] 0/sec
PyMOL>_

Prokaryote vs. Eukaryote

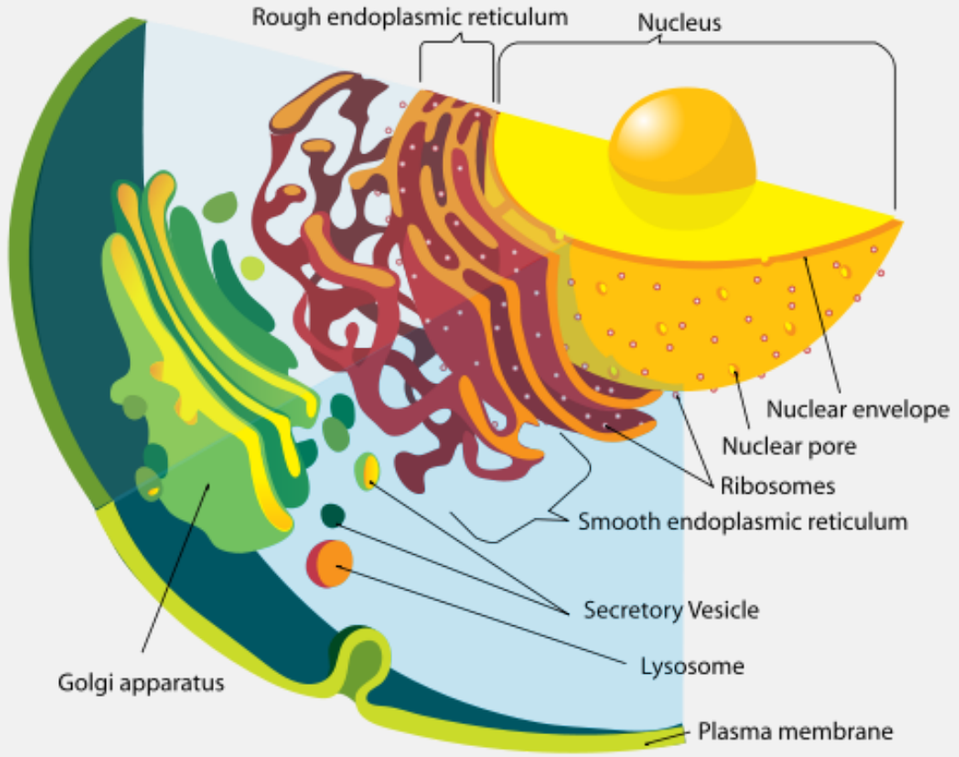
Prokaryote

- pro-karyon (before kernel).
- The prokaryote are a group of organisms that lack a cell nucleus , or any other *membrane-bound* organelles.



Eukaryote

- eu-karyon (good kernel).
- A eukaryote is an organism whose cells contain complex structures enclosed within membranes.



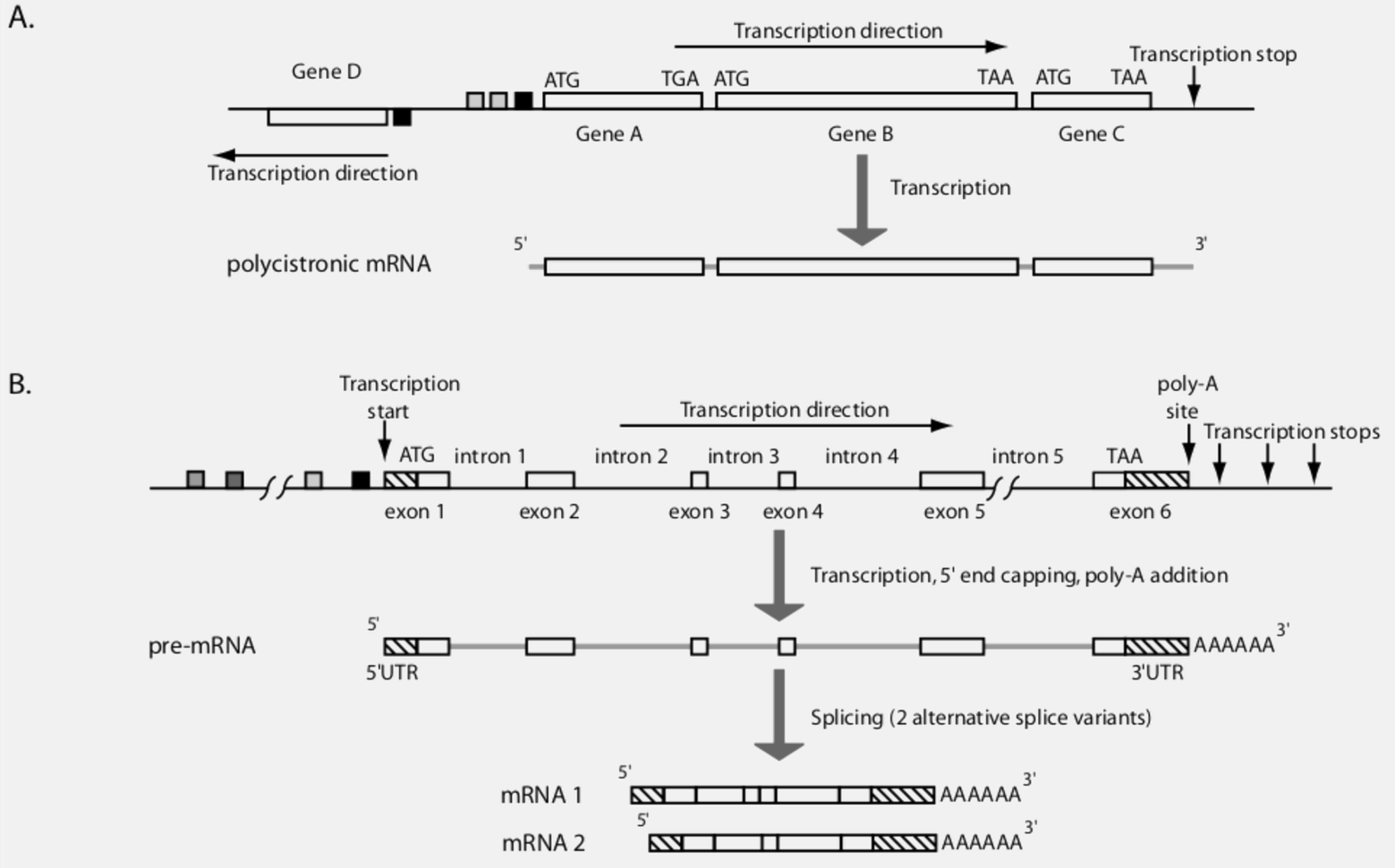
Prokaryote

- Genes in bacteria are normally close together.
- May be organized as operons.
- Or may be individually transcribed.
- The mRNA product of the transcription is ready for immediate translation.

Eukaryote

- Has transcription factor binding sites upstream of the promoter
- Promoter regions may be extensive, and introns may also be long.
- First and last exons contains regions that will not be translated (5’ UTR, 3’ UTR).
- The immediate product of transcription is a pre-mRNA that is modified at the 5’ and 3’ ends.
- The poly-A tail is added to the end of the message after transcription has occurred.
- Splicing removes the introns to produce mature mRNA species that are ready for translation
- Alternative splicing may or may not occur.

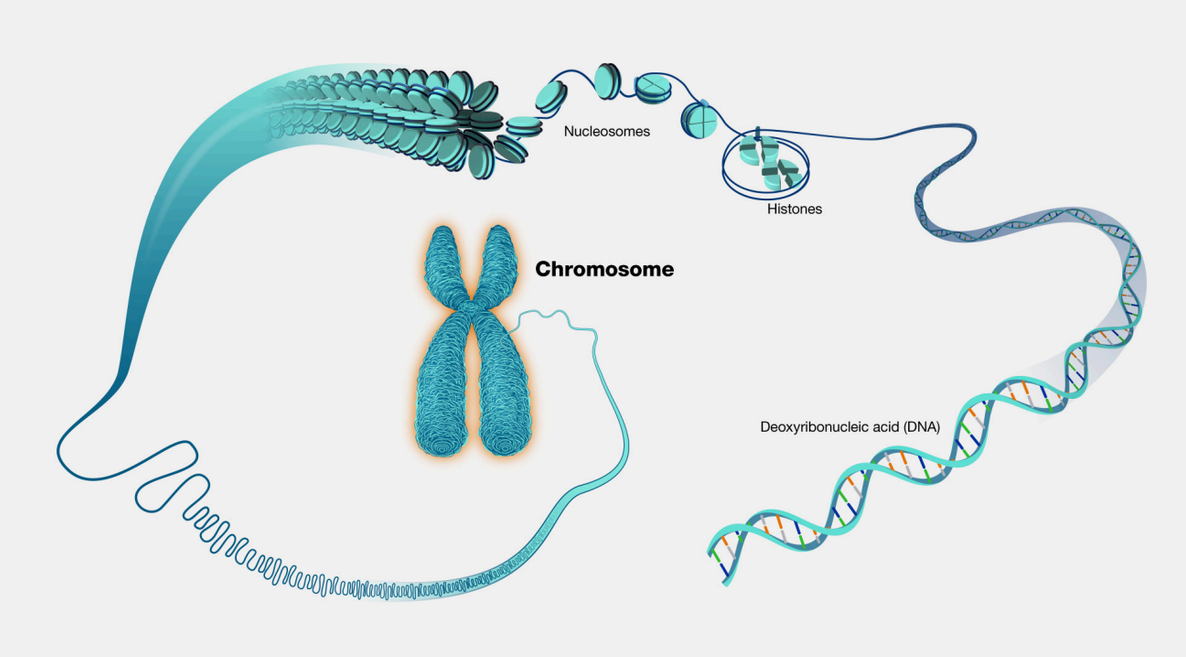
Prokaryote vs. Eukaryote



Coding sequence and ORF

- A **coding sequence (CDS)** specifies the nucleotide sequence translated into a polypeptide.
- For an intron-containing eukaryotic gene, the CDS is assembled from the coding portions of its exons.
- Untranslated regions belong to the transcript but not to the CDS.
- An ORF is a sequence property. A CDS is an annotation supported by biological evidence or prediction.

DNA packaging: nanometres to metres



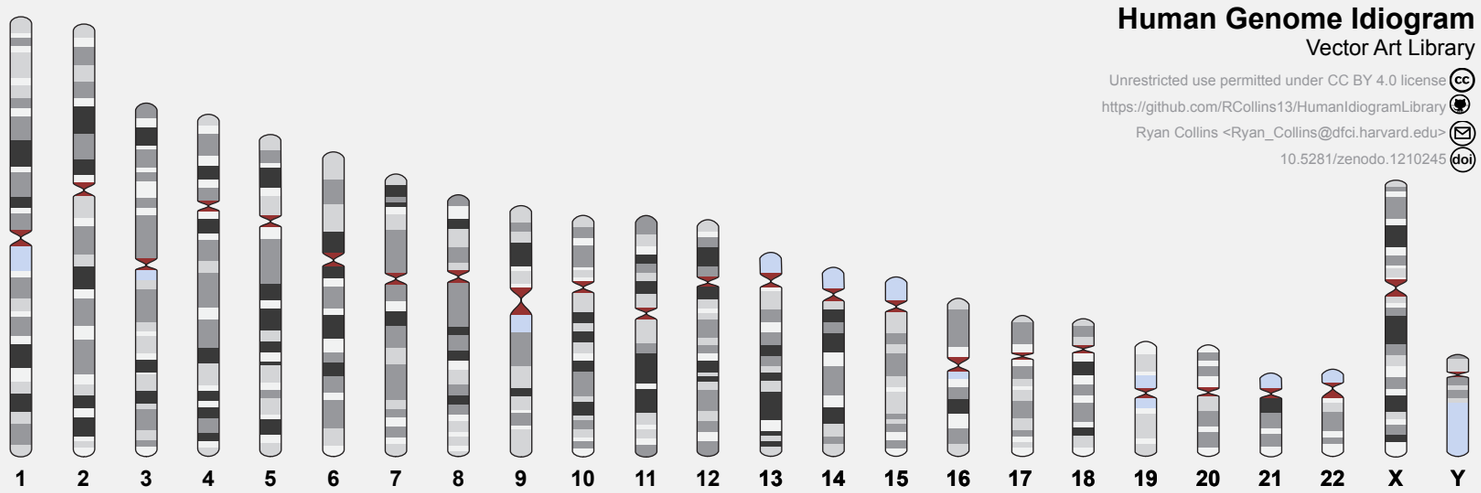
Structure	Approximate scale
DNA double helix	**2 nm** wide
Nucleosome core	**11 nm** across
DNA around one core	**147 bp**, ~1.7 turns
Human cell nucleus	**6 µm** across*

1 µm = 1,000 nm

About 2 metres of DNA fit inside one nucleus. A typical diploid cell before DNA replication contains ~6 billion bp; 6×10^9 bp × 0.34 nm/bp ≈ 2 m.

Figure: NIH / NHGRI. *Nuclear size varies by cell type. Schematic, not a fixed folding ladder: chromatin organisation is heterogeneous. The X-shaped example is a replicated chromosome with two sister chromatids.

