When AI meets life sciences
当人工智能遇上生命科学

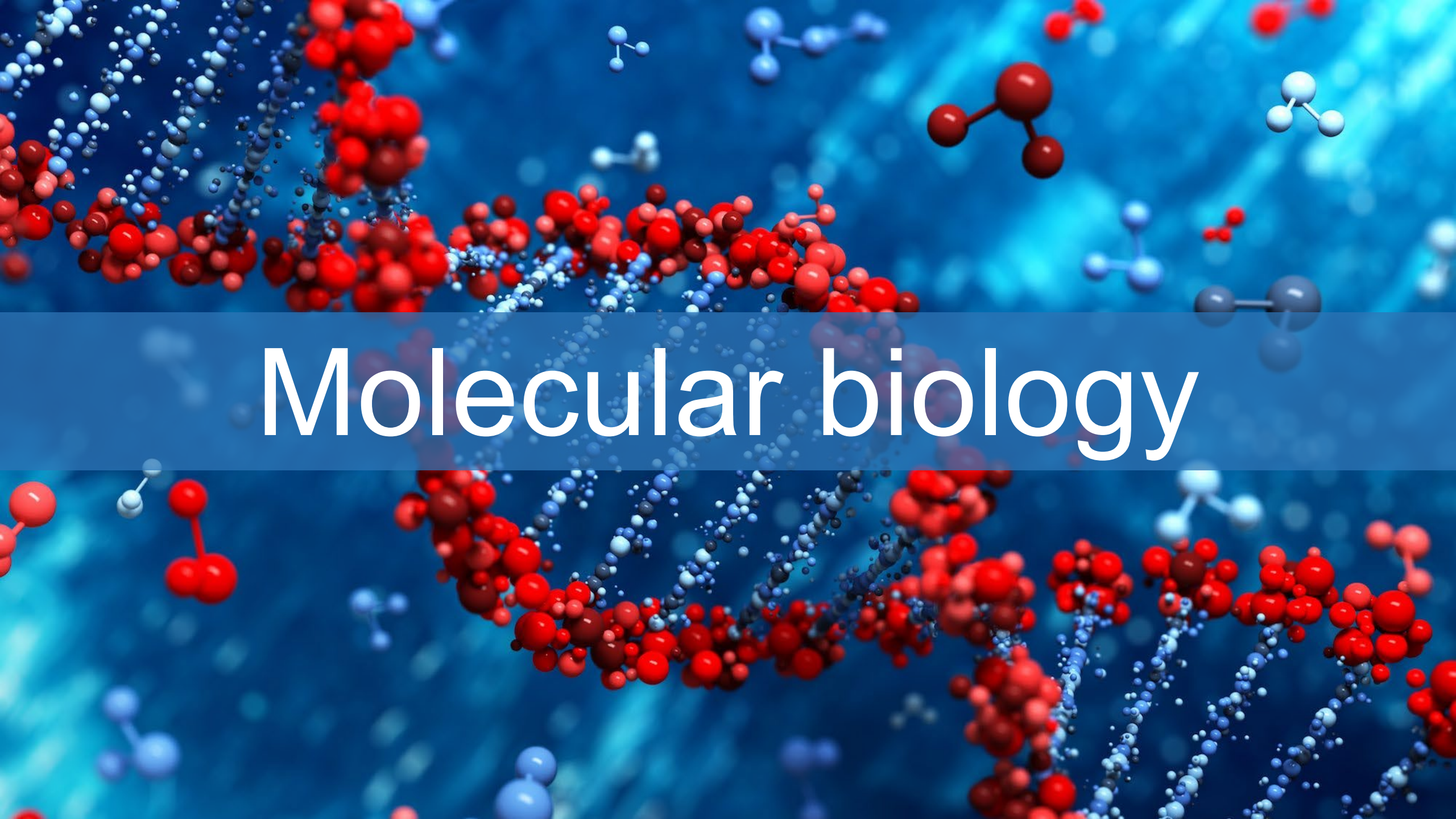
2025 清华大学 AI创新实践课程

张强锋，清华大学

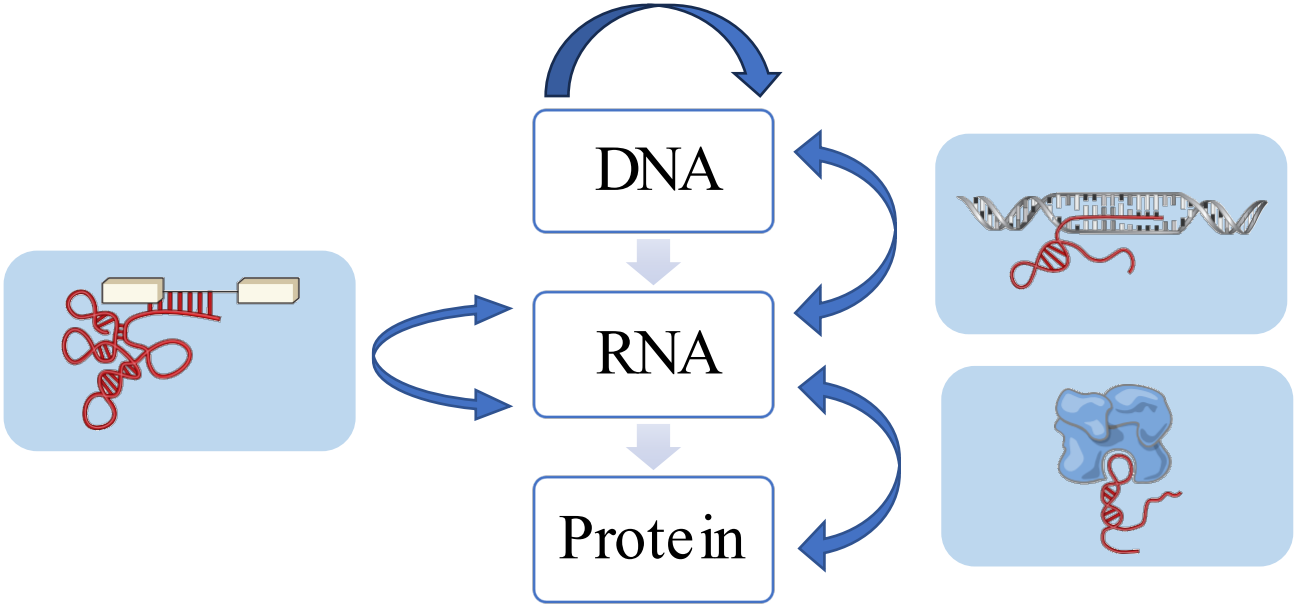
Nov 3, 2025

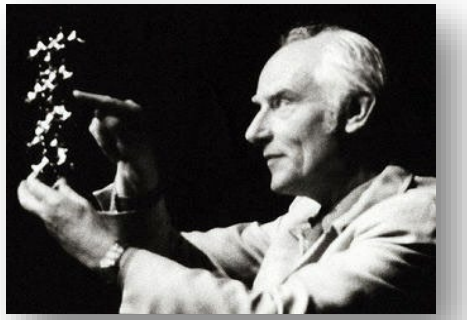
Outline

- **Molecular Biology**: the central dogma, regulation of gene expression, functional diversity of RNA
- **Genomics**: evolution of sequencing technologies, genome annotation, epigenetic regulation, comparative genomics
- **Structural Biology**: techniques for determining 3D structures of biomolecules (NMR spectroscopy, X-ray crystallography, Cryo-EM/ET, MD simulation, AI Prediction)
- **Cell and Systems Biology**: signal transduction pathways, dynamic modeling of metabolic pathways

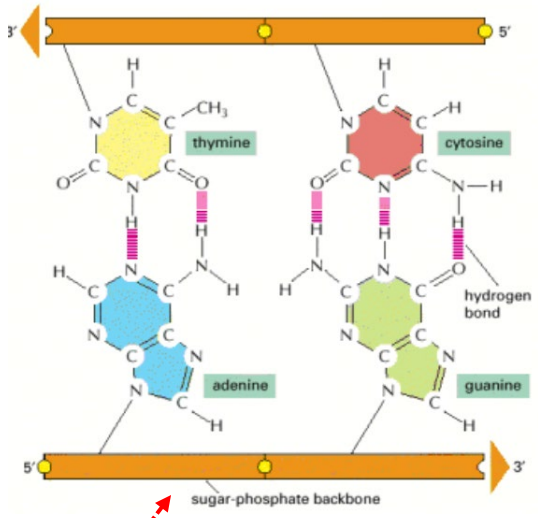

Molecular biology

The Centre Dogma

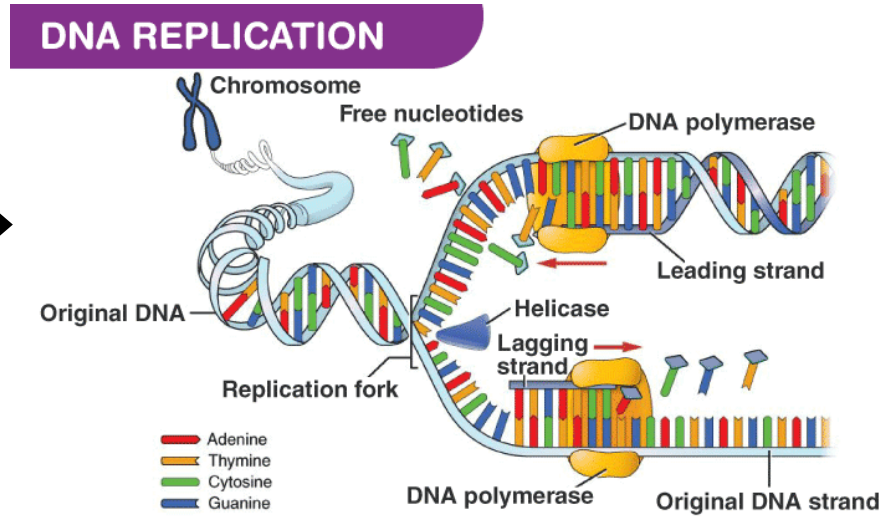


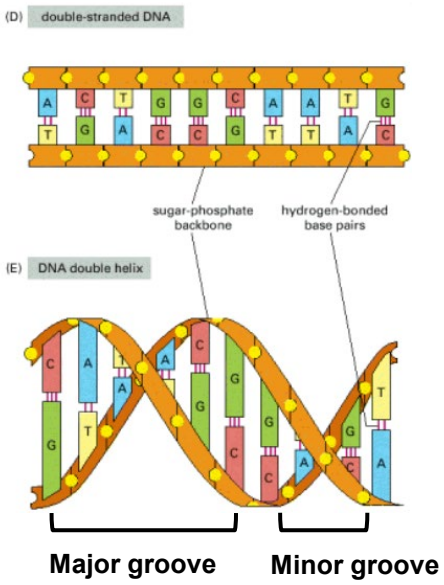


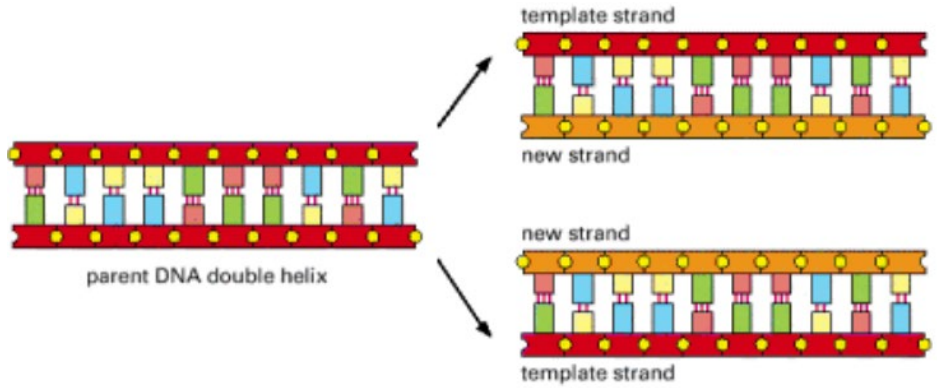
DNA and its replication



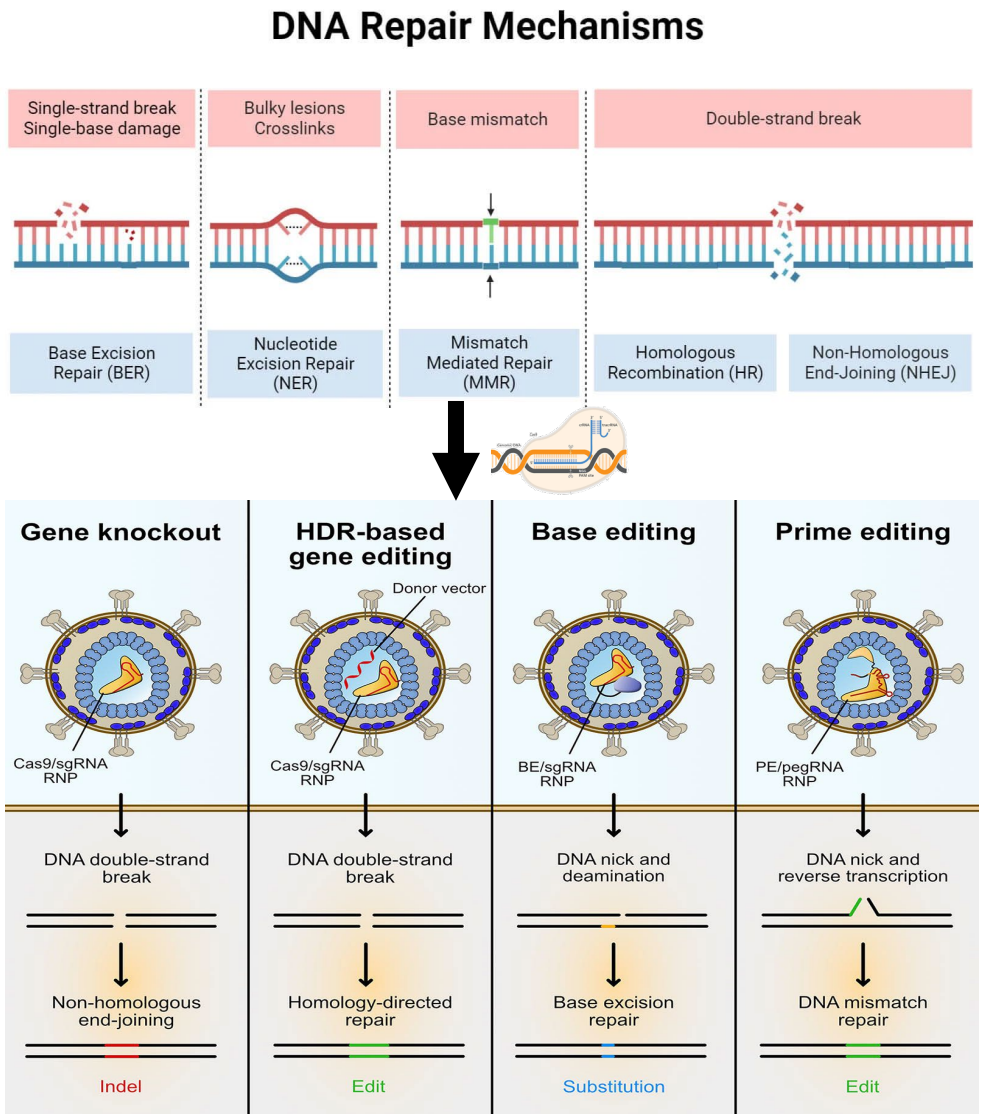
DNA copies its own information through replication.







From DNA repair to gene editing

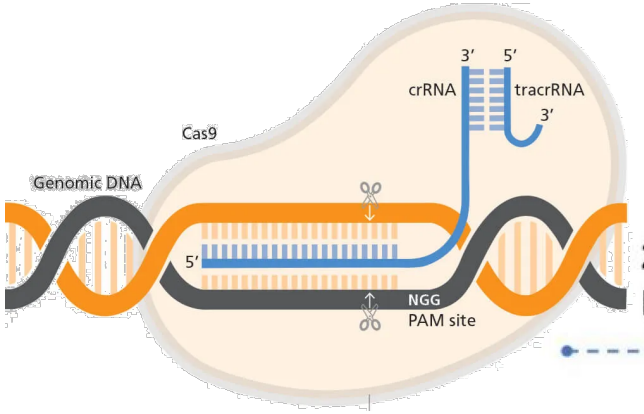


- **DNA damage:** Cells are continuously faced with endogenous stress (for example, during replication) and exogenous stress (for example, during exposure to ultraviolet radiation) that can ultimately lead to DNA damage.

- **DNA repair mechanisms:** To preserve genomic integrity, cells have an arsenal of repair proteins that engage the appropriate repair pathway or, if damage is irreparable, induce cell cycle arrest and/or apoptosis.

- **From DNA repair to therapy:** The CRISPR/Cas system creates a targeted double-strand break in the DNA, which co-opts the cell's innate repair mechanisms—either NHEJ for gene knockouts or HR for precise gene insertion—to achieve gene editing.

Gene editing, "a manual regulation"



2019
Breakthrough in CRISPR: The emergence of Prime Editor

2017
Breakthrough in CRISPR: The emergence of Cpf1...

2016
Breakthrough in CRISPR: The emergence of Base Editor

2015
CRISPR first applied to gene editing in human embryos

The fourth stage
The era of CRISPR technology

2012
The emergence of CRISPR gene editing technology

2010
The emergence of TALEN gene editing technology

1996
The emergence of zinc finger nuclease technology

1994
Meganuclease first applied to mammal

1988
Meganuclease first applied to yeast

The third stage
The era of gene editing

1982
Electroporation technology first applied to transgenic

1973
The emergence of the first transgenic animals

The second stage
The era of GMO

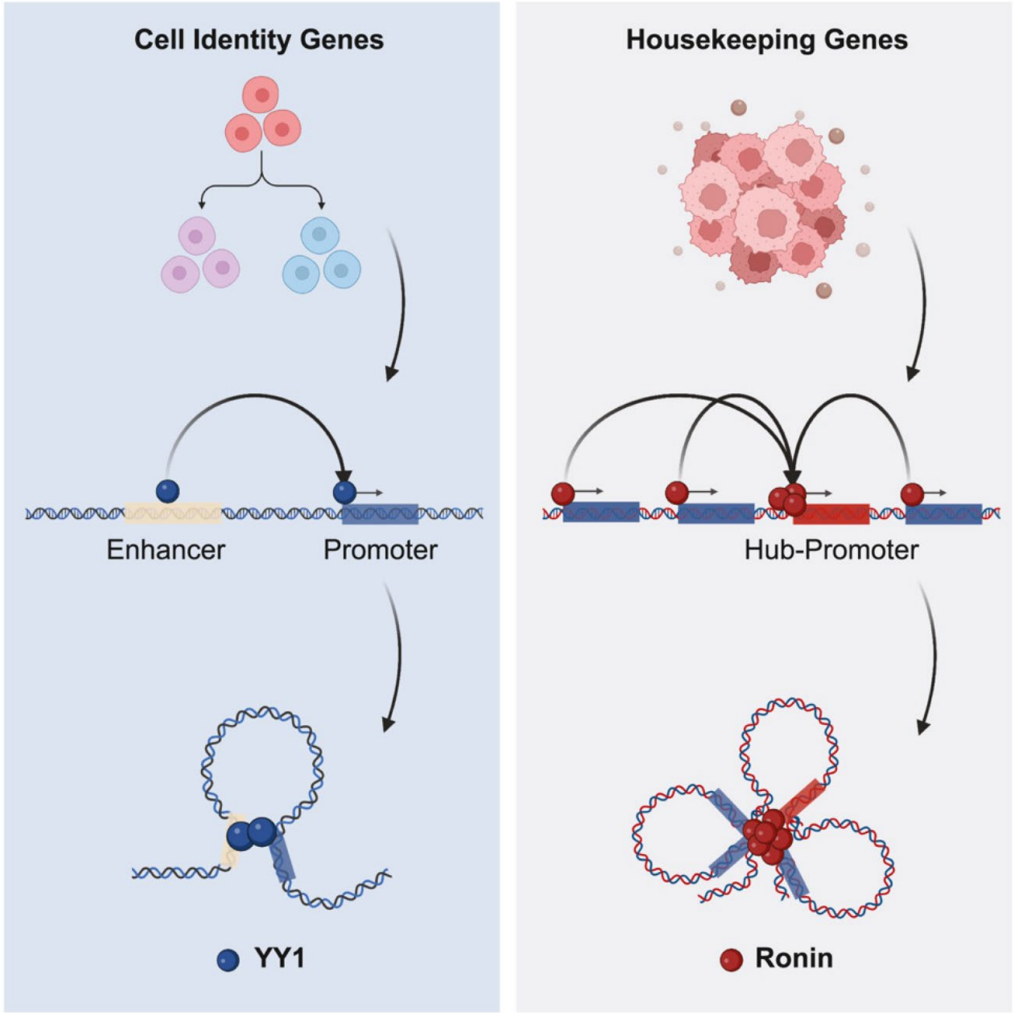
1952–1953
The discovery of DNA as genetic material and its double helix structure

The first stage
The era of molecular genetics

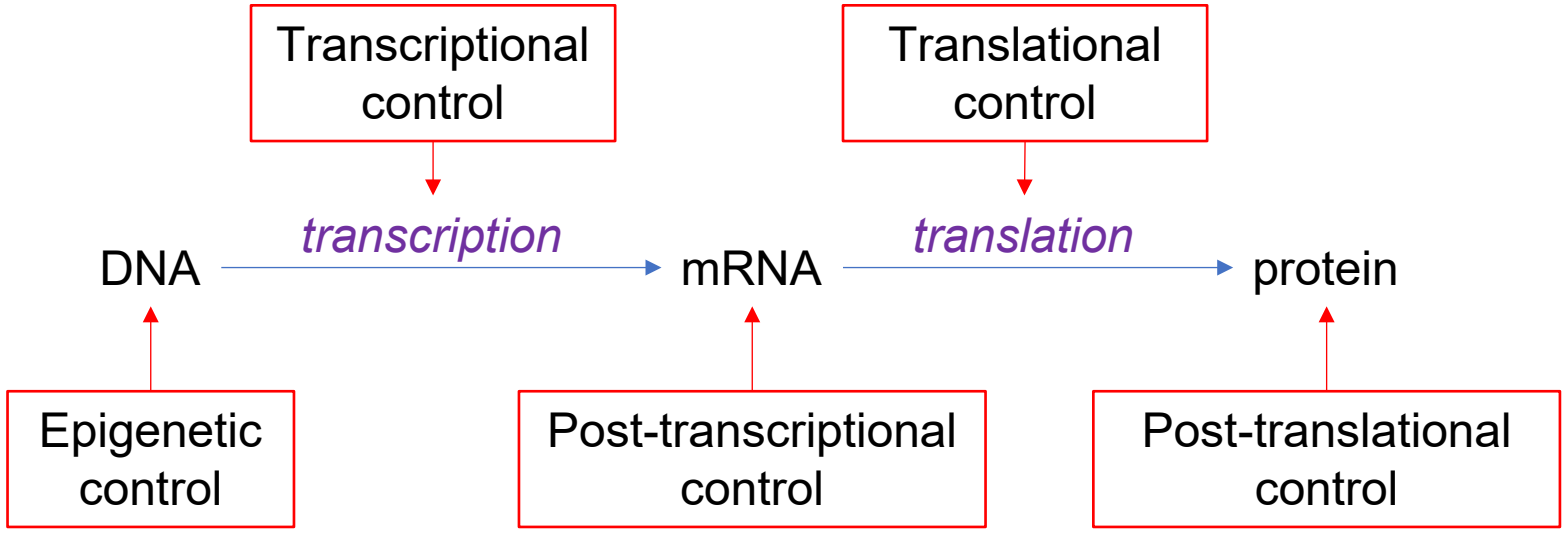
Gene expression

- **Housekeeping gene**
  - ➢ Constantly expressed in approximately all cells

- **Regulated gene**
  - ➢ Levels of the gene product rise and fall with the needs of the organism.
  - ➢ Such genes are inducible (able to be turned on)
  - ➢ Such genes are also repressible (able to be turned off)

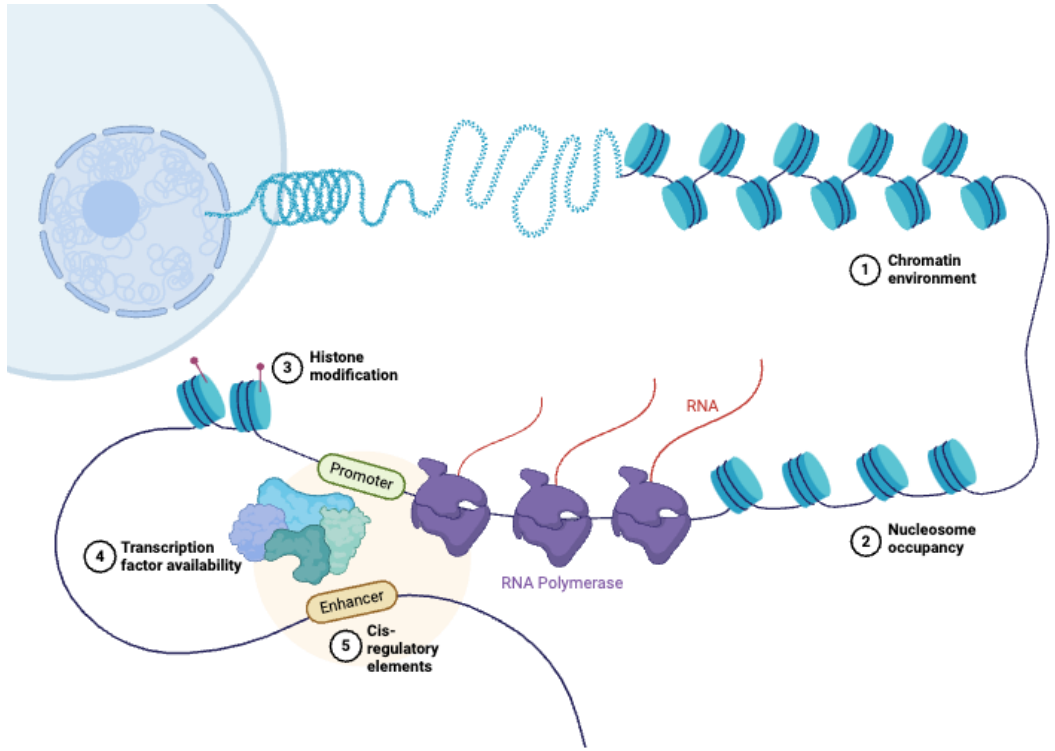


Gene expression regulation



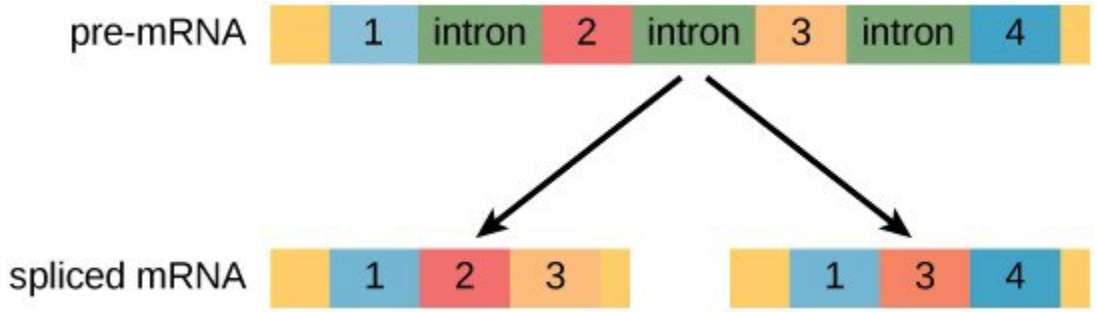
Epigenetic and transcriptional regulation

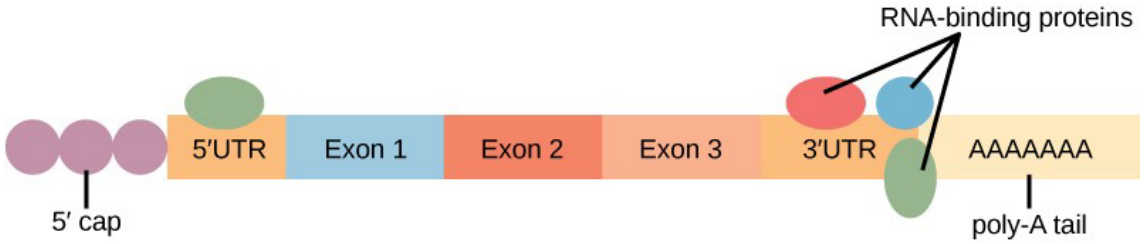
- **Epigenetic Regulations**: As relating to or arising from non-genetic influences on gene expression, including DNA modifications (methylation) , histone modifications (acetylation, methylation, phosphorylation), and RNA interference (RNAi).
- **Promoters and Enhancers**: Transcription begins with the binding of transcription factors to specific DNA sequences known as promoters and enhancers. Promoters are located near the transcription start site, while enhancers can be located at varying distances from the gene.
- **RNA Polymerase**: Transcription factors recruit RNA polymerase, the enzyme responsible for synthesizing RNA from a DNA template. Different RNA polymerases transcribe different types of RNA, such as mRNA, tRNA, and rRNA.



