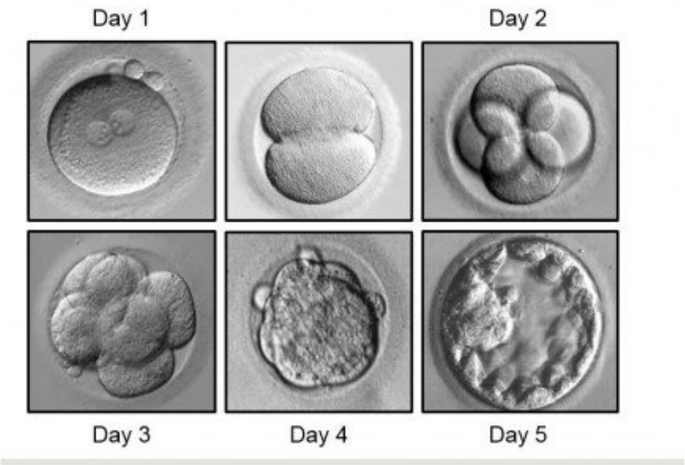
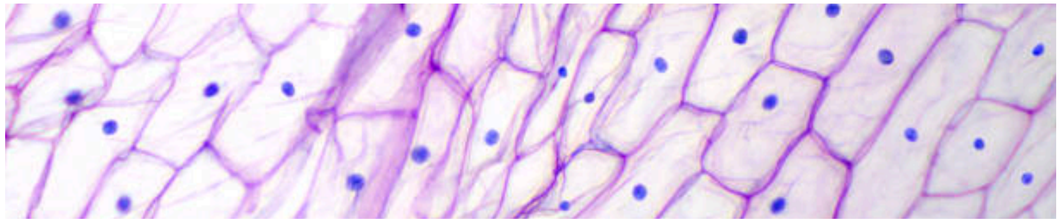
Genomic Big Data

Introduction

Why Genomics?

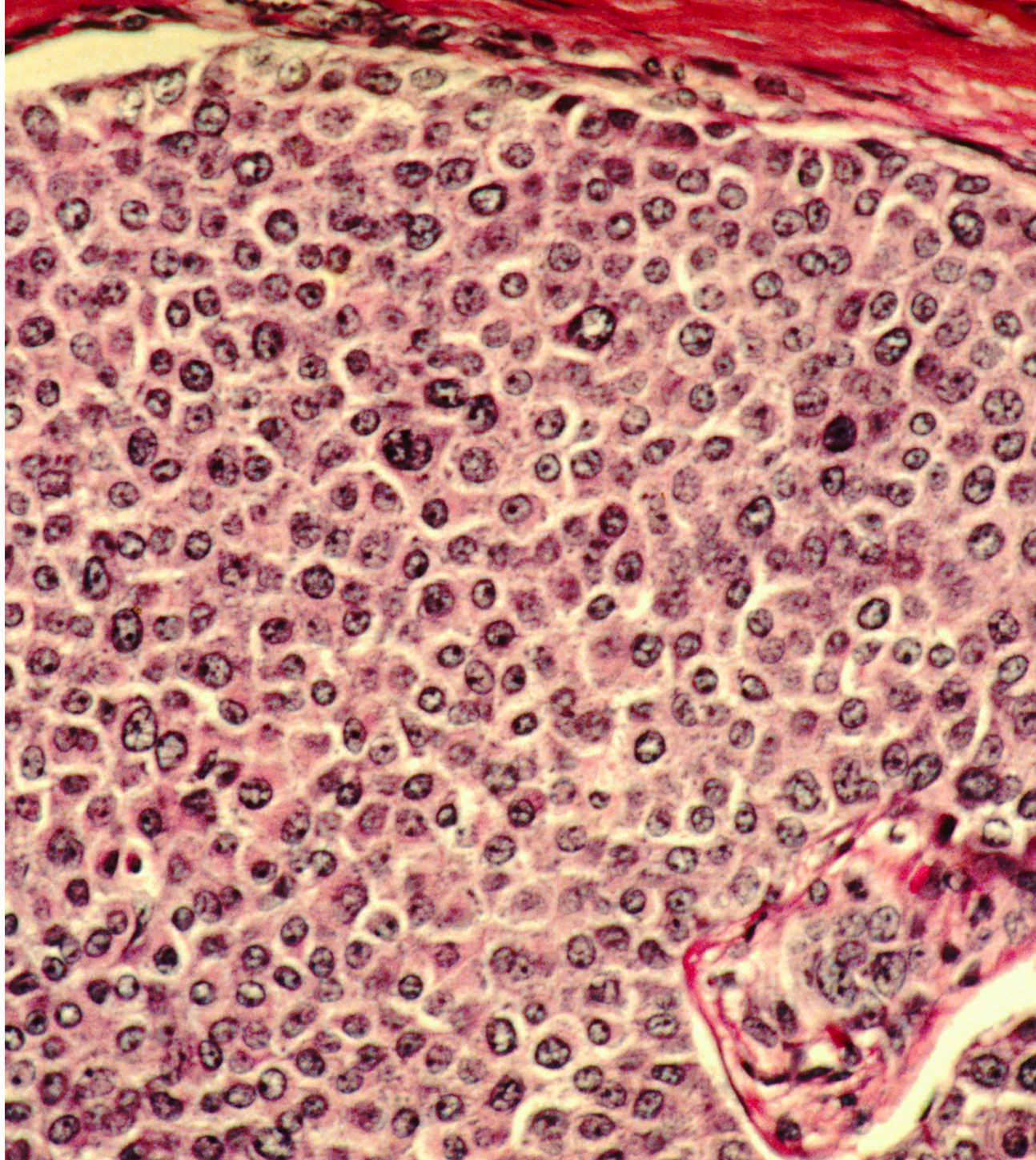
- Genome is a program
- It is copied from parents
- How does the cells differentiate?
- How is the organs formed?
- How are neuron cells fromed?
- How are they different from skin cells?





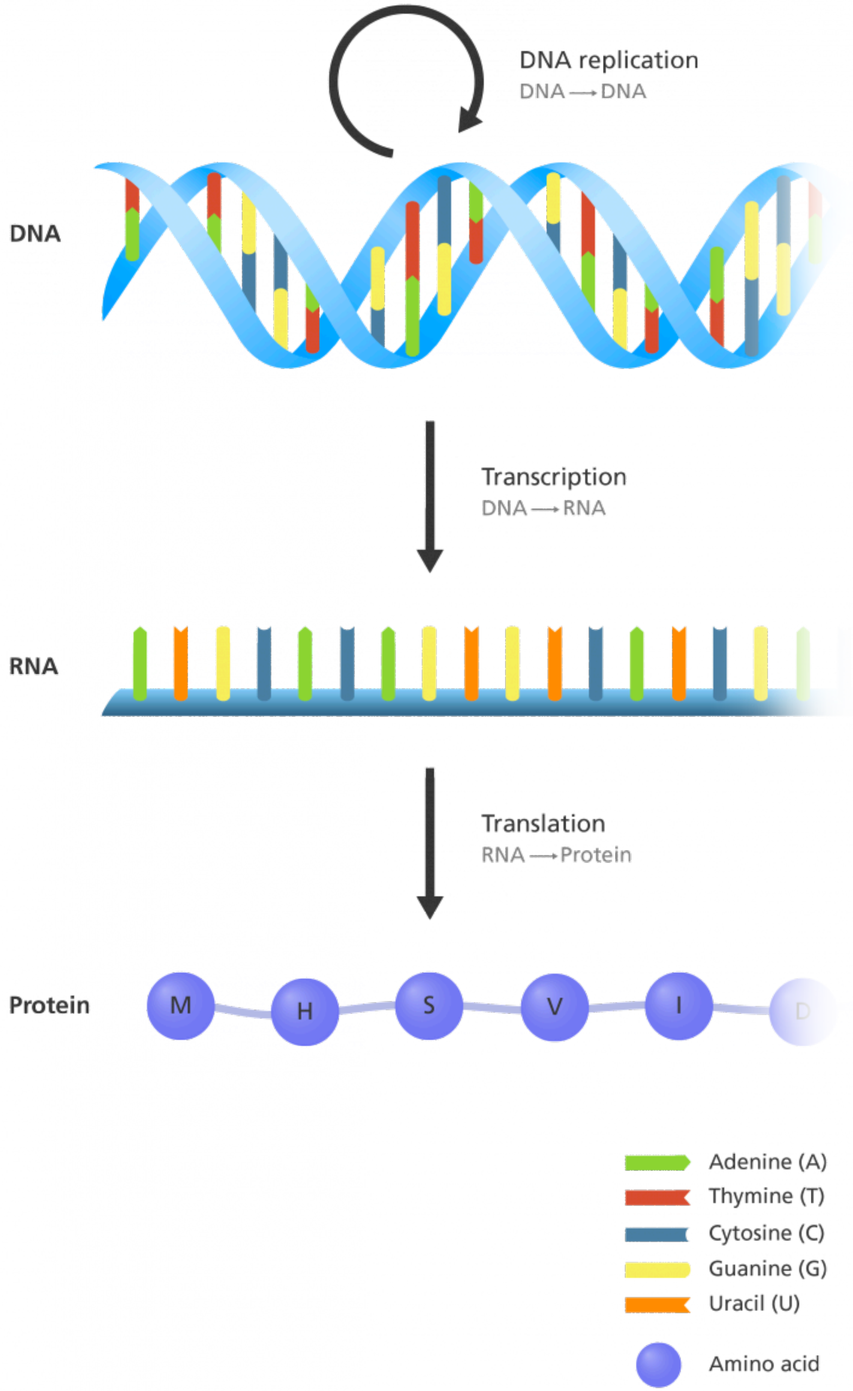
Cancer Cells

- Cancer is a genetic Disease
- They divide in an uncontrolled way
- Mutation in a change in genome
  - Accident in replication
  - 1 to 3 errors per cell division
- Mutation might affect a gene which might fail to control the cell division



Central Dogma

- Proposed by Francis Crick
- It is not a dogma
- Information flows in a single direction
  DNA -> RNA -> Protein
- An exon is a region of DNA that is transcribed into mRNA
  - Coding and
  - Non-Coding
- There are DNA binding proteins
  - Epigenetics

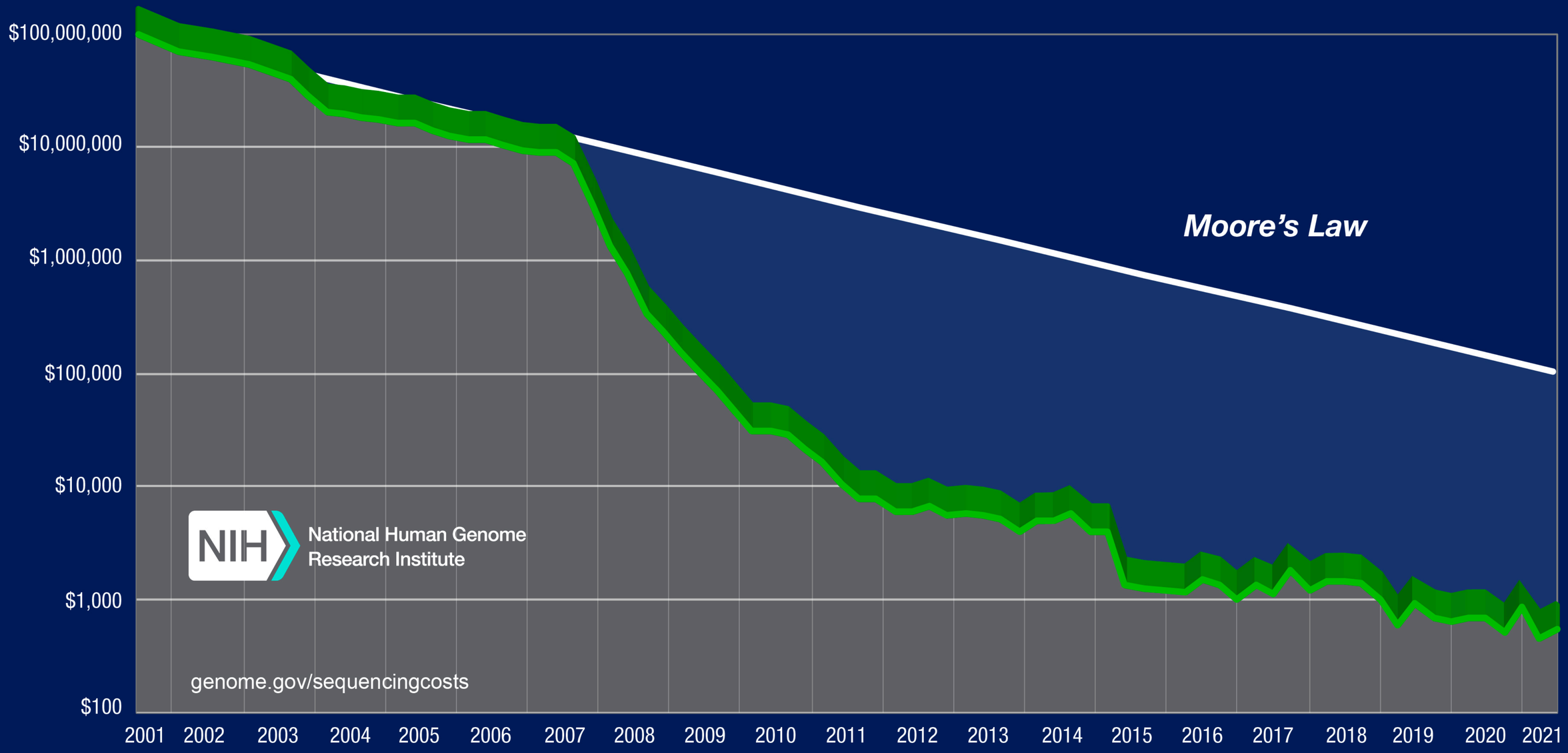


Genome Technology

- Sequencing machines
  - Single run trillion nucleotides
- Human Genome started in 1989 with a target of 15 years
- It was complete in 2001
- Joint and Collaborative Effort


```
@HWI-EAS121:4:100:1783:550#0/1
CGTTACGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACGGATCTCGTATGCGGTCTGCTGCGTGACAAGACAGGGG
+HWI-EAS121:4:100:1783:550#0/1
aaaaa`b_aa`aa`YaX]aZ`aZM^Z]YRa]YSG[[ZREQLHESDHNDDHNMEEDDMPENITKFLFEEDDDHEJQMEDDD
@HWI-EAS121:4:100:1611#0/1
GGGTGGGCATTTCCACTCGCAGTATGGGTTGCCGCACGACAGGCAGCGGTCAGCCTGCGCTTTGGCCTGGCCTTCGGAAA
+HWI-EAS121:4:100:1783:1611#0/1
a```^\__`_```^a``a`^a_^__]a_]\]`a______`_^^`]X]_]XTV_\]]NX_XVX]]_TTTTG[VTHPN]VFDZ
@HWI-EAS121:4:100:1783:322#0/1
CGTTTATGTTTTTGAATATGTCTTATCTTAACGGTTATATTTTAGATGTTGGTCTTATTCTAACGGTCATATATTTTCTA
+HWI-EAS121:4:100:1783:322#0/1
abaa`^aaaaabbbaabaabbbbbb`bbbb_bbbbbbbb`bbbaV^_a``a``]``aT]a__V\]]_]^a`]a_abbaV__
@HWI-EAS121:4:100:1783:1394#0/1
GGGTCTTTATTGGTCTGGTGATCCCCCATATTCTCCGGTTGTGTGGTTTAACCGATCATCGCGCATTACTTCCCGGCTGC
+HWI-EAS121:4:100:1783:1394#0/1
```[aa\b^^[]aabbb][`a_abbb`a``bbbbbabaabaaaab_VZa_^___bab_X`[a\HV_[_][^_X\T_VQQ
```
```