The human genome: chromosome ideogram

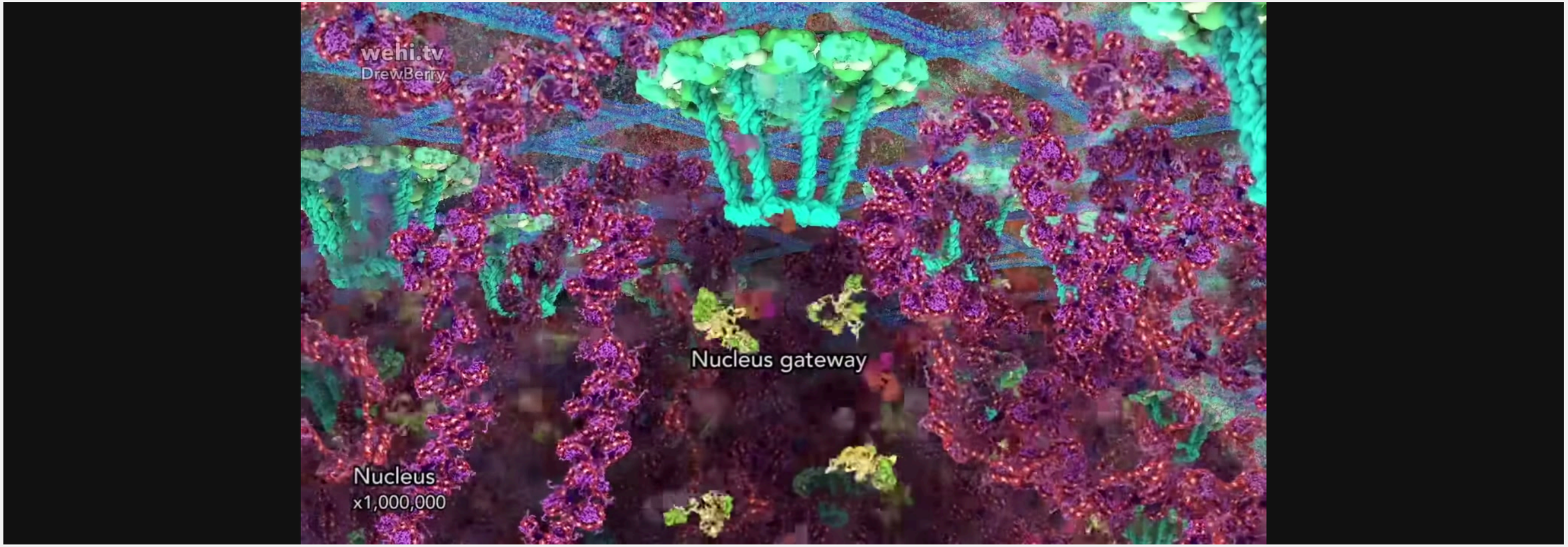


Chromosome complement	DNA amount	Sequence / extended DNA
23 chromosomes per typical gamete	**~3 billion bp** per haploid set	**Chr 1:** 249 Mb ≈ 8.5 cm
46 in a typical diploid cell	**~6 billion bp** before replication	**Chr 21:** 46.7 Mb ≈ 1.6 cm

24 types are shown (1–22, X, Y), not the chromosomes of one cell. Bands are cytogenetic landmarks, not individual genes. **1 Mb = 1 million bp.**

Figure: Ryan L. Collins, Human Idiogram Library, CC BY 4.0, unchanged. Relative lengths/bands are schematic. Numerical lengths: NCBI / GRC, GRCh38.

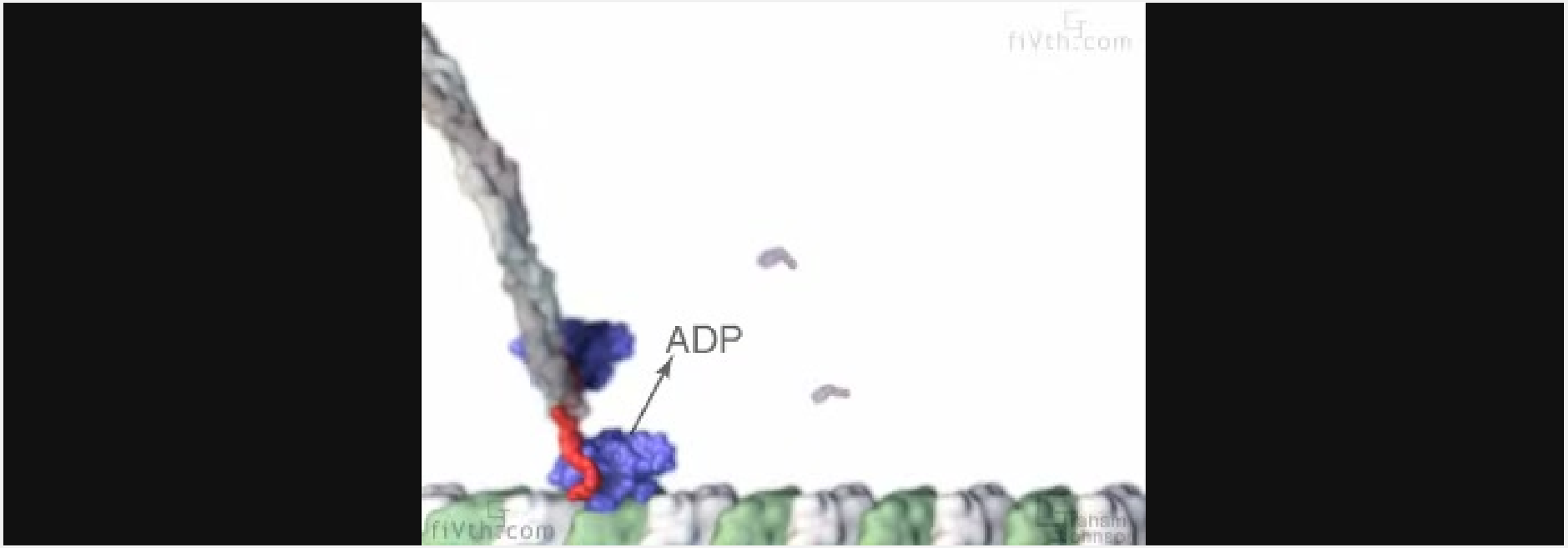
DNA at molecular scale



Watch how the view moves from the nucleus to DNA and its molecular machinery.

WEHImovies · 7:19 · No narration; original English labels. Soundtrack retained.

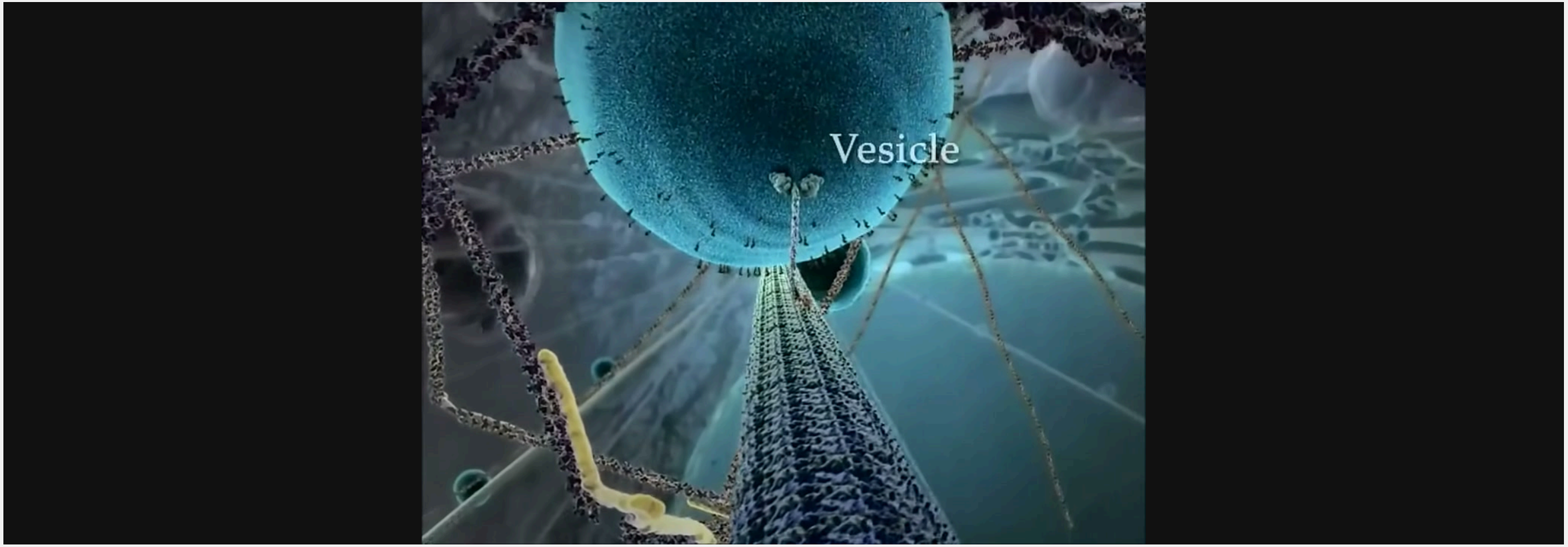
Kinesin: how a molecular motor walks



Follow the two heads and the ATP → ADP cycle.

Graham Johnson · 2:09 · English subtitles: automatic transcript with terminology and segmentation corrected.

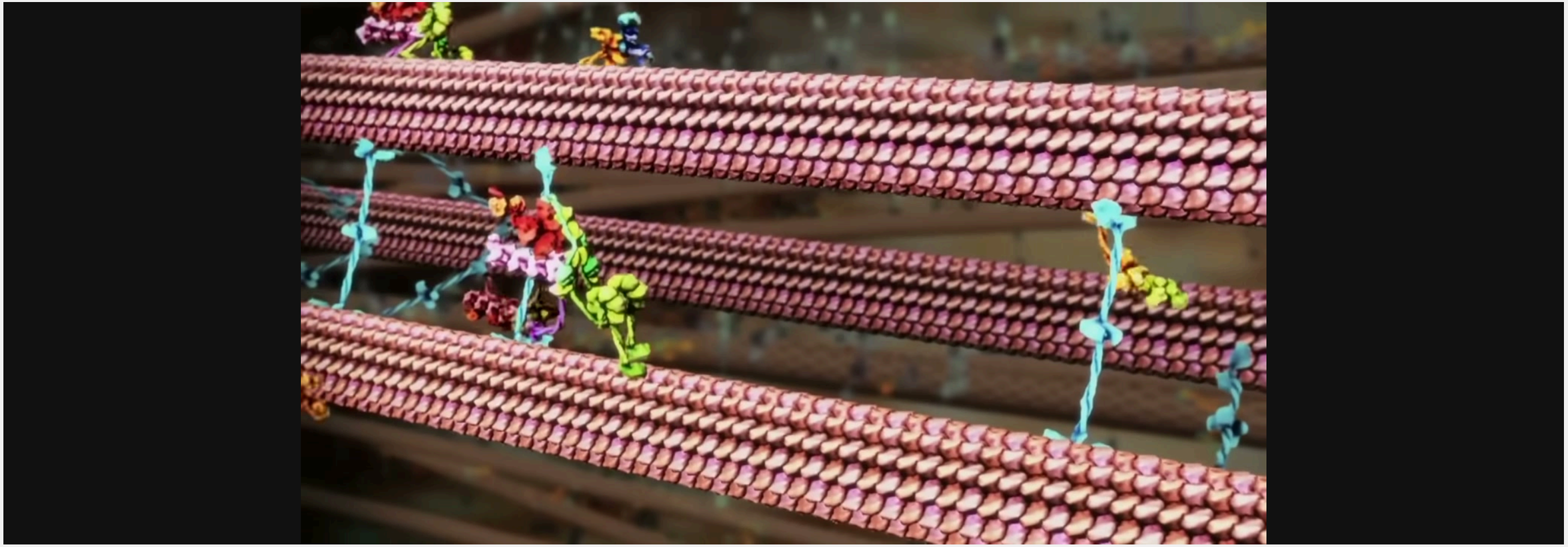
Kinesin: transporting cargo



Connect molecular stepping to transport at the scale of a cell.

em2134x · 0:22 · English viewing-guide captions (not a spoken transcript).

Your body’s molecular machines



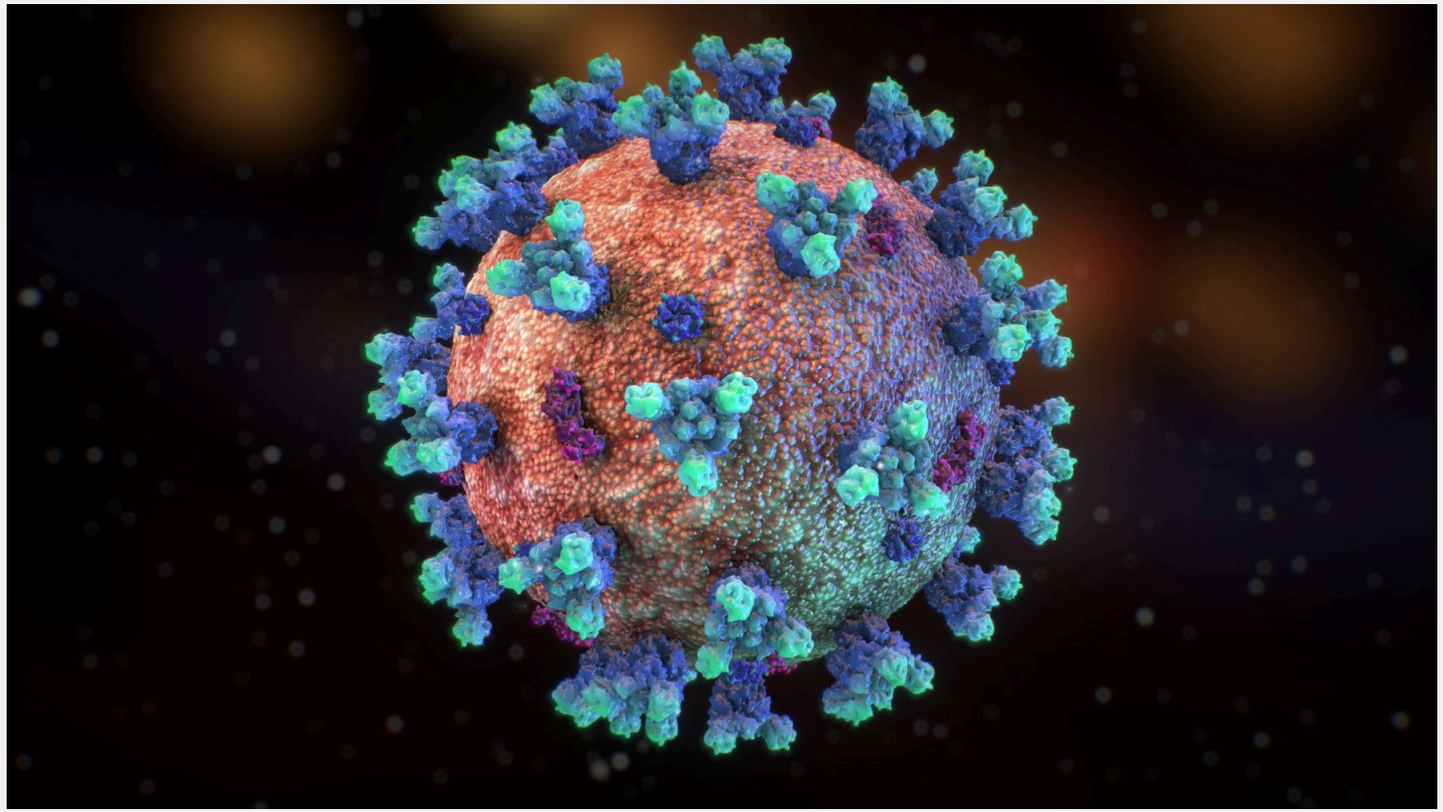
Watch how molecular machines cooperate during cell division.

