Introduction to Bioinformatics

Intro to biology & genetics

Martin Blažek

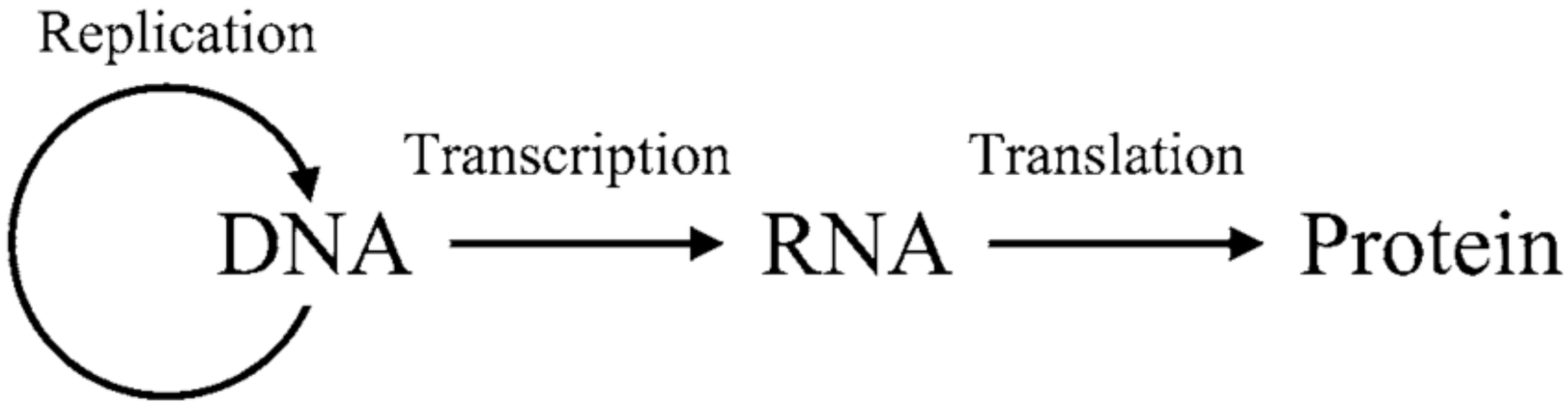
17.02.2026

What is bioinformatics?

- Informatics + molecular biology
- Algorithms
- Statistics
- Big data

What is molecular biology?

- Study of biological processes at the molecular level, especially nucleic acids and proteins
- Focused on the mechanisms of replication, transcription, translation, and regulation of gene expression, using knowledge from genetics, biochemistry, and physics



Nucleic acids

- Contain and transmit genetic information
- Present in all bacteria (including mitochondria), viruses, plants, and animals
- **DNA** (deoxyribonucleic acid):
  - template
  - contains coding, non-coding and regulatory regions
- **RNA** (ribonucleic acid):
  - transcribed DNA containing only coding regions
  - a recipe for proteins the body needs in the moment

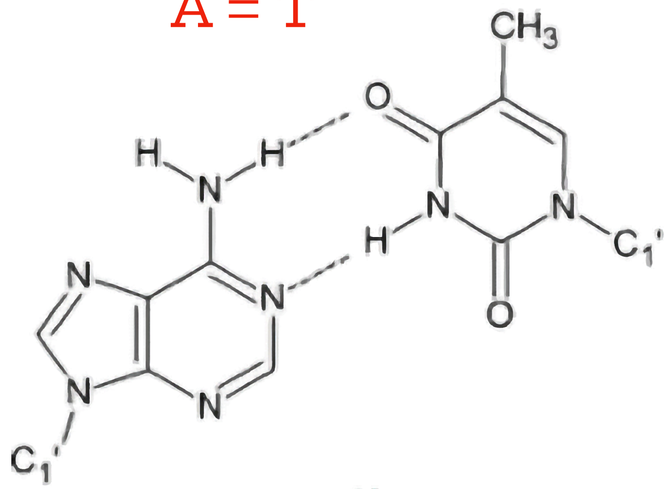
Nucleic acids

- **Structure**: composed of H_3PO_4, sugar, and **nitrogenous base**
- **Arrangement of a single strand**:
  - **Sugar-phosphate**: forms a backbone
  - **Nitrogenous base**: sticks out from the backbone
- **Nucleotide**: The portion of nucleic acid containing 1 **nitrogenous base**
- **Codon(Triplet)**: The portion of nucleic acid containing 3 **nucleotides**