Post-transcriptional regulation

- **Alternative RNA splicing** is a mechanism that allows different combinations of introns, and sometimes exons, to be removed from the primary transcript. This allows different protein products to be produced from one gene.



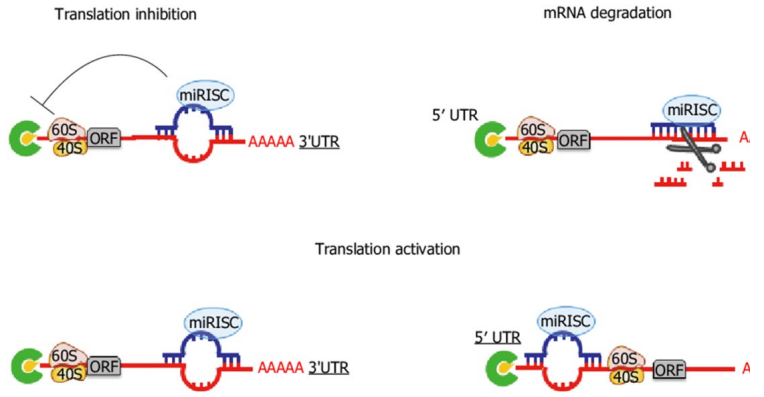


- **RNA Stability:** The stability of mRNA molecules is regulated. Some sequences in the mRNA affect its half-life, and specific proteins can bind to these sequences to regulate stability.

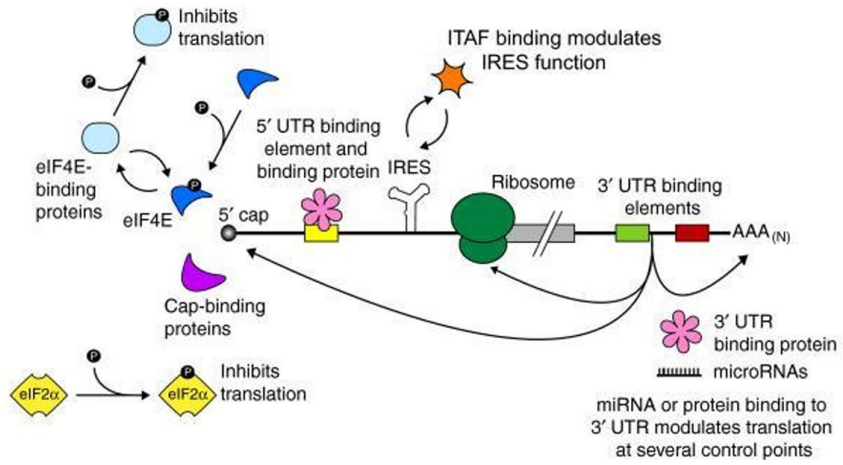
Translational regulation



- **MicroRNAs (miRNAs):** These small RNA molecules can bind to specific mRNA sequences, leading to mRNA degradation or inhibiting translation.

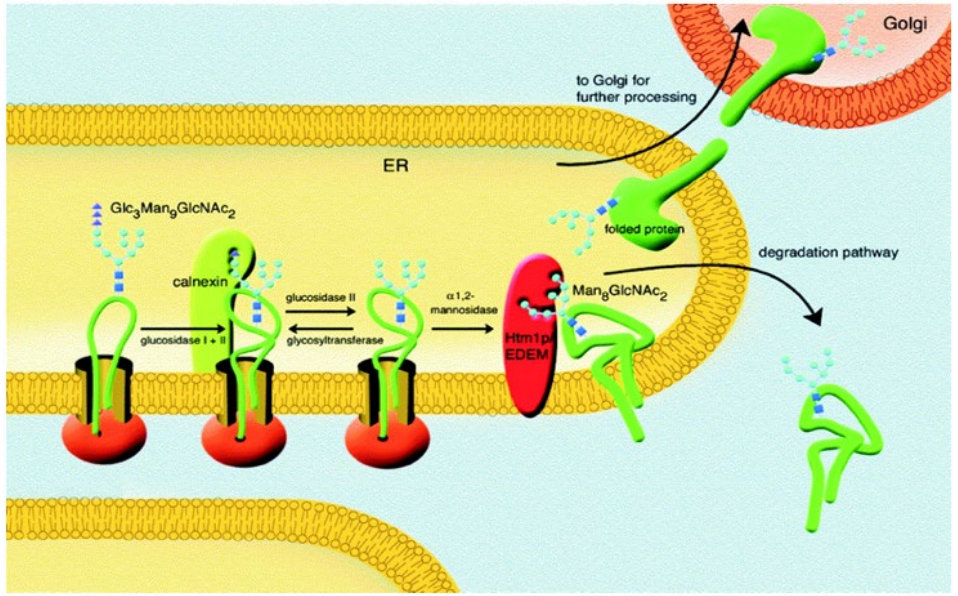


- **Initiation, Elongation, and Termination**: The translation of mRNA into proteins involves three main stages: initiation, elongation, and termination. Each stage can be regulated by various factors, including initiation factors, ribosome-binding proteins, and release factors.

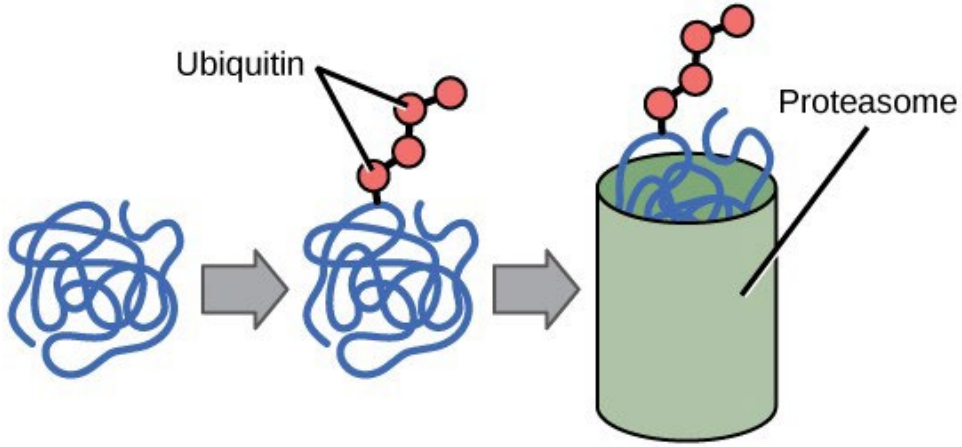


Post-translational regulation

- **Protein Folding and Modification:** Newly synthesized proteins undergo folding and may require post-translational modifications (e.g., phosphorylation, glycosylation) to become functional.

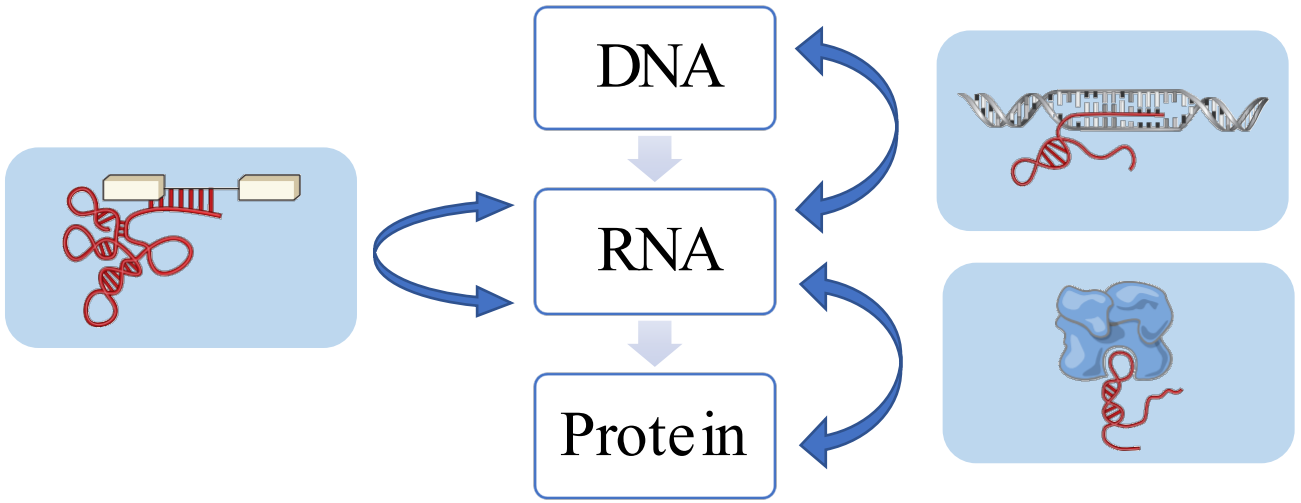


- **Ubiquitination and degradation:** Ubiquitin molecules can be attached to proteins, targeting them for degradation by the proteasome. This process plays a role in regulating protein levels within the cell.

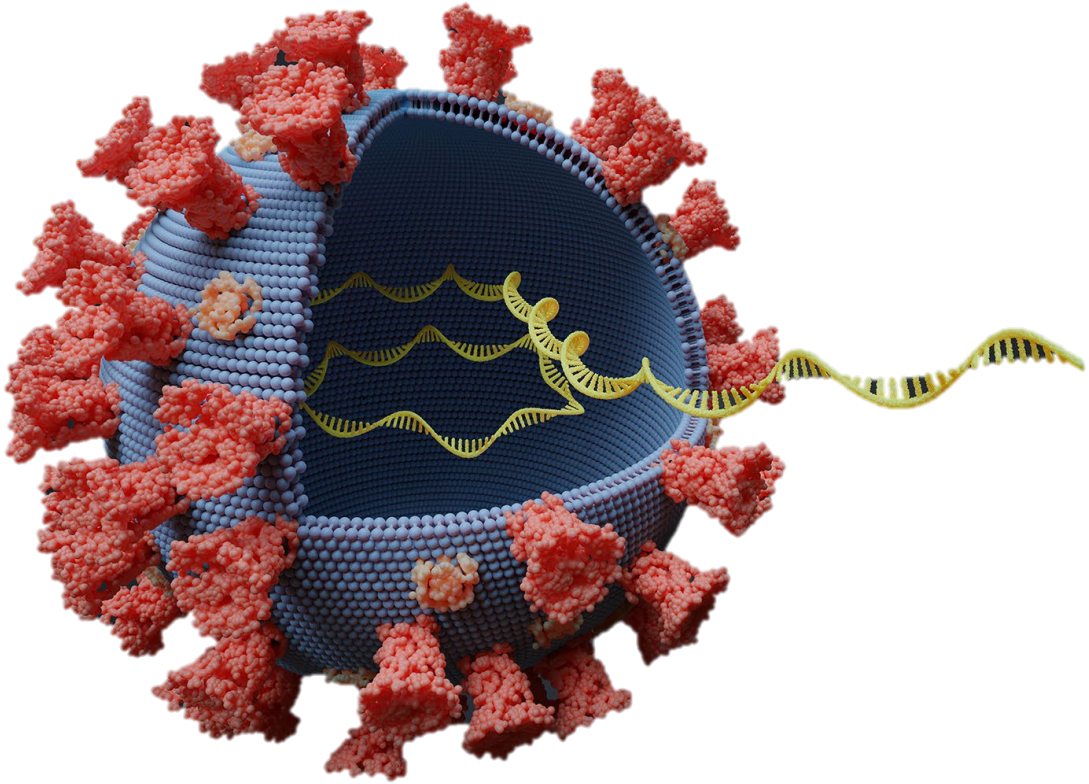


Functional diversity of RNA

- RNA processes both catalytic property and stores genetic information.
- The "**RNA world hypothesis**" proposes that RNA was the central biopolymer in the first life on the Earth and DNA and protein evolved from it.



RNA can encode genetic information

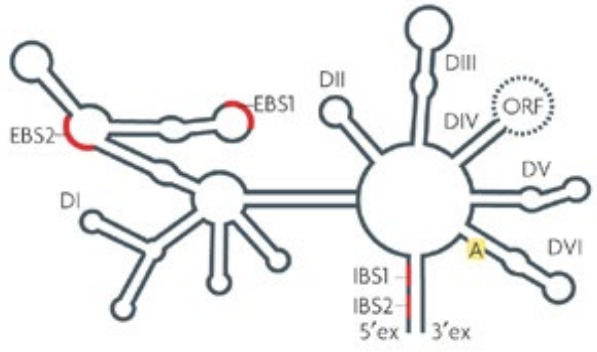


RNA viruses:

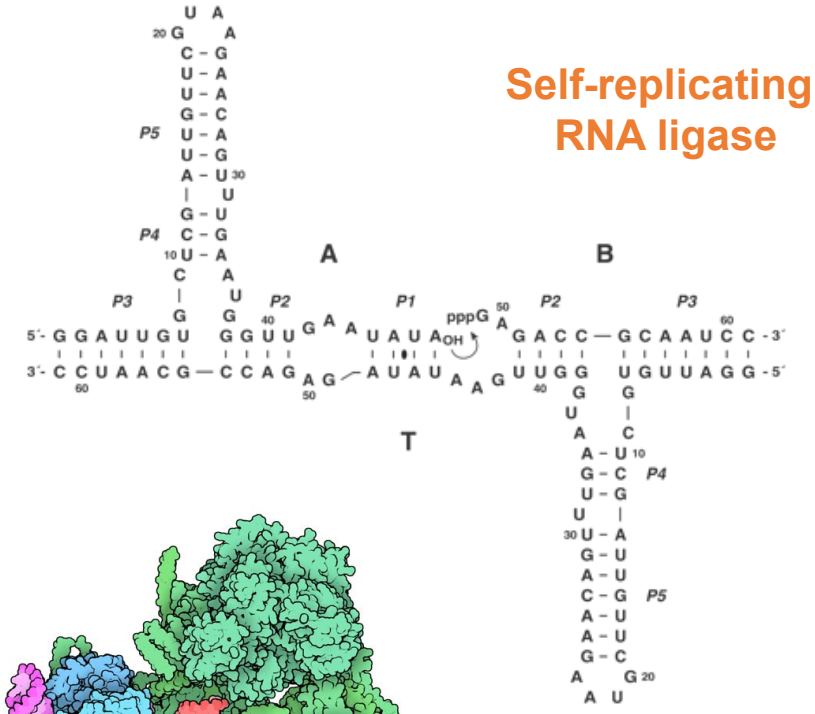
- Influenza virus
- Hepatitis C virus
- Dengue virus
- Zika virus
- HIV
- Ebola virus
- SARS virus
- SARS-CoV-2!