Veritasium · 6:20 · English subtitles from the published caption track. Starts at 0:02.

Case study: SARS-CoV-2, a virus written in the same code

- A coronavirus is an RNA virus: its genome is a single positive-sense RNA strand of about 30,000 nucleotides — read directly by the host ribosome, no DNA step needed.
- Its ~30 genes encode the components you can see in the next 3D models: **spike (S)**, **envelope (E)** and **membrane (M)** proteins stud the lipid envelope; inside sits the **nucleocapsid (N)** wrapped around the RNA.
- Every structural protein is specified by an open reading frame — the same reading-frame logic we just saw applies to a virus.
- Sequencing this genome — and its mutations — is what made diagnostics, vaccines and variant tracking possible in weeks.

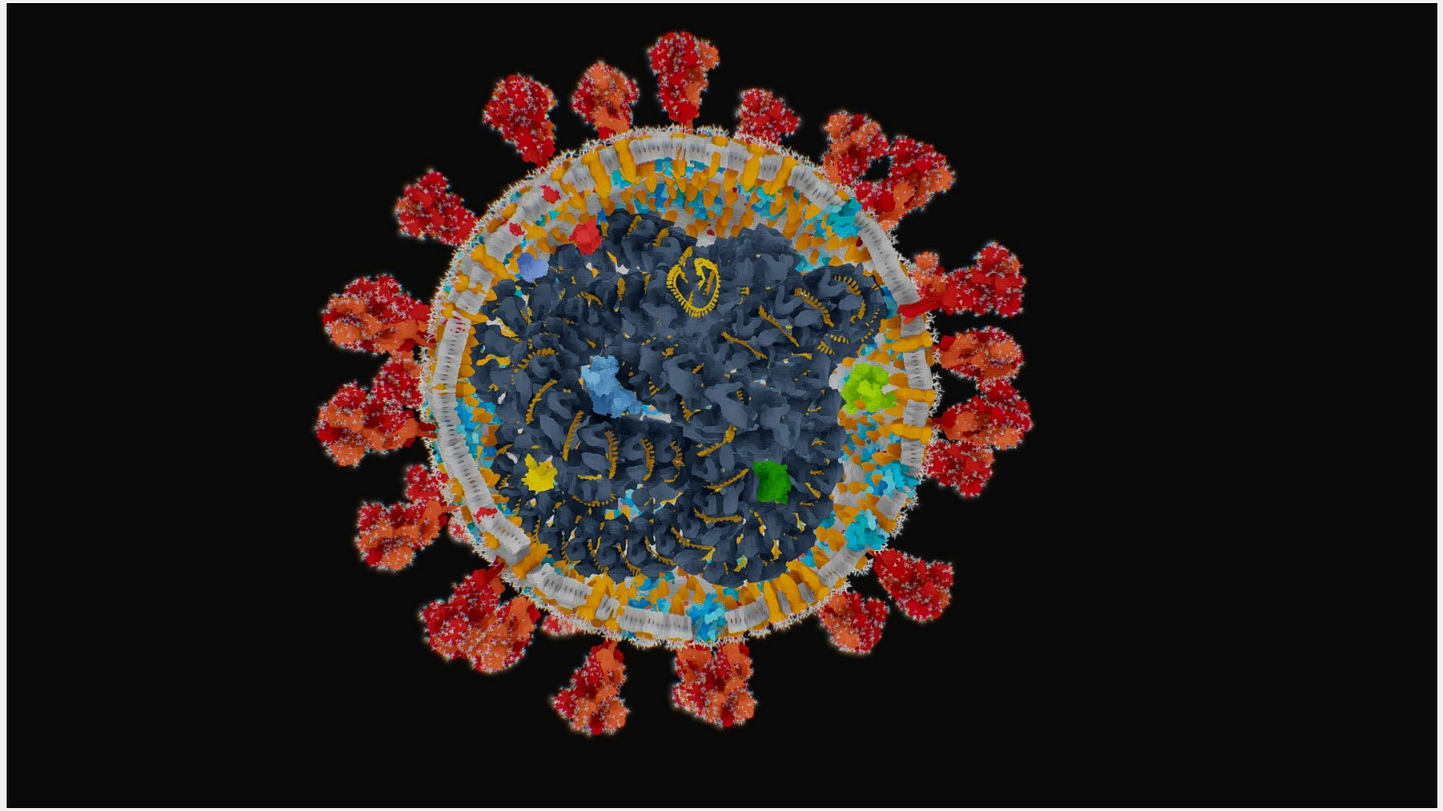
Explorable: the virion assembled from PDB structures



Rotate the virion; the protein distribution is research-based but stylised, assembled from published PDB structures of the S, E and M proteins.

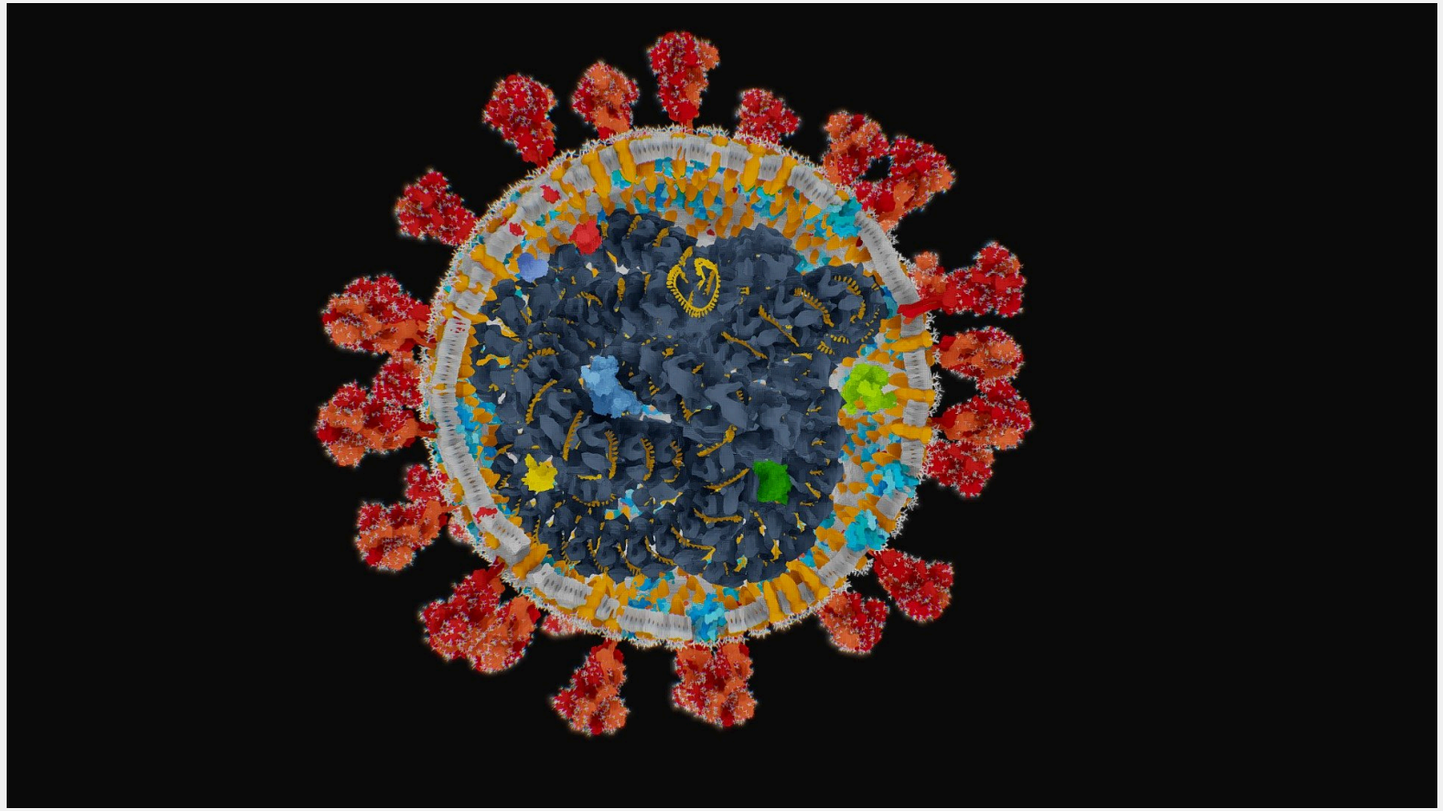
3D model: SARS-CoV-2 (COVID-19) · Fusion Medical Animation · S protein 6CRV, E protein 5X29, M protein 3I6G · via the Sketchfab viewer.

Anatomy of a coronavirus particle



3D model: corona virus (SARS-COV-2) · eMirage · Sketchfab · render still, unchanged. Interactive 3D version on the next slides.

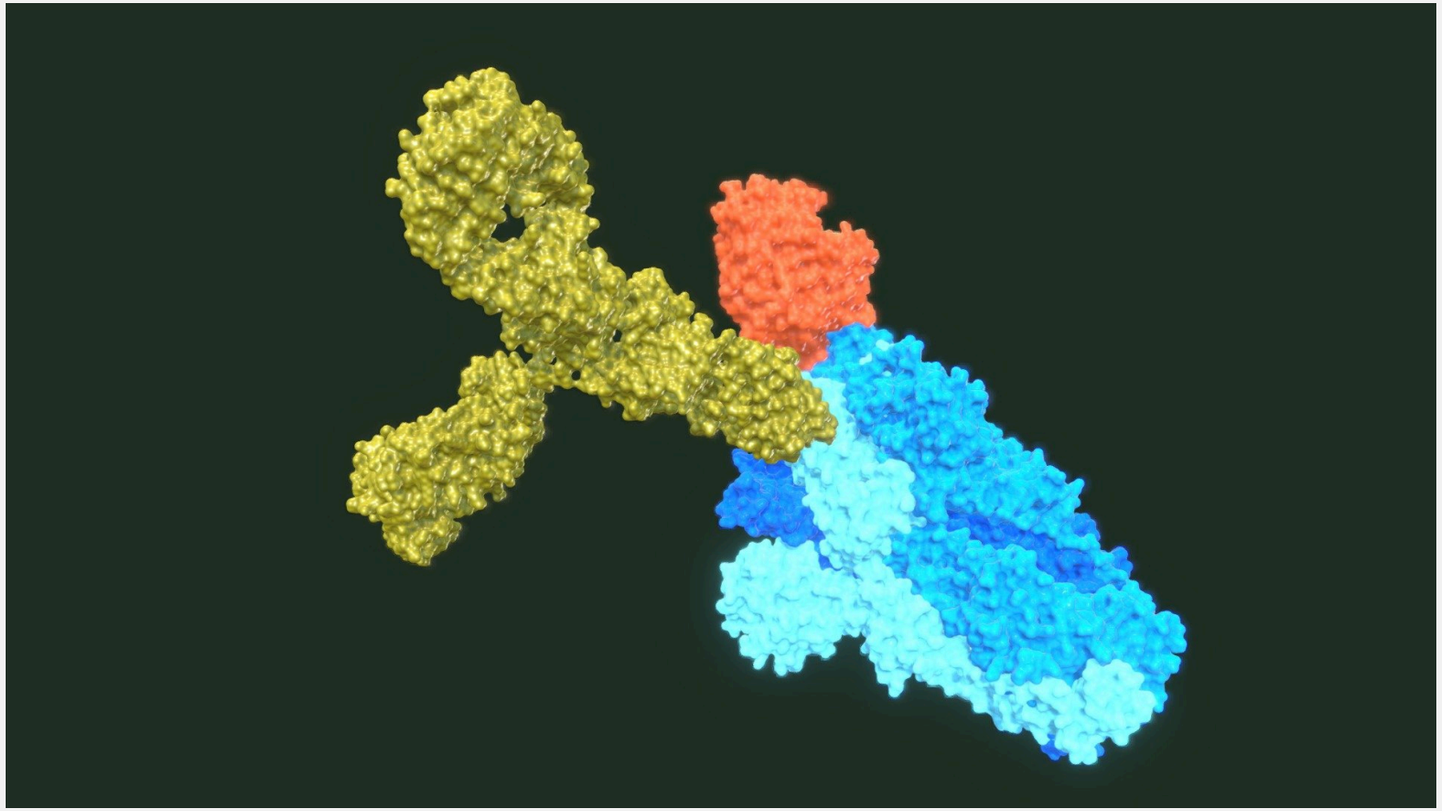
Explorable: corona virus (SARS-COV-2)



Rotate, zoom and cut into the particle with the mouse or trackpad. Double-click a protein to isolate it; press the camera icons to tour the annotated parts.