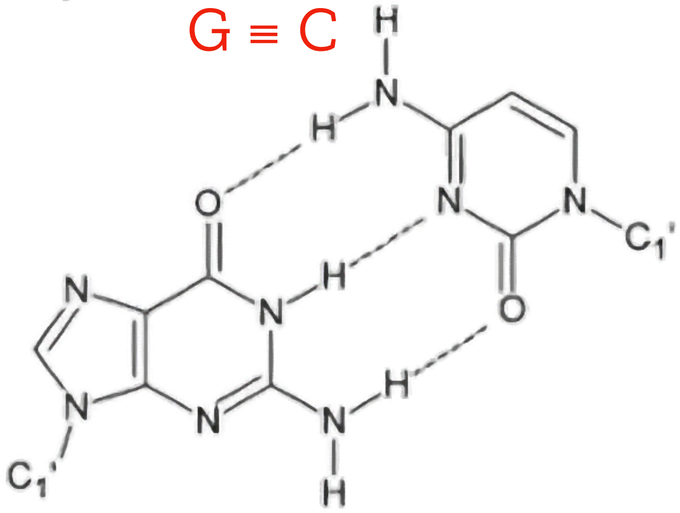
Grammar of nucleic acid

- **Nitrogenous base** (or simply **base**):
  - Purine: **A** - Adenine, **G** - Guanine
  - Pyrimidine: **T** - Thymine **C** - Cytosine (& **U** - Uracil)
- **DNA complementarity**: **A** pairs only to **T**, **C** pairs only to **G**
- **T** is in DNA, in RNA it is replaced by **U**
- Bases form triplets which encode **amino acids**
- Amino acids form **proteins**
- **Proteins** are workhorses of the organism