Cost per Human Genome



Growth of Sequences

National Library of Medicine
National Center for Biotechnology Information

Log in

SRA

SRA

Search

Advanced

Help

SRA - Now available on the cloud

Sequence Read Archive (SRA) data, available through multiple cloud providers and NCBI servers, is the largest publicly available repository of high throughput sequencing data. The archive accepts data from all branches of life as well as metagenomic and environmental surveys. SRA stores raw sequencing data and alignment information to enhance reproducibility and facilitate new discoveries through data analysis.

Getting Started

- How to Submit
- How to search and download
- How to use SRA in the cloud
- Submit to SRA

Tools and Software

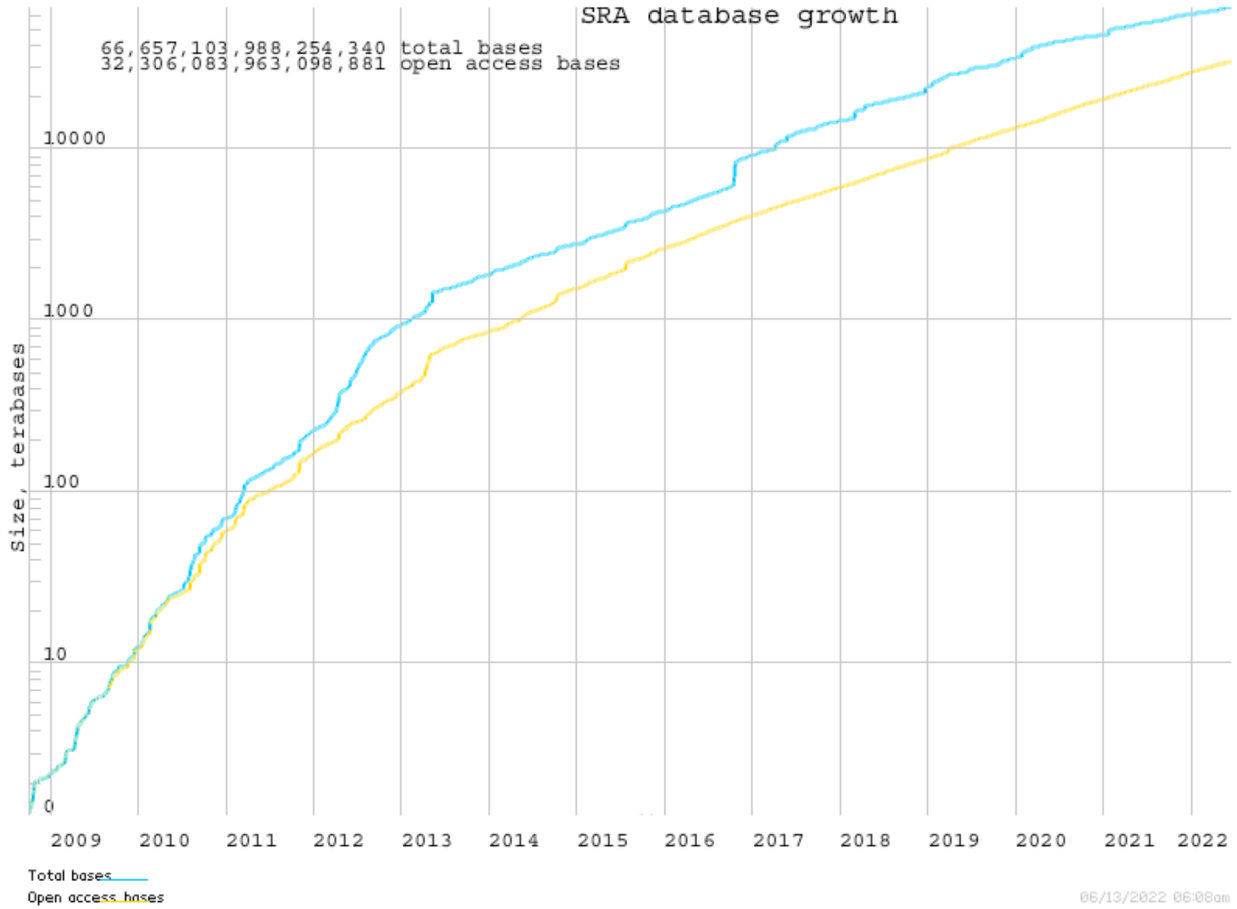
- Download SRA Toolkit
- SRA Toolkit Documentation
- SRA-BLAST
- SRA Run Browser
- SRA Run Selector

Related Resources

- Submission Portal
- dbGaP Home
- BioProject
- BioSample

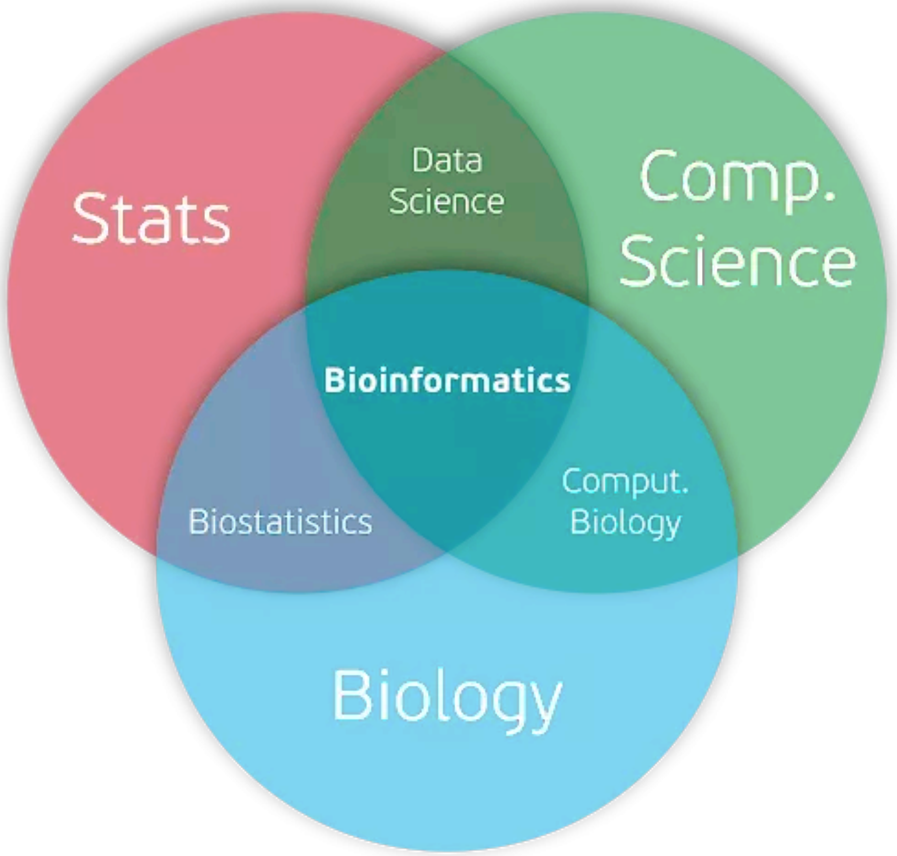
SRA data base

- Downloadable
- Growing very fast
- Genotyping is faster than phenotyping
- A lots of studies are needed



What is Genomics?

- Genomics is an interdisciplinary field of biology focusing on the structure, function, evolution, mapping, and editing of genomes.
- Structure
  - Sequence of neucleotides , 3 billion - A, C, G, T
  - 23 chromosome pairs, 22 Autosome pairs - identical copies
- Function - What does all DNA do?
  - Organs, tissues, respiration, metabolism
- Evolution of genomes over time
- Where are the genes?



genomics

in British English

(dʒɪˈnɒmɪks)

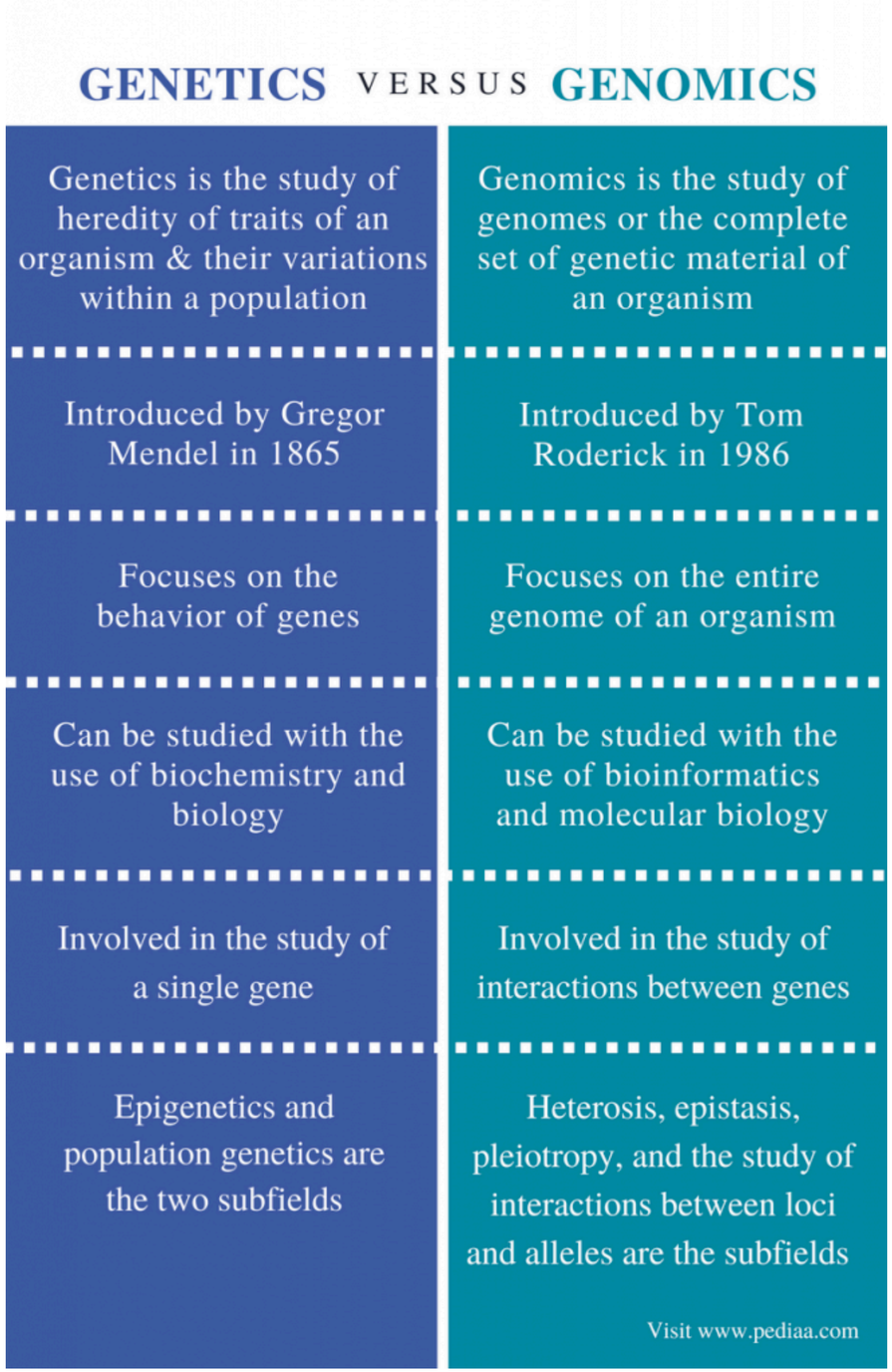
NOUN (*functioning as singular*)

the branch of molecular genetics concerned with the study of genomes, specifically the identification and sequencing of their constituent genes and the application of this knowledge in medicine, pharmacy, agriculture, etc