3D model: corona virus (SARS-COV-2) · eMirage via the Sketchfab viewer.

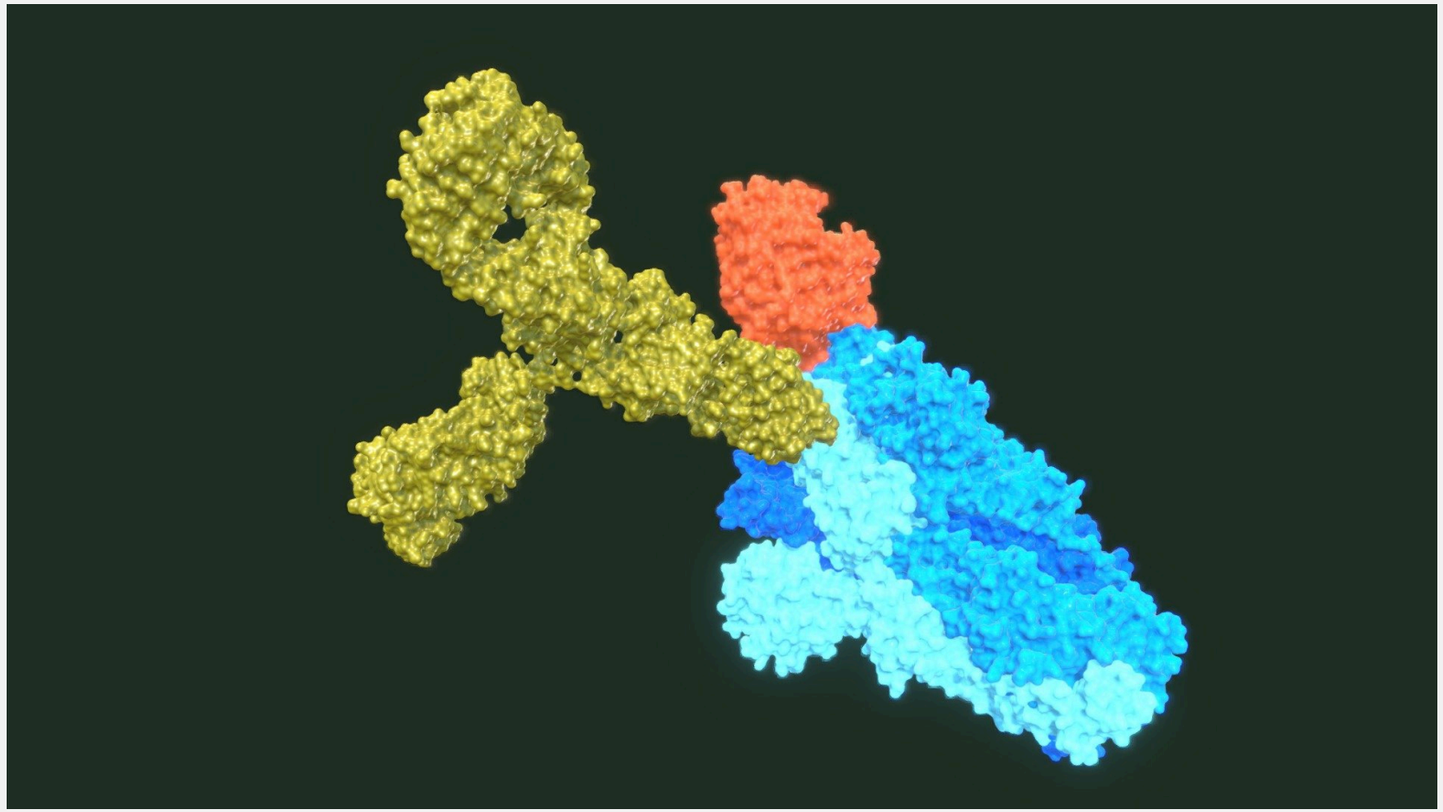
How the virus enters a cell — and how antibodies block it



- **Infection:** spike binds the human **ACE-2** receptor; ACE-2 is pulled inside (endocytosis) and viral replication begins.
- **Protection:** vaccine- or infection-elicited **antibodies** coat the spike and physically block the ACE-2 contact.

3D model: Coronavirus spike + ACE-2 receptor + antibody · axonology · Sketchfab (CC BY-NC-SA 4.0) · render still, unchanged. Raw files by Darren Gowers, molecularmodels.co.uk.

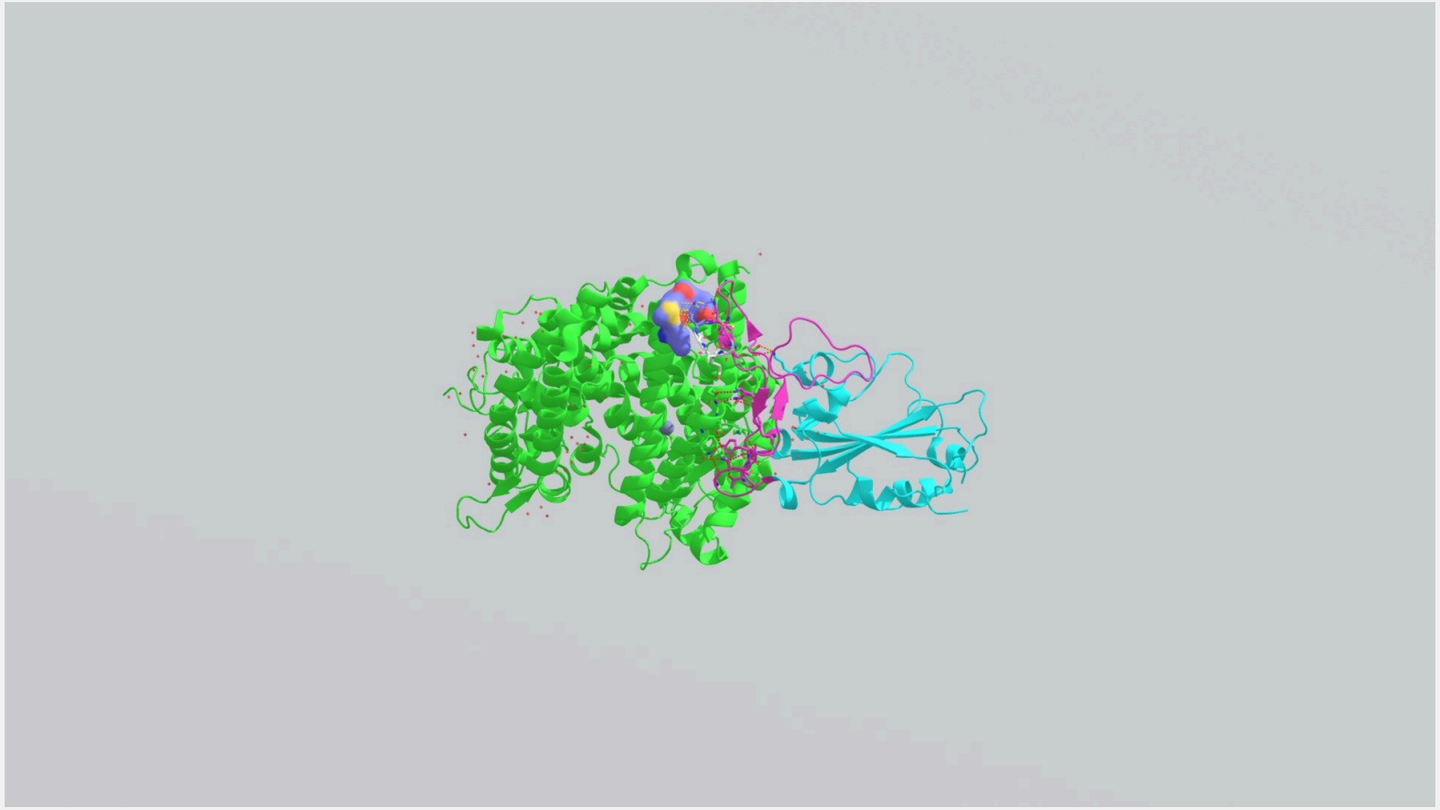
Explorable: spike + ACE-2 receptor + antibody



Rotate the spike trimer, dock it against ACE-2, then bring in the antibody to see what “blocking” means in molecular terms.

3D model: Coronavirus spike + ACE-2 receptor + antibody · axonology (CC BY-NC-SA 4.0) via the Sketchfab viewer.

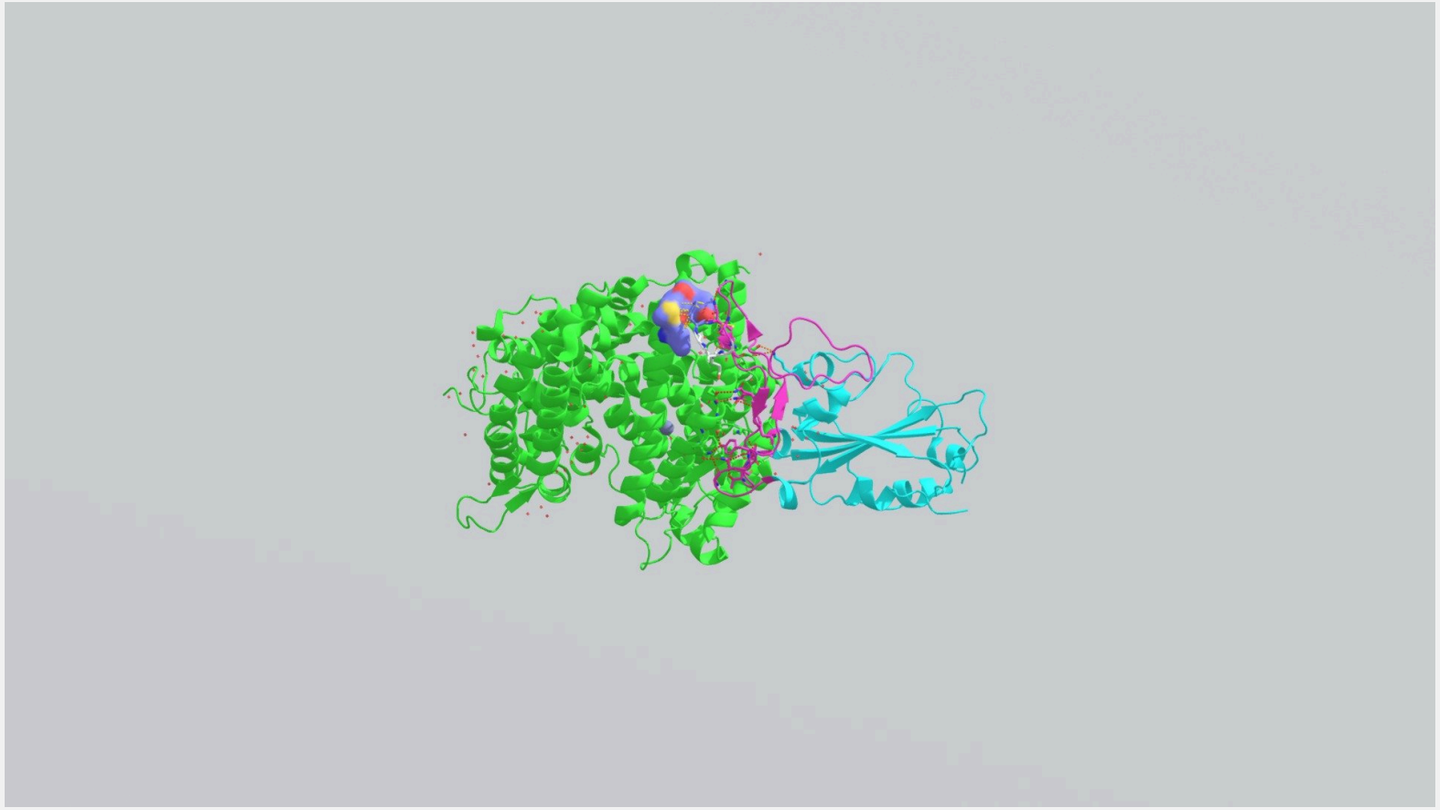
Molecular binding close-up: the RBD–ACE2 interface



- The **receptor-binding domain (RBD)** of the spike docks onto the ACE-2 peptidase domain.
- The magenta **receptor-binding motif (RBM)** forms most of the contact — a few substitutions here (e.g. N501Y) change how tightly the virus binds, which is why mutations matter.

3D model: Binding of SARS-CoV-2 to ACE2 · AlizaK · Sketchfab · render still, unchanged.

Explorable: binding of SARS-CoV-2 to ACE2



Inspect the contacts residue by residue with the mouse; note the magenta RBM, and N-glycans (NAGs) are hidden by the author for clarity.

3D model: Binding of SARS-CoV-2 to ACE2 · AlizaK via the Sketchfab viewer.