RNA can catalyze biochemical reactions

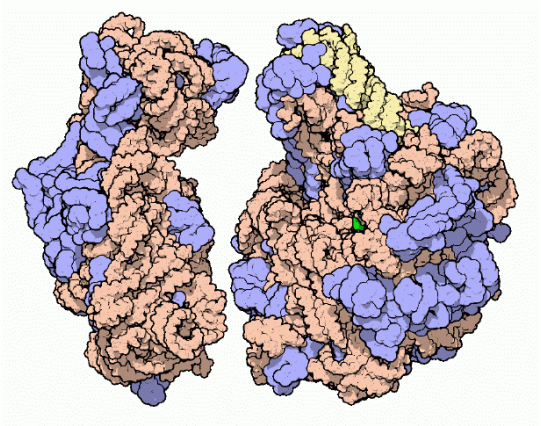
Group II intron

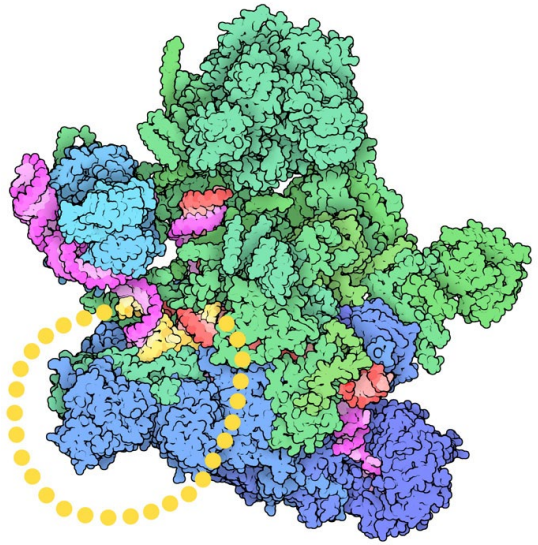


Self-replicating RNA ligase



Ribosome

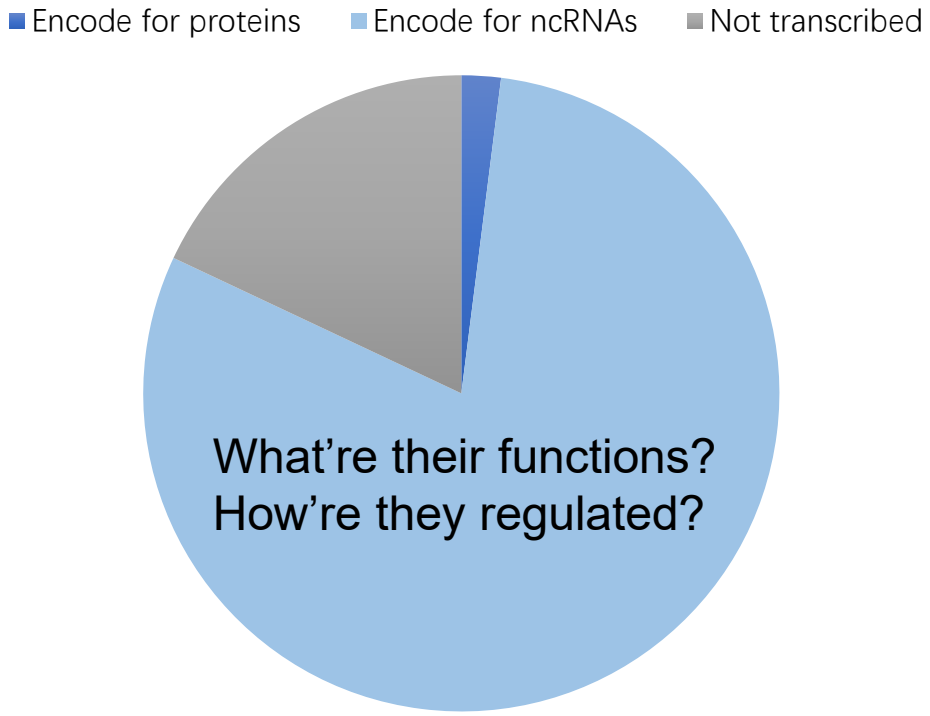




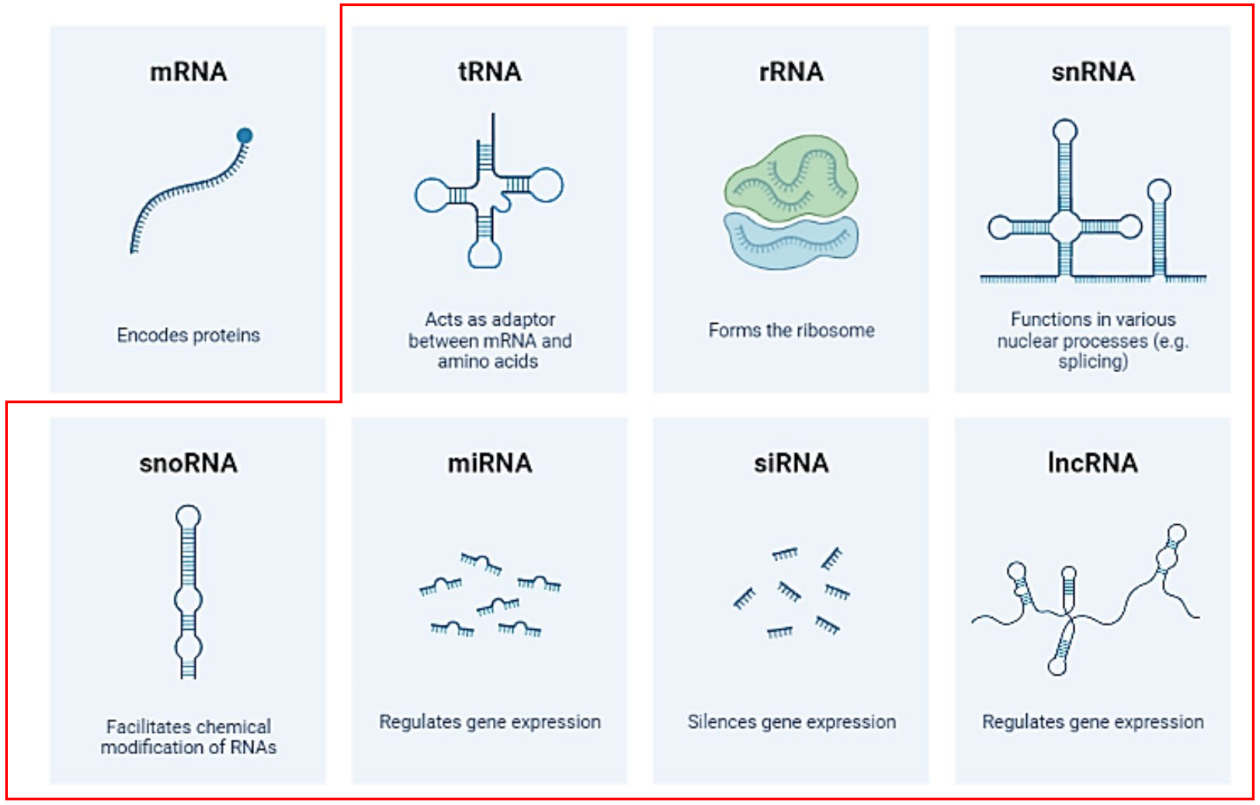
Spliceosome

Dark matter in the human genome: Non-coding RNA

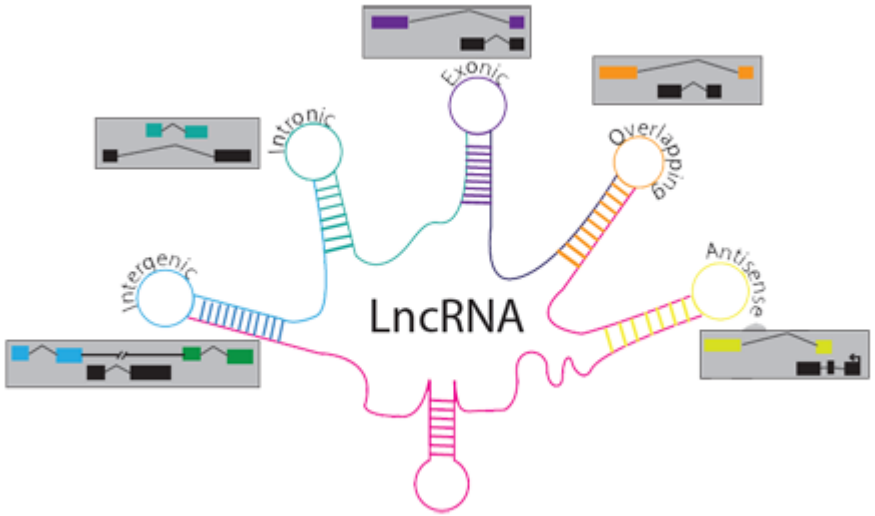
The human genome



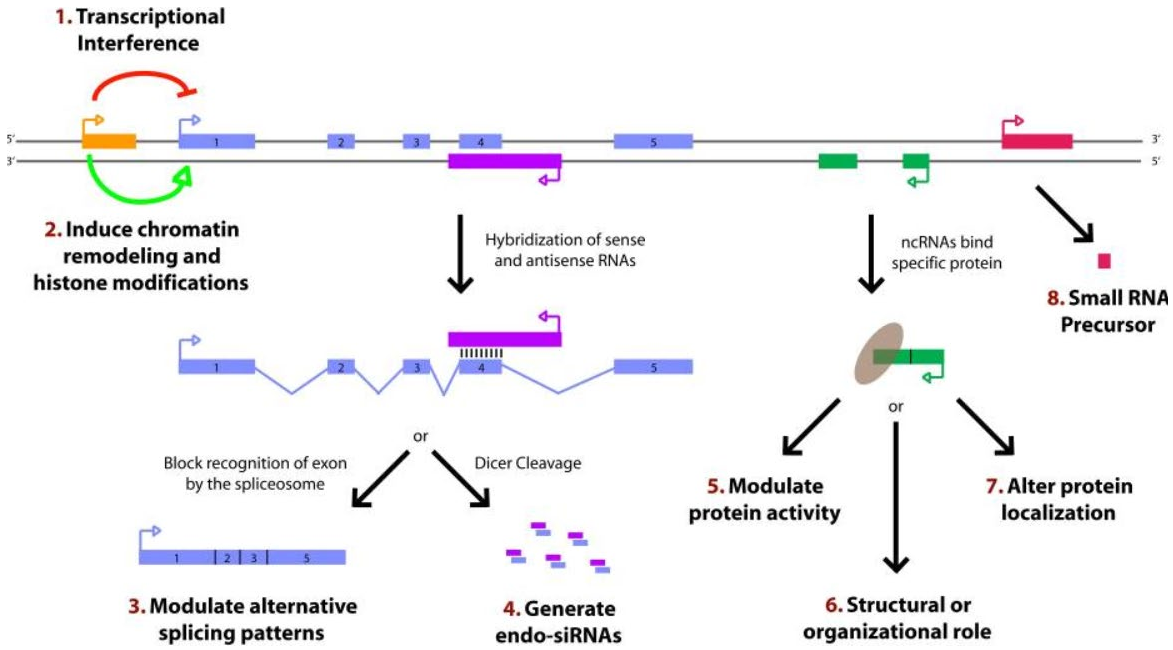
Types of RNA produced in cells

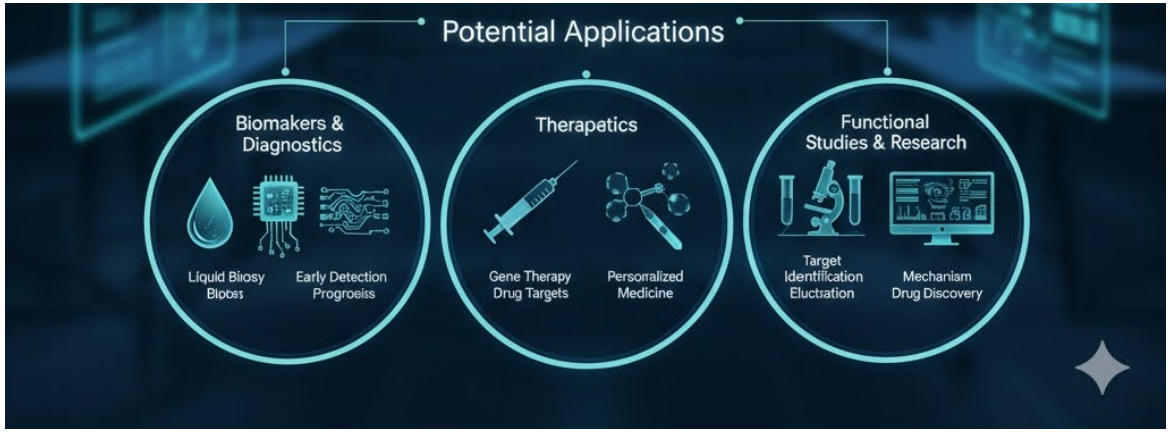


lncRNA

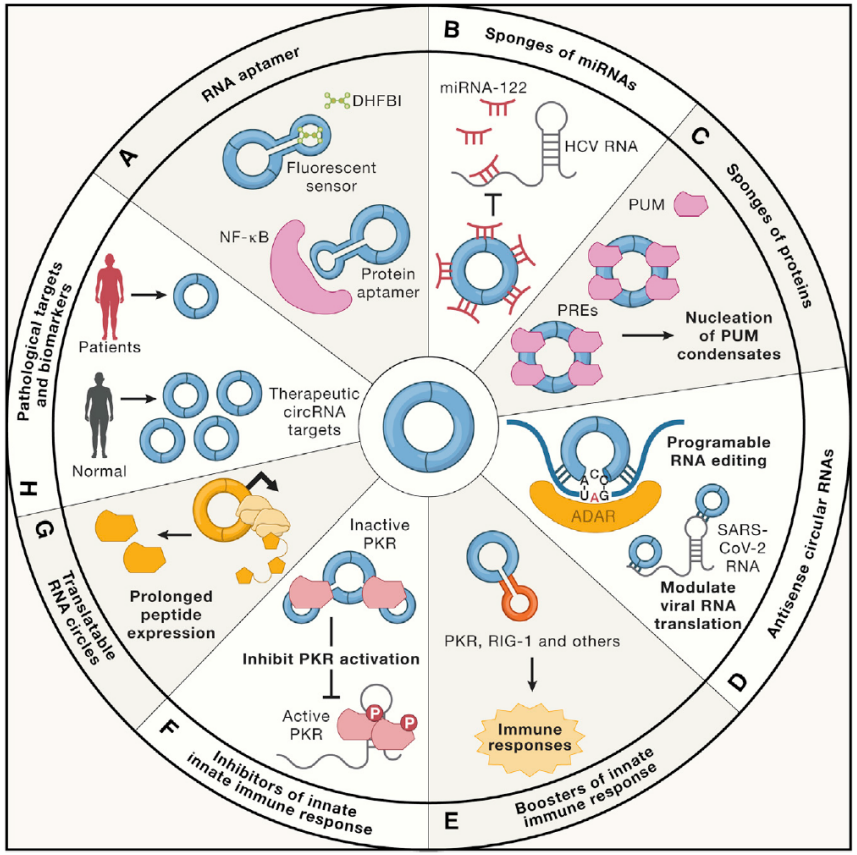


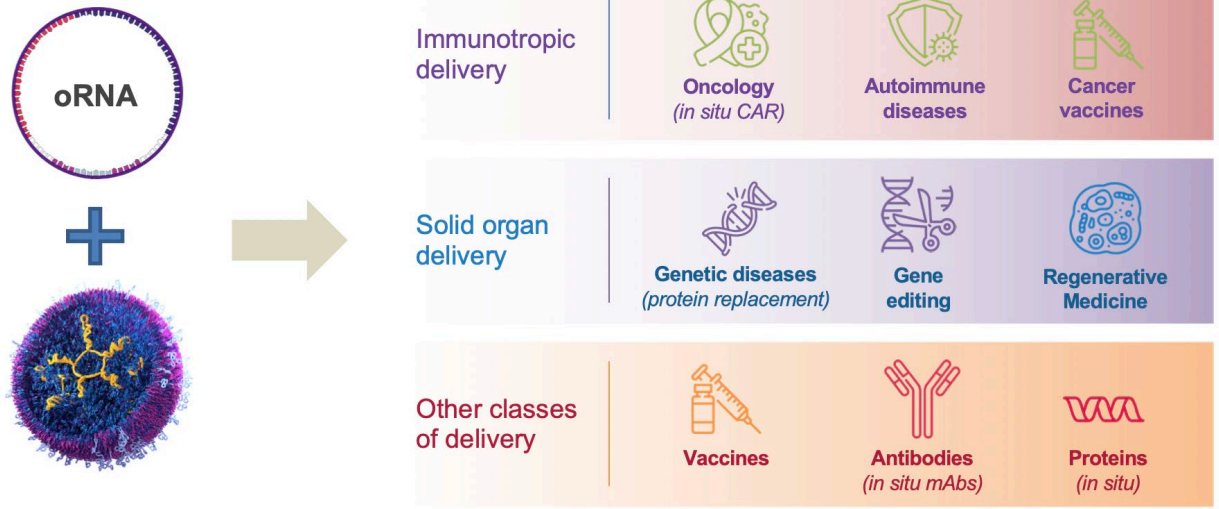
- Intergenic lncRNA
- Intronic lncRNA
- Sense lncRNA (exonic)
- Antisense lncRNA
- Bidirectional lncRNA





The function and application of circular RNAs (circRNAs)

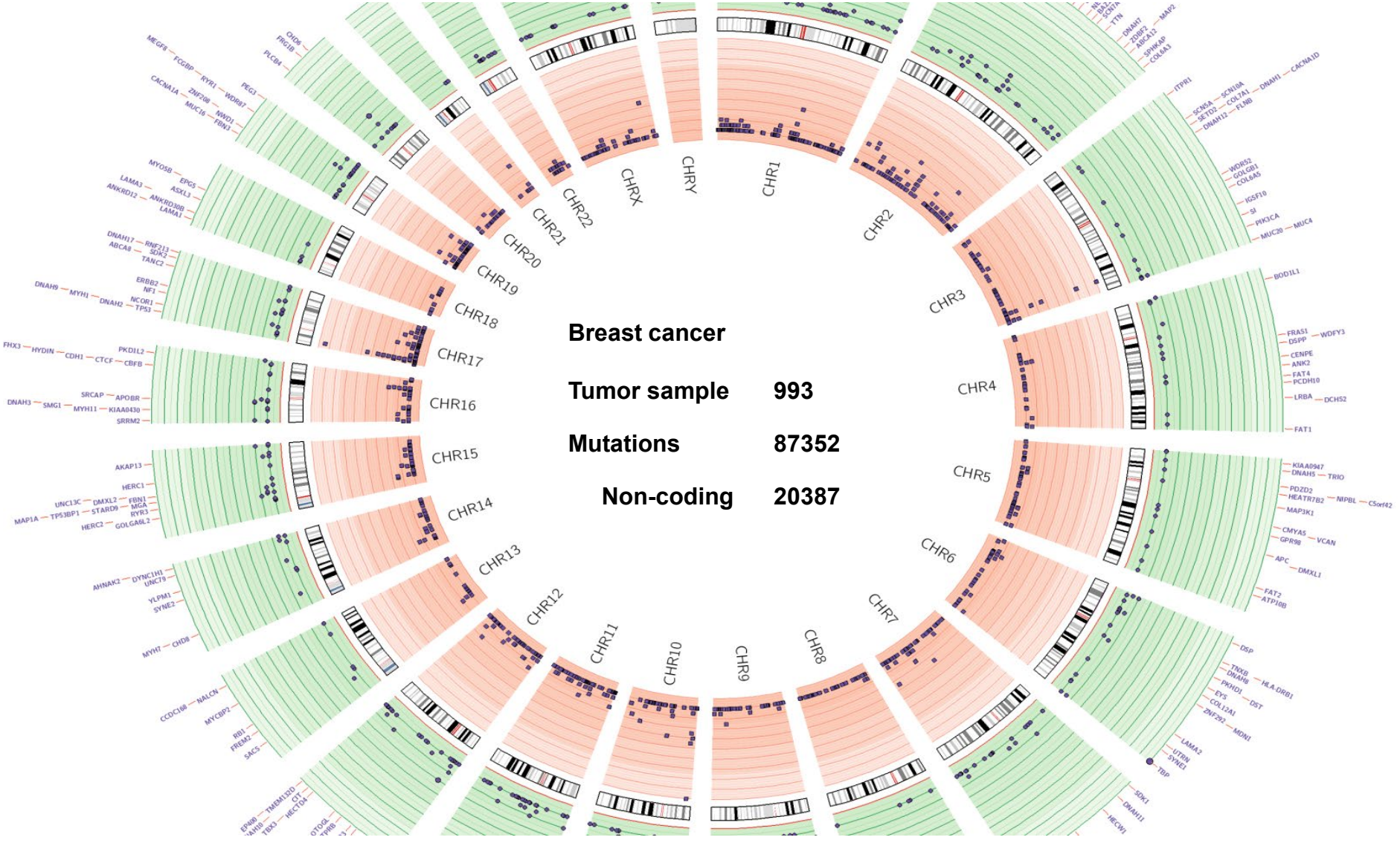




circRNAs are functional: circRNAs mediate complex biological processes by functioning as miRNA sponges, RNA aptamers, and regulators of transcription and translation.

The application of circRNAs: circRNAs hold immense therapeutic and commercial potential, with emerging applications in cancer and autoimmune disease treatment, vaccine development, protein replacement therapy, and gene editing.

"Silent" mutations in the cancer genome



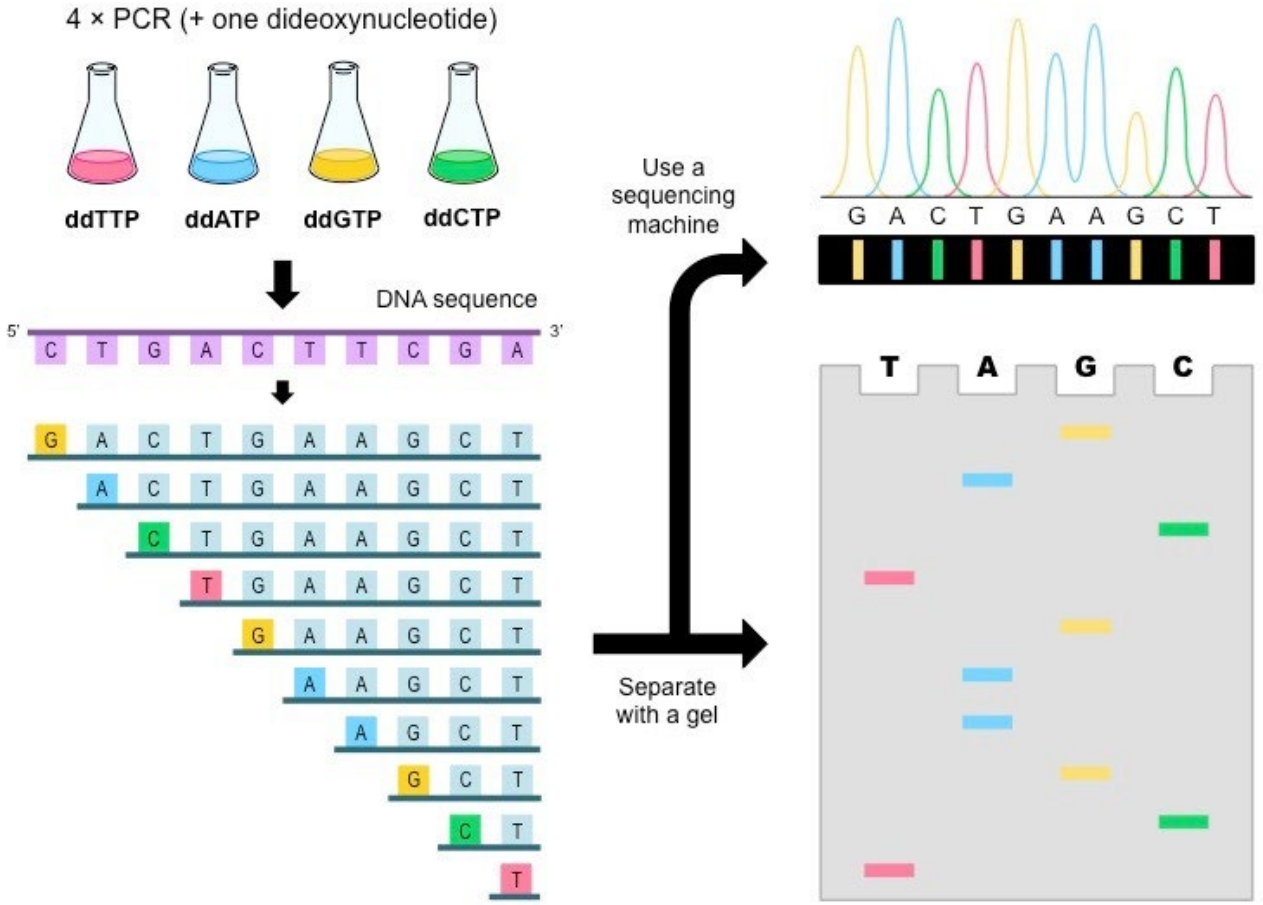

Genomics

Evolution of sequencing technologies



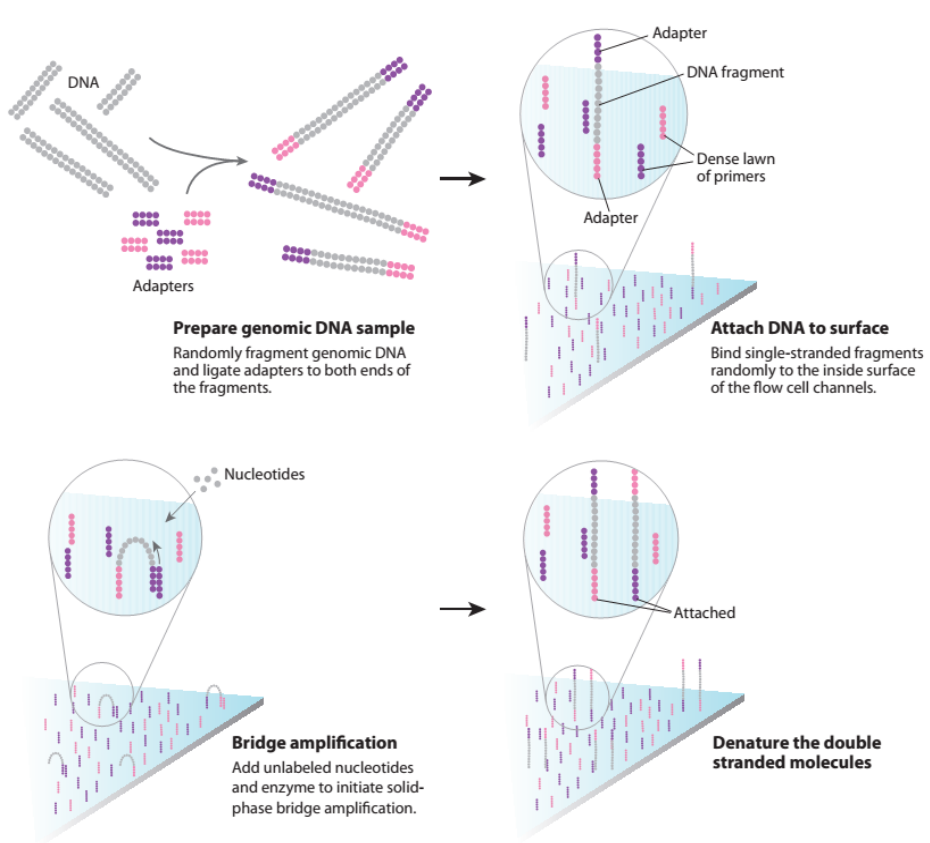
First generation sequencing — Sanger sequencing

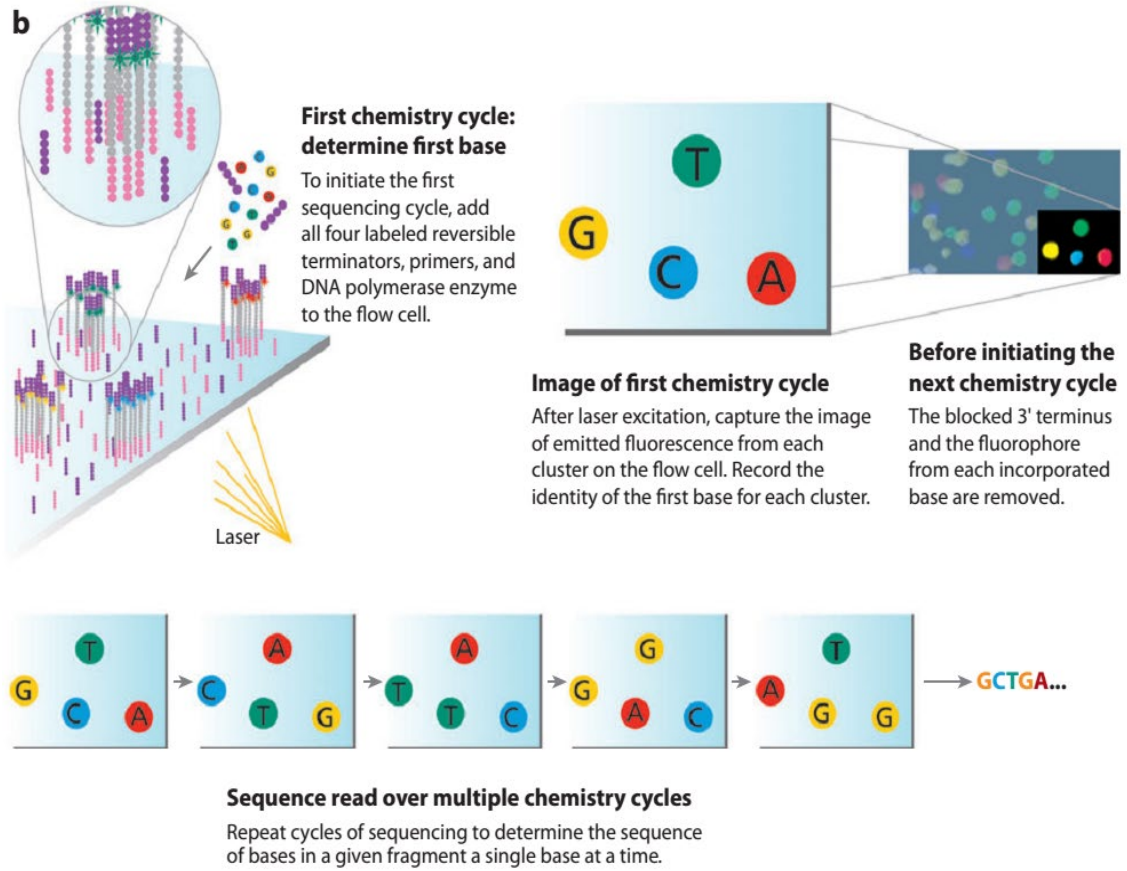
- The principle of Sanger sequencing is based on the termination of DNA strand elongation by ddNTPs
- 500-1,000 bp fragments
- Served as the bedrock of the Human Genome Project



Second generation sequencing — Illumina

- Allows for sequencing at very low cost
- Typically produce reads of ~50–500 bp in length