Word Frequency

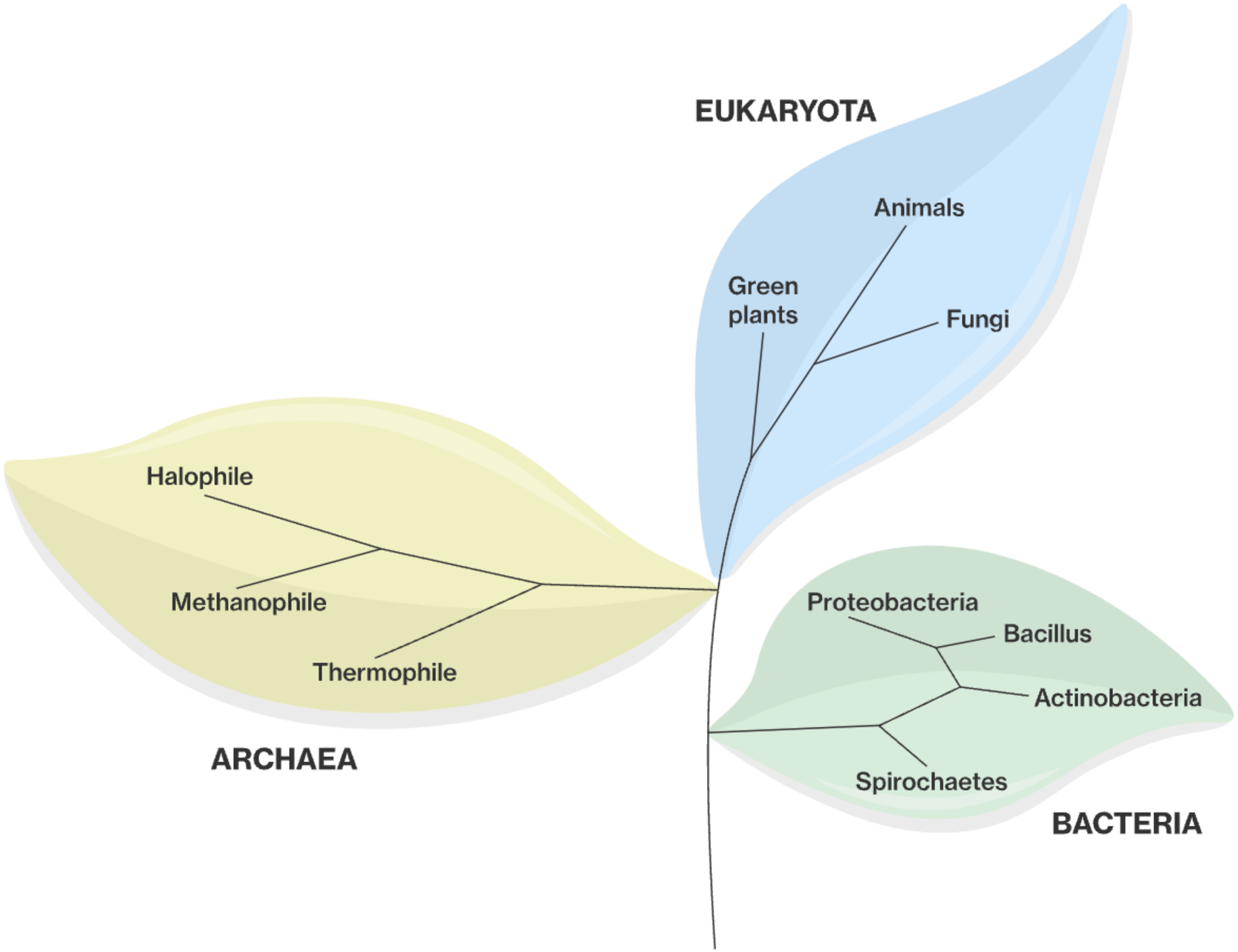
Genetics vs Genomics

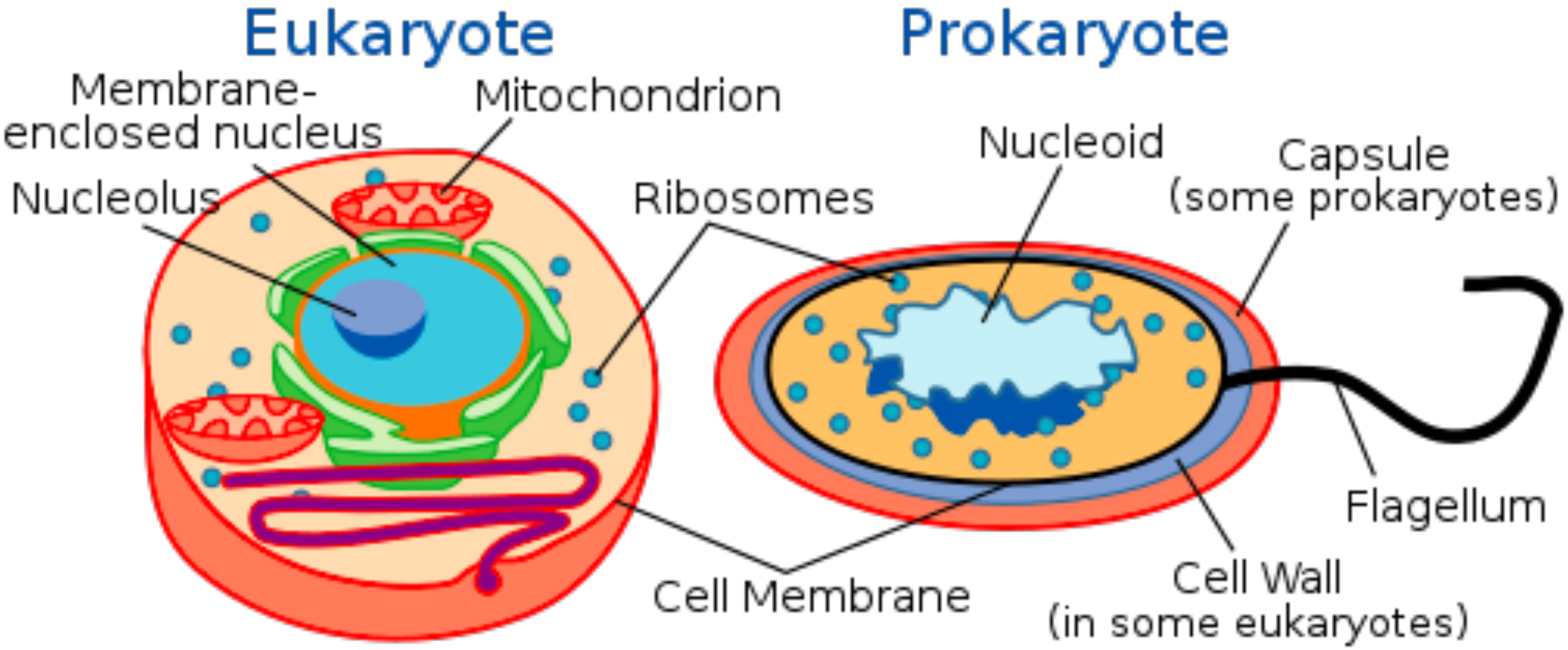
- The main difference between genomics and genetics is that genetics scrutinizes the functioning and composition of the single gene where as genomics addresses all genes and their inter relationships in order to identify their combined influence on the growth and development of the organism.

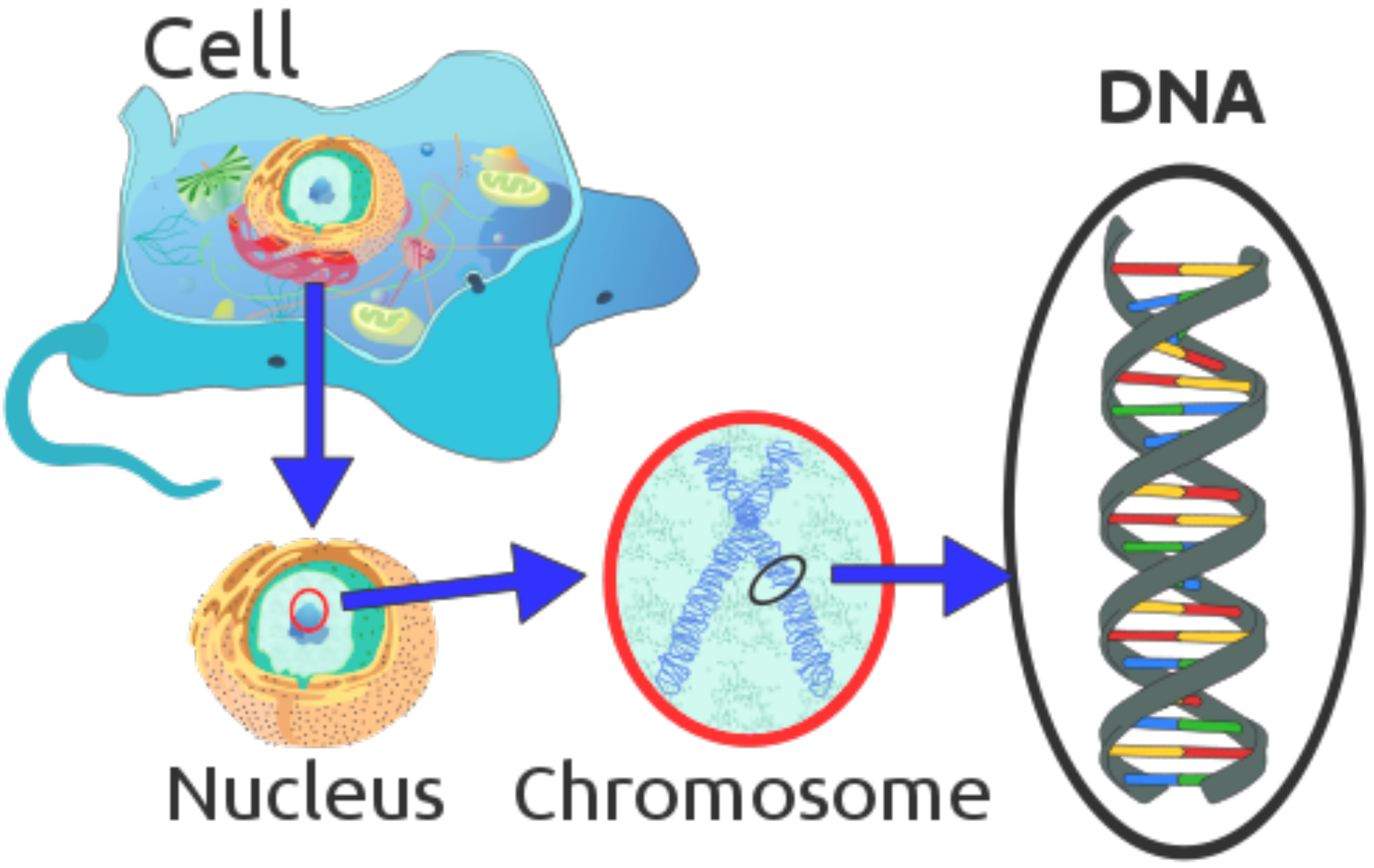


Cell Biology

- Three domains of
  - Prokaryotes
    - No Cell nucleu
  - Eukaryotes

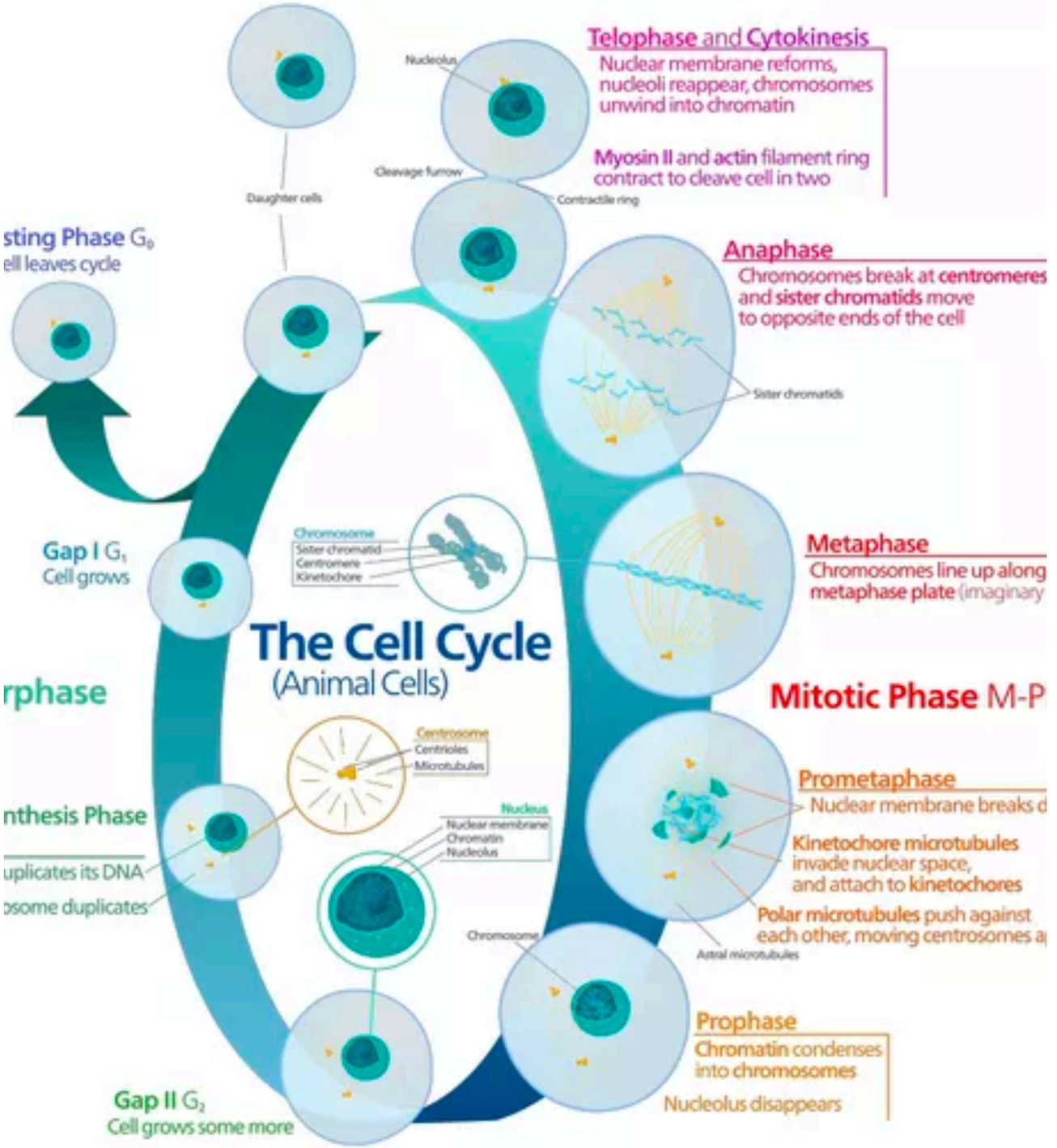



Eukaryote
Prokaryote
Membrane-
enclosed nucleus
Mitochondrion
Nucleolus
Ribosomes
Nucleoid
Capsule
(some prokaryotes)
Flagellum
Cell Membrane
Cell Wall
(in some eukaryotes)