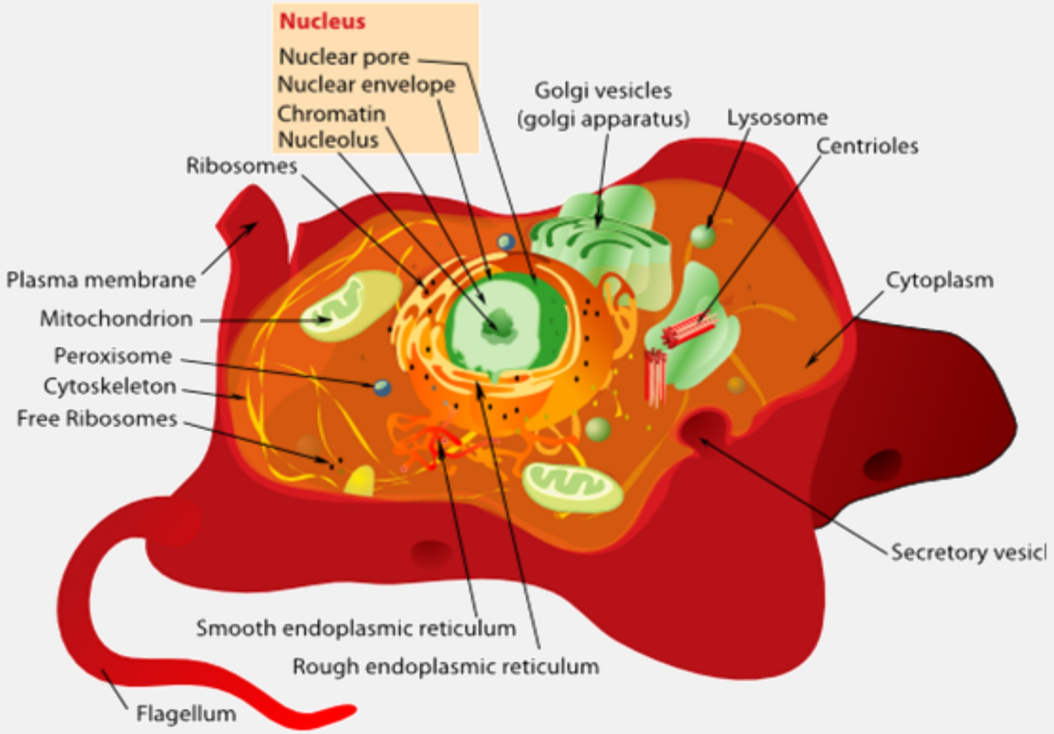
Each cell

Despite its complexity, there exists some fundamental principles common to all cell types.

1. Information. e.g. Required information to create a replica of itself.
2. Machinery. e.g. collect materials to manufacture its own components, replicate information and boot the new cell.

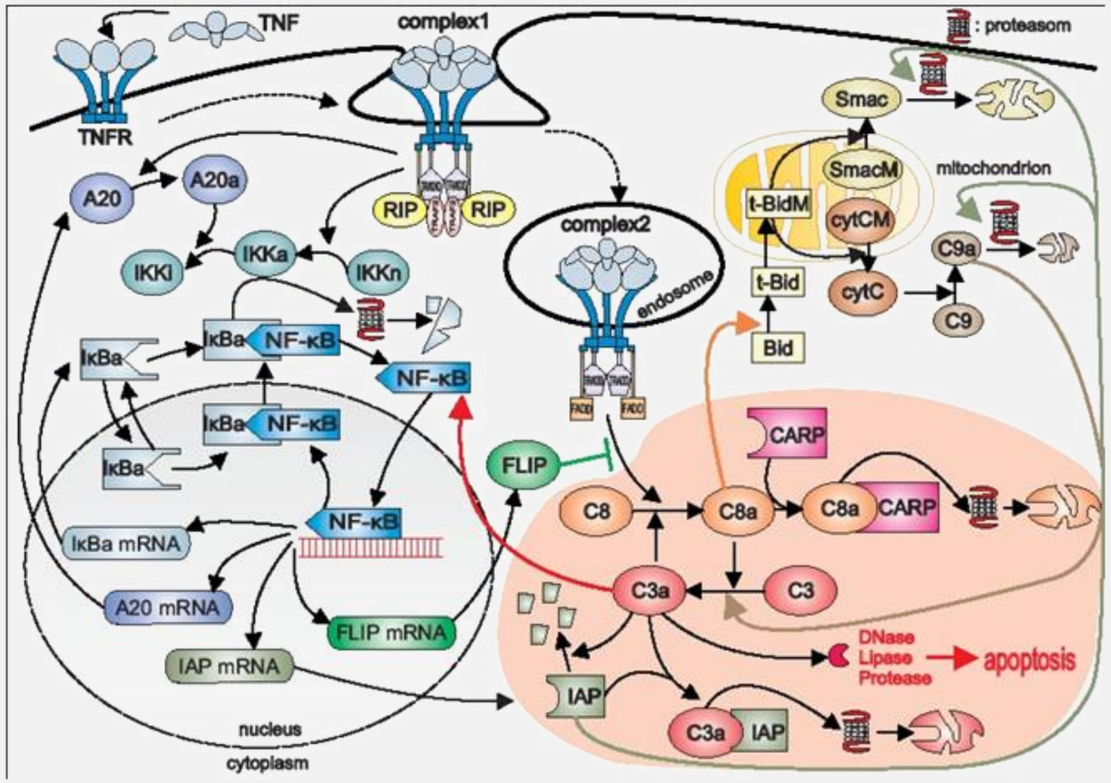
Cells

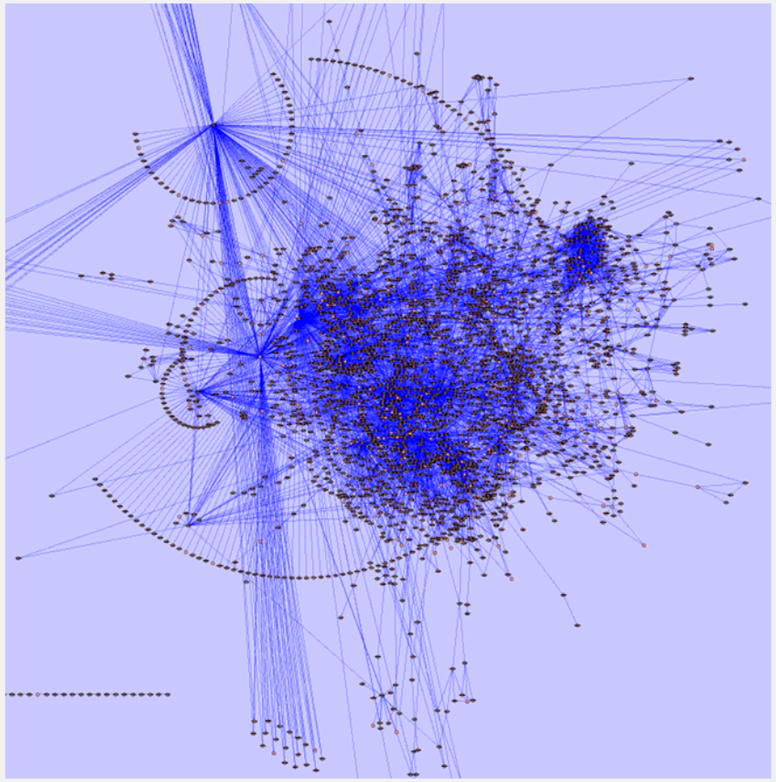
- Life-cycle! (born, eat, reproduce and die)
- Very controlled processes
- e.g. A cell only replicates if there are sufficient resources (nutrients) to complete the process
- How do they decide?
- Hey, they have no brain!
- They take decisions through pathway networks.



Pathway Networks

- Synthesize materials.
- Break compounds.
- Signal information (apoptosis).
- Complex reaction networks.





Genotype and phenotype

Genotype describes an organism's genetic constitution, either at particular loci or more broadly.

Phenotype comprises observable or measurable characteristics.

Phenotypes arise through genetic effects, environmental influences, development, and their interactions. "Genotype + environment" is a conceptual shorthand, not a universal additive equation.

articles

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

* *A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.*

The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome. We also present an initial analysis of the data, describing some of the insights that can be gleaned from the sequence.

The rediscovery of Mendel's laws of heredity in the opening weeks of the 20th century[1–3] sparked a scientific quest to understand the nature and content of genetic information that has propelled biology for the last hundred years. The scientific progress made falls naturally into four main phases, corresponding roughly to the four quarters of the century. The first established the cellular basis of heredity: the chromosomes. The second defined the molecular basis of heredity: the DNA double helix. The third unlocked the informational basis of heredity, with the discovery of the biological mechanism by which cells read the information contained in genes and with the invention of the recombinant DNA technologies of cloning and sequencing by which scientists can do the same.

The last quarter of a century has been marked by a relentless drive to decipher first genes and then entire genomes, spawning the field of genomics. The fruits of this work already include the genome sequences of 599 viruses and viroids, 205 naturally occurring plasmids, 185 organelles, 31 eubacteria, seven archaea, one

coordinate regulation of the genes in the clusters.

- There appear to be about 30,000–40,000 protein-coding genes in the human genome—only about twice as many as in worm or fly. However, the genes are more complex, with more alternative splicing generating a larger number of protein products.
- The full set of proteins (the 'proteome') encoded by the human genome is more complex than those of invertebrates. This is due in part to the presence of vertebrate-specific protein domains and motifs (an estimated 7% of the total), but more to the fact that vertebrates appear to have arranged pre-existing components into a richer collection of domain architectures.
- Hundreds of human genes appear likely to have resulted from horizontal transfer from bacteria at some point in the vertebrate lineage. Dozens of genes appear to have been derived from transposable elements.
- Although about half of the human genome derives from transposable elements, there has been a marked decline in the overall