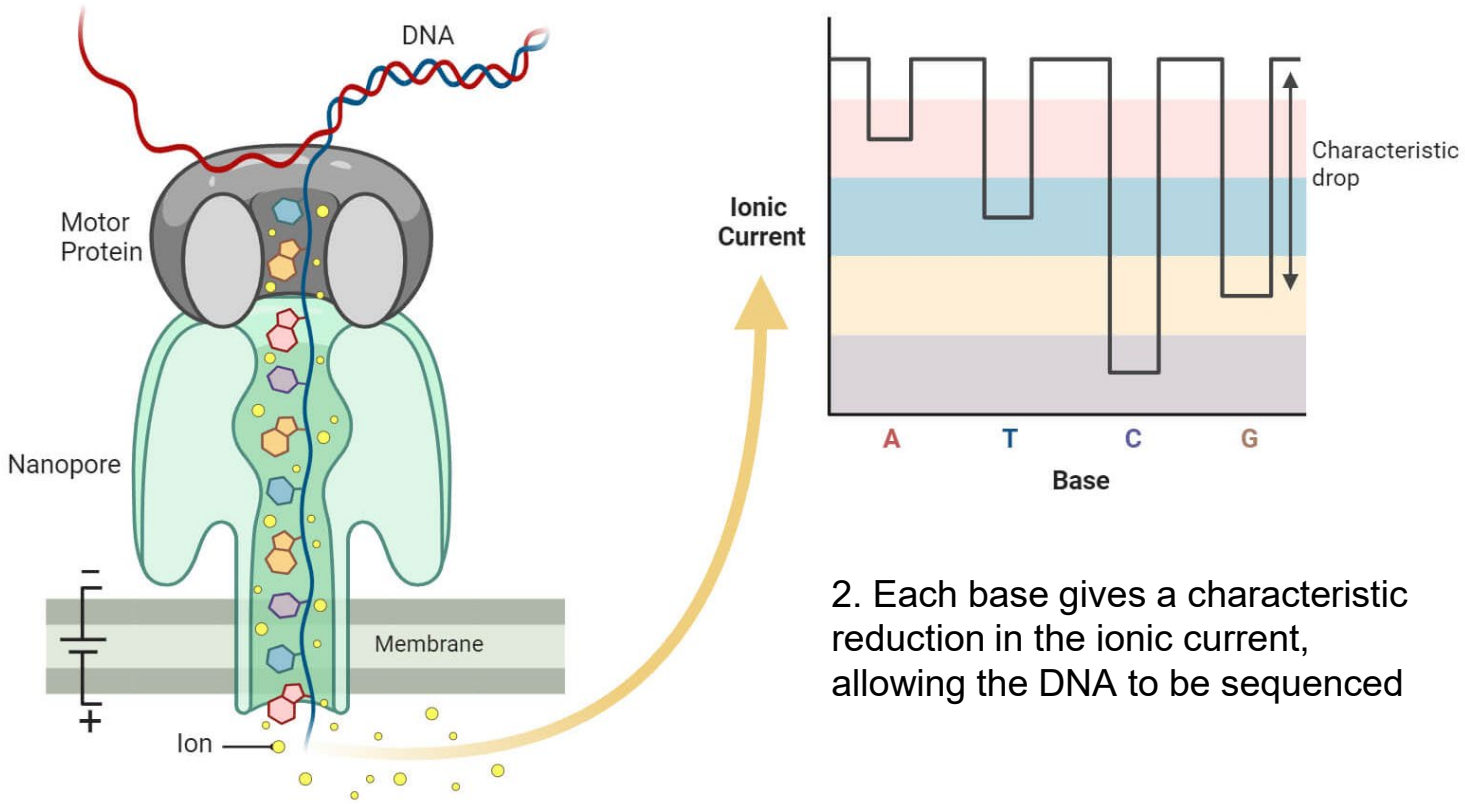




Third generation sequencing — Nanopore

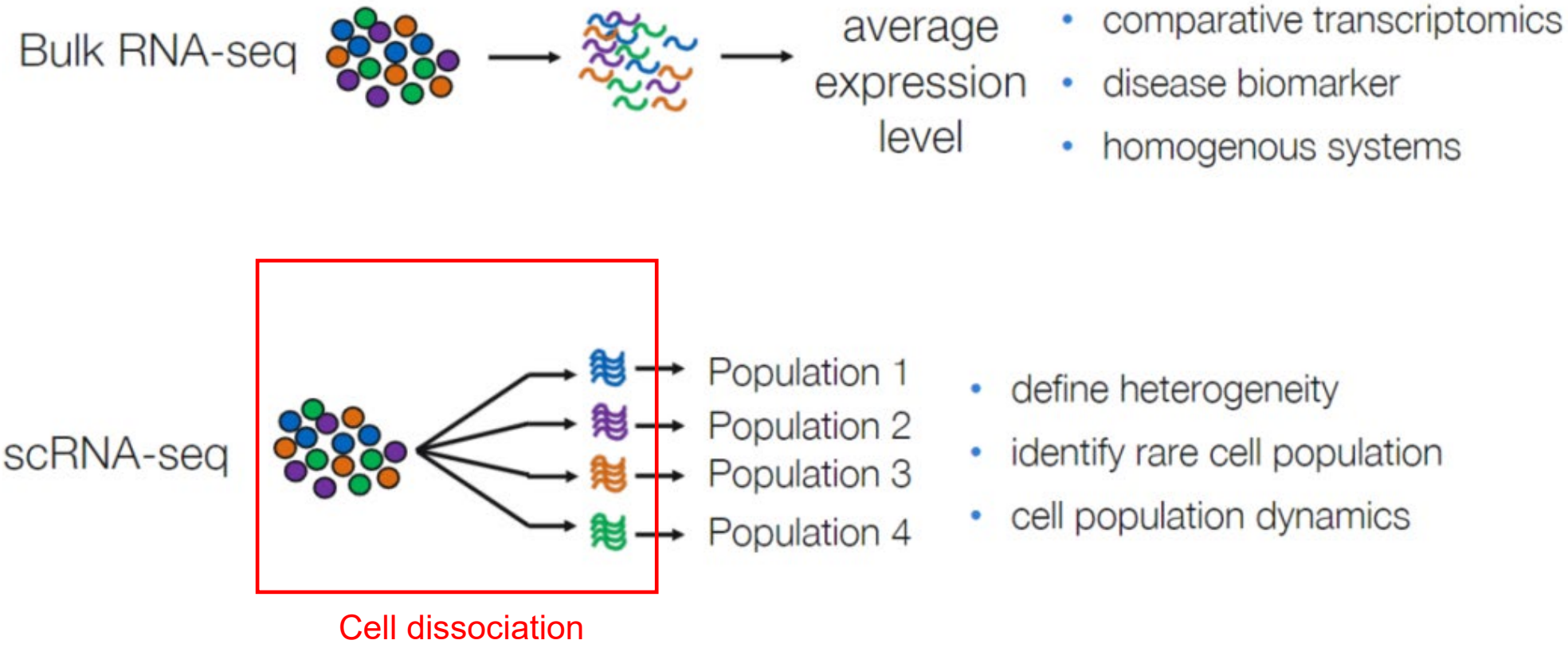
- Long-read sequencing using pore-forming proteins and electrical detection
- Tens of kb fragments

1. DNA is unwound by the motor protein and one strand is translocated through the pore to the +ve side of the membrane

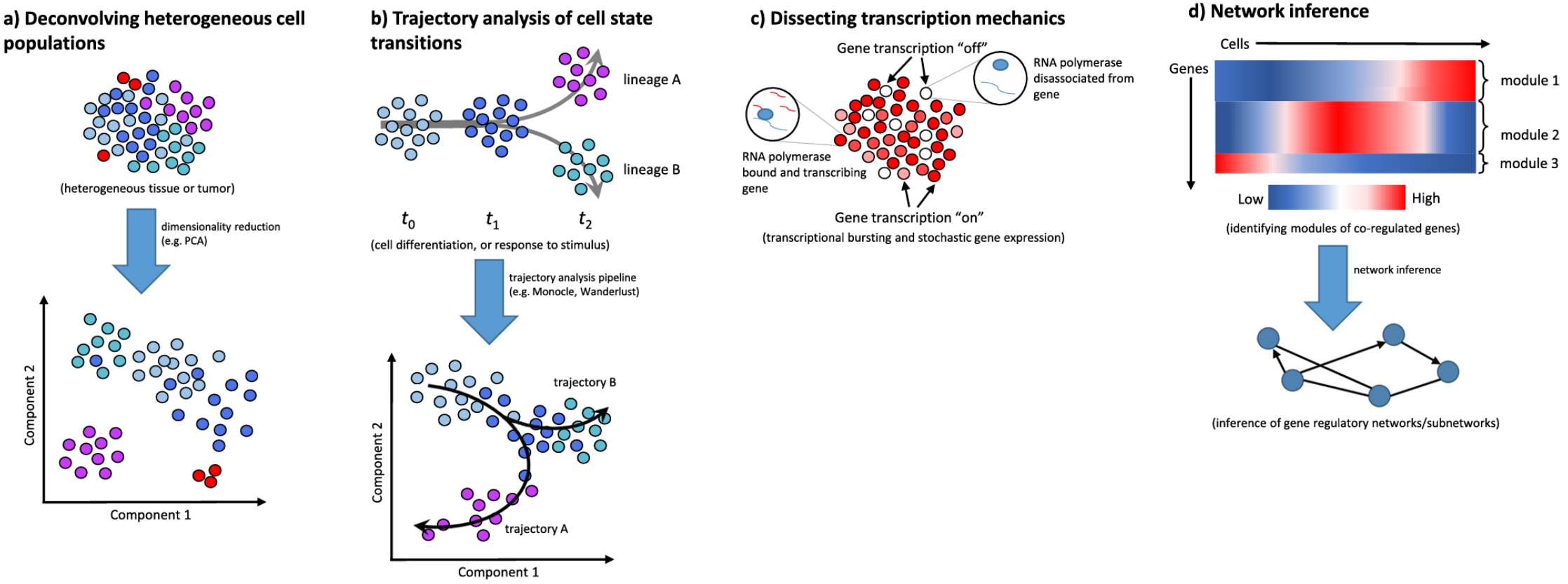


2. Each base gives a characteristic reduction in the ionic current, allowing the DNA to be sequenced

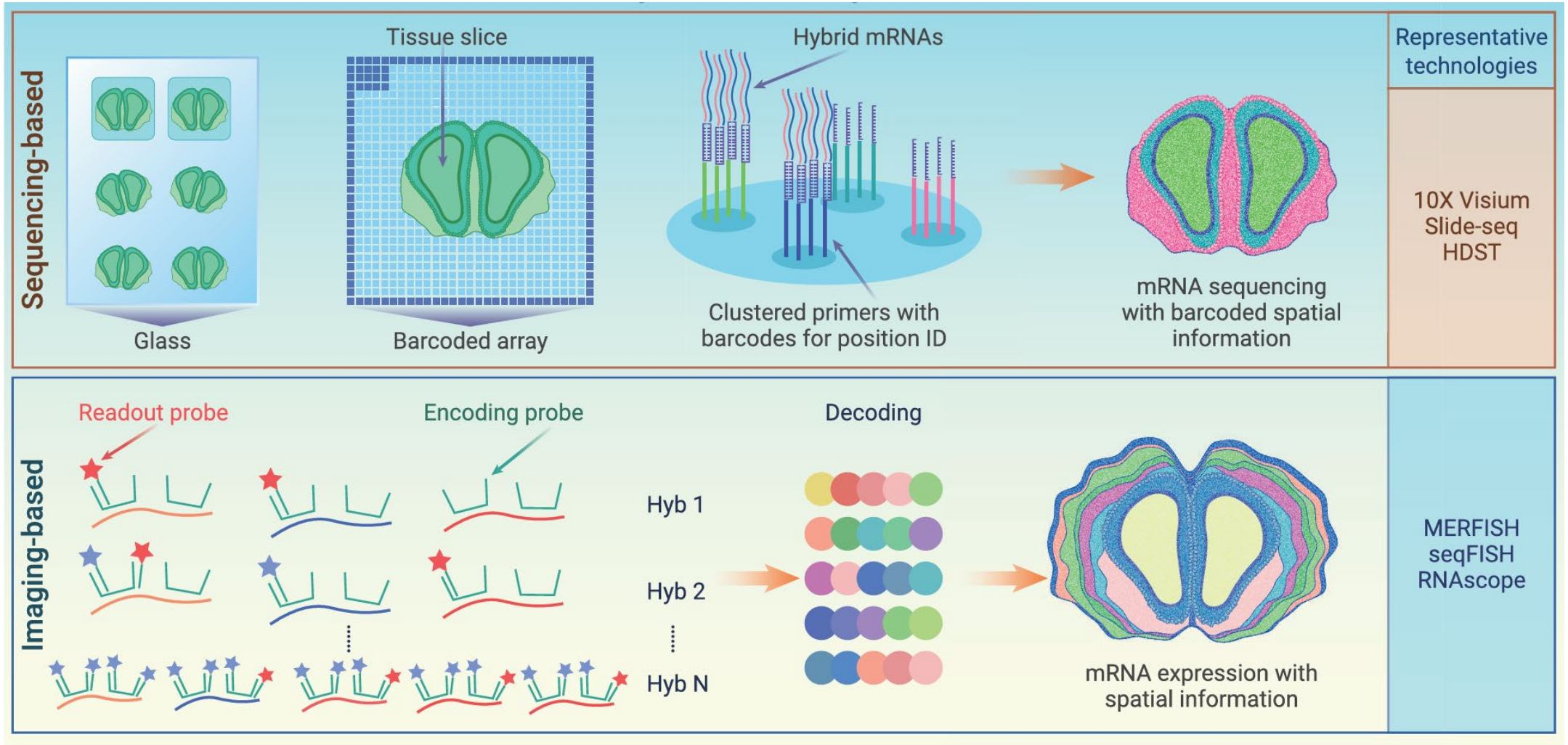
From bulk to single cell RNA-seq (scRNA-seq)



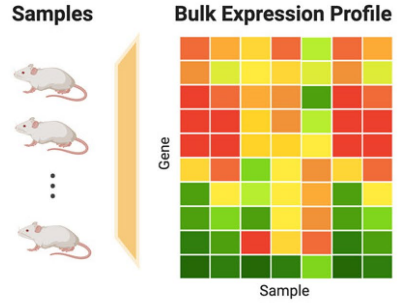
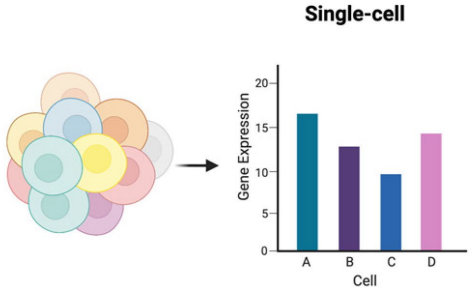
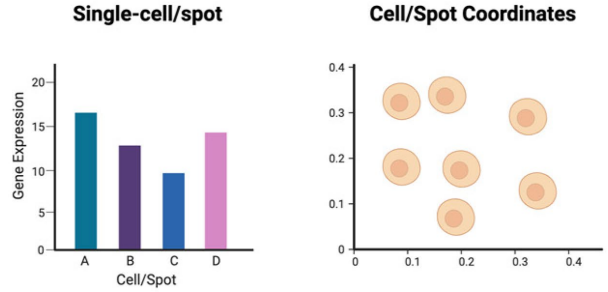
Common applications of scRNA-seq



Spatial transcriptomics preserves the spatial location information of tissues

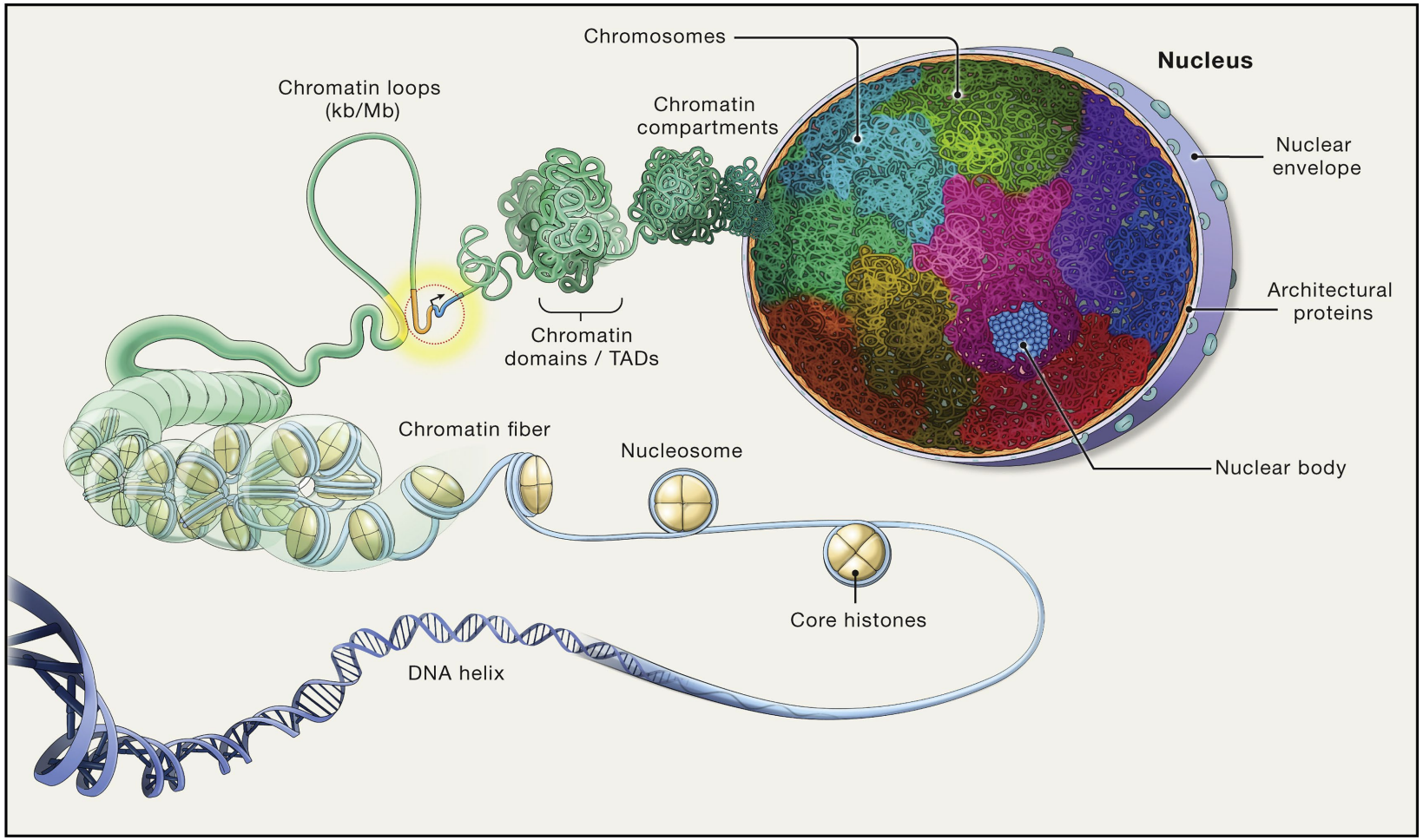


Comparison of bulk RNA-seq, scRNA-seq, and spatial transcriptomics technologies

	Bulk RNA-Seq	Single-cell RNA-seq	High-throughput Spatial Transcriptomics
Level			
Data Structure	Subject x Gene Expression Count Data	Cell x Gene Expression Count Data	Cell/Spot x Gene Expression Count Data Cell/Spot 2-dimensional Coordinates
Detection Target	Differentially Expressed Genes	Differentially Expressed Genes Cell Sub-populations	Spatially Variable Genes Tissue Architecture Cell-Cell Communication

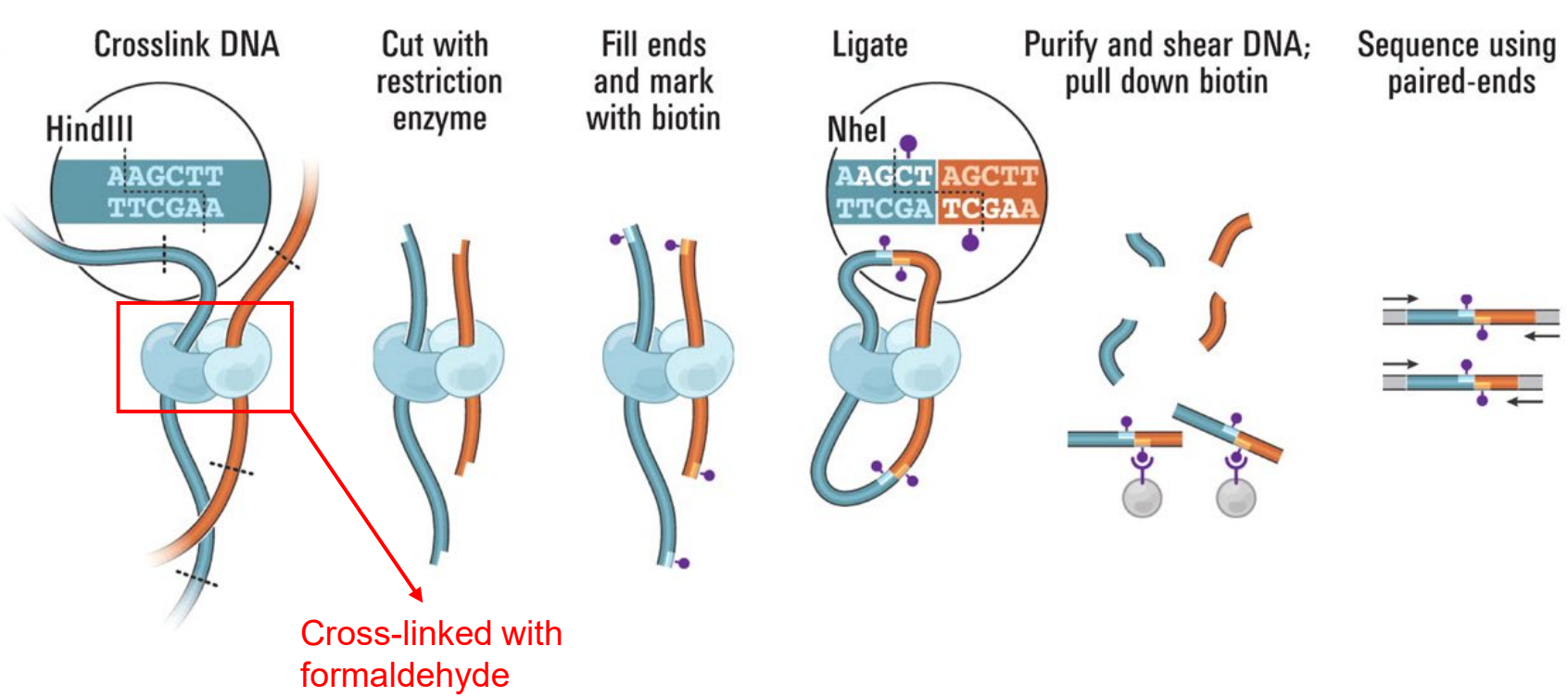
The organization of the eukaryotic genome

- The relatively stable high-order structure of DNA is maintained by **proteins**.