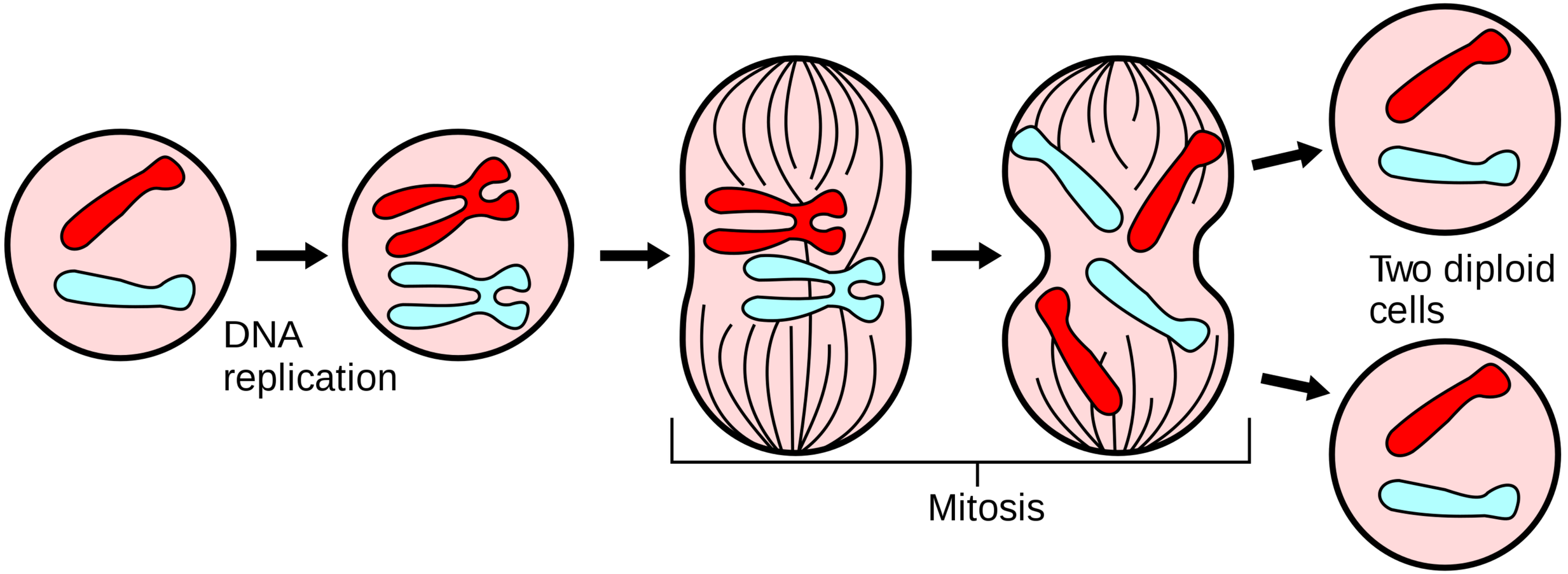

Cell
DNA
Nucleus
Chromosome

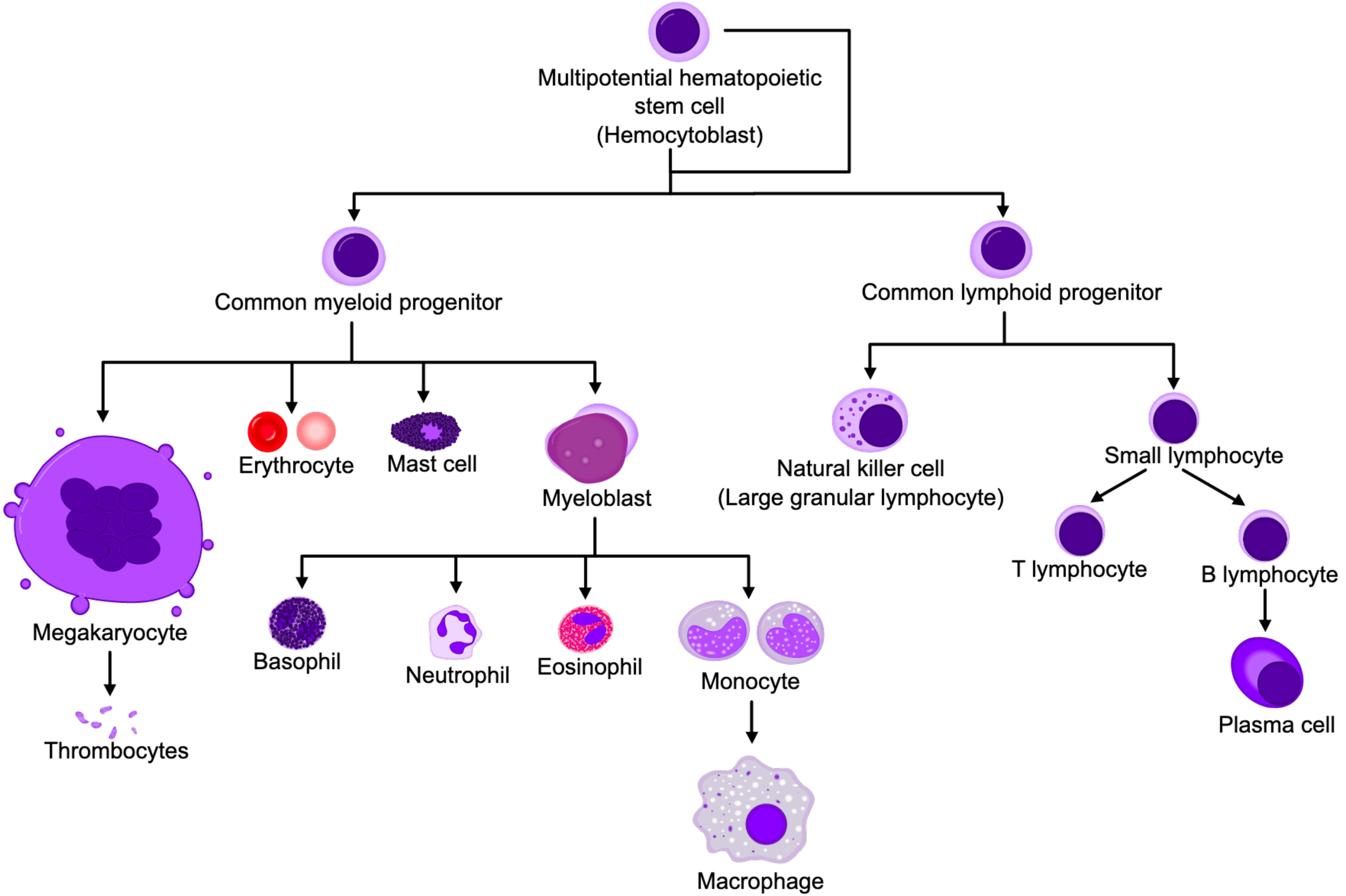
Cell Cycle

- Cells have many cycles
- Most critical one
  - Cell Division
- Cells are dying and replaced
  - Mitosis helps the process

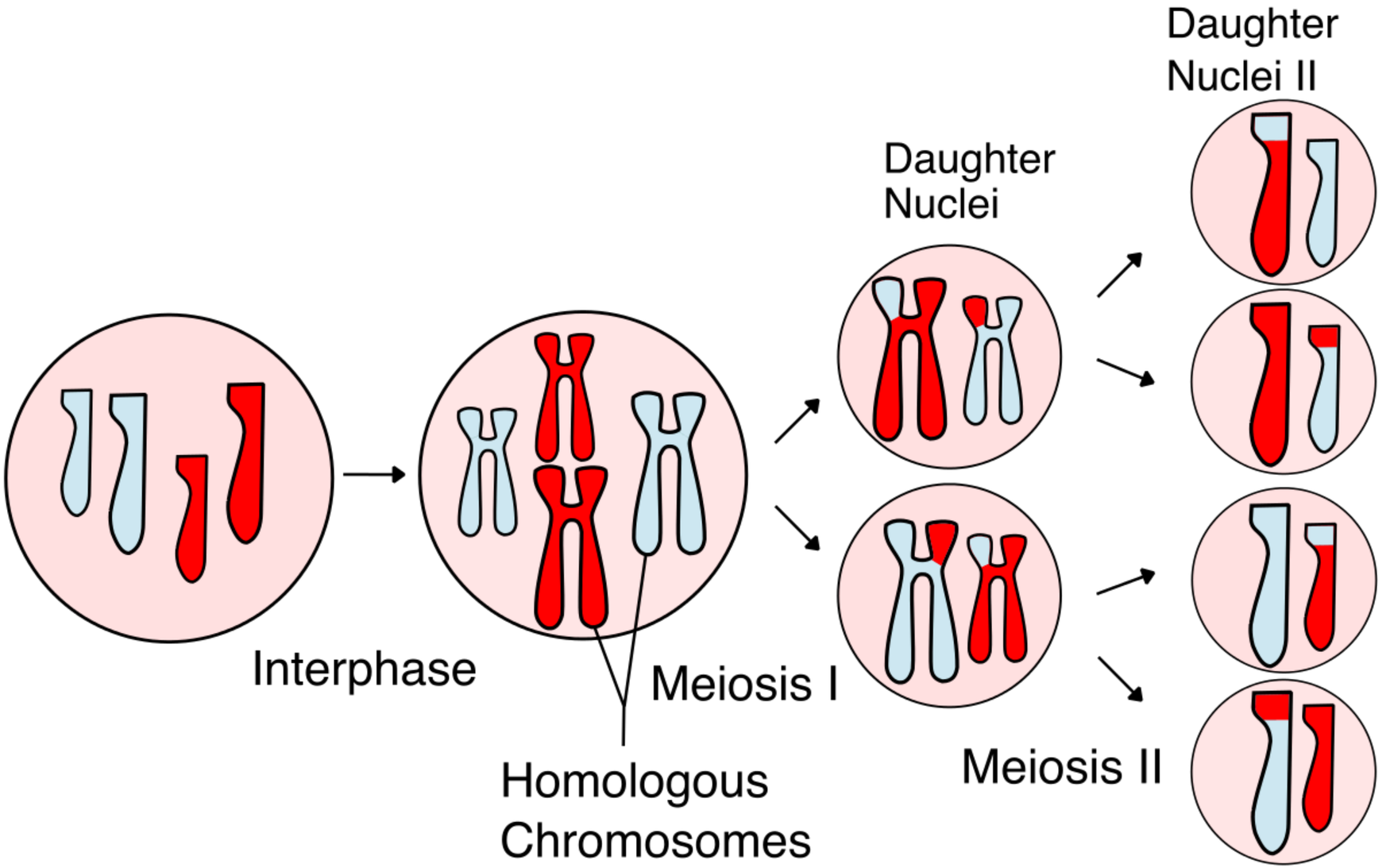


Mitosis

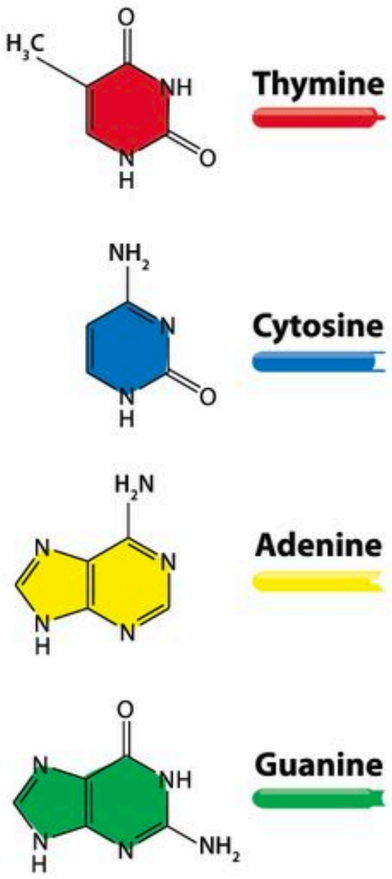



Multipotential hematopoietic stem cell (Hemocytoblast)
Common myeloid progenitor
Common lymphoid progenitor
Erythrocyte
Mast cell
Myeloblast
Megakaryocyte
Thrombocytes
Basophil
Neutrophil
Eosinophil
Monocyte
Macrophage
Natural killer cell (Large granular lymphocyte)
Small lymphocyte
T lymphocyte
B lymphocyte
Plasma cell