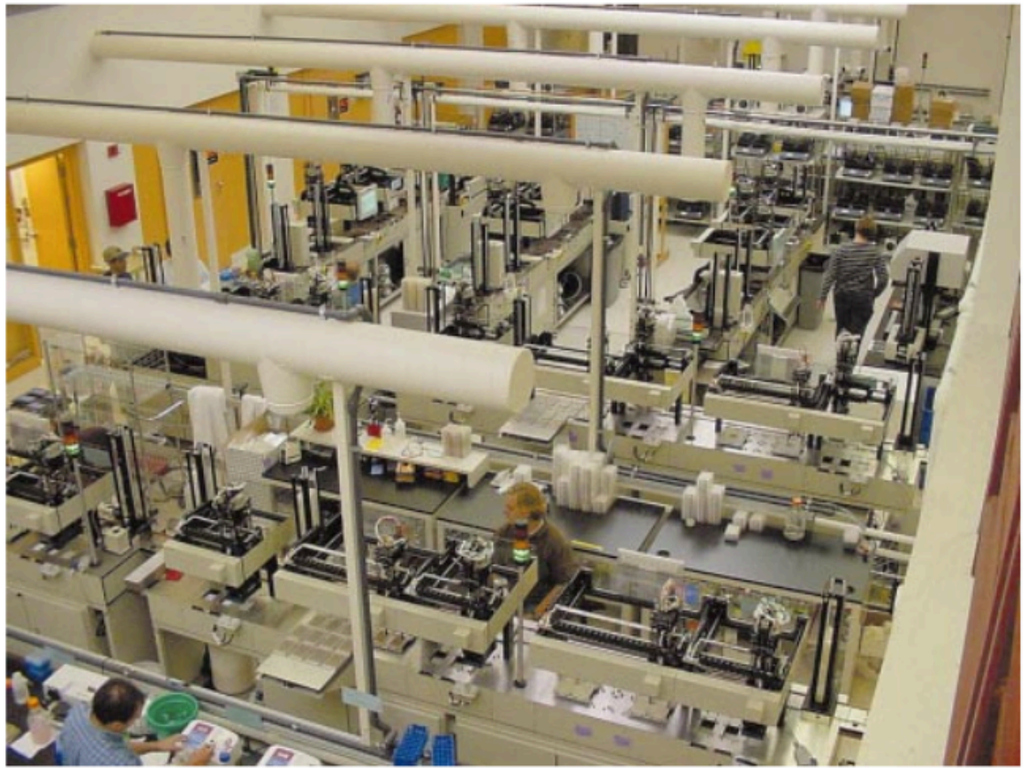
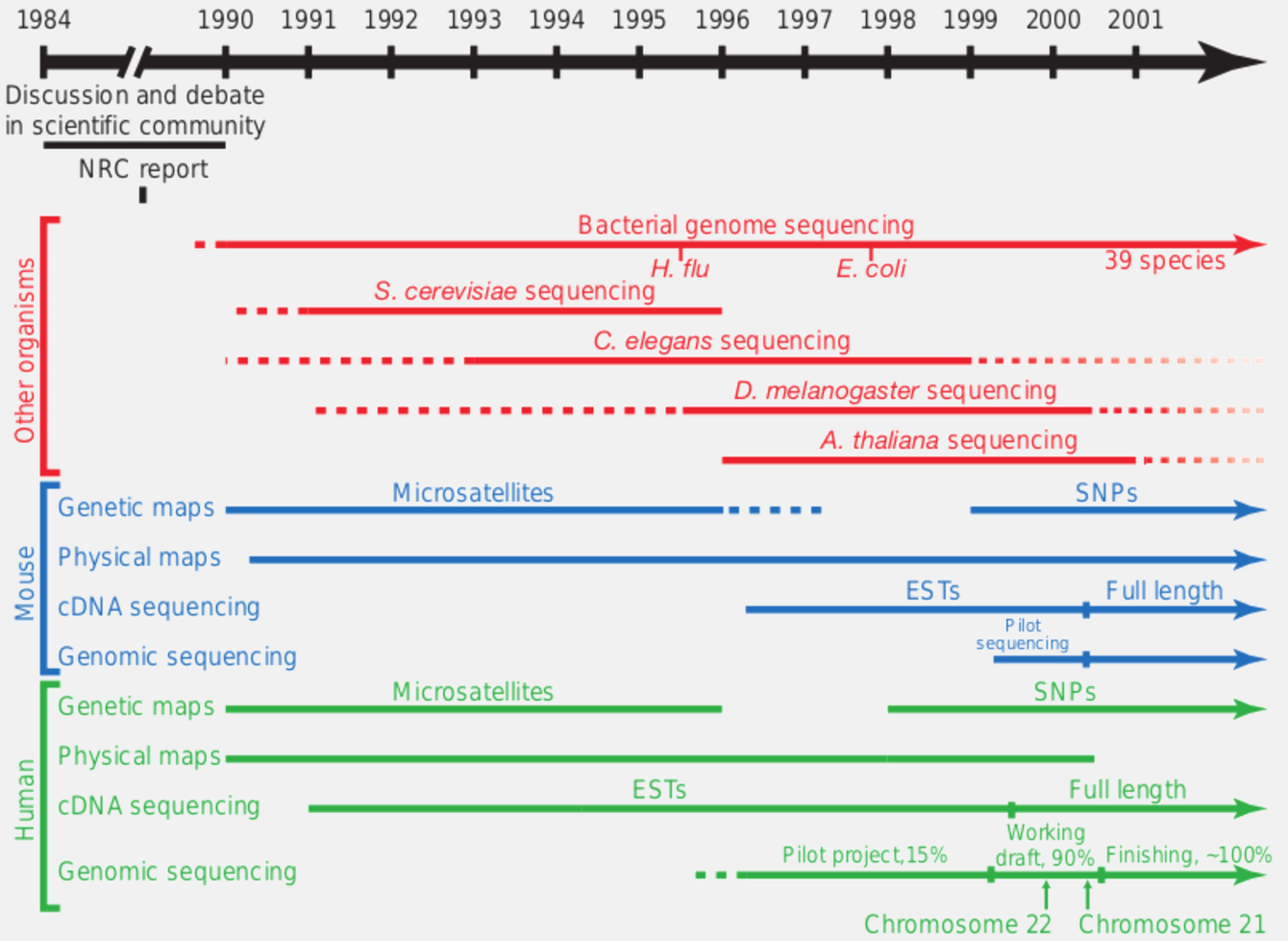
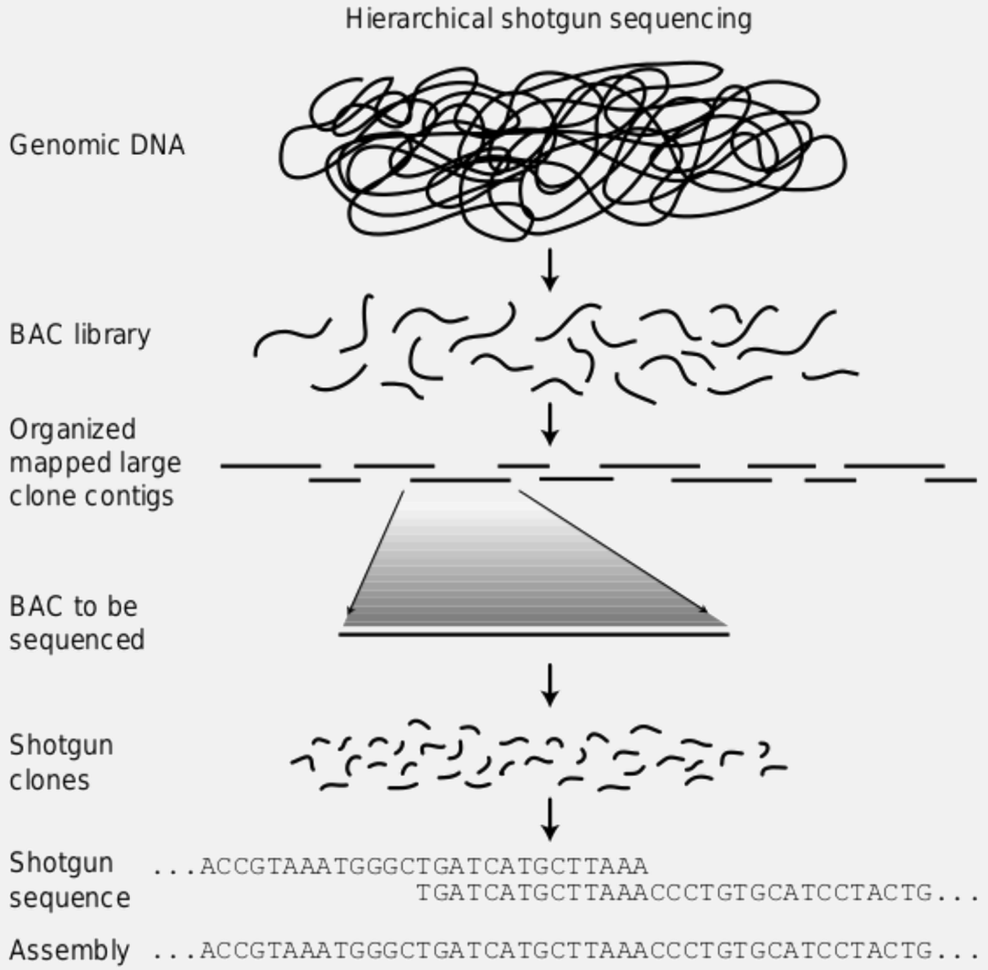


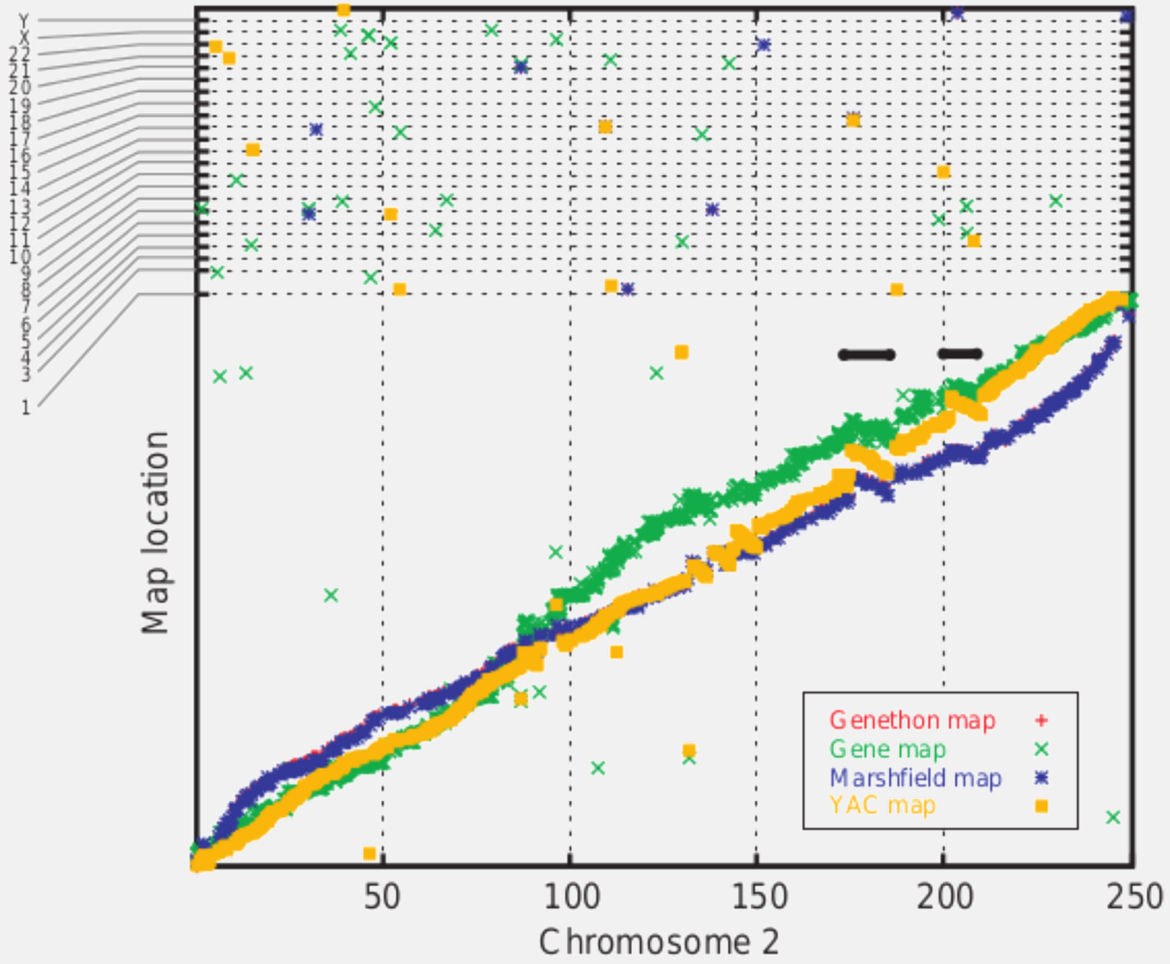
Figure 3 The automated production line for sample preparation at the Whitehead Institute, Center for Genome Research. The system consists of custom-designed factory-style conveyor belt robots that perform all functions from purifying DNA from bacterial cultures through setting up and purifying sequencing reactions.

Some notes on the Human Genome Project



Some notes on the Human Genome Project




Y
X
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
Map location
50
100
150
200
250
Chromosome 2
Genethon map
Gene map
Marshfield map
YAC map

The draft genome sequence makes it possible to explore the variation in GC content in a direct and global manner. Visual inspection (Fig. 9) confirms that local GC content undergoes substantial long-range excursions from its genome-wide average of 41%. If the genome were drawn from a uniform distribution of GC content, the local GC content in a window of size n bp should be $41 \pm \sqrt{((41)(59)/n)}\%$. Fluctuations would be modest, with the standard deviation being halved as the window size is quadrupled—for example, 0.70%, 0.35%, 0.17% and 0.09% for windows of size 5, 20, 80 and 320 kb.

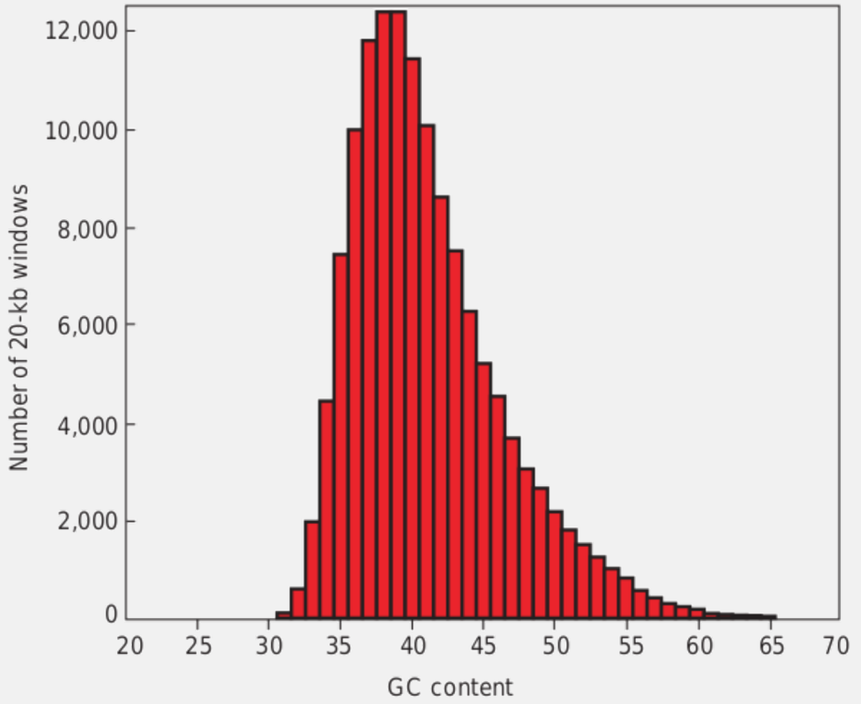


Figure 12 Histogram of GC content of 20-kb windows in the draft genome sequence.

Genetic variation

- Mutation introduces new sequence variants; recombination reshuffles existing variation.
- At a typical diploid autosomal locus, an individual carries two allele copies.
- A population can have more than two alleles at the same locus.
- Rare variants remain biologically relevant; a 1% frequency threshold is a historical convention for the word “polymorphism”.

Genetic Variability

In spots where *no genetic variability exists*, both alleles will be equal.

In spots where *genetic variability exists*, the alleles could be different

- Homozygote: Both alleles are equal
- Heterozygote: alleles are different

Mutation types: the DNA change and its consequence

A DNA change and a protein consequence describe different things.

DNA change	Possible consequence in a coding region
Single-base substitution	Synonymous, missense or stop-gain
Insertion or deletion	In-frame change or frameshift
Larger rearrangement	Changes in dosage, gene structure or regulation

The consequence depends on **where** the change occurs and which transcript is affected.

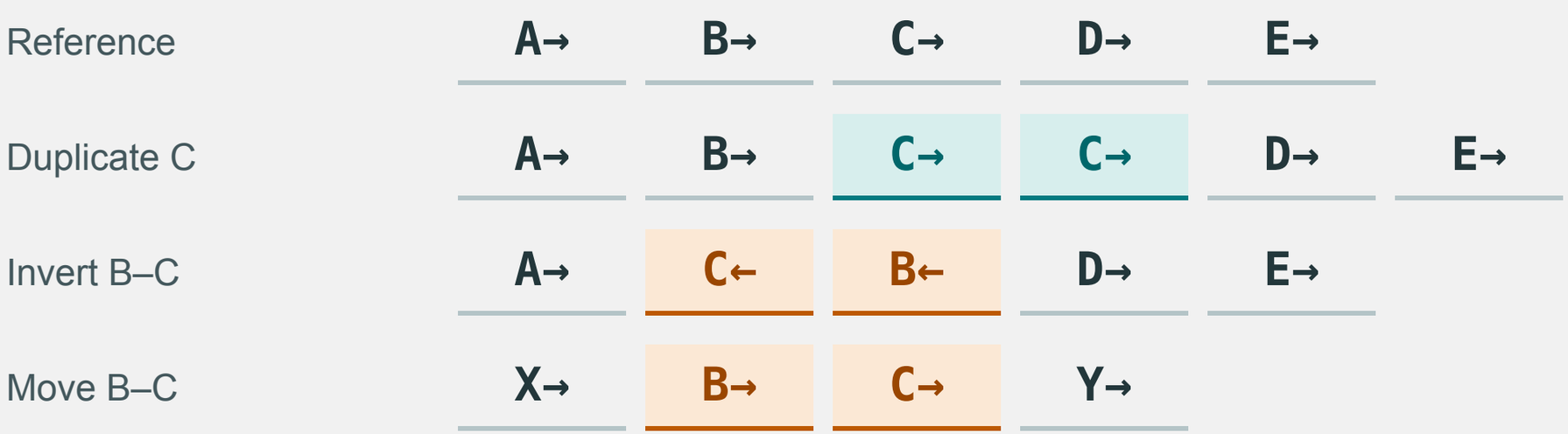
Small DNA changes: substitution, insertion and deletion



> A substitution replaces a base. An insertion adds sequence. A deletion removes sequence. ∅ marks the deleted position.