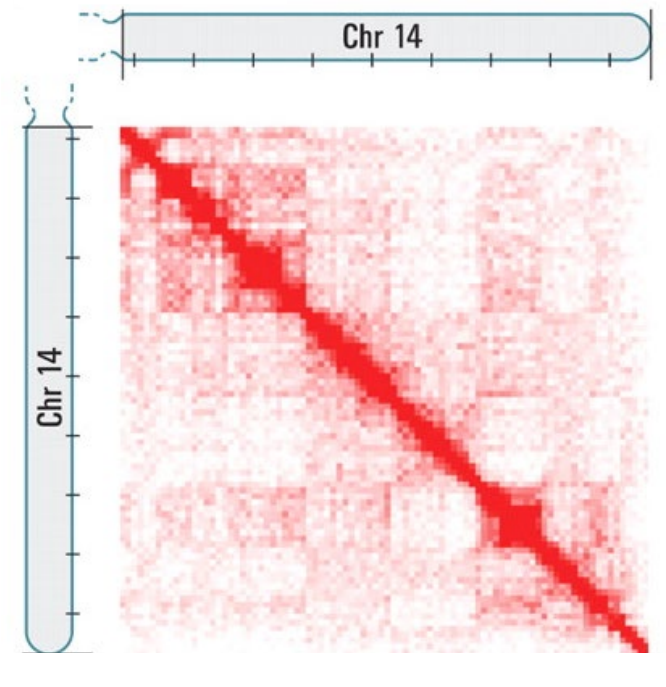


Hi-C reveals the dynamic conformations of whole genomes

Overview of Hi-C



Hi-C produces a genome-wide contact matrix

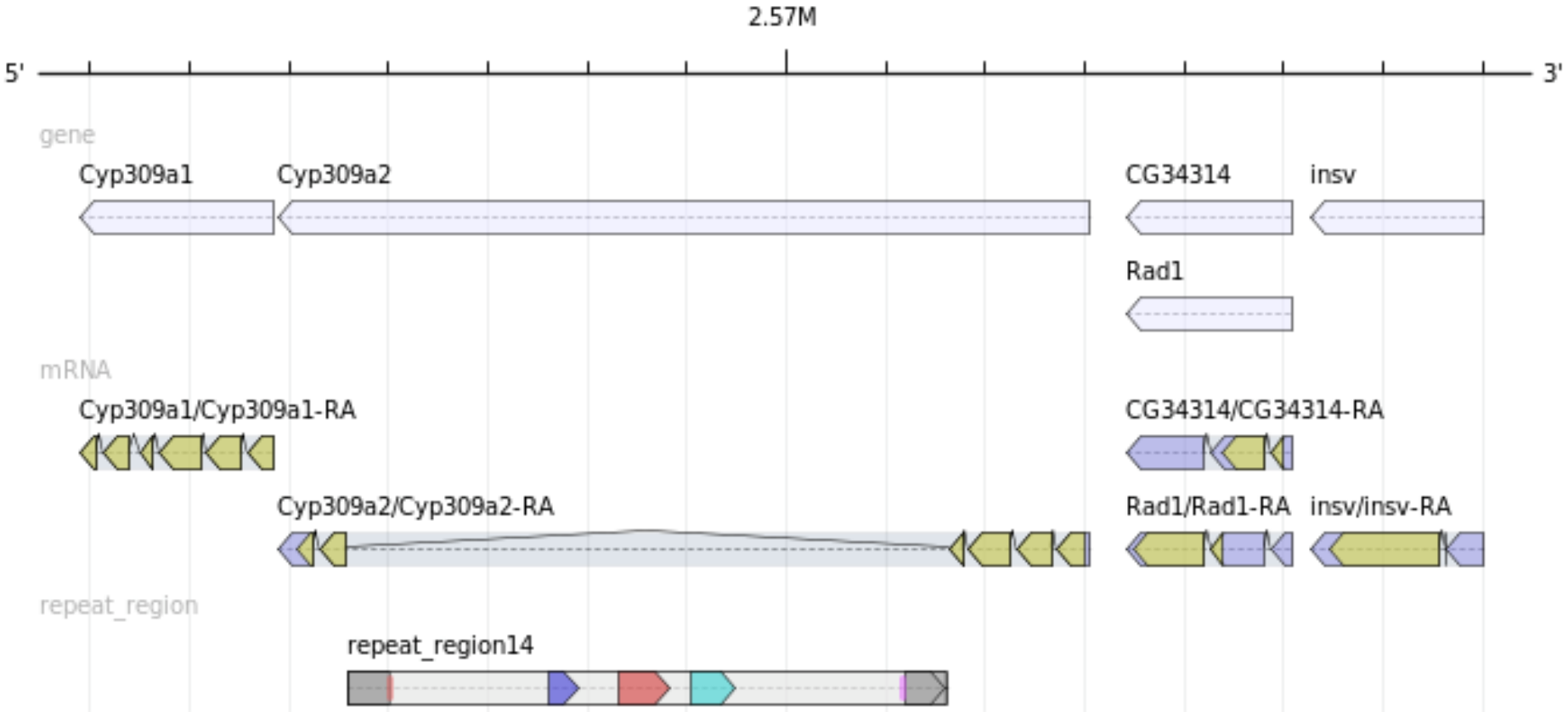


Genome annotation

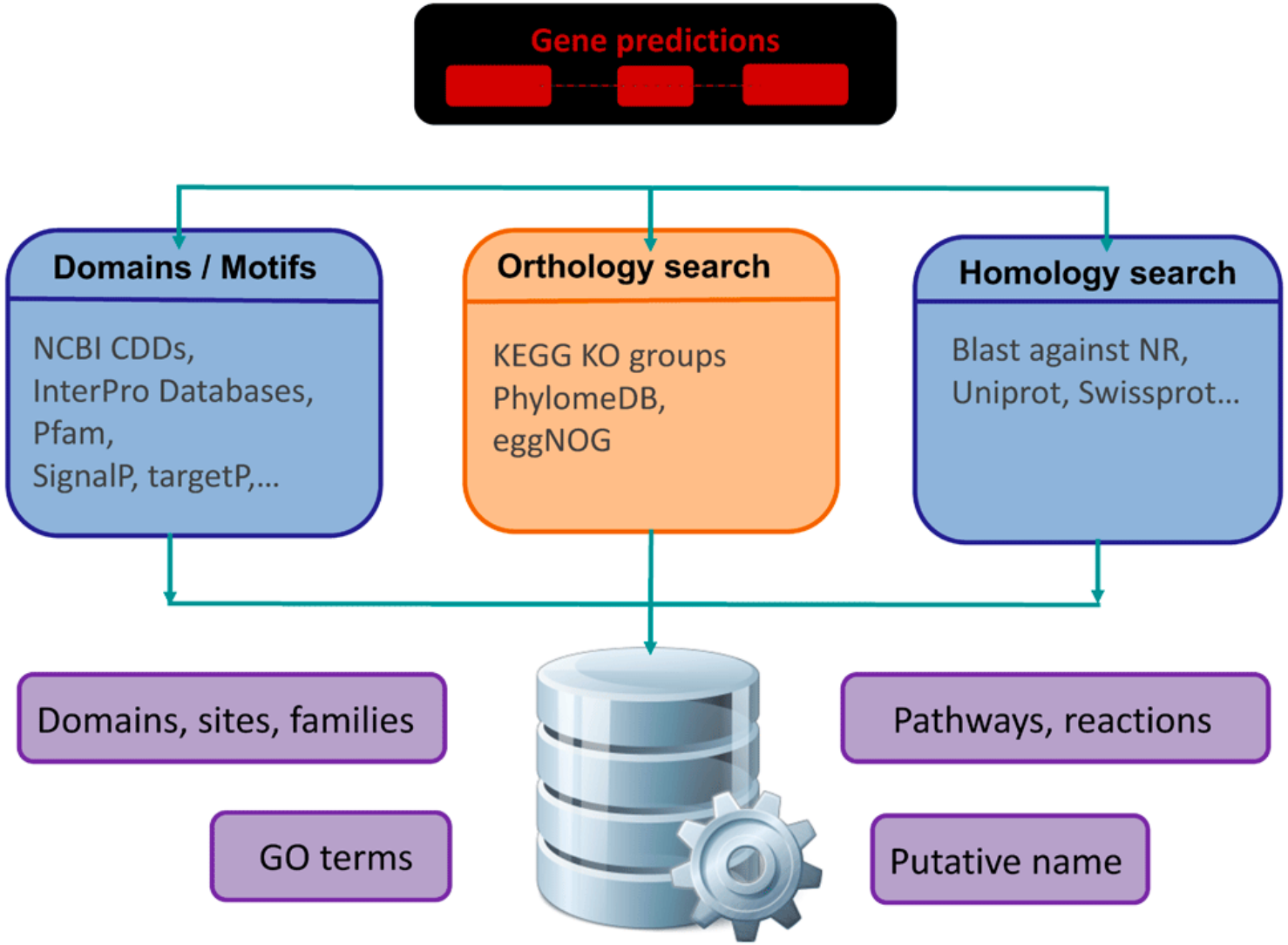
- Purpose and definition
  - ➢ “Structural” genome annotation is the process of identifying genes and their intron–exon structures.
  - ➢ “Functional” genome annotation is the process of attaching meta-data such as gene ontology terms to structural annotations.



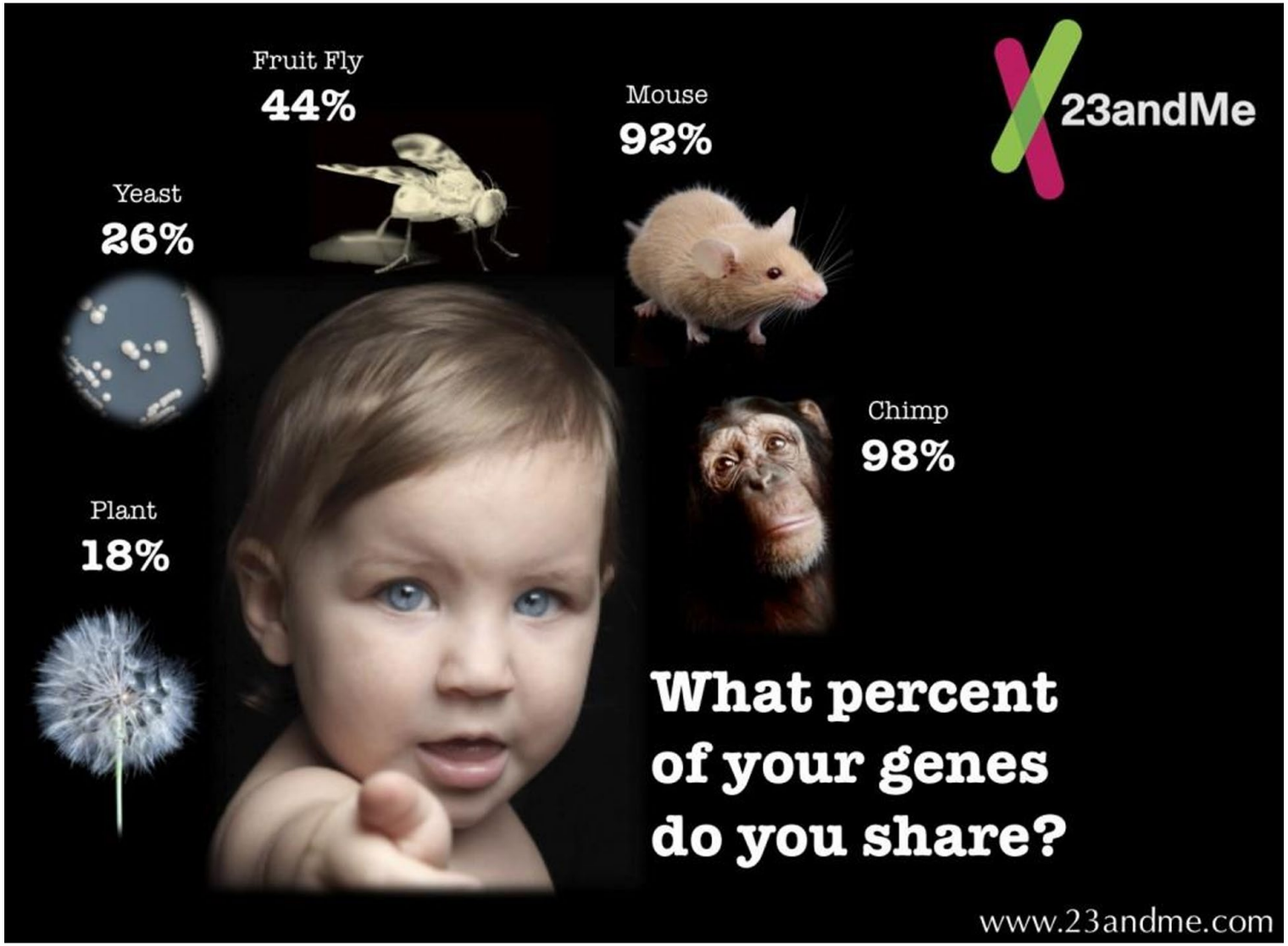
Structural genome annotation



Functional genome annotation

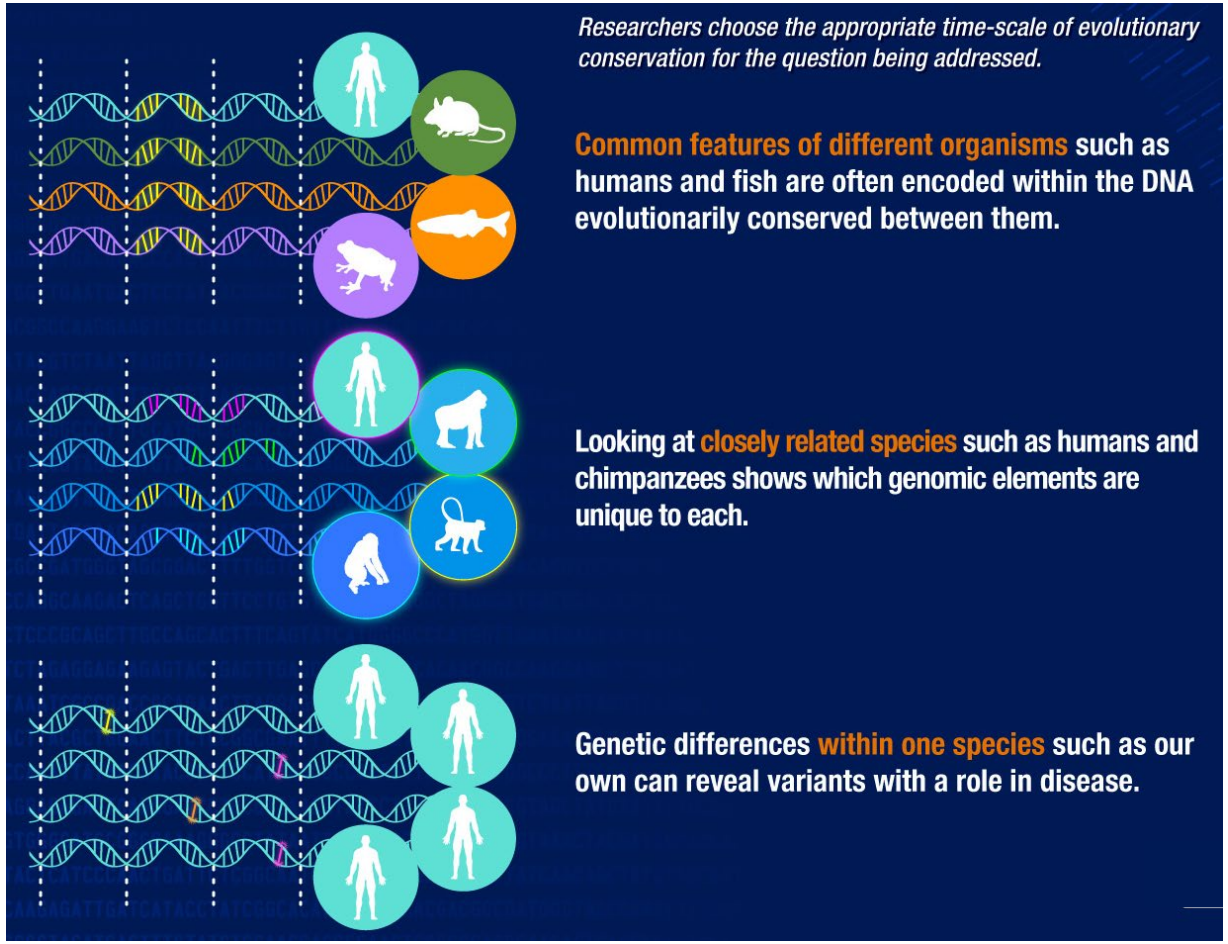


How do different species differ?

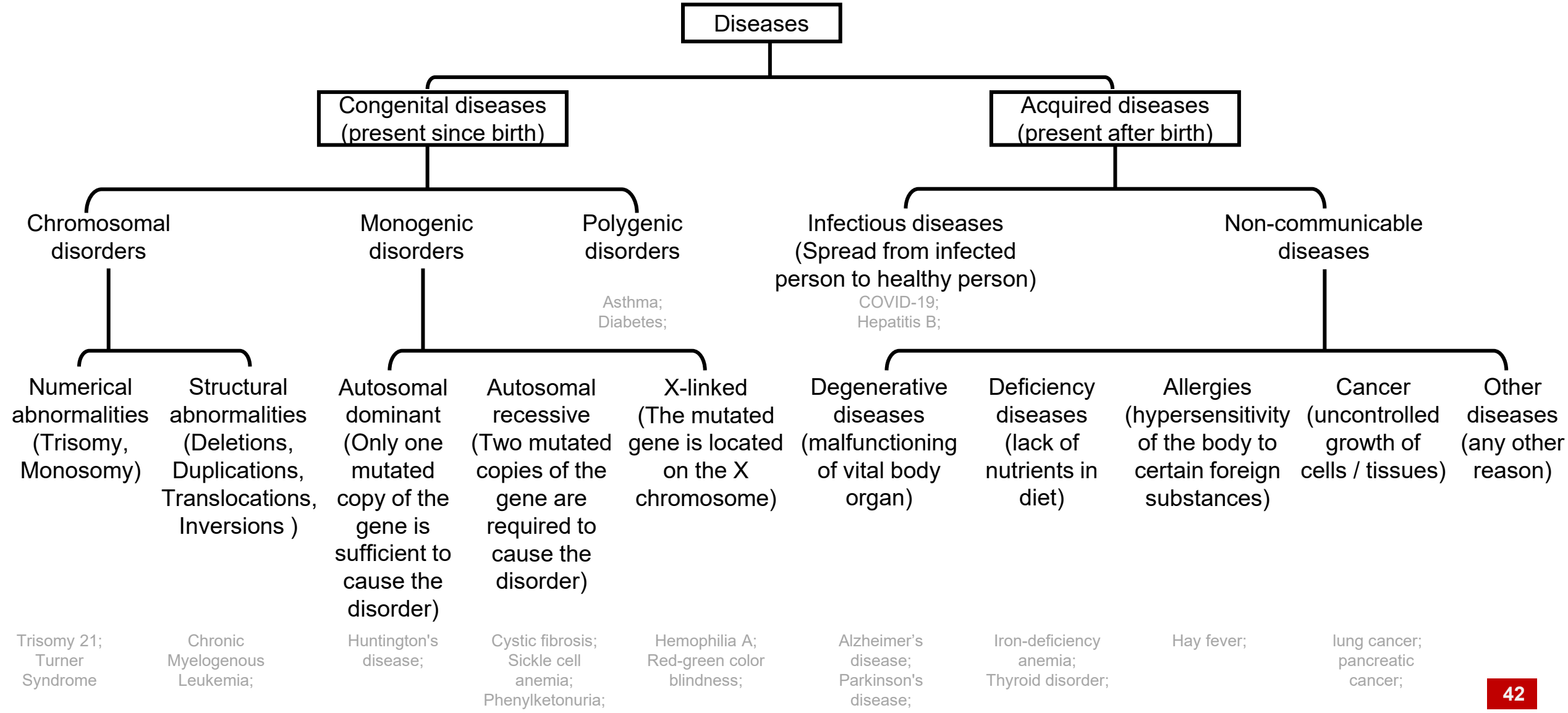


Comparative genomics

- **Comparative genomics** is a field of biological research in which researchers use a variety of tools to compare the complete genome sequences of different species.
- **Interspecies comparison**: By comparing the genomic sequences of species with different genetic relationships, evolutionary conserved regions, species-specific regions, and evolutionary relationships among species are studied.
- **Intraspecies comparison**: Identifying variations and polymorphisms within the genome of the same species, and studying individual and population susceptibility to diseases as well as differential responses to drugs and environmental factors based on these variations.

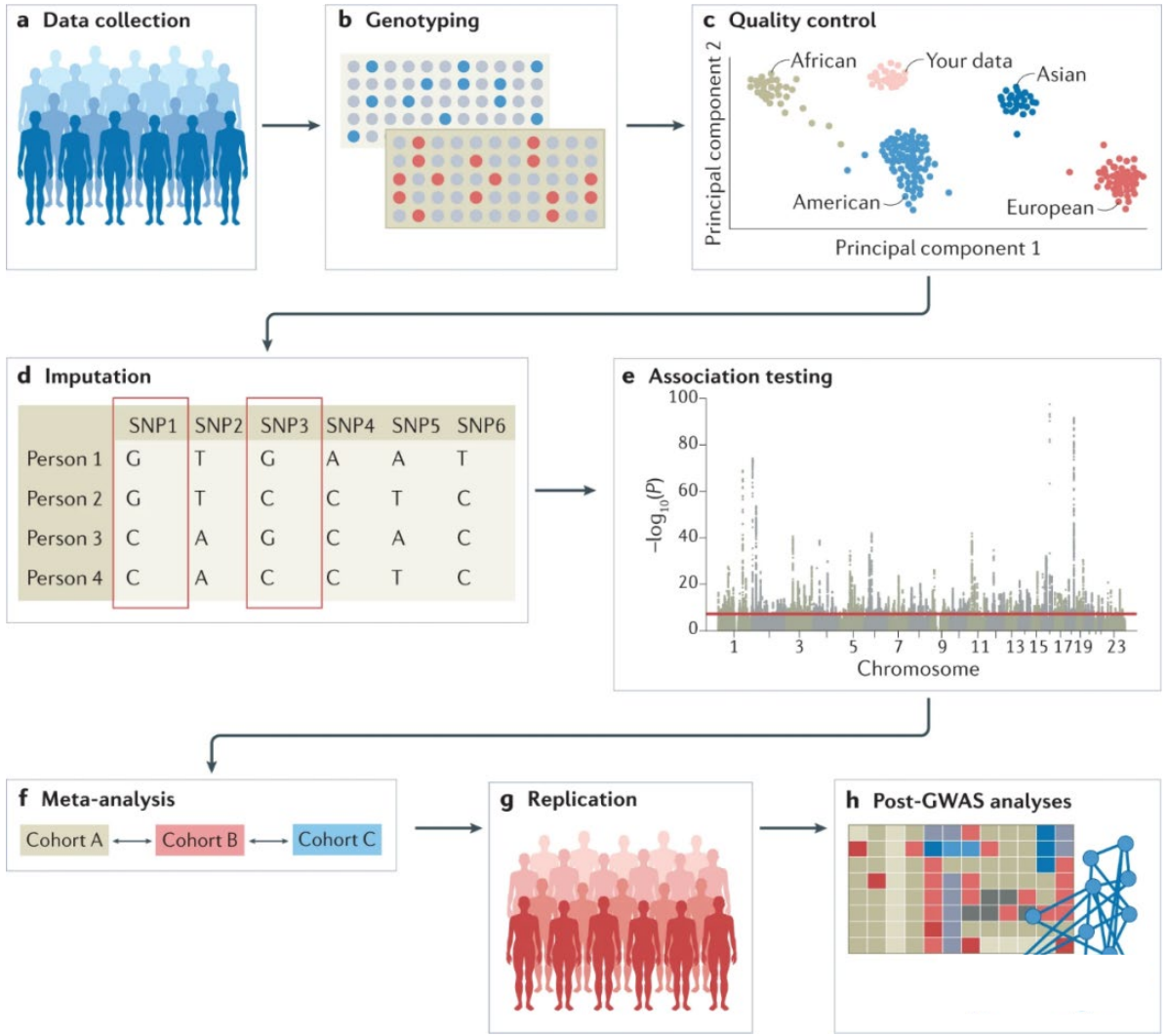


Diseases and their genetic basis



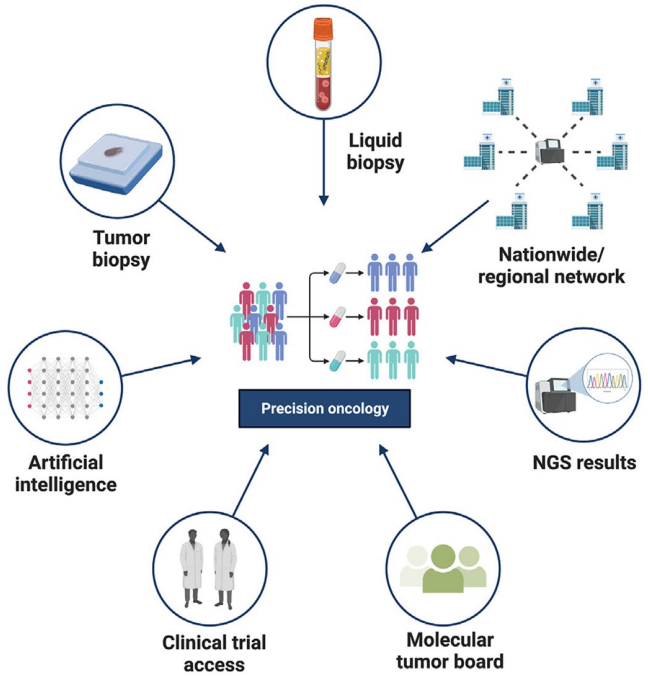
Monogenic disorders and GWAS analysis

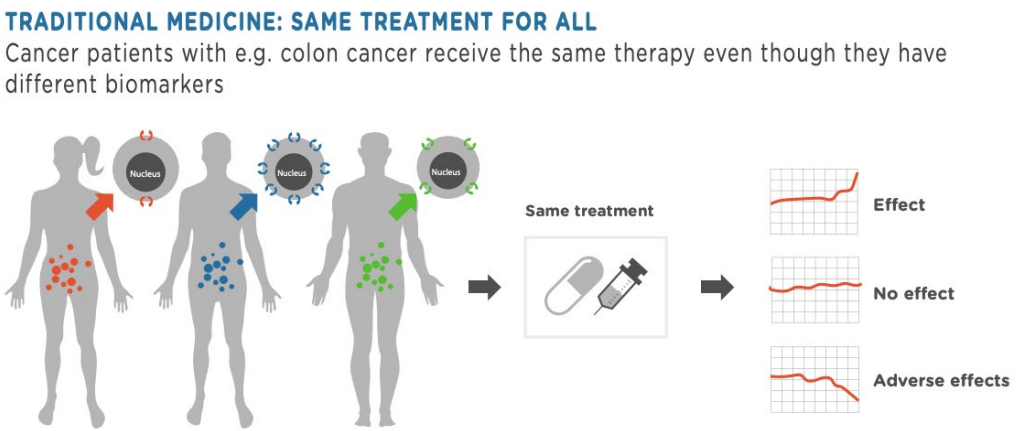
Mendelian disease (monogenic disorder) is the result of a single mutated gene. It is identifiable through linkage analysis, whereas GWAS is designed to detect the collective contribution of numerous common, low-penetrance variants that underlie **polygenic complex diseases**.