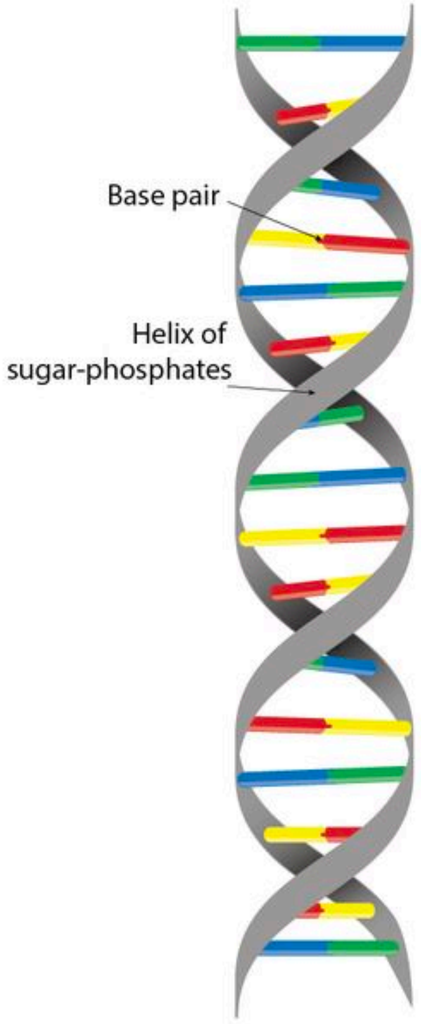
Meiosis


Daughter
Nuclei II
Daughter
Nuclei
Interphase
Meiosis I
Homologous
Chromosomes
Meiosis II

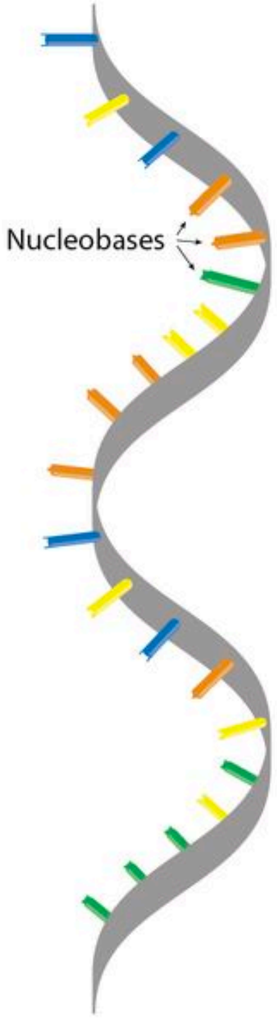
DNA and RNA



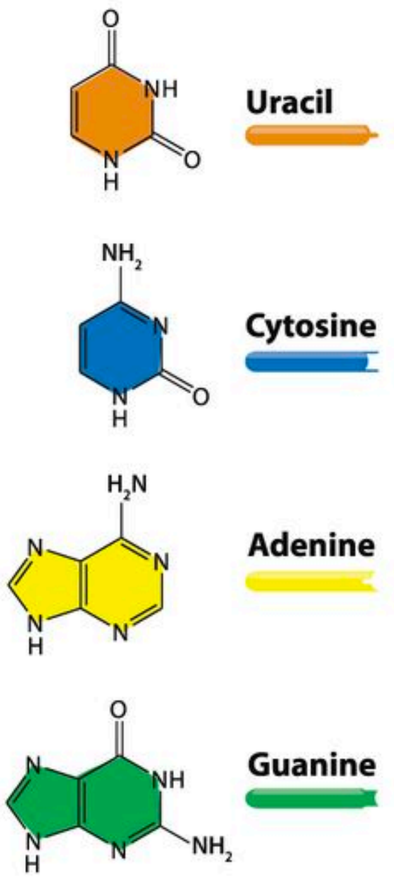
Nucleobases of DNA



DNA
Deoxyribonucleic acid



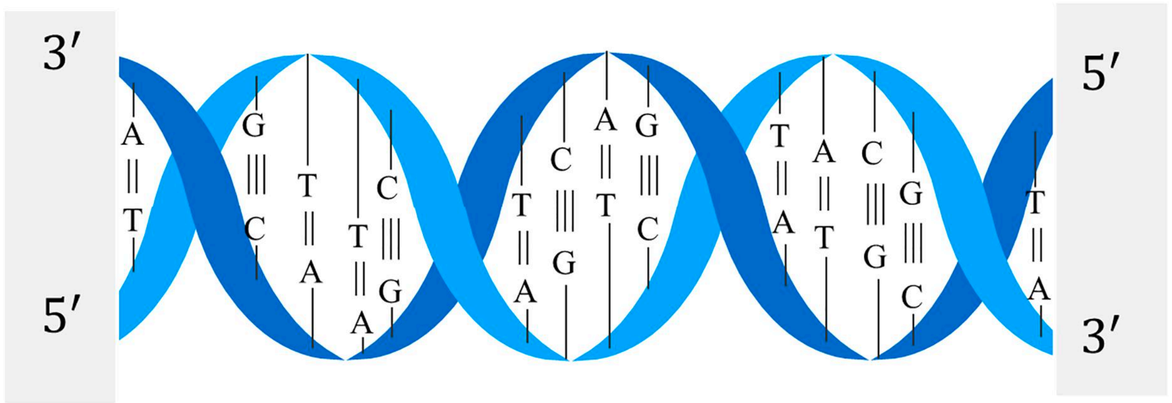
RNA
Ribonucleic Acid

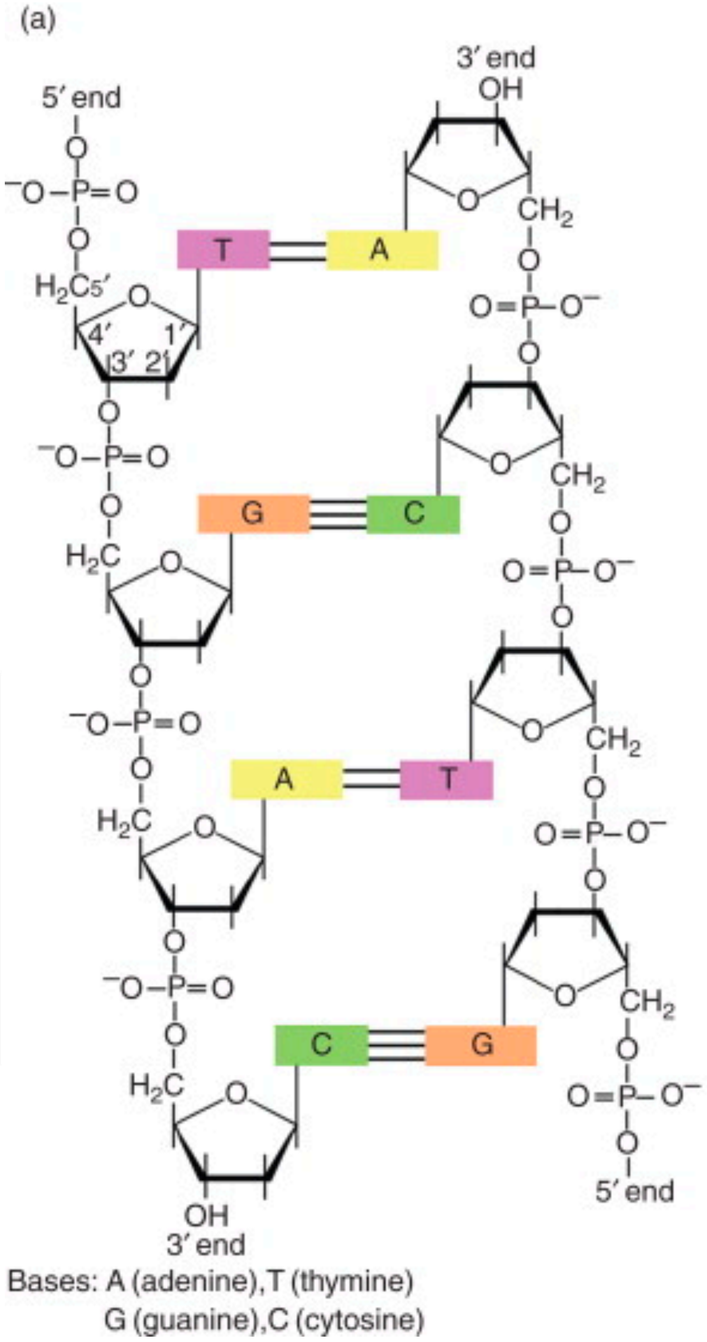


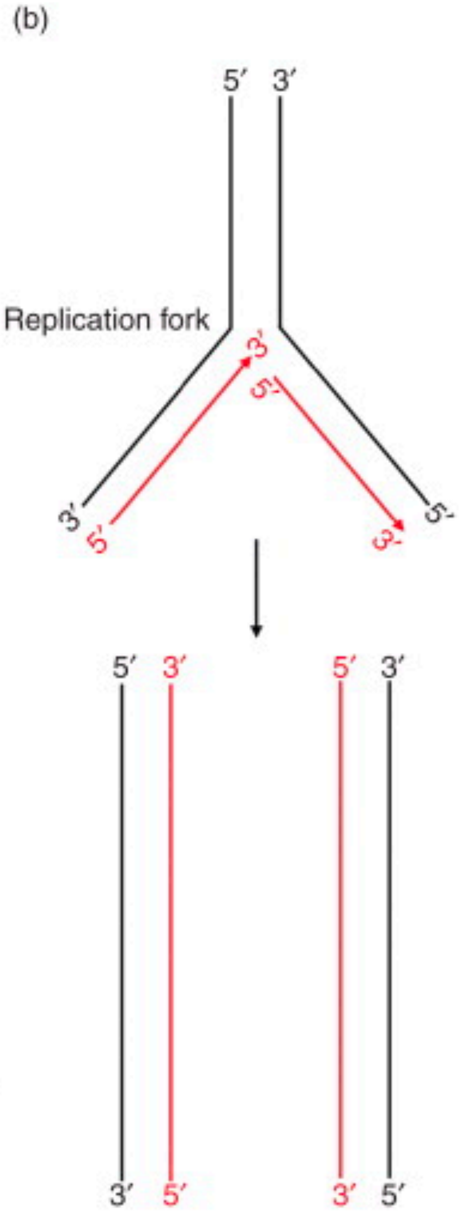
Nucleobases of RNA

DNA Strands

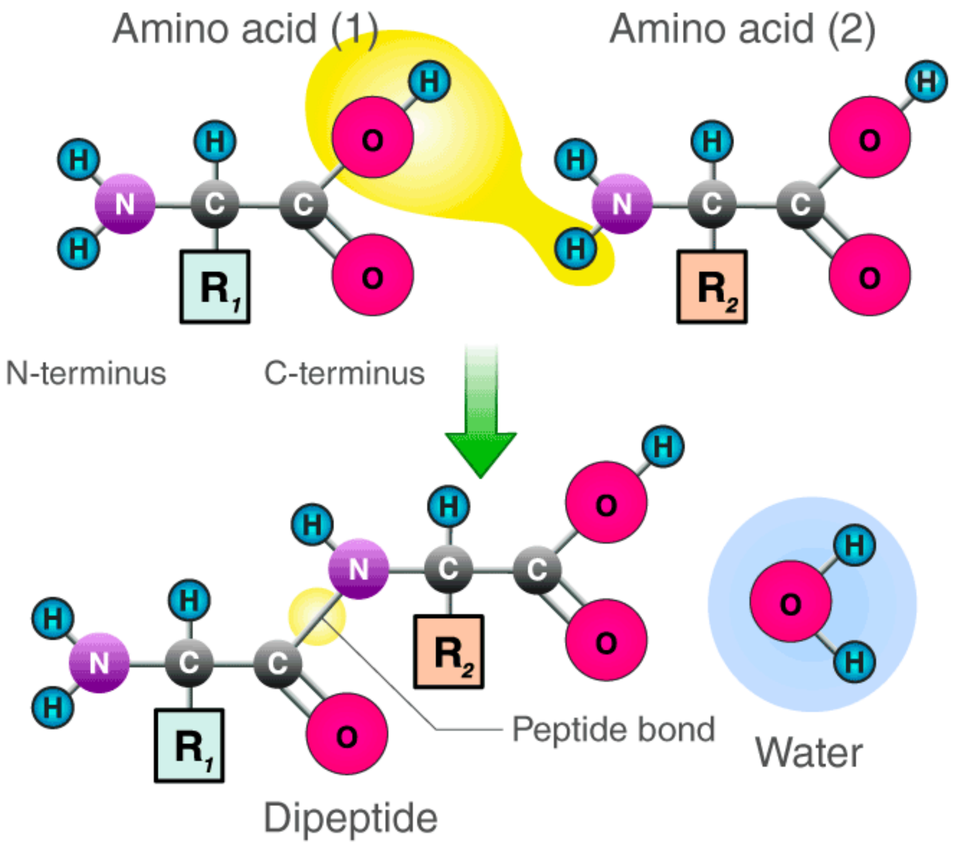
- The 5' and 3' designations refer to the number of carbon atom in a deoxyribose sugar molecule to which a phosphate group bonds.