Larger changes: copy number, orientation and location

Letters below represent chromosome segments, not individual bases.



> A translocation moves a segment to a new chromosomal context. An inversion reverses its orientation.

Variant names describe different properties

Term	What it describes
SNV	A difference at one nucleotide position
SNP	A single-nucleotide polymorphism in a population
Indel	An insertion or deletion of sequence
CNV	A difference in the copy number of a genomic segment

An SNV can be synonymous, missense or stop-gain in a coding region.

An indel can affect coding sequence, splicing or regulation.

Mutation types: a map from DNA change to disease

Classify **what changes in DNA** first; then ask what happens to RNA, protein or gene dosage.

1 · Point substitutions

Synonymous · Missense · Nonsense

Replace one base; examine the codon outcome.

2 · Small insertions / deletions

In-frame · Frameshift

Does the coding-length change divide by three?

3 · Within one chromosome

Inversion · Deletion · Relocation

Change segment orientation, amount or position.

4 · Between chromosomes

Interchromosomal translocation

Move or exchange segments between chromosomes.

5 · Repeat expansion

More copies of a short repeat

Repeat length and genomic context matter.

These are teaching groups, not mutually exclusive mechanisms. Here, “point mutation” means a single-base substitution; some definitions also include one-base indels. Splicing and regulatory effects can occur across groups.

How to read mutation notation

Reference : level . position change → `NM_000518.5:c.20A>T`
HBB transcript/version · coding DNA · base 20 · A replaced by T

DNA: what changed?

Example[1]	Read it as…
`c.20A>T`	Substitute A with T at base 20
`c.1521_1523del`	Delete bases 1521–1523 (CFTR)
`c.76_77insA`	Insert A **between** bases 76 and 77
`c.76_78dup`	Copy bases 76–78 next to themselves
`g.100_200inv`	Invert genomic bases 100–200
`CAG[40]`	40 repeats; locus/coordinates also needed

Protein: what is the consequence?

Example[1]	Read it as…
`p.Gly608=`	Glycine unchanged: synonymous
`p.Glu7Val` / `E7V`	Glutamate → valine at residue 7
`p.Gln40Ter` / `Q40*`	Glutamine → STOP: nonsense
`p.Phe508del` / `ΔF508`	Lose phenylalanine 508: in-frame
`p.Arg97ProfsTer23`	Arg97 → Pro; frame shifts; STOP is codon 23 of the new frame

Levels: `g.` genomic DNA · `c.` coding DNA · `r.` RNA · `p.` protein. **Chromosomes:** `t(9;22)` means a translocation involving chromosomes 9 and 22 (cytogenetic shorthand, not HGVS).

[1] Abbreviated examples; a full HGVS report needs a reference accession **and version**. Parentheses mark a prediction: `p.(Glu7Val)`. Legacy HbS **E6V** = HGVS **p.Glu7Val** because the numbering conventions differ.

Further reading: Khan Academy: mutation notations · HGVS guide · Frameshift notation

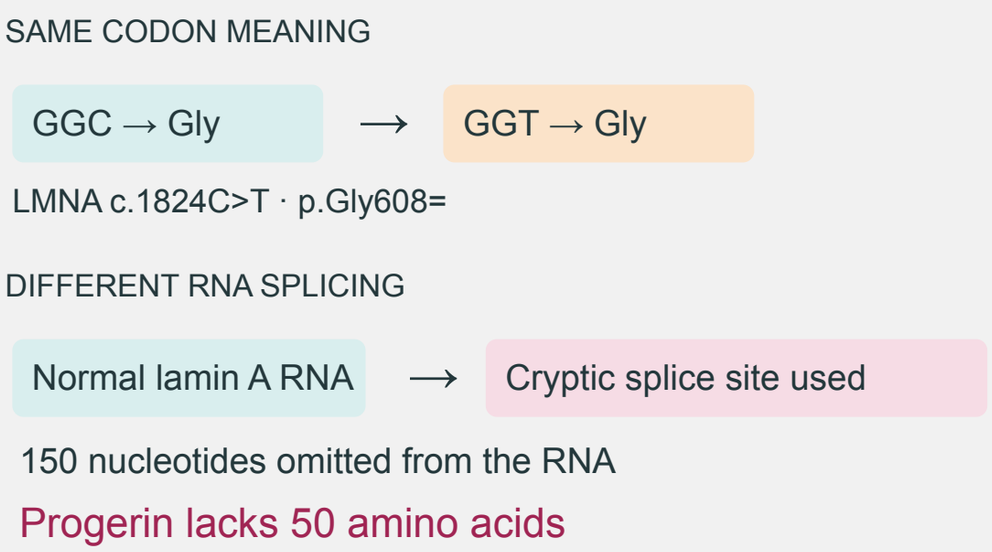
One reference sequence for the coding examples

An **illustrative sequence**, not a patient allele. Read the **coding DNA strand 5′ to 3′** in groups of three bases.

Coding DNA	**ATG**	**GAA**	**TTT**	**GGC**	**TAA**
mRNA	**AUG**	**GAA**	**UUU**	**GGC**	**UAA**
Protein · N → C	**Met**	**Glu**	**Phe**	**Gly**	**STOP**

> The coding strand matches the mRNA sequence, with T replaced by U. STOP ends translation and is not an amino acid.

Synonymous: same amino acid, possible RNA effect



Codon meaning is unchanged. This does **not** guarantee unchanged splicing or expression.

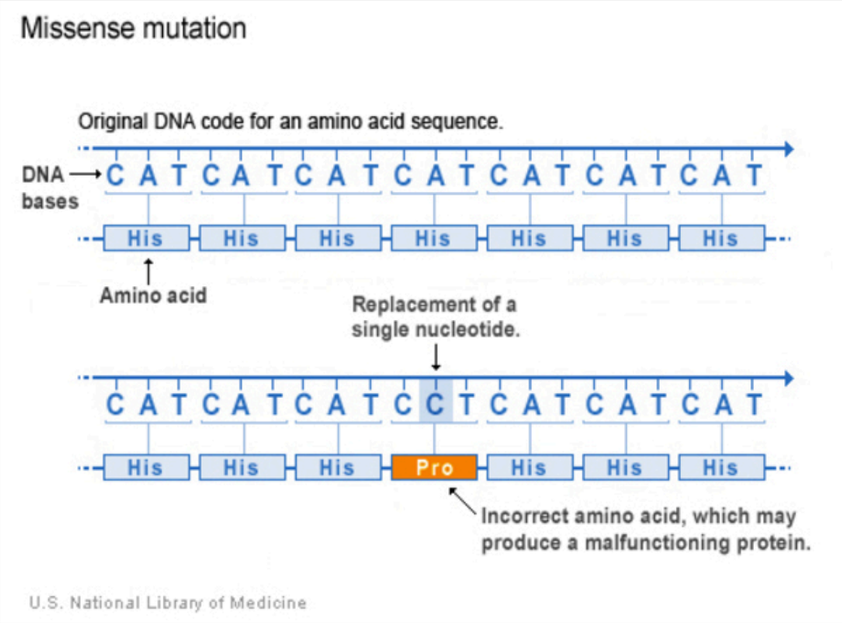
Hutchinson–Gilford progeria syndrome

- **LMNA c.1824C>T (p.Gly608=)**: GGC → GGT.
- A cryptic splice site is activated; the RNA loses 150 nucleotides.
- The resulting **progerin** lacks 50 amino acids.

> A synonymous DNA change can have a major effect through RNA splicing.

Disease paper: Eriksson et al., Nature (2003).
Figure: Original vector schematic based on Eriksson et al. (2003).

Missense: one amino acid is replaced



One codon changes its amino-acid meaning; the reading frame stays intact.

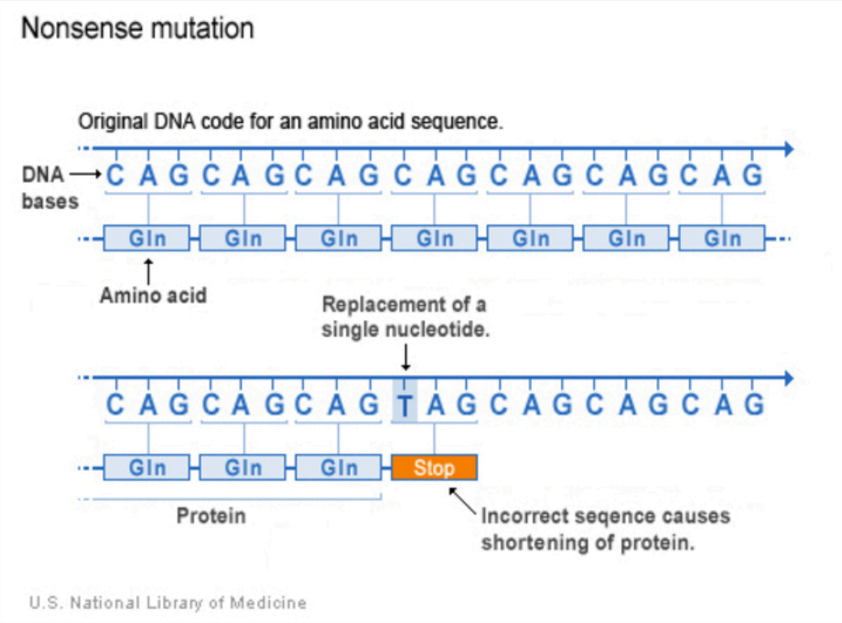
Sickle cell disease

- **HBB c.20A>T**: GAG → GTG; glutamate → valine.
- HbS can polymerise when deoxygenated, promoting red-cell sickling.
- Disease usually requires HbS plus another disease-causing HBB allele.

> One altered residue can change a protein's interactions.

Disease paper: Ingram, Nature (1956).
Figure: U.S. National Library of Medicine, MedlinePlus Genetics, unchanged.

Nonsense: a premature stop codon



A sense codon becomes STOP. The RNA may be degraded by nonsense-mediated decay.

Disease paper: Trecartin et al., J Clin Invest (1981).
Figure: U.S. National Library of Medicine, MedlinePlus Genetics, unchanged.

β^0-thalassaemia

- **HBB codon 39** in traditional β-globin numbering: CAG → TAG.
- Glutamine becomes STOP; the affected allele produces no functional β-globin.
- Two severe HBB alleles can cause severe anaemia.

> Do not assume that a stable shortened protein always accumulates.

In-frame indel: the downstream frame is retained

ILLUSTRATIVE CODING DNA · 5′ → 3′



Met – Glu – Phe – Gly → Met – Glu – Gly

Real example: CFTR p.Phe508del

One residue lost; downstream frame retained.

A net change of **3n coding bases** preserves the downstream reading frame.

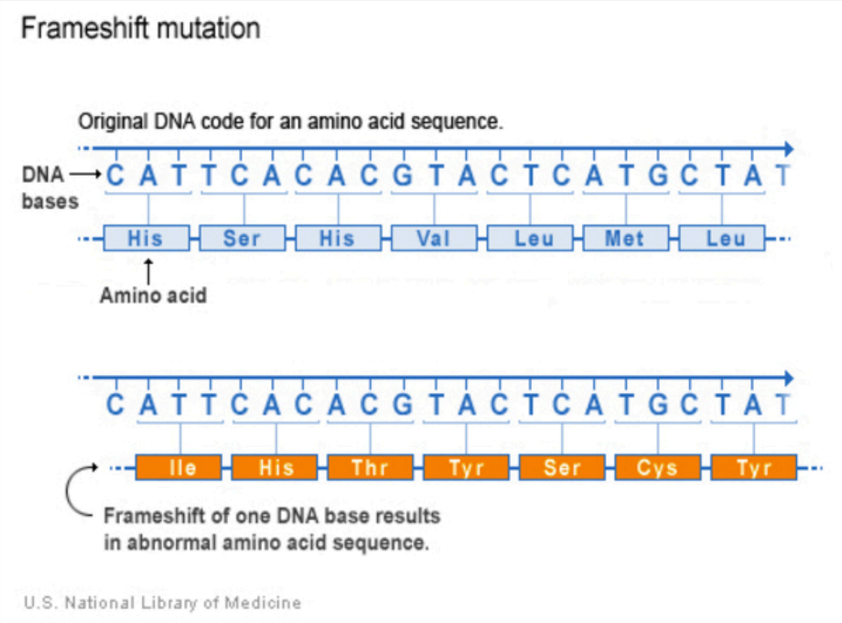
Cystic fibrosis

- **CFTR c.1521_1523delCTT → p.Phe508del** (F508del).
- Three DNA bases are lost; phenylalanine 508 is absent.
- This impairs CFTR folding/processing; CF is typically recessive.

> Preserving the reading frame does not mean preserving function.

Disease paper: Riordan et al., Science (1989).
Figure: Original vector teaching schematic; clinical example from Riordan et al.