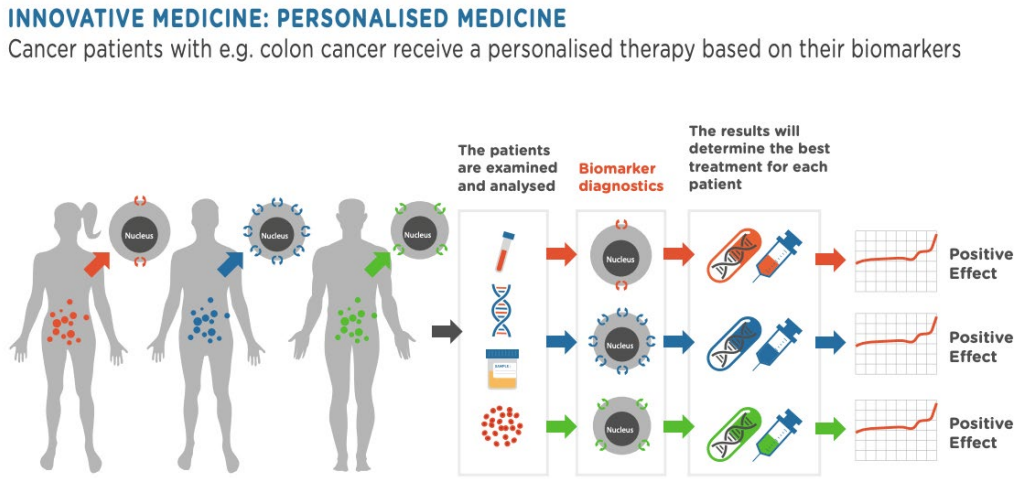


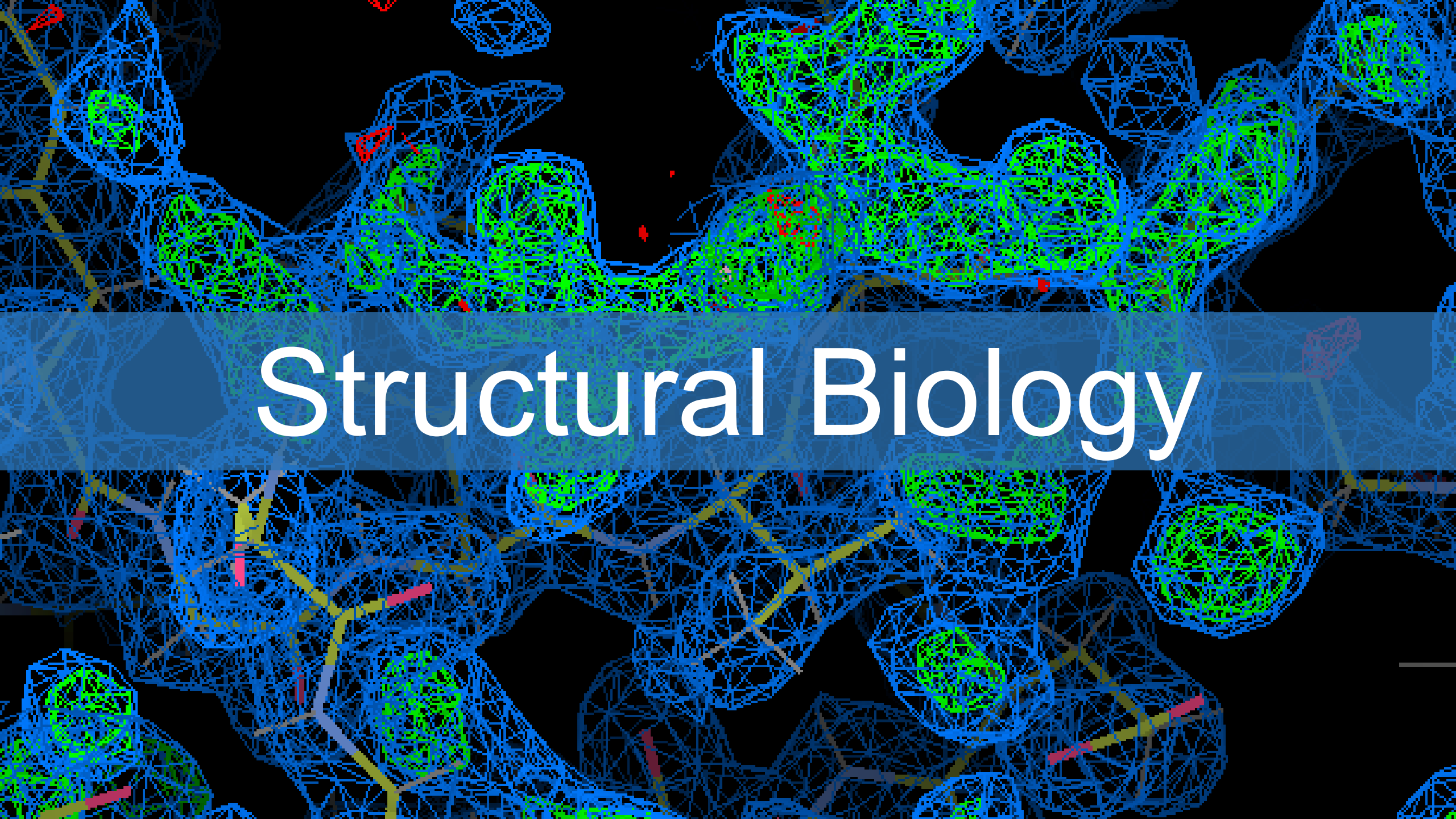
Precision medicine

- **Precision medicine**, also referred to as **personalized medicine**, is a healthcare approach that utilizes molecular information (e.g. genomic, transcriptomic, proteomic, metabolomic), phenotypic and health data from patients to generate care insights to prevent or treat human disease resulting in improved health outcomes.



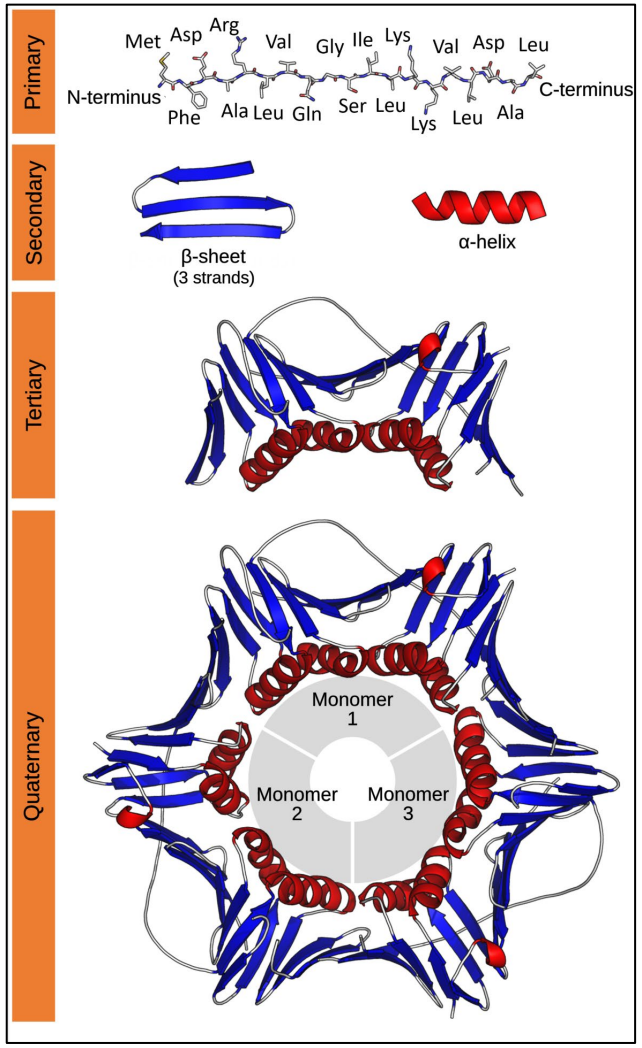




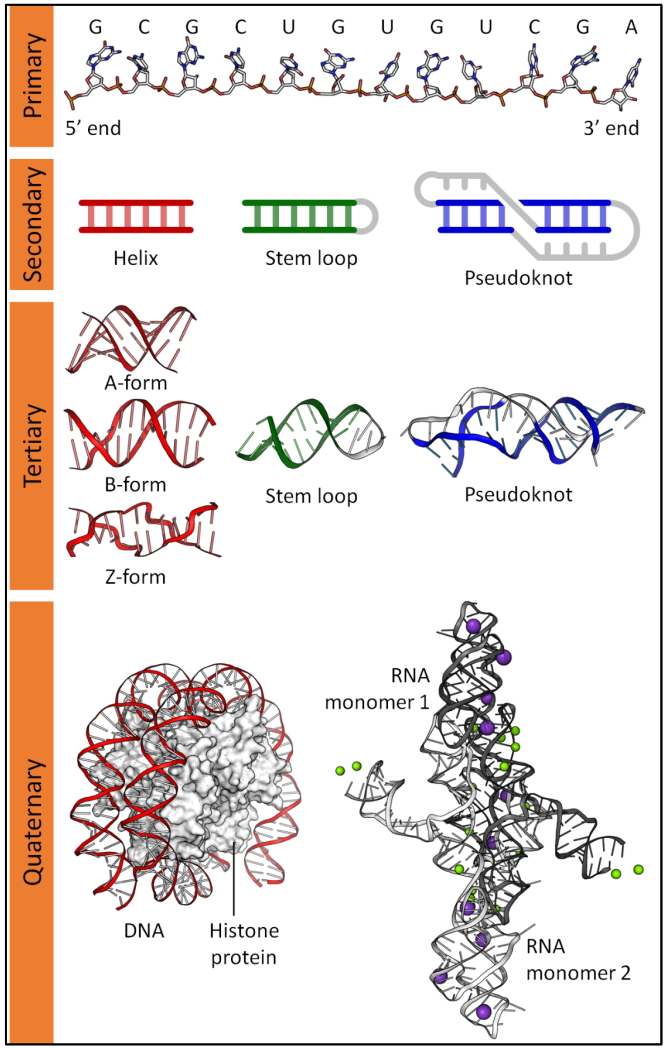

Structural Biology

Hierarchical organization of biomolecule structure

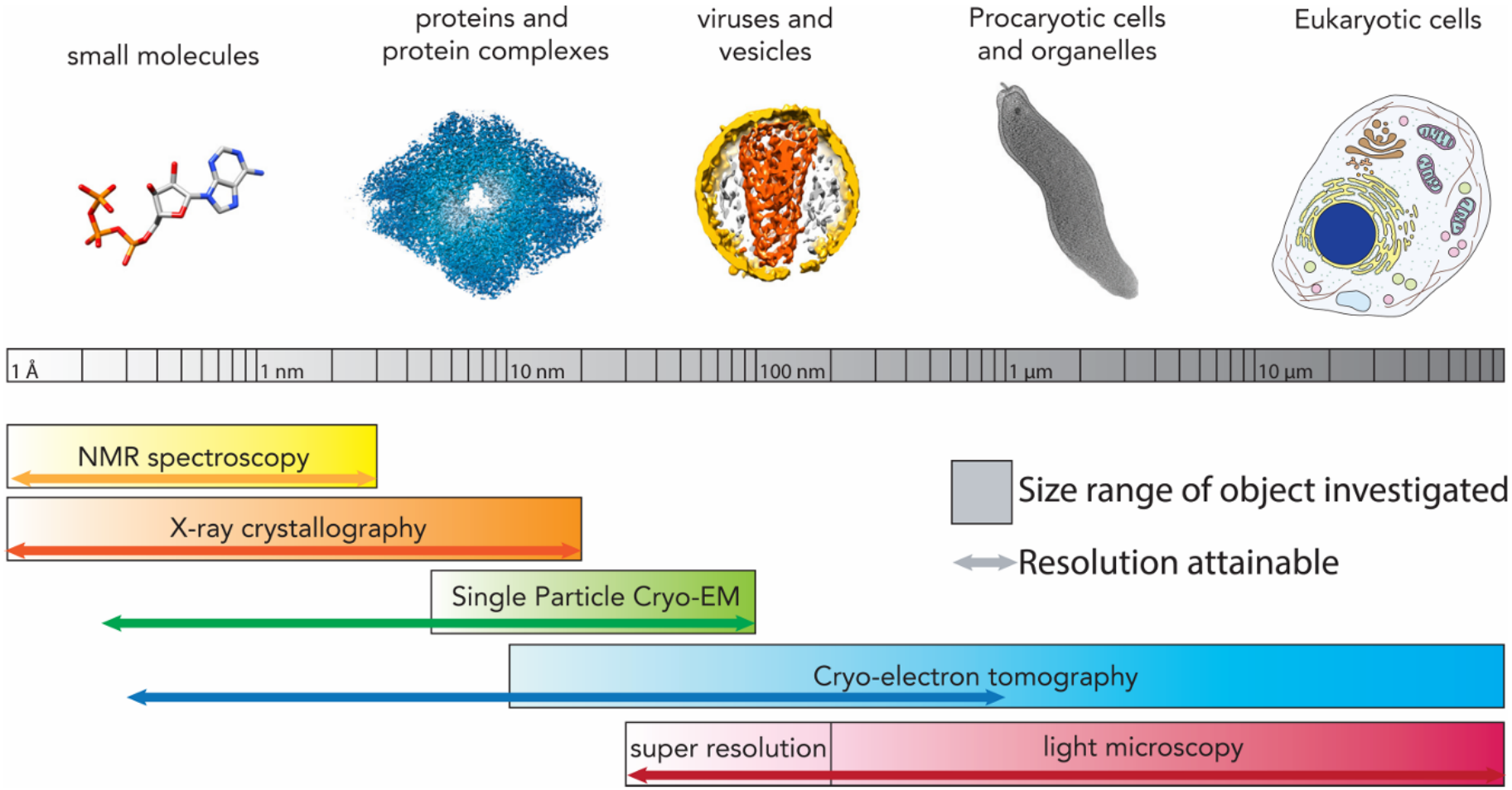
Proteins



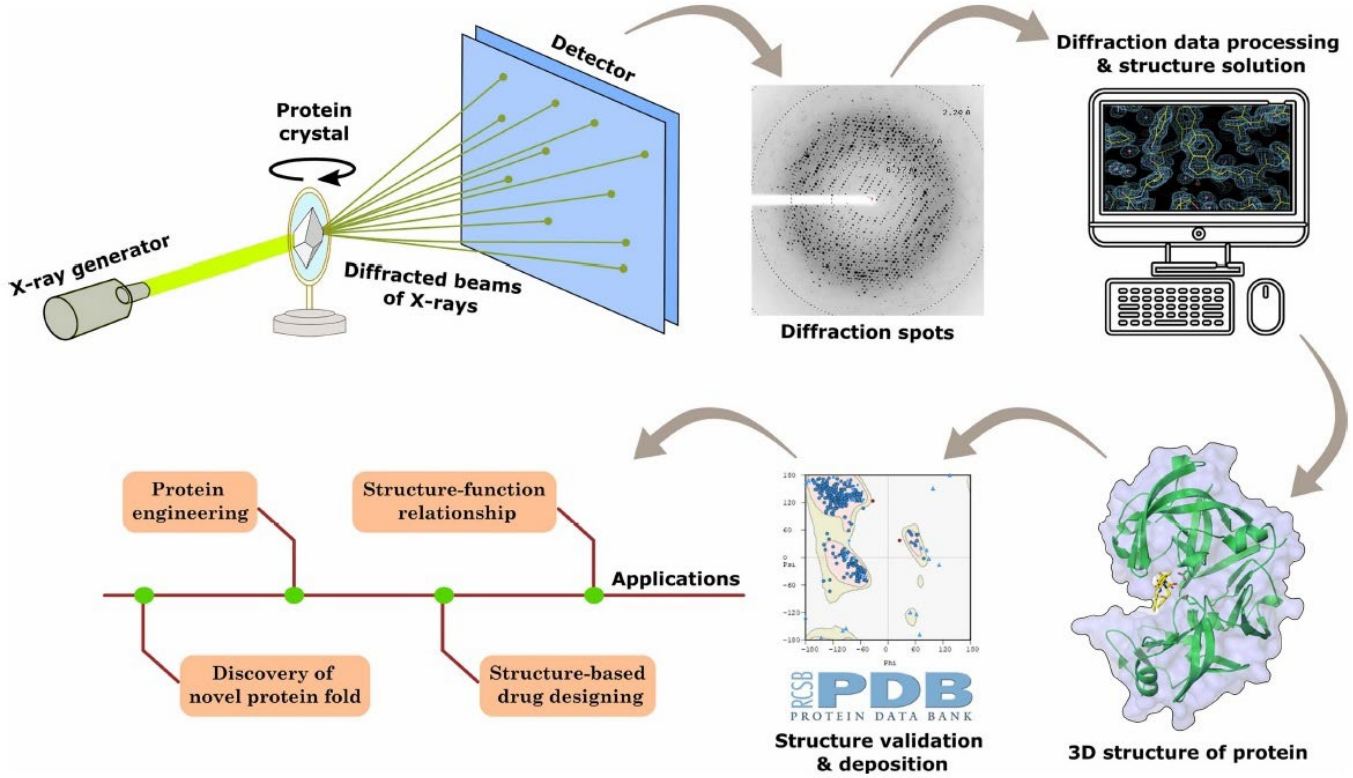
Nucleic Acids



Techniques for determining biomolecule structures



X-ray crystallography

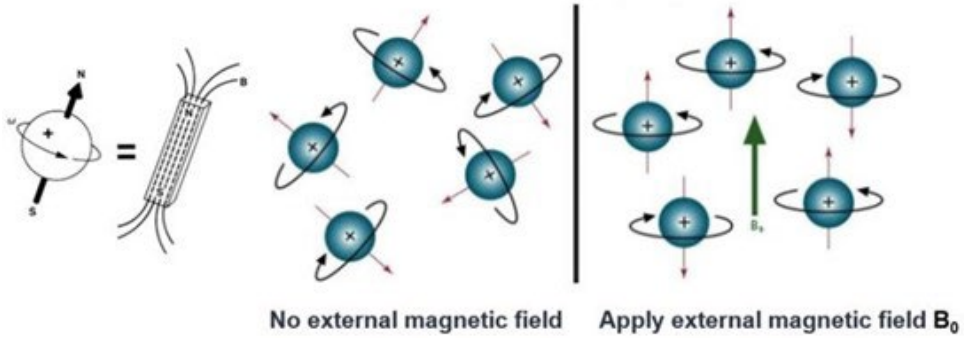


- Diffraction happens when waves encounter periodic structures, causing them to scatter in specific directions.
- In X-ray crystallography, the crystal lattice serves as such a periodic structure, causing the X-rays to undergo diffraction.
- The resulting diffraction pattern carries valuable information about the arrangement of atoms within the crystal

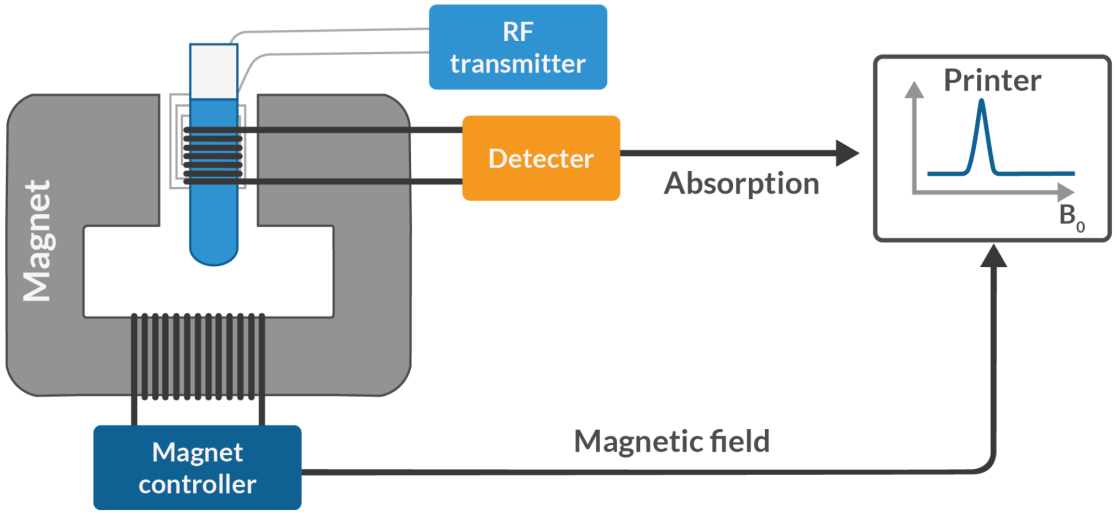
NMR spectroscopy

- Many nuclei have spin and all nuclei are electrically charged. If an external magnetic field is applied, an energy transfer is possible between the base energy to a higher energy level (generally a single energy gap).
- The energy transfer takes place at a wavelength that corresponds to radio frequencies and when the spin returns to its base level, energy is emitted at the same frequency.
- The signal that matches this transfer is measured in many ways and processed in order to yield an NMR spectrum for the nucleus concerned.

Nuclear spin without and with an external magnetic field

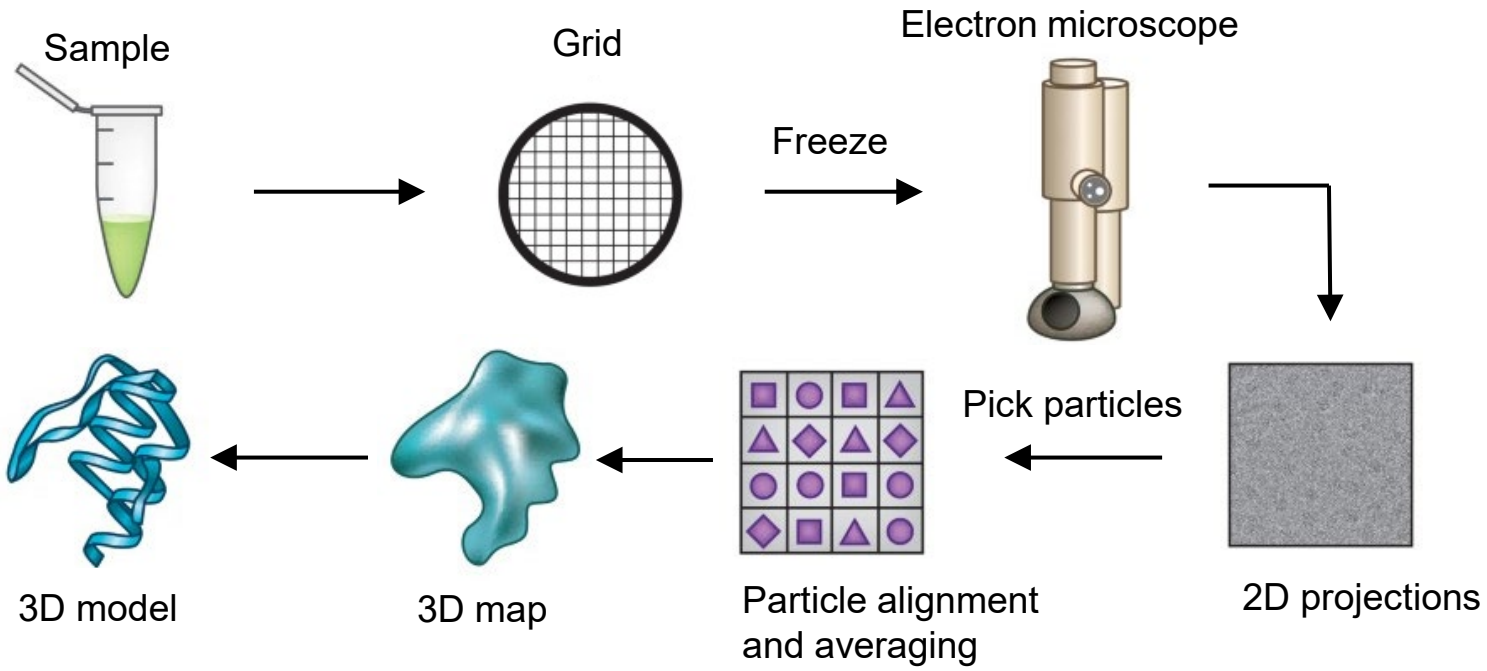


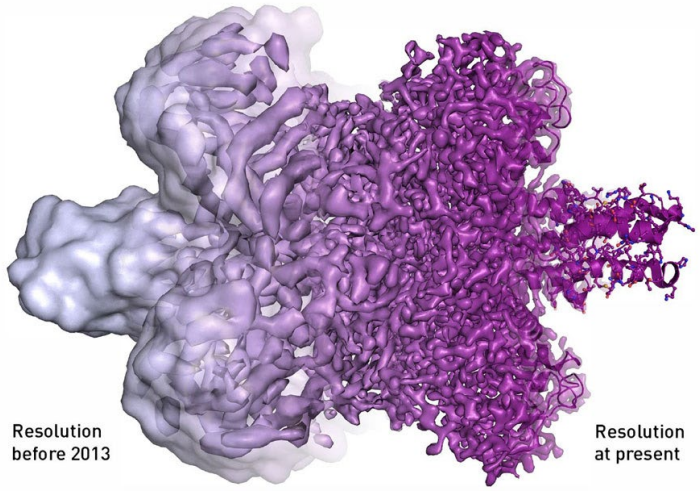
NMR spectroscopy instrumentation



Cryo-EM

- When the sample is rapidly frozen to temperatures below -150 °C, it would be trapped in **vitreous ice**.
- The sample is then imaged from different angles using an electron microscope, where a beam of electrons passes through the specimen, resulting in a set of two-dimensional (2D) projections.
- Advanced computational algorithms are then used to reconstruct a 3D model of the sample from these projections.