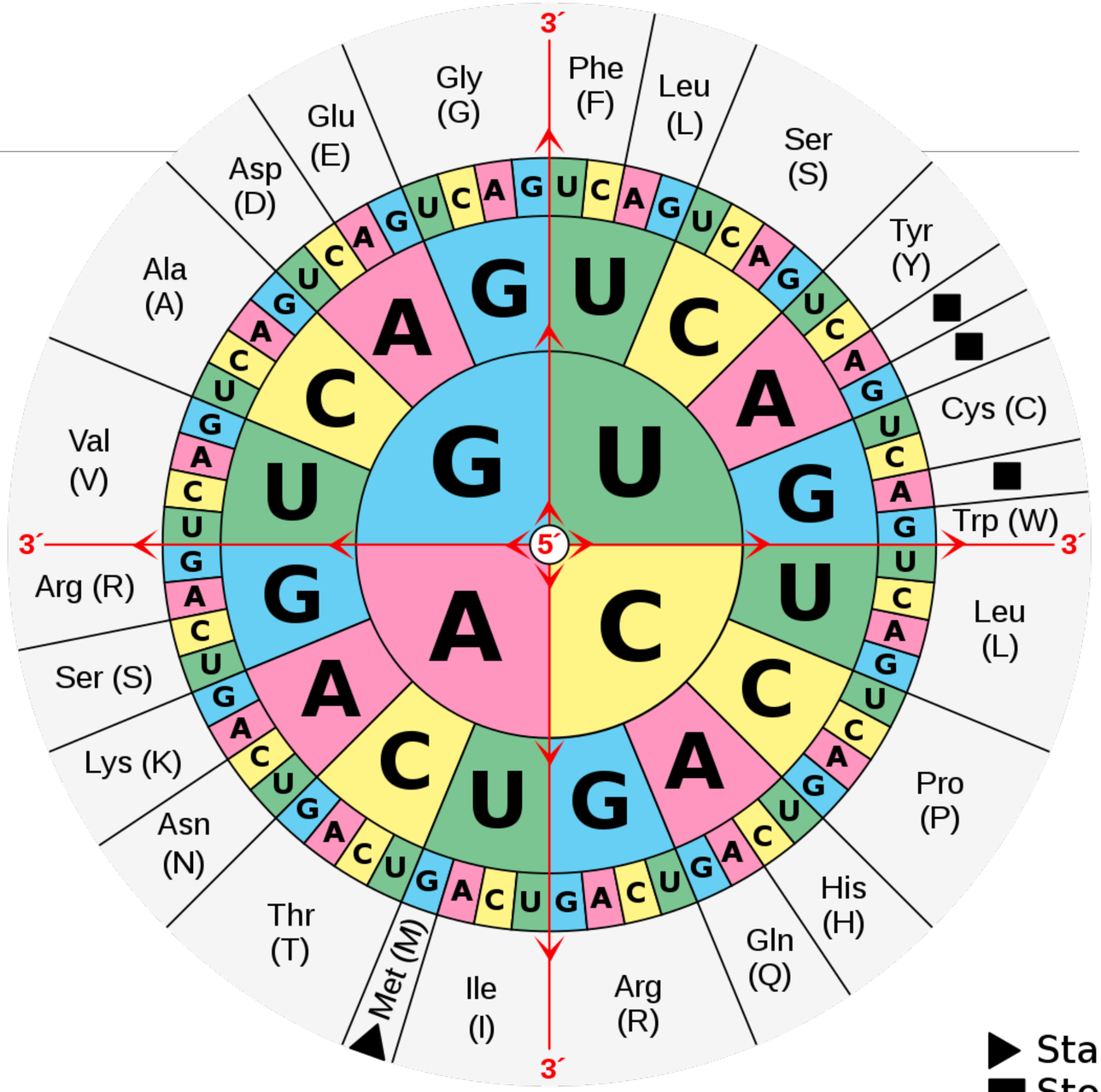






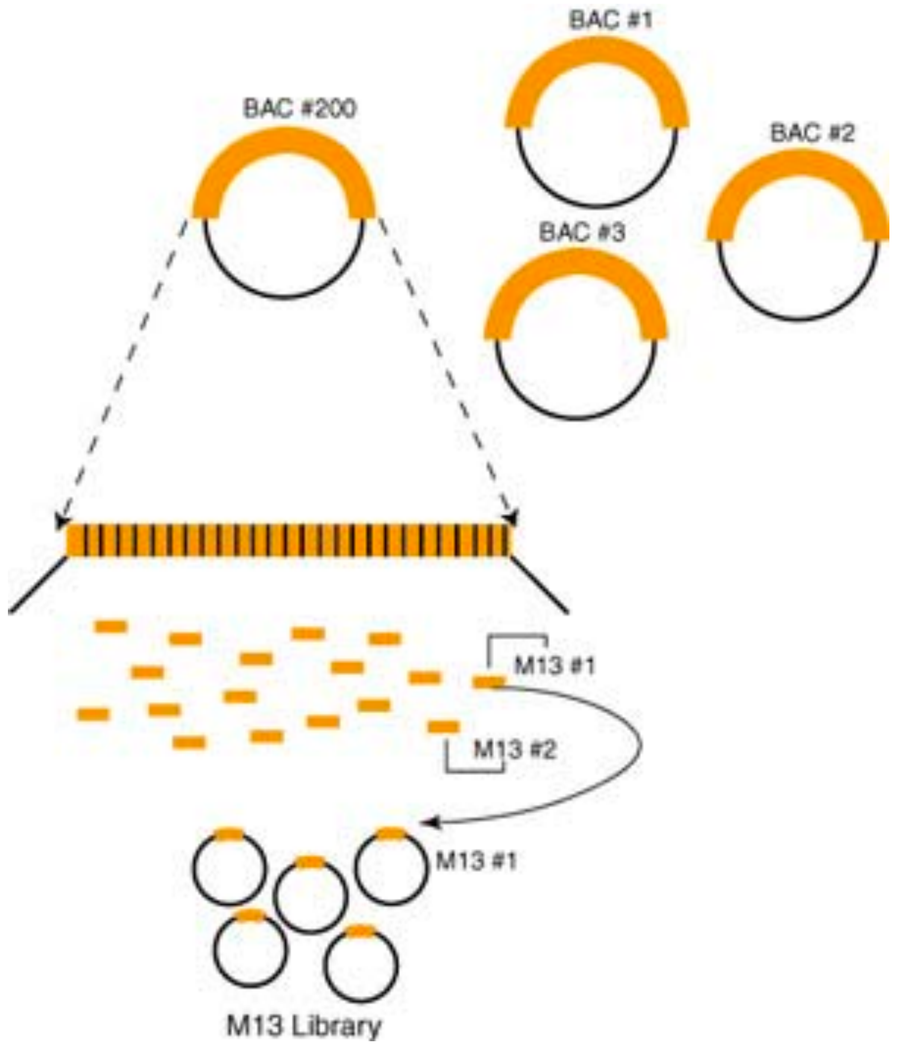
DNA -> RNA -> Protein





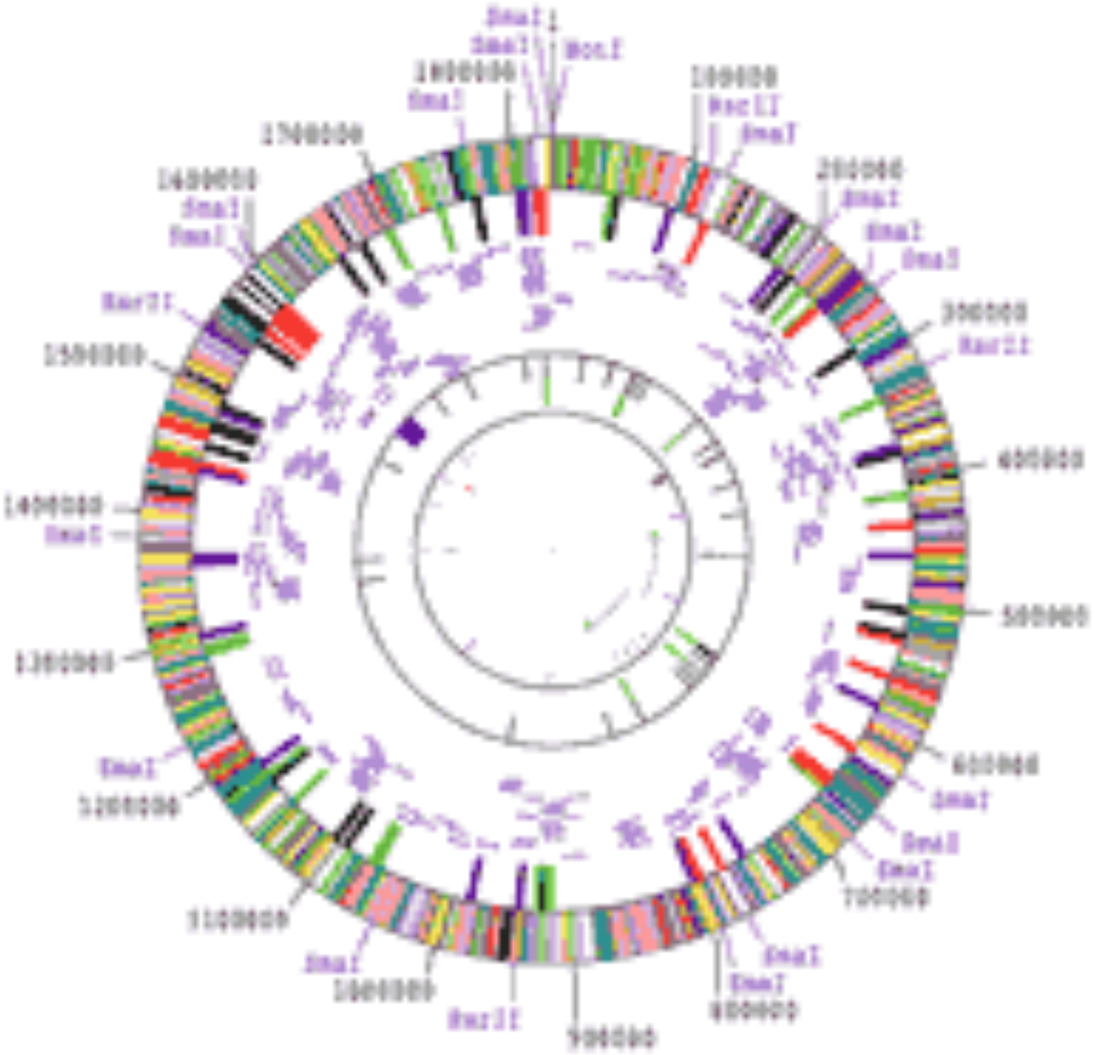
The Human Genome Project

- Biology’s Manhattan Project
- Started in 1989
- Sanger Center in the UK
- 10$ per base / 3 Billion base pairs
- By 2005 -> $1 per base
- 1990s
  - Small of large pieces of DNA called maps and put them somewhere on the genome
  - Bacterial Artificial Chromosomes (BAC)



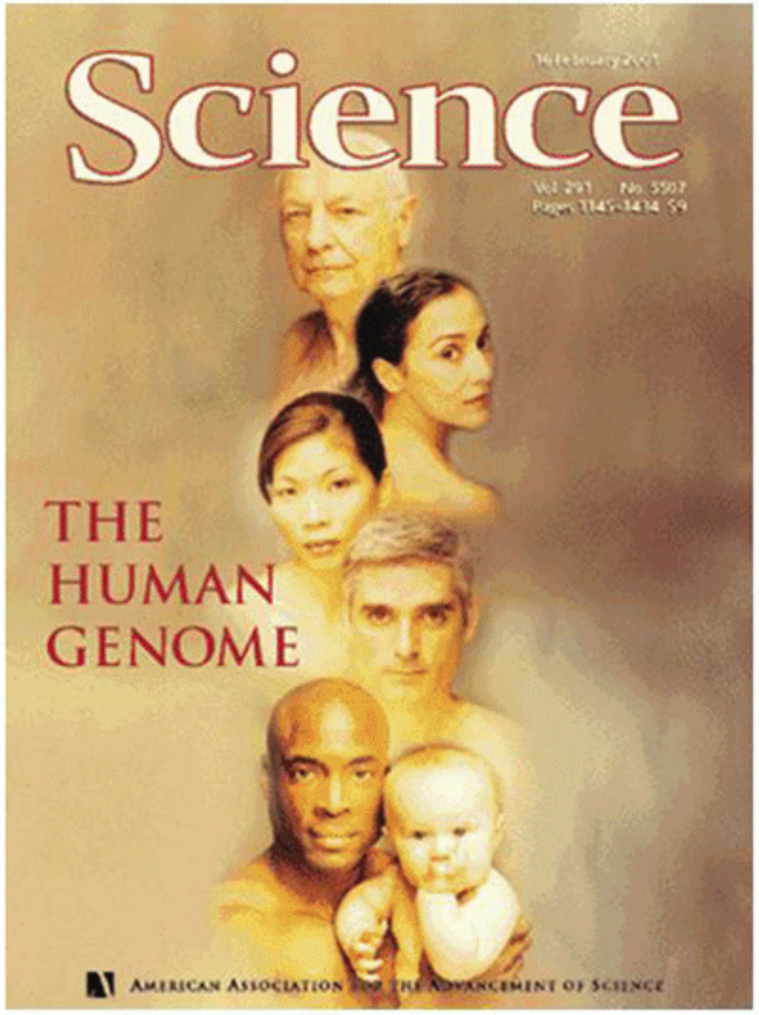
The first genome

- TIGR - 1995
  - *Haemophilus influenza*
  - 1.8 million bases
  - 1742 genes
  - Whole genome sequencing
- 1998
  - Applied Biosystems
  - Celera Genomics
- Sanger and NIH started acceleration



https://www.yourgenome.org/facts/timeline-organisms-that-have-had-their-genomes-sequenced

The Human Genome Project



Number of Genes?