Frameshift indel: downstream codons are regrouped



Adding or deleting a number of coding bases **not divisible by 3** changes the reading frame.

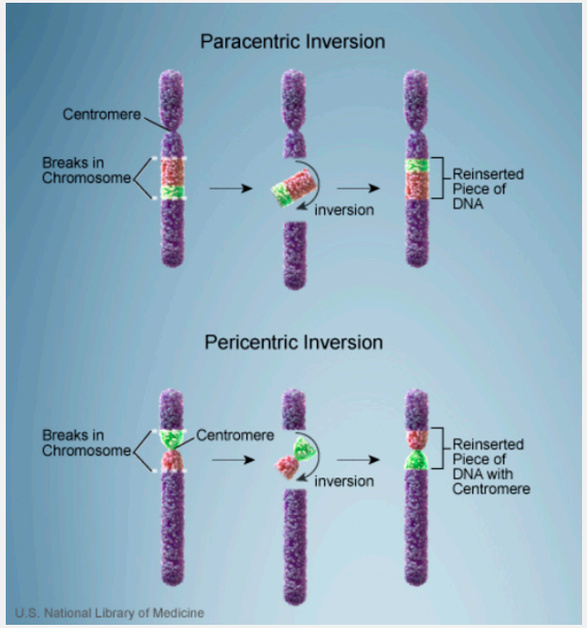
Tay–Sachs disease

- A known **HEXA exon 11 four-base insertion** shifts the frame.
- Loss of hexosaminidase A activity impairs GM2 breakdown.
- With pathogenic variants in both copies, GM2 accumulates and neurons are damaged.

> 4 mod 3 = 1: a four-base insertion is a frameshift too.

Disease paper: Myerowitz & Costigan, J Biol Chem (1988).
Figure: U.S. National Library of Medicine, MedlinePlus Genetics, unchanged.

Inversion: a segment reverses orientation



A segment flips. In double-stranded DNA, this reverses orientation: the reference-strand sequence becomes its **reverse complement**.

Disease paper: Lakich et al., Nature Genetics (1993).
Figure: U.S. National Library of Medicine, MedlinePlus Genetics, unchanged.

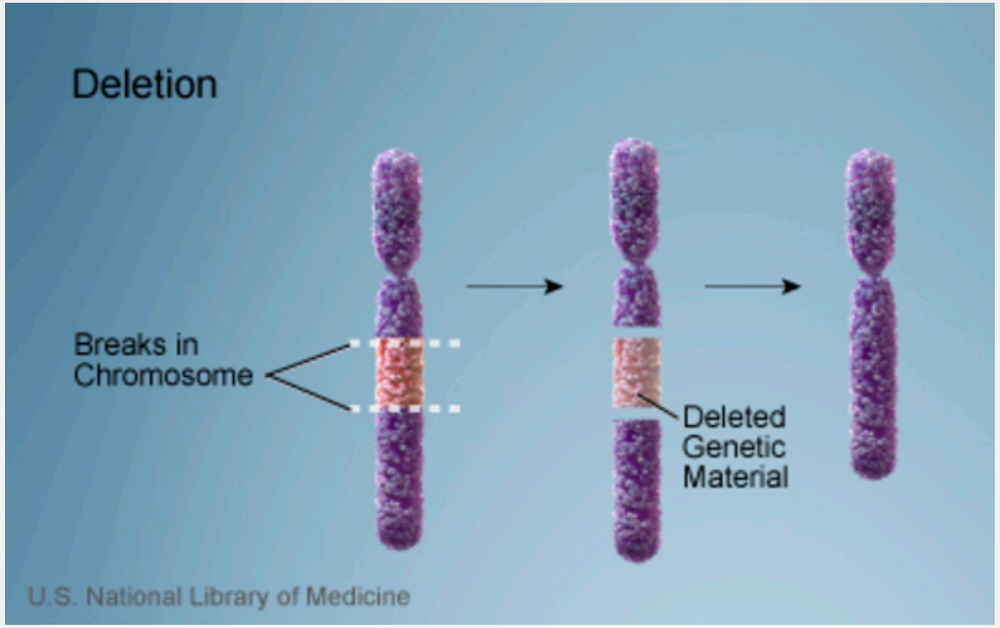
Severe haemophilia A

- Recombination involving **F8 intron 22 repeats** can invert an intervening segment on chromosome X.
- The rearrangement disrupts the F8 gene.
- Factor VIII deficiency impairs blood clotting.

> Copy number can be unchanged while a gene is disrupted.

Chromosomal deletion: a segment is lost

Mutation map → 3 · Within one chromosome



A deletion removes DNA and reduces gene dosage. **Diagram:** interstitial loss; cri-du-chat often involves **terminal 5p loss**.

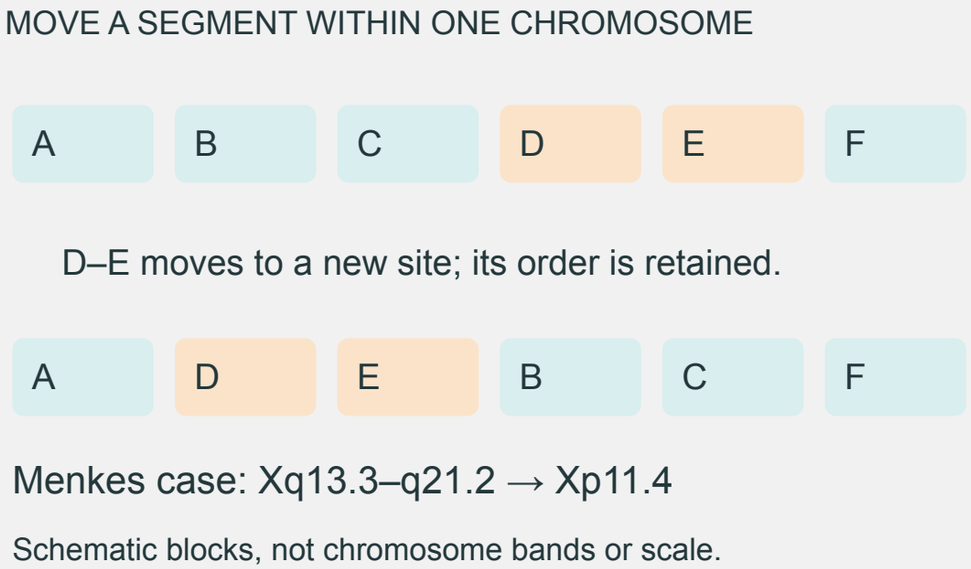
Disease paper: Mainardi et al., J Med Genet (2001).
Figure: U.S. National Library of Medicine, MedlinePlus Genetics, unchanged.

Cri-du-chat syndrome (5p deletion)

- Loss of part of the **short arm of chromosome 5**.
- Clinical features can include a characteristic infant cry and developmental disability.
- The size and position of the deletion vary between people.

> A chromosome deletion affects gene dosage, not just codon grouping.

Intrachromosomal relocation: same chromosome, new site



Often described as an **intrachromosomal insertion**; “single-chromosome translocation” is a broader teaching label.

A reported Menkes disease case

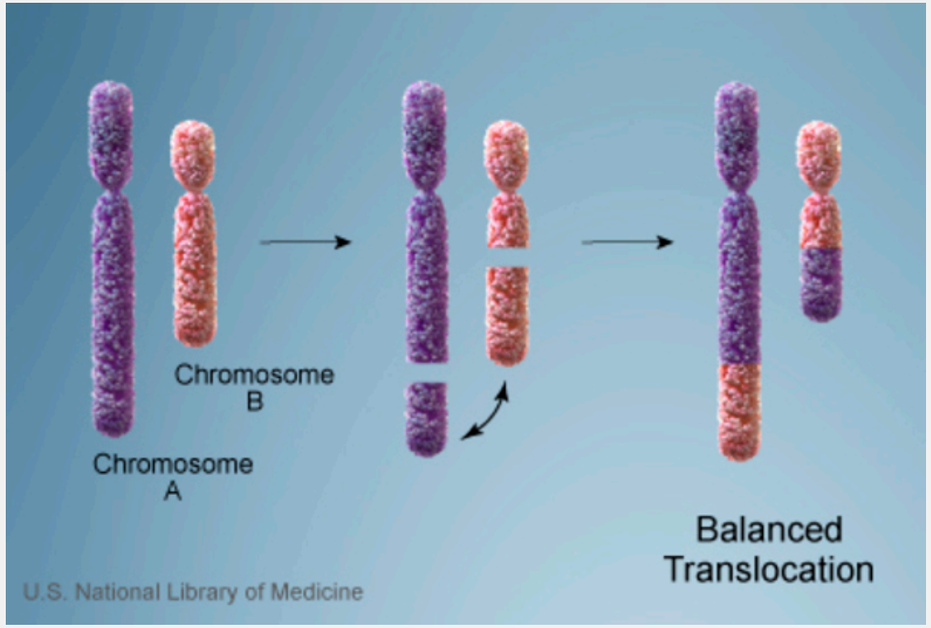
- An X-chromosome segment **Xq13.3–q21.2** was inserted at **Xp11.4**.
- The report described a boy with Menkes disease: **46,XY,ins(X)(p11.4q13.3q21.2)**.
- This is a rare structural example, not the usual mutation in Menkes disease.

> Movement and inversion are different: ask whether the segment changes orientation.

Disease paper: Human Genetics (1992), original case report.
Figure: Original vector schematic based on the reported intrachromosomal insertion.

Interchromosomal translocation: two chromosomes

Mutation map → 4 · Between chromosomes



Segments move between chromosomes. A reciprocal exchange can be cytogenetically balanced yet create a gene fusion.

Chronic myeloid leukaemia (CML)

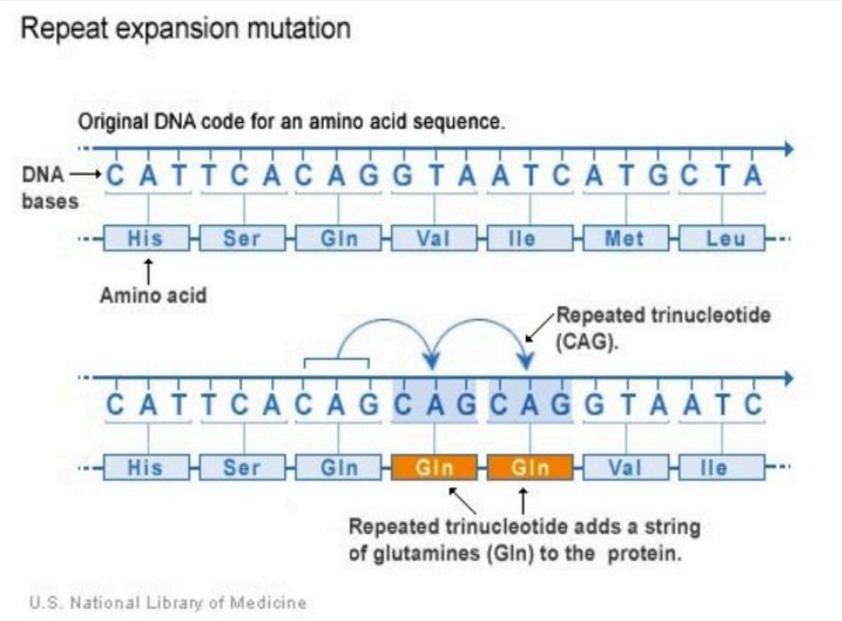
- **t(9;22)** creates the **Philadelphia chromosome**, a derivative chromosome 22.
- **BCR::ABL1** encodes an abnormally active tyrosine kinase.
- This is an acquired mutation in the leukaemic cell lineage.

> “Balanced” describes DNA amount; it does not guarantee a harmless outcome.

Disease paper: Rowley, Nature (1973); Shtivelman et al., Nature (1985).
Figure: U.S. National Library of Medicine, MedlinePlus Genetics, unchanged.

Repeat expansion: more copies of a short motif

Mutation map → 5 · Repeat expansion



The repeated motif stays recognisable, but the number of copies increases. Effects depend on its genomic context.

Huntington disease

- Expansion of the **CAG repeat in HTT** lengthens its polyglutamine tract.
- **36–39 repeats:** reduced penetrance. **≥40:** full-penetrance range over a typical lifespan.
- Autosomal dominant; repeat length can change between generations.

> Adding CAG triplets preserves the reading frame but can still cause disease.

Disease paper: Huntington's Disease Collaborative Research Group, Cell (1993).
Figure: U.S. National Library of Medicine, MedlinePlus Genetics, unchanged.

Compare mechanisms, not just names

DNA change	What to ask next	Disease example
Synonymous / missense / nonsense	Splicing? Residue change? Premature stop?	Progeria / sickle cell / β-thalassaemia
In-frame / frameshift indel	Is the coding-length change divisible by 3?	Cystic fibrosis / Tay–Sachs
Inversion / deletion / relocation	Orientation? Dosage? Position?	Haemophilia A / cri-du-chat / Menkes case
Interchromosomal translocation	Which genes meet at the breakpoints?	CML: BCR::ABL1
Repeat expansion	Which motif, where, and how many copies?	Huntington disease

> The **molecular change**, its **functional consequence**, and its **clinical effect** are three different descriptions.

Check your classification

1. A synonymous change activates a cryptic splice site. Is it necessarily harmless?
2. A coding deletion removes six bases. Does it necessarily shift the reading frame?
3. A rearrangement has no visible loss of DNA. Can it still cause disease?
4. One chromosome segment moves elsewhere on that same chromosome. Is that interchromosomal?

> **1 No:** LMNA/progeria. **2 No:** six is divisible by three, assuming unchanged splicing. **3 Yes:** F8 disruption or BCR::ABL1 fusion. **4 No:** it is intrachromosomal relocation.

The End