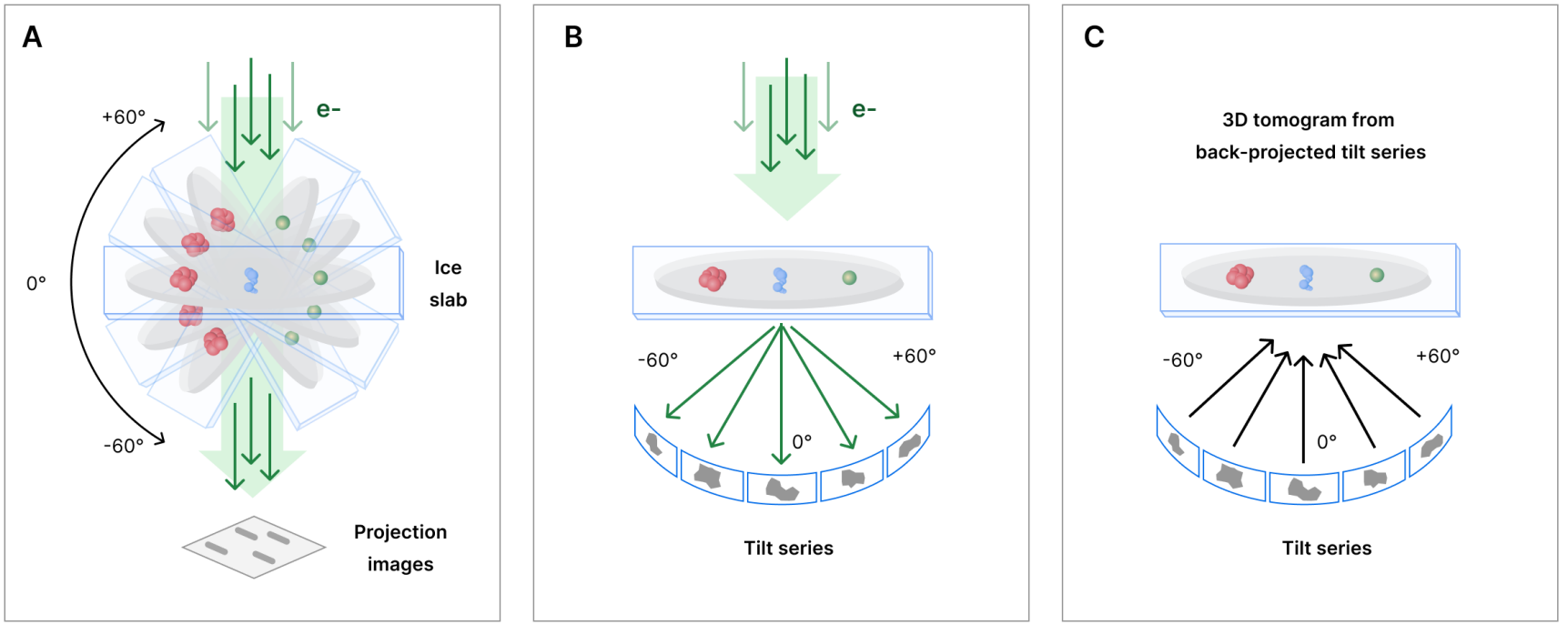


Comparation of the three techniques

Characteristics	X-ray crystallography	NMR spectroscopy	Cryo-EM
Resolution	Ultra-high (≤ 1 Å)	Moderate (~2-3 Å)	High (~2-3 Å)
Sample state	Crystalline form	Solution state	Vitrified film (cryogenic state)
Sample requirements	Crystallizable	High purity, low molecular weight	High purity, homogeneously dispersed
Applicable molecular weight range	Small-to-medium and medium-to-large molecules	<50 kDa	Small-to-medium to very large complexes
Dynamic information	Static	Dynamic information resolvable	Capable of capturing dynamic conformations
Cost and instrumentation requirements	Moderate	Moderate-to-low	Very high

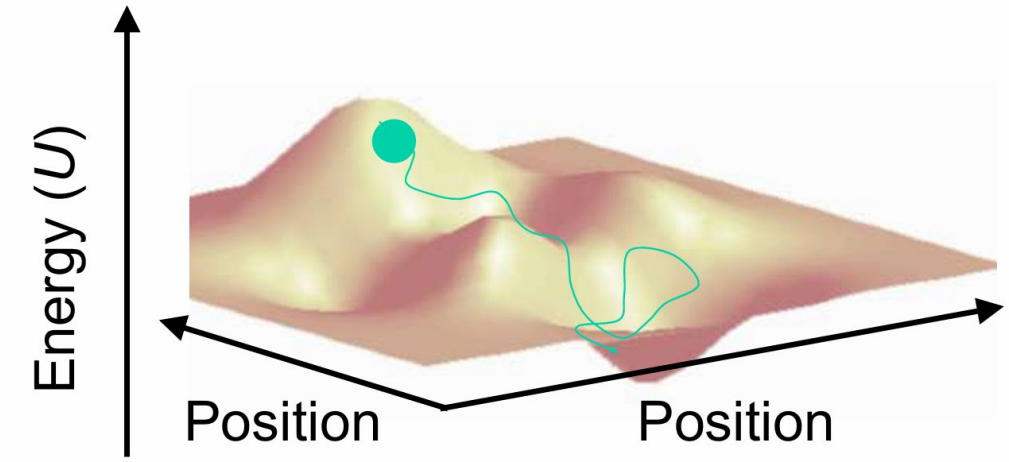
Cryo-ET

- Cryo-electron tomography (Cryo-ET) is a technique that provides researchers with a 3D view of cells, subcellular components (e.g., organelles), viruses, and proteins.
- Thanks to Cryo-ET, researchers can now bridge the gap between techniques that enable them to study cells at the micron scale (e.g., light microscopy) and molecular structures at the atomic-level (e.g., X-ray crystallography).



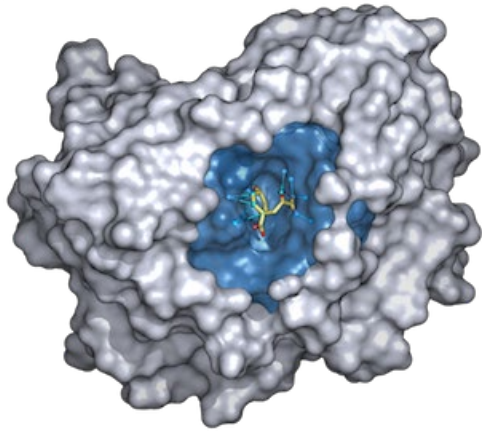
Molecular dynamics simulation

- Basic idea: mimic what atoms do in real life, assuming a given potential energy function.
  - The energy function allows us to calculate the force experienced by any atom, given the positions of the other atoms
  - Newton's laws tell us how those forces will affect the motions of the atoms
- Limitations
  - Short-timescale simulation
  - Molecular mechanics force fields are inherently approximations
  - Covalent bonds cannot break or form during standard MD simulations (required, e.g. disulfide bonds between cysteines)

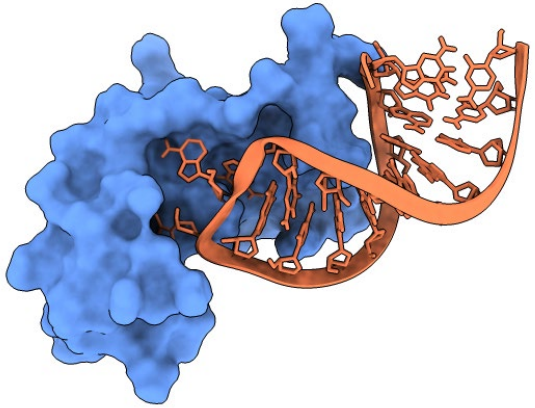


Protein complexes and biomolecular interactions

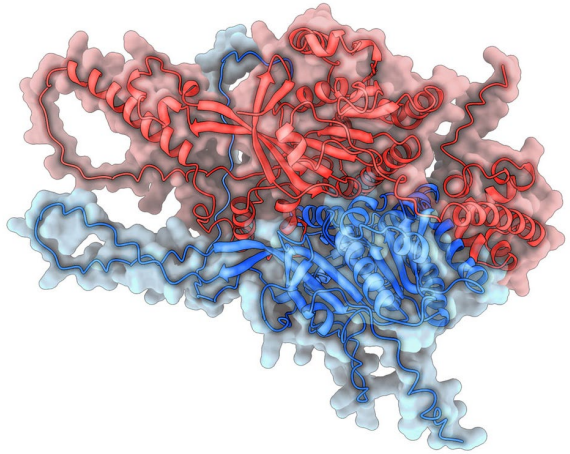
protein-ligand



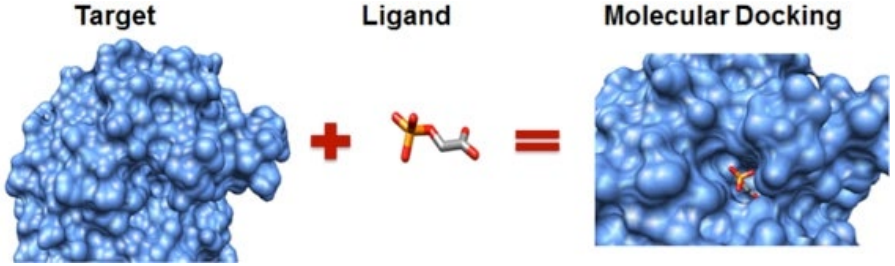
protein-nucleotide



protein-protein

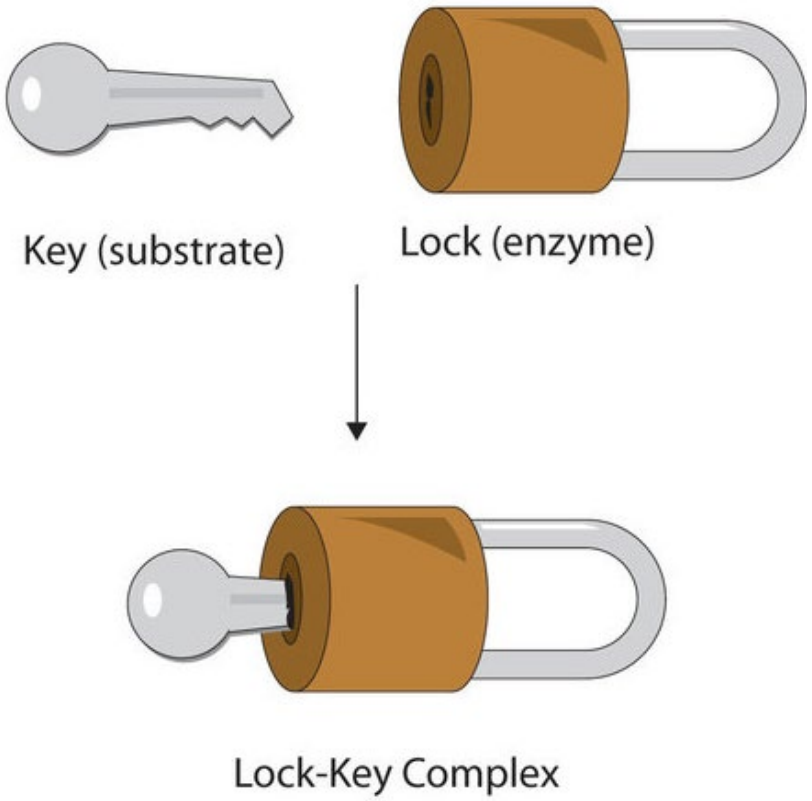


- Biomolecules do not operate in isolation. They engage in numerous interactions that facilitate cellular function and communication.
- Types of interactions
  - Hydrophobic interactions
  - Hydrogen bonds
  - Ionic interactions
  - Van der Waals forces
  - Disulfide bonds
  - …
- **Molecular docking:** Computational simulation of a candidate ligand binding to a receptor.

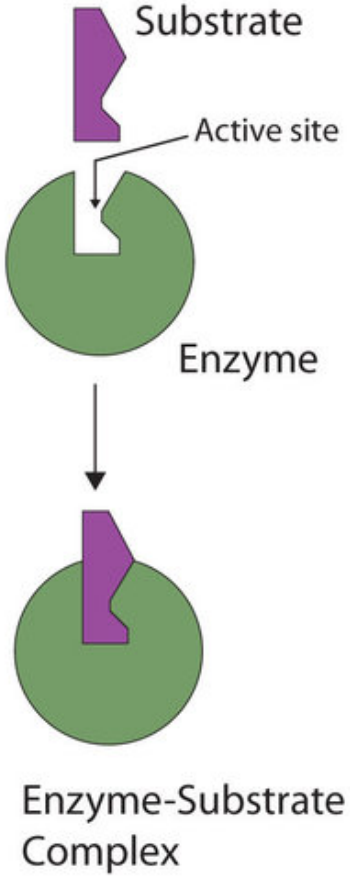


Two molecular models used for molecular docking

Lock-and-key model (for rigid docking)

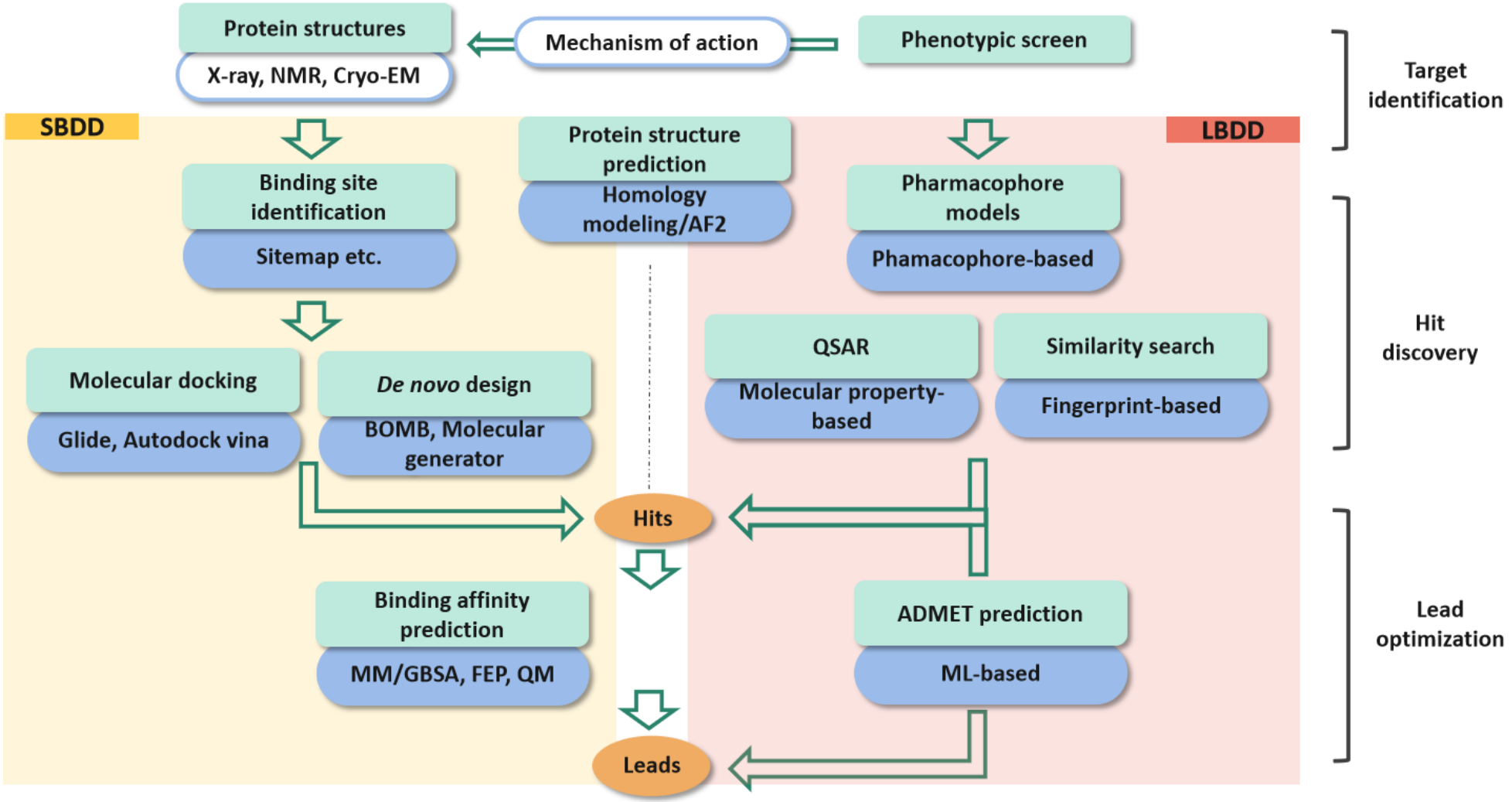


Induced-fit model (for flexible docking)



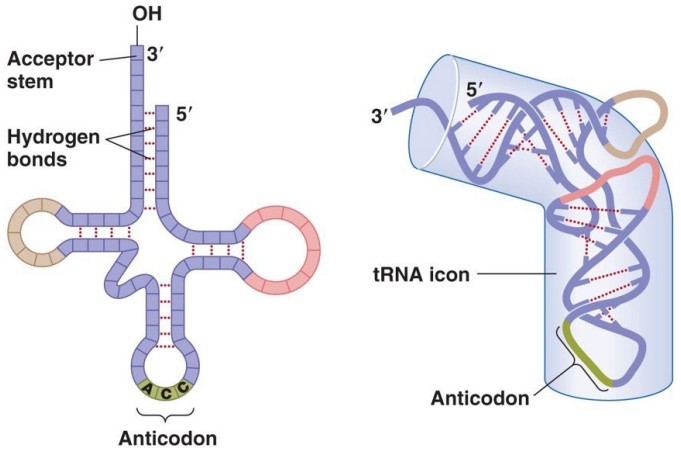
- Rigid docking is a molecular docking model that treats the ligand and target as **rigid objects**, most commonly used for protein-to-protein docking.
- Flexible docking is a type of molecular docking that models changes in the internal geometry of interacting partners when a complex is formed. Allows conformational changes during the docking process.

Computer-assisted drug design (CADD)