Published: 22 February 1964

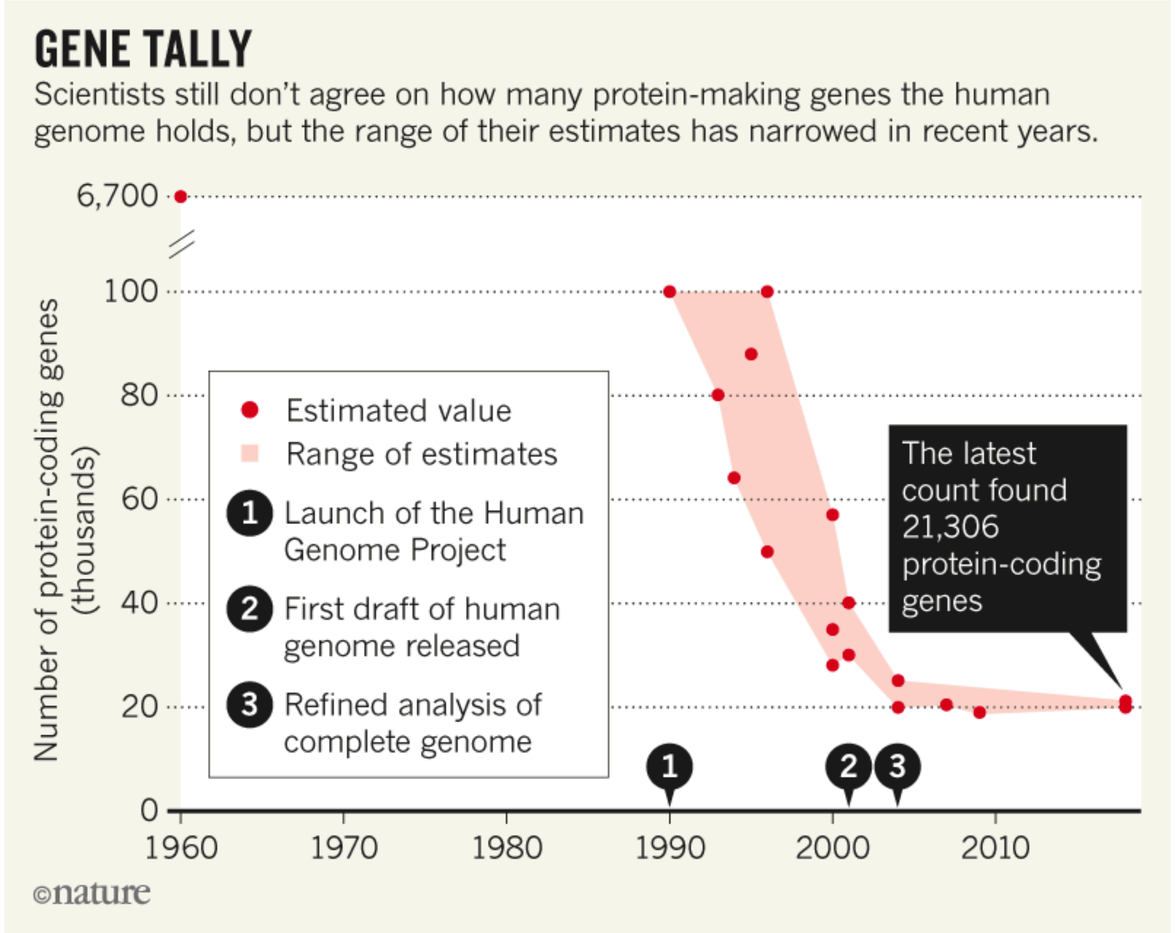
A Preliminary Estimate of the Number of Human Genes

F. VOGEL

Nature **201**, 847 (1964) | Cite this article

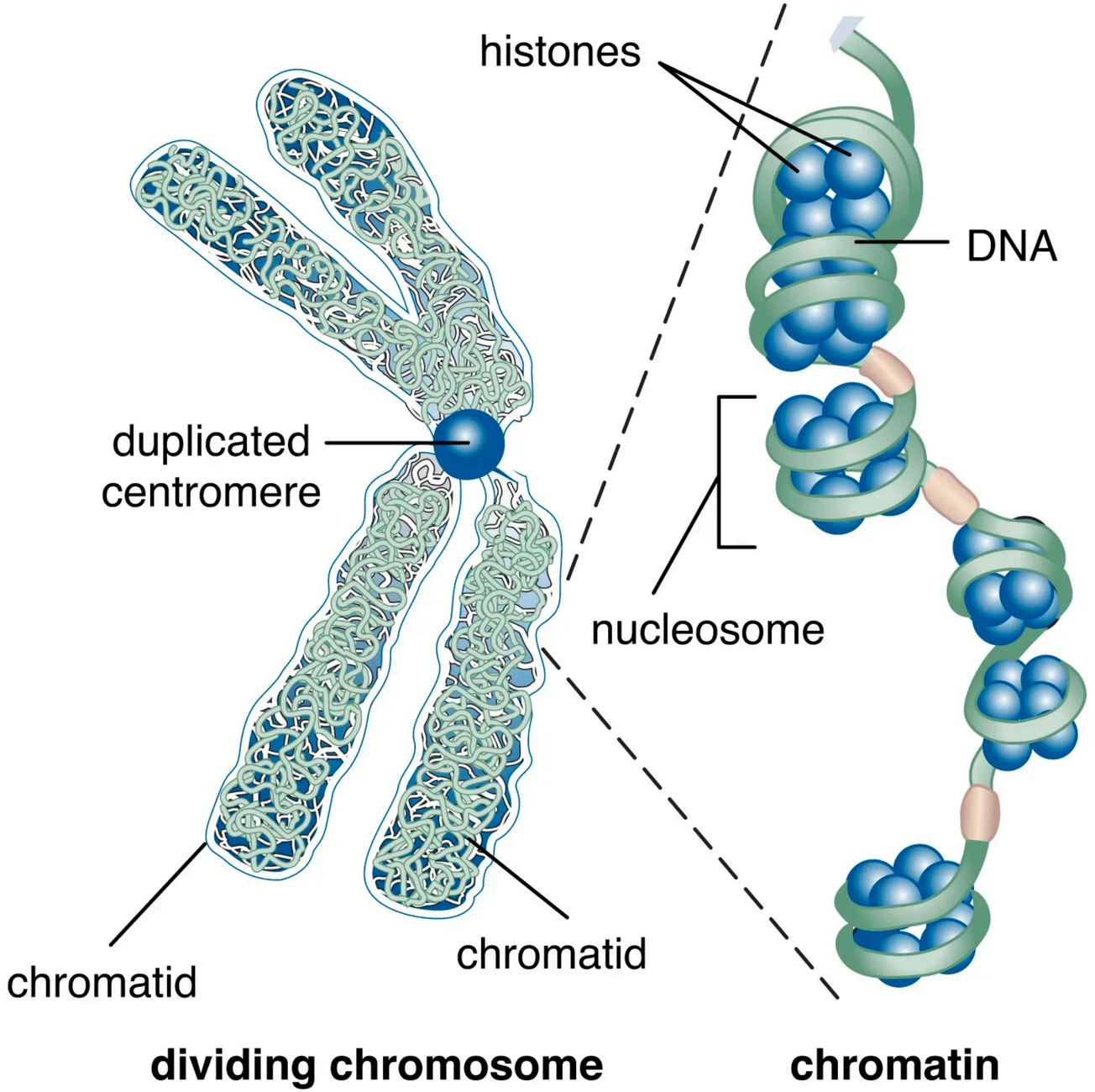
Abstract

RECENT results of molecular genetics enable us to estimate the number of human genes, if certain assumptions are made. The following data are available: (1) The α-chain of human hæmoglobin contains 141, the β-chain contains 146 amino-acids, corresponding to a molecular weight of about 17,000 each[1]. Assuming a triplet code[2,3] this means that the α- and β-chains are determined by 423 and 438 nucleotide pairs, respectively. According to 'Svedberg's law'4, many proteins consist of sub-units of the same order of magnitude (molecular weight of about 17,500). Hence, the assumption seems to be warranted that one average structural geno might have a length of about 450 nucleotide pairs. (2) The weight of one haploid human chromosome set in human spermatozoa is about 2.72×10^{-12} g. Granulocytes contain about 6.23×10^{-12} g; lymphocytes contain about $5.84 \times {}^{-12}$ g (ref. 5). Extensive examinations have shown that the DKA. content is constant in all resting cells of one species, which have the same number of chromosome sets, and depends on the degree of



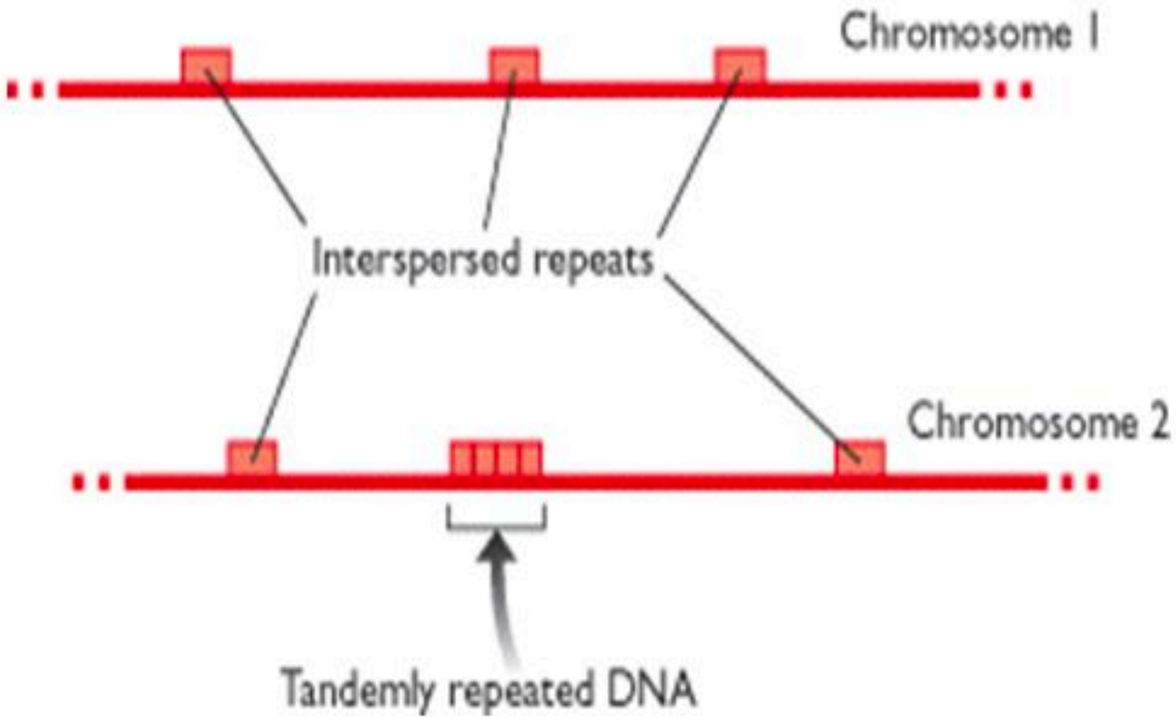
Structure of DNA

- DNA itself, if you stretched it out, would be about two meters long inside each cell.
- In order to fit into a cell it has to be wrapped up in very, a very efficient way.
- So it's actually wrapped around other molecules called histones in what you can see as sort of this beads on a string.
- And those histones are wrapped together in these organized slightly longer structures. And those together, are, are then coiled up and super coiled in, in even bigger structures which eventually form the chromosomes.
- Now for DNA to be transcribed and translated, it has to unwrap itself a bit so we'll talk a little bit more about that next.



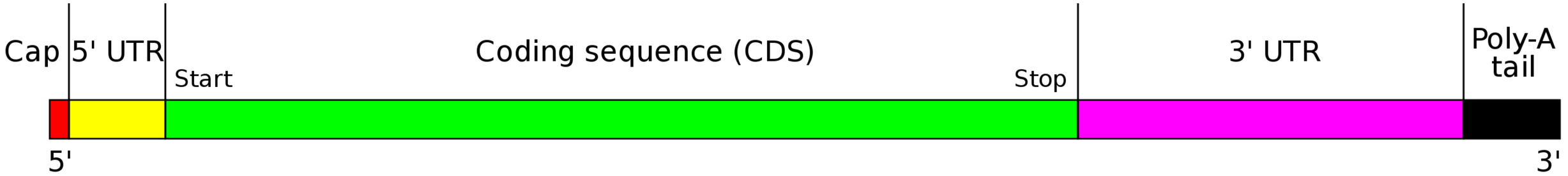
Repeat

- Tandem Repeats
- Interspaced Repeats
- Repeats cause trouble in sequencing
  - You don’t know where they come from

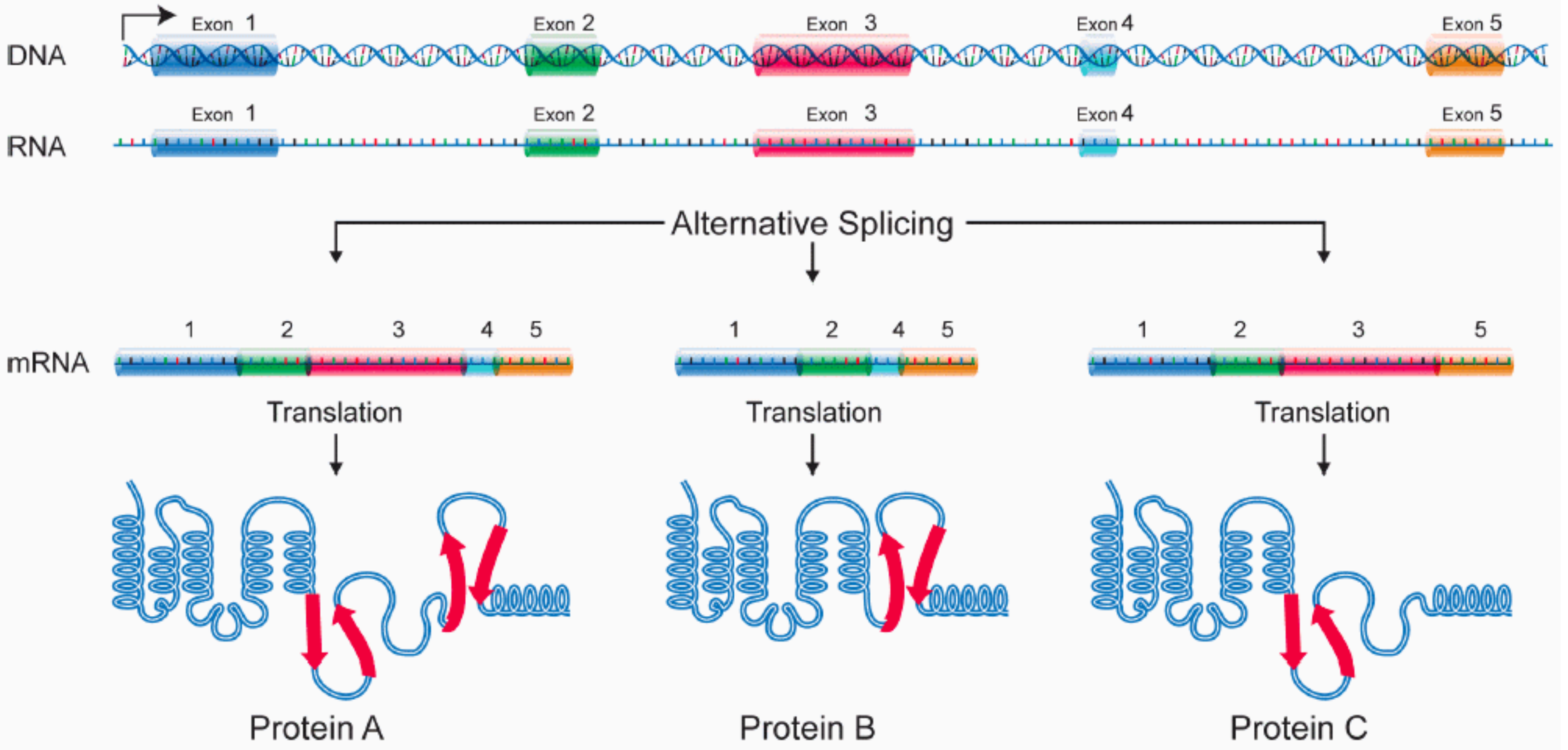


RNA

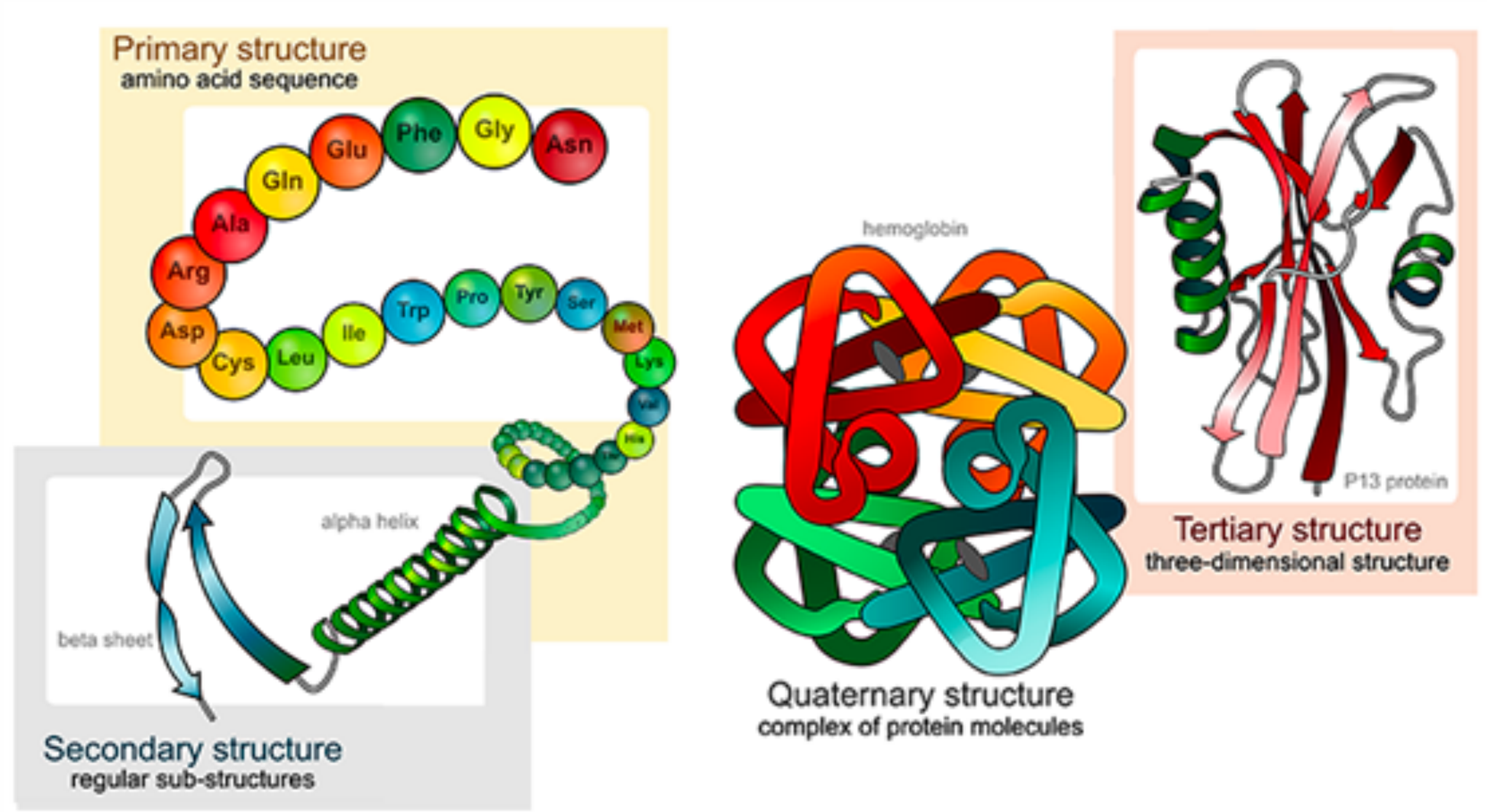
The structure of a typical human protein coding mRNA including the untranslated regions (UTRs)



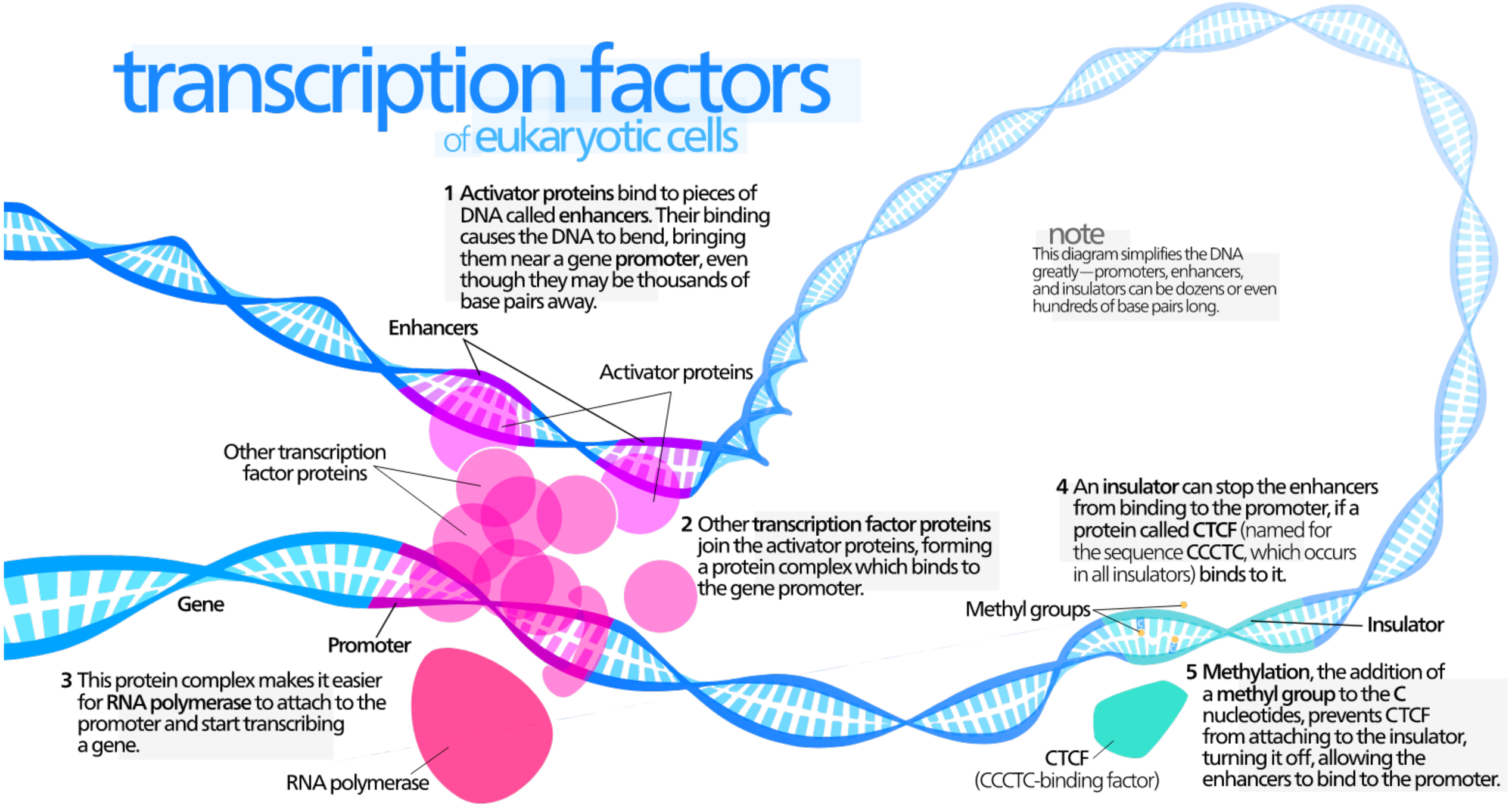
RNA

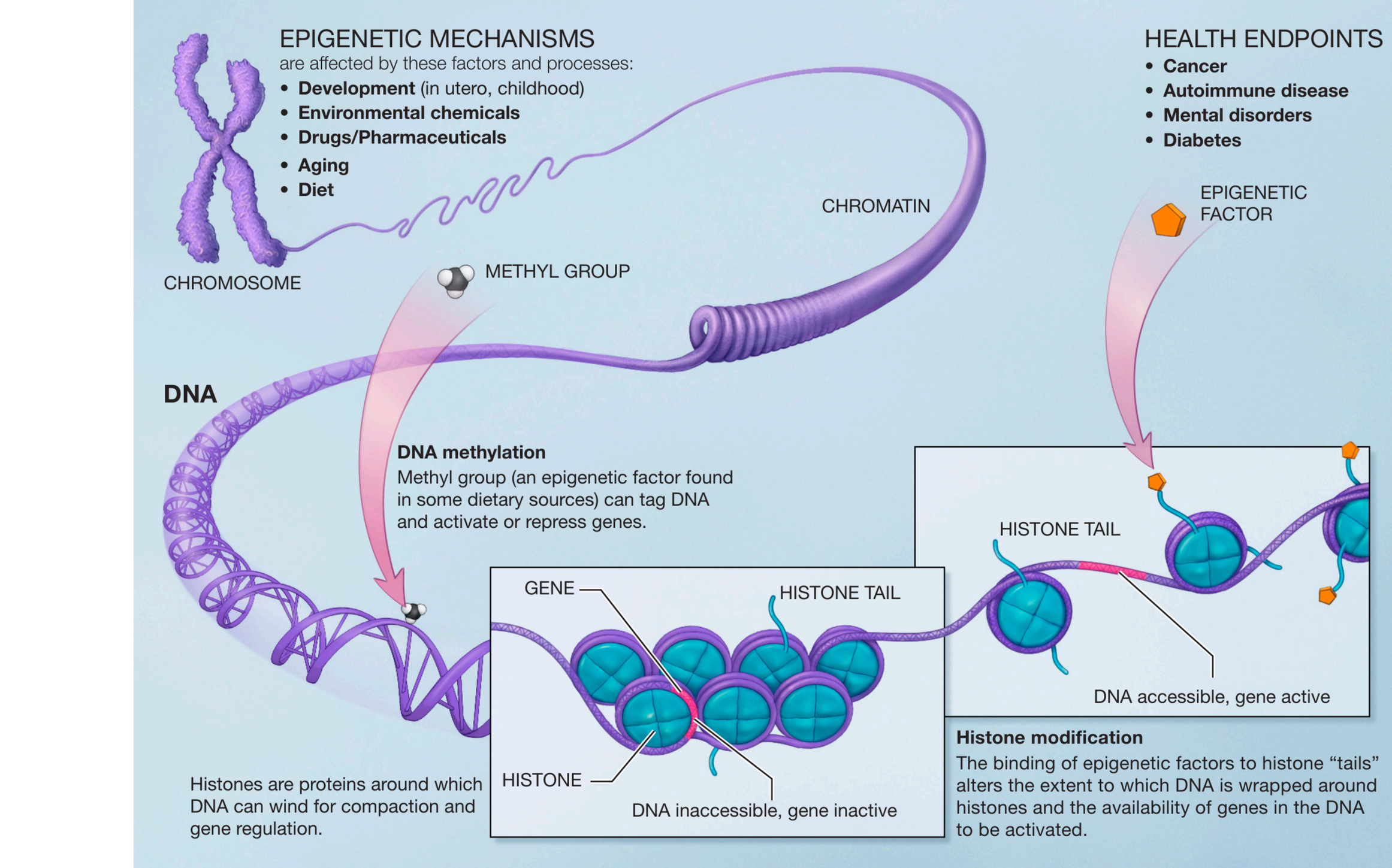

DNA
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
RNA
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
Alternative Splicing
mRNA
1 2 3 4 5
1 2 4 5
1 2 3 5
Translation
Translation
Translation
Protein A
Protein B
Protein C

Protein


Primary structure
amino acid sequence
Phe
Gly
Asn
Glu
Gln
Ala
Arg
Asp
Cys
Leu
Ile
Trp
Pro
Tyr
Ser
Met
Lys
Val
His
alpha helix
beta sheet
Secondary structure
regular sub-structures
hemoglobin
Quaternary structure
complex of protein molecules
P13 protein
Tertiary structure
three-dimensional structure

transcription factors
of eukaryotic cells




EPIGENETIC MECHANISMS
are affected by these factors and processes:
• Development (in utero, childhood)
• Environmental chemicals
• Drugs/Pharmaceuticals
• Aging
• Diet
HEALTH ENDPOINTS
• Cancer
• Autoimmune disease
• Mental disorders
• Diabetes
CHROMOSOME
CHROMATIN
METHYL GROUP
EPIGENETIC FACTOR
DNA
DNA methylation
Methyl group (an epigenetic factor found in some dietary sources) can tag DNA and activate or repress genes.
GENE
HISTONE TAIL
HISTONE
DNA inaccessible, gene inactive
HISTONE TAIL
DNA accessible, gene active
Histones are proteins around which DNA can wind for compaction and gene regulation.
Histone modification
The binding of epigenetic factors to histone "tails" alters the extent to which DNA is wrapped around histones and the availability of genes in the DNA to be activated.