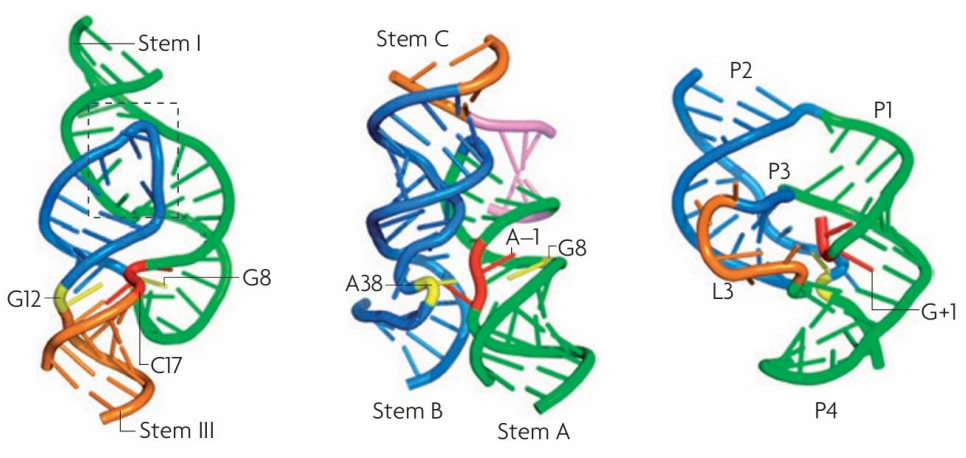


RNA structures

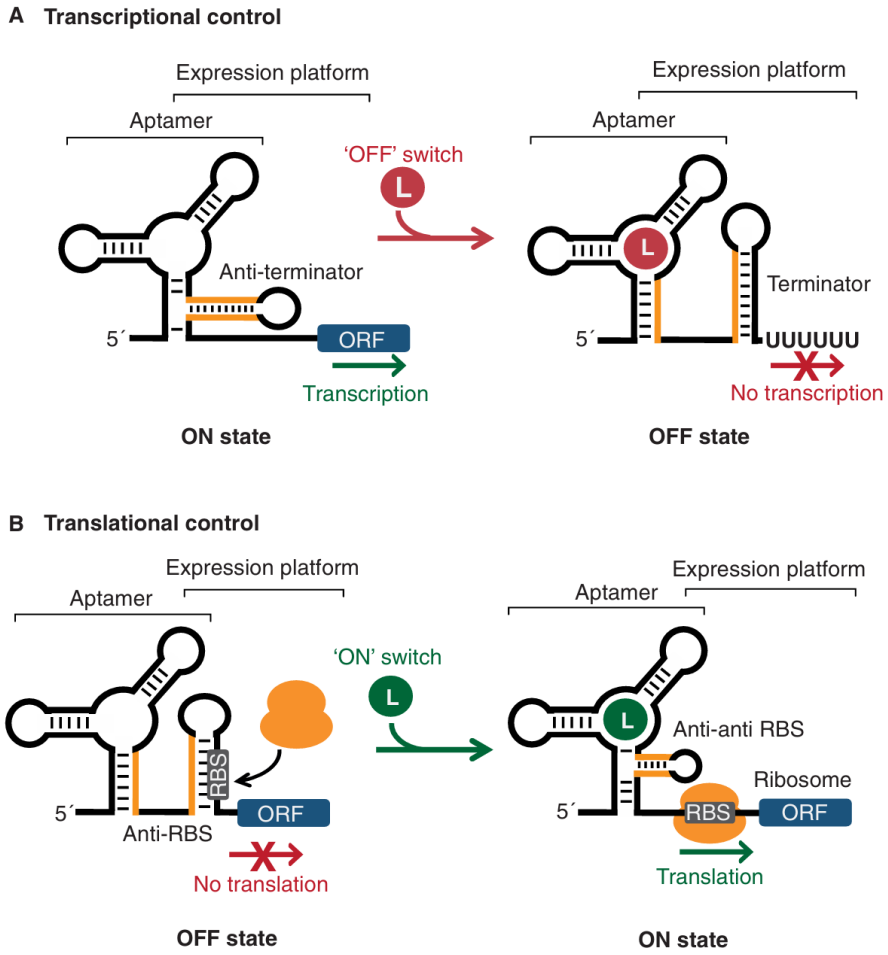
- tRNA

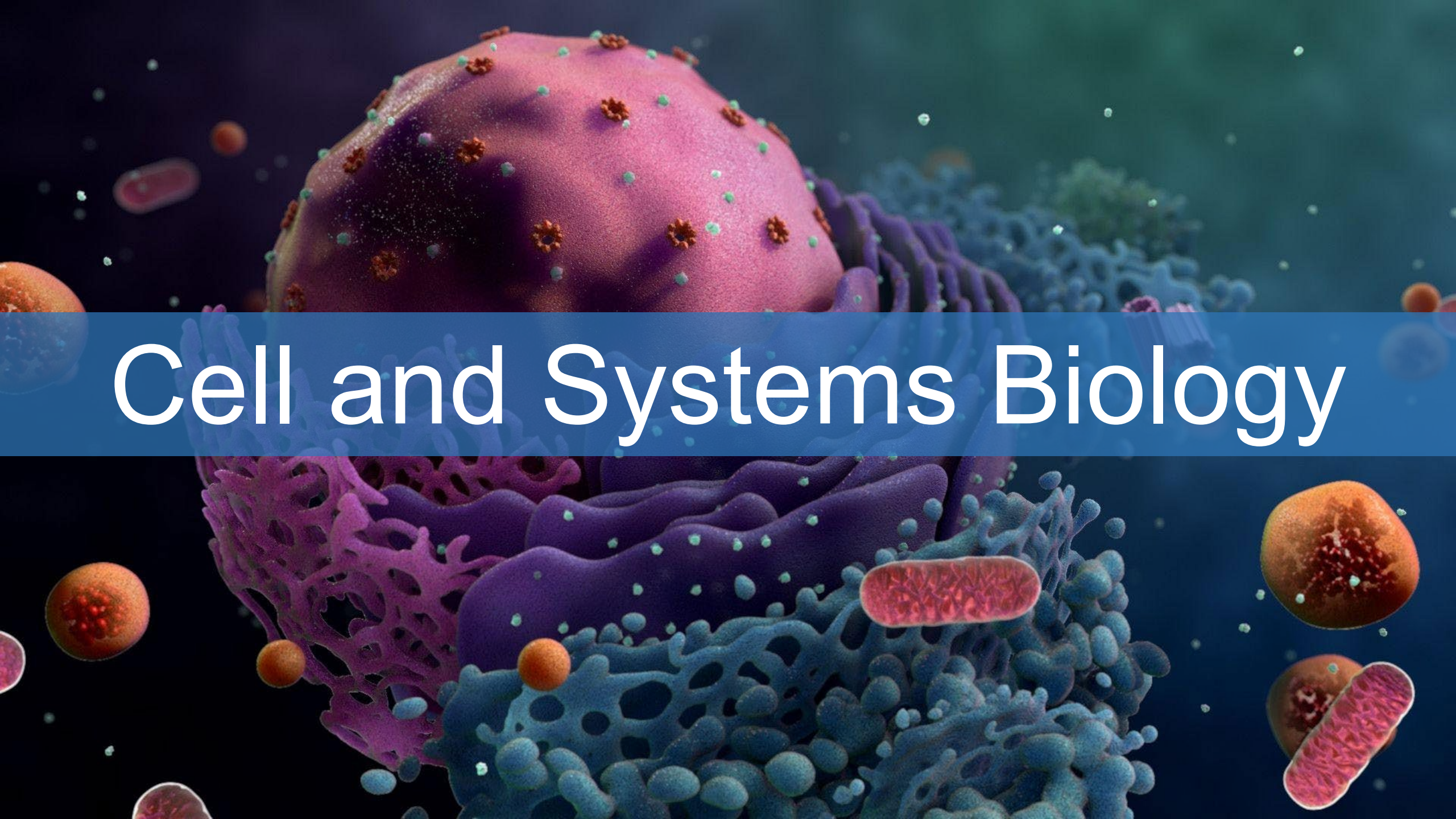


- ribozyme



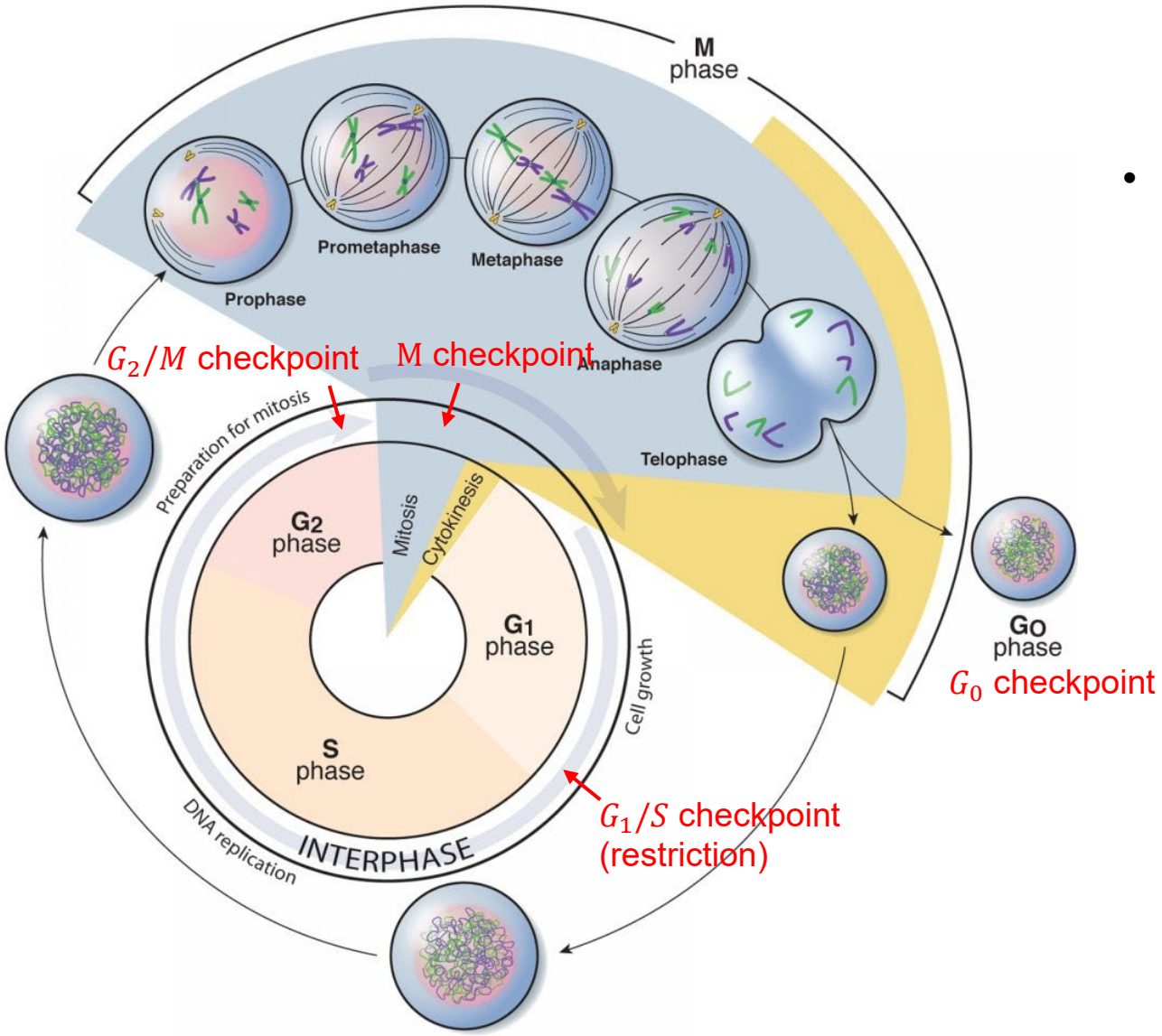
- riboswitch





Cell and Systems Biology

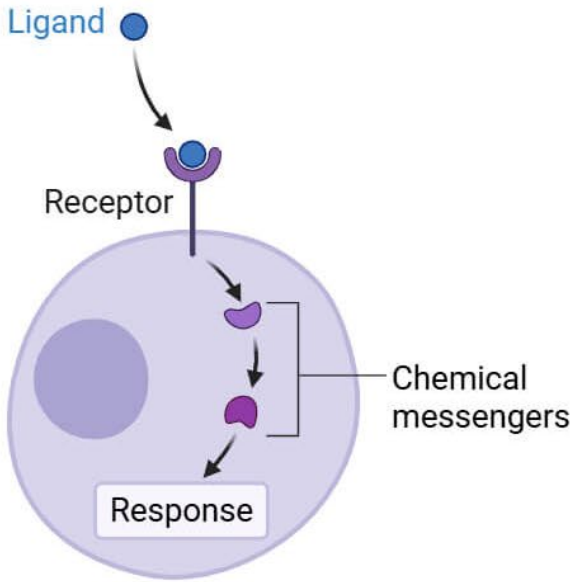
Cell cycle



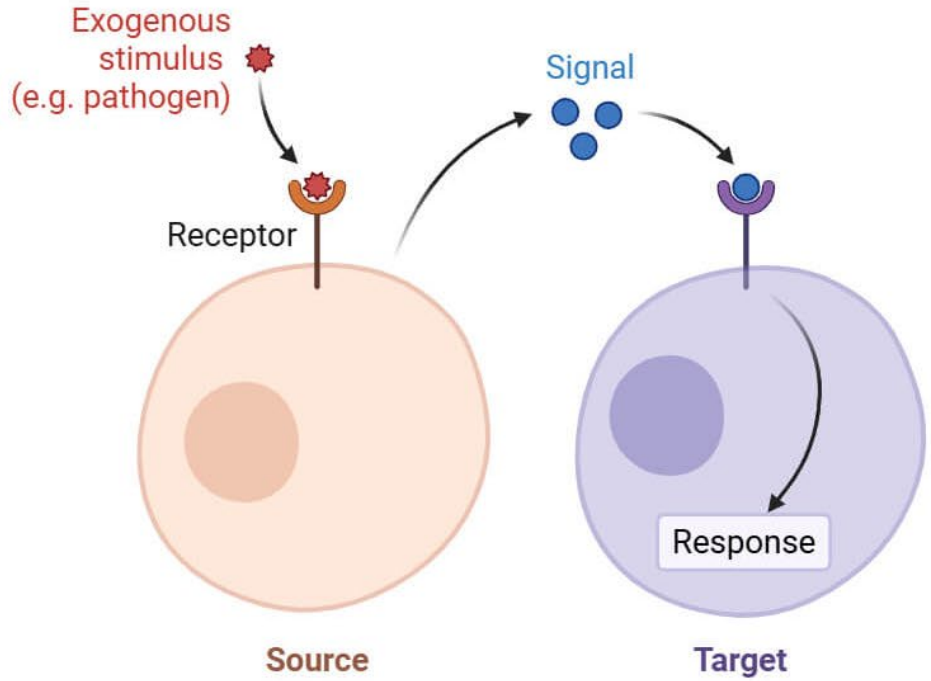
- **Cell cycle:** Cell-division cycle is the sequential series of events that take place in a cell that causes it to divide into two daughter cells. These events include:
  - G_0: A phase where the cell has left the cycle and has stopped dividing.
  - G_1: Cell growth. The G_1 checkpoint ensures that everything is ready for DNA synthesis.
  - S : DNA replication.
  - G_2: Growth and preparation for mitosis. The G_2 checkpoint ensures that everything is ready to enter the M (mitosis) phase and divide.
  - M : Cell division occurs. The metaphase checkpoint ensures that the cell is ready to complete cell division.

Cell signaling

- Cell signaling, or intracellular signal transduction, is a series of chemical reactions that work together to control cellular functions by activating a chain of biochemical events.



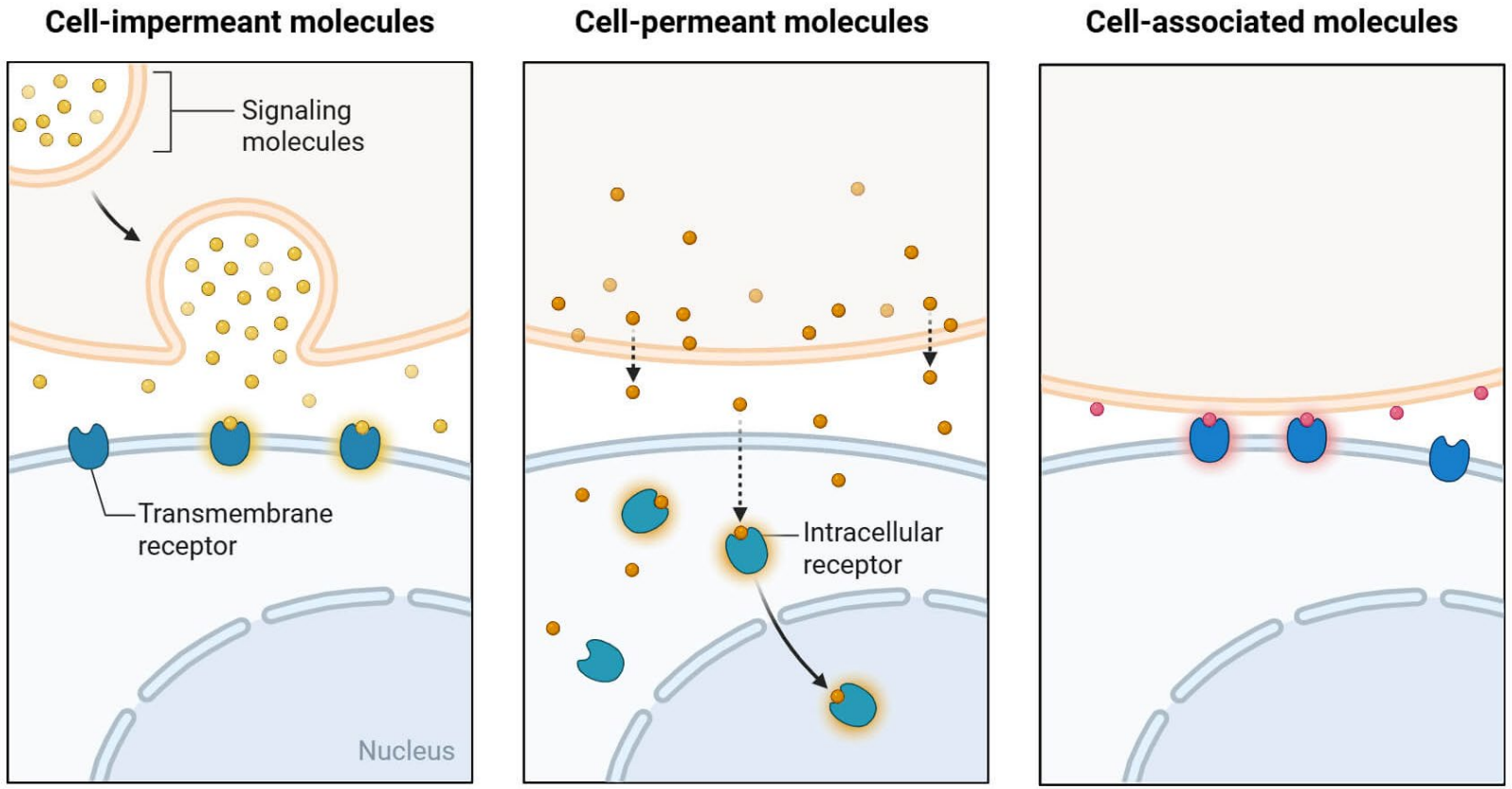
Chemical messengers relay the signal from the ligand and receptor



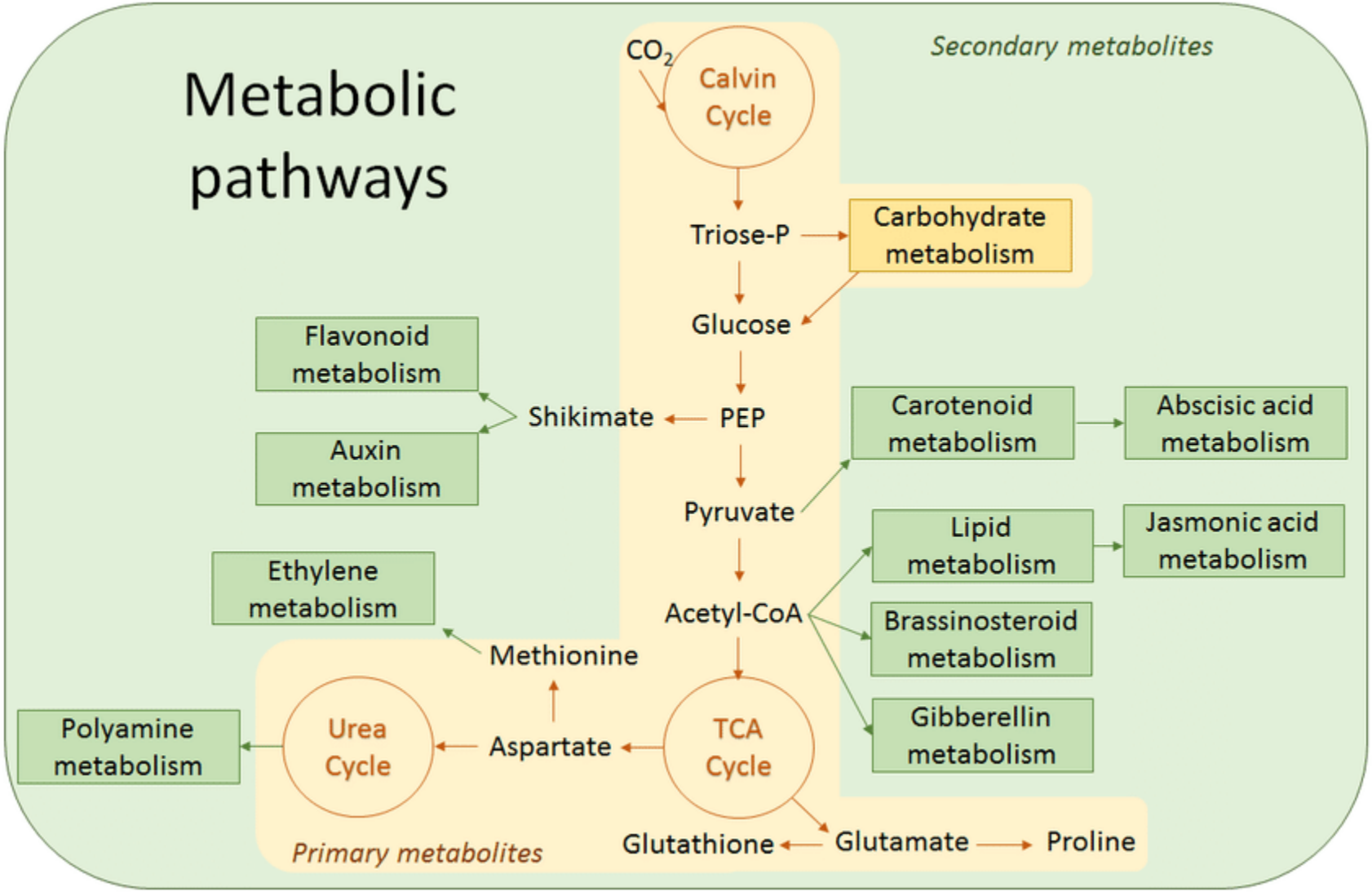
Cell signaling leads to the production of signals that serve as inputs for other cells

Classes of signaling molecules

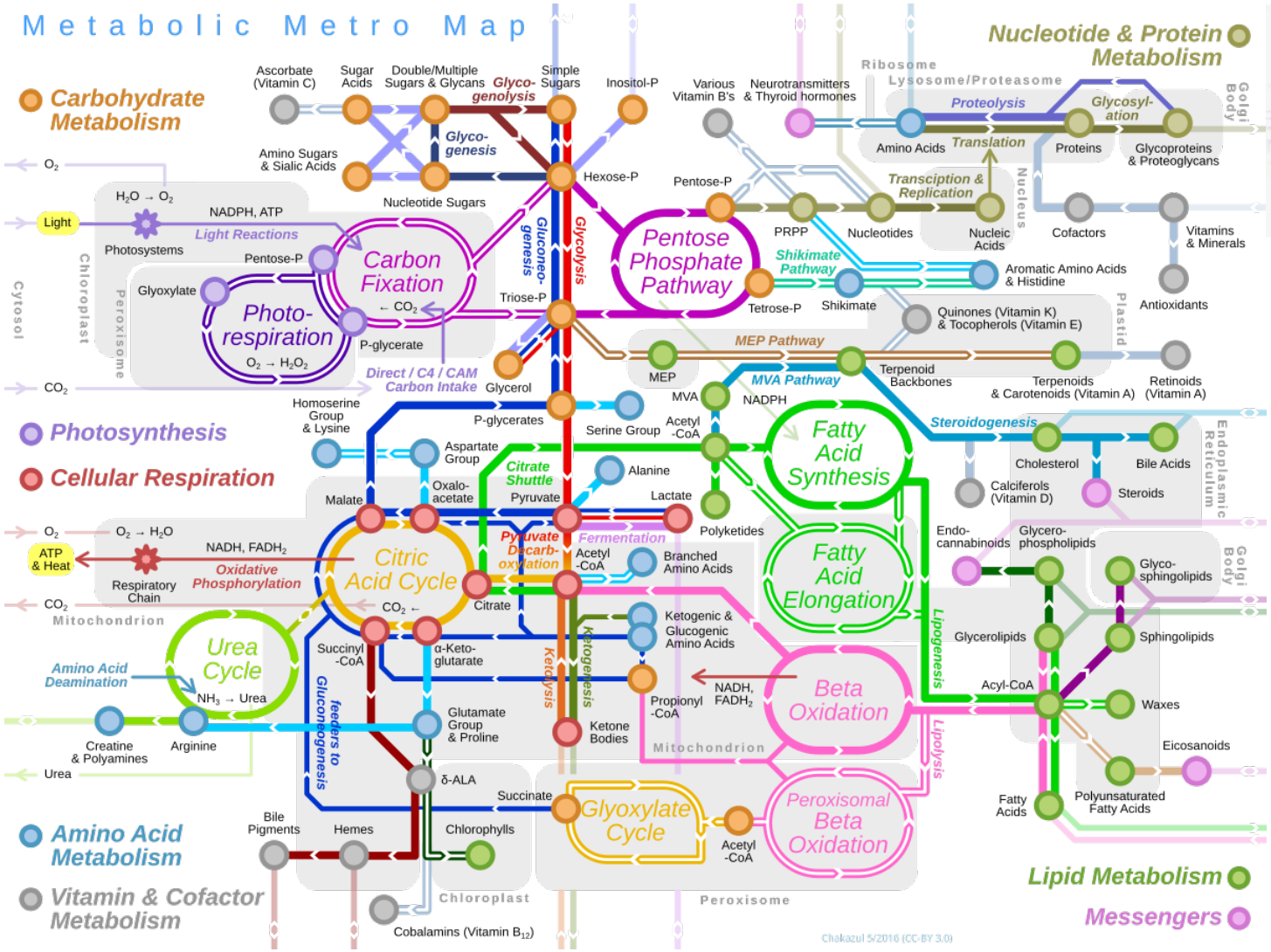
- Signaling molecules are mostly hydrophilic and are unable to directly cross the plasma membrane, therefore, they bind to cell-surface receptors.
- Smaller signal molecules diffuse across the plasma barrier and bind to intracellular receptors inside the target cell in the cytosol or the nucleus.

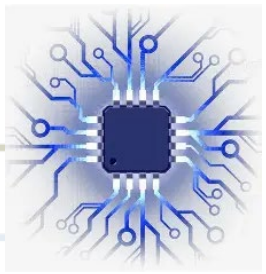


Metabolic pathways



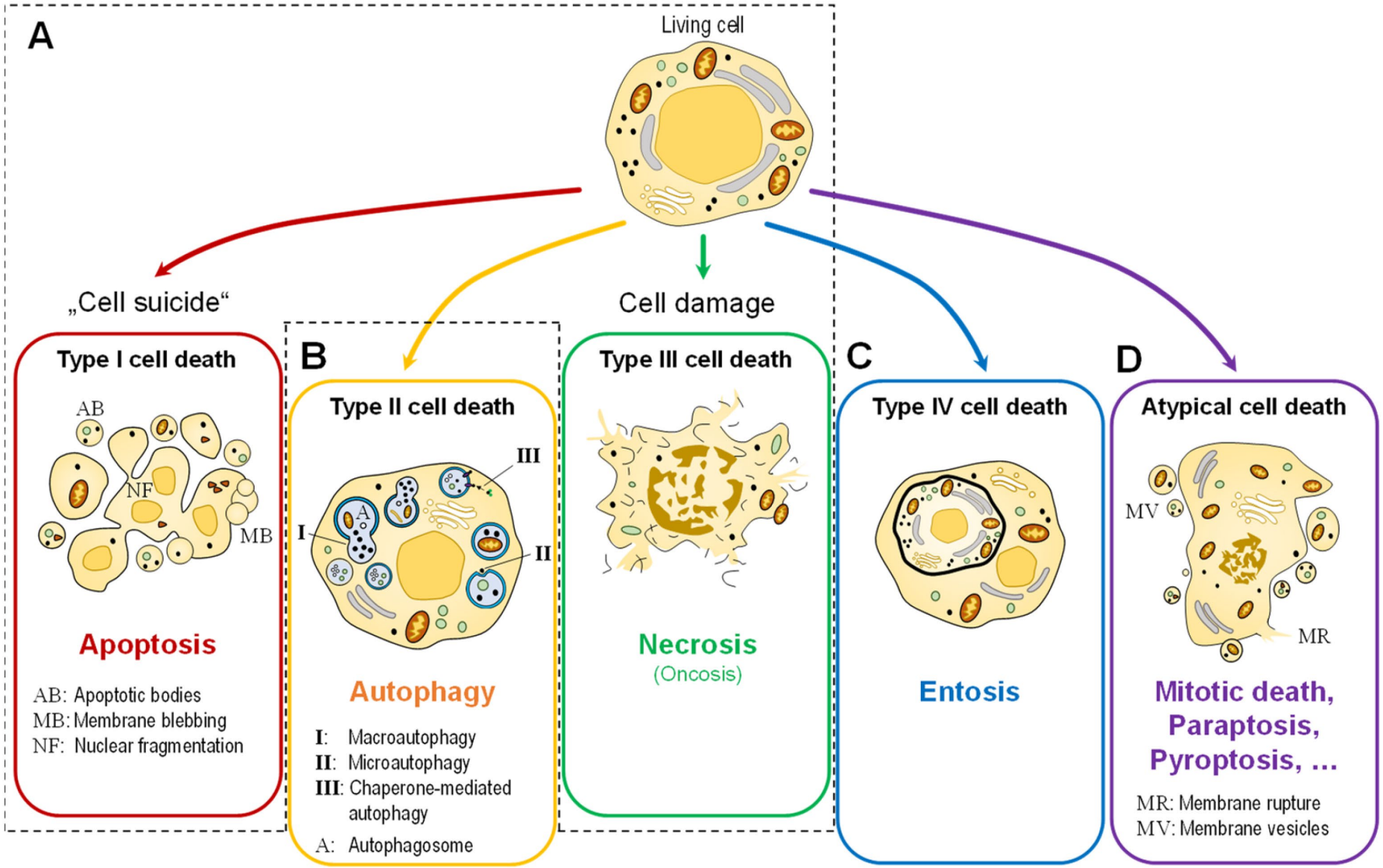
Metabolic networks





Metabolic pathways are the fundamental components that, when interconnected, form the complex system of a **metabolic network**.

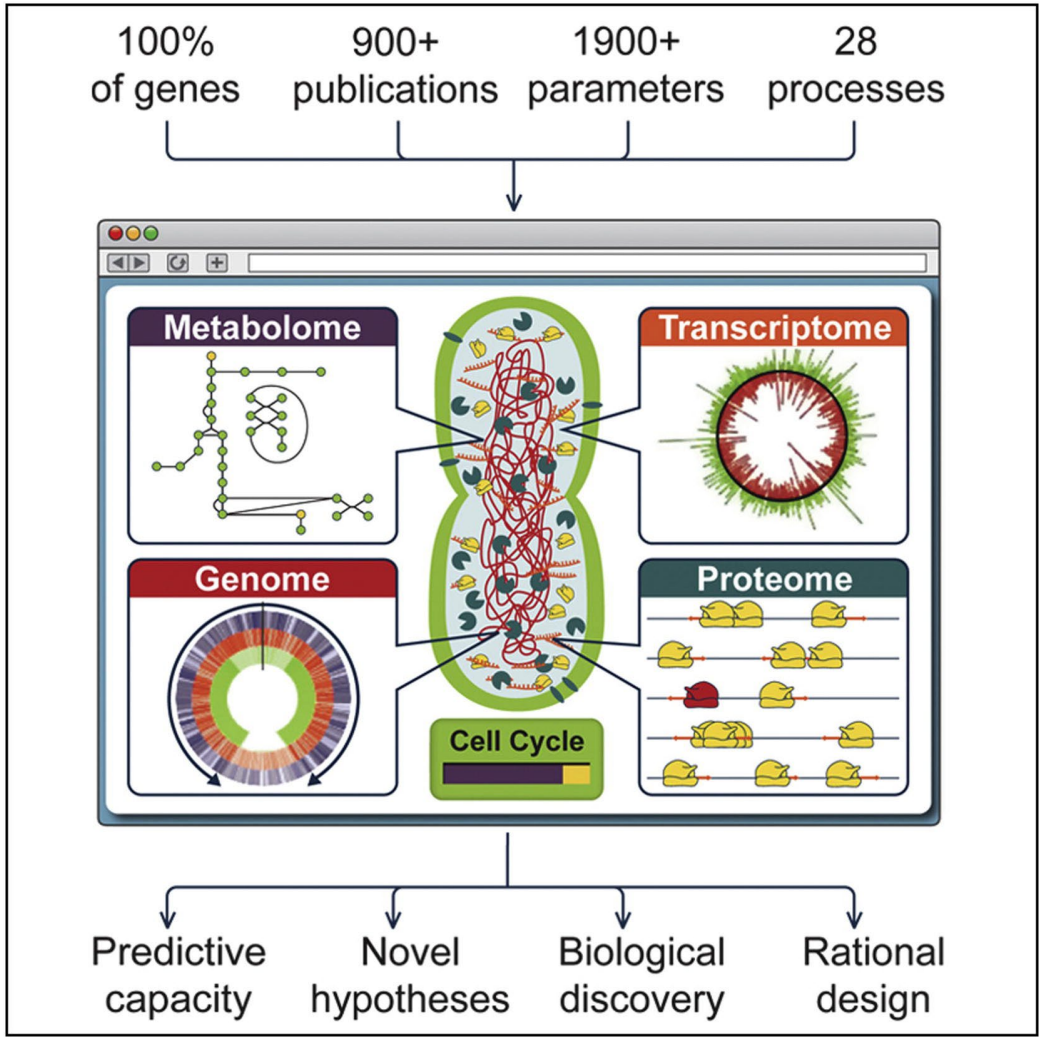
Cell death


A
Living cell
„Cell suicide“
Cell damage
Type I cell death
AB
NF
MB
Apoptosis
AB: Apoptotic bodies
MB: Membrane blebbing
NF: Nuclear fragmentation
B
Type II cell death
III
I
A
II
Autophagy
I: Macroautophagy
II: Microautophagy
III: Chaperone-mediated autophagy
A: Autophagosome
Type III cell death
Necrosis
(Oncosis)
C
Type IV cell death
Entosis
D
Atypical cell death
MV
MR
Mitotic death, Paraptosis, Pyroptosis, ...
MR: Membrane rupture
MV: Membrane vesicles

Systems biology: emergent properties



System biology for dynamic modeling of living cells



- An entire organism is modeled in terms of its molecular components
- Complex phenotypes can be modeled by integrating cell processes into a single model
- Unobserved cellular behaviors are predicted by model of *M. genitalium*
- New biological processes and parameters are predicted by model of *M. genitalium*

(Karr et al., *Cell*, 2012)

Q&A