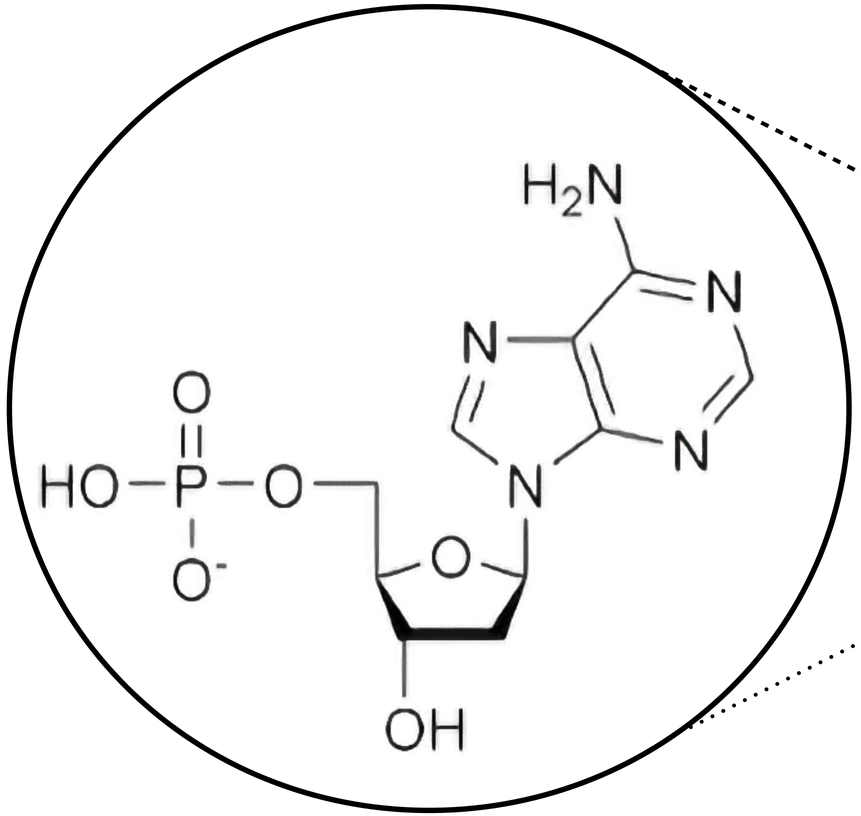
A = T

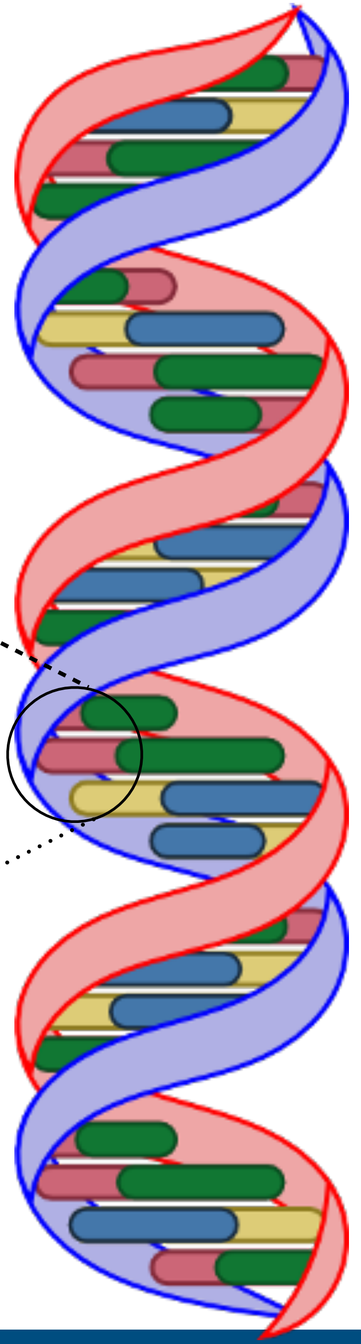


G ≡ C



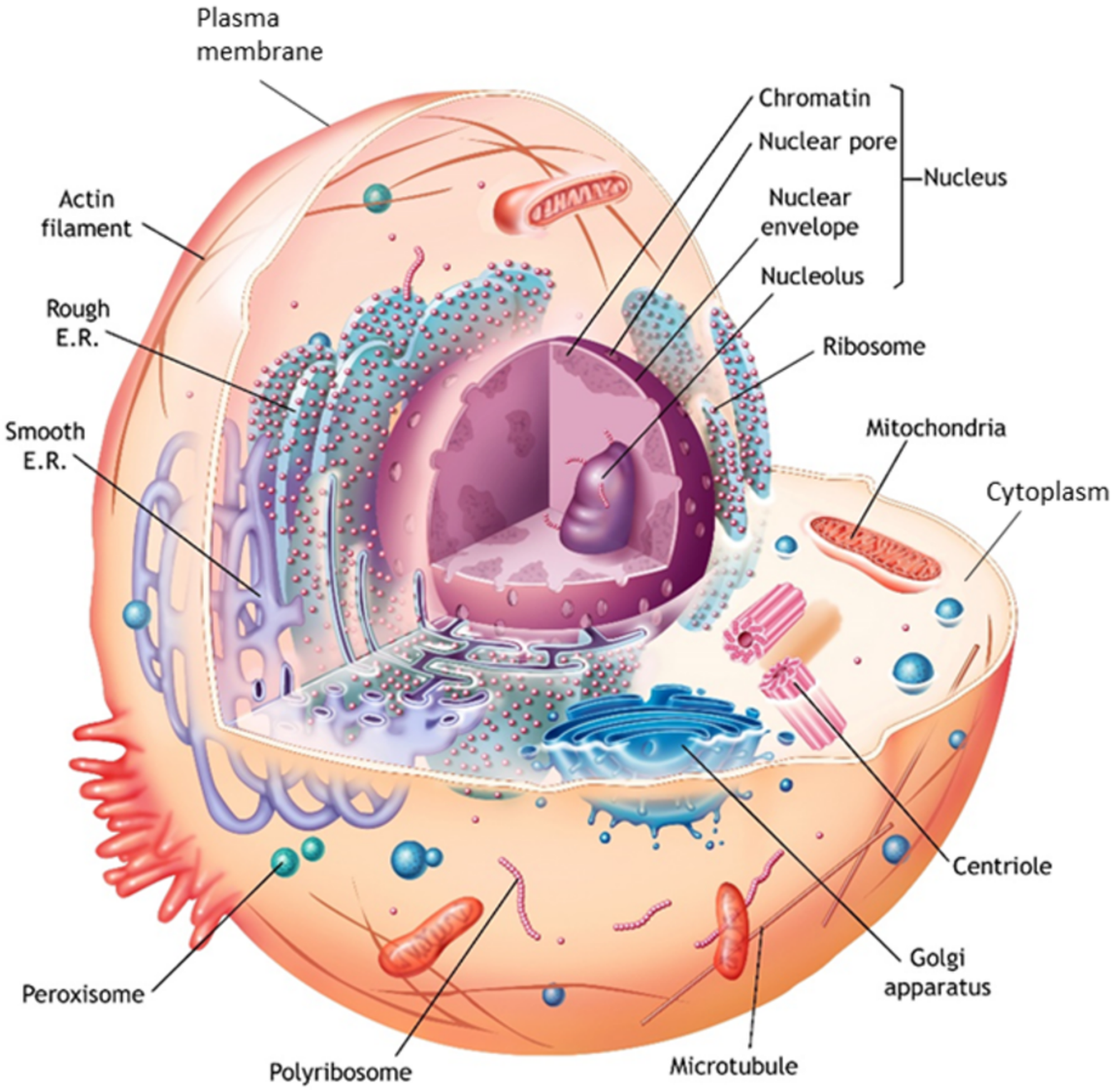
Grammar of nucleic acid





The building blocks of human body are cells

- **Nucleus**: houses the cell's genetic material, **DNA**, organised in **chromosomes**
  - **Chromosome**: organises **DNA** divided into smaller parts
  - **Nucleolus**: produces **ribosomes**
- **Cytoplasm**: fluid filling the interior of all cells
  - **Mitochondria**: generates ATP, source of chemical energy
  - **Ribosome**: translates **RNA** into amino acids, which form **Proteins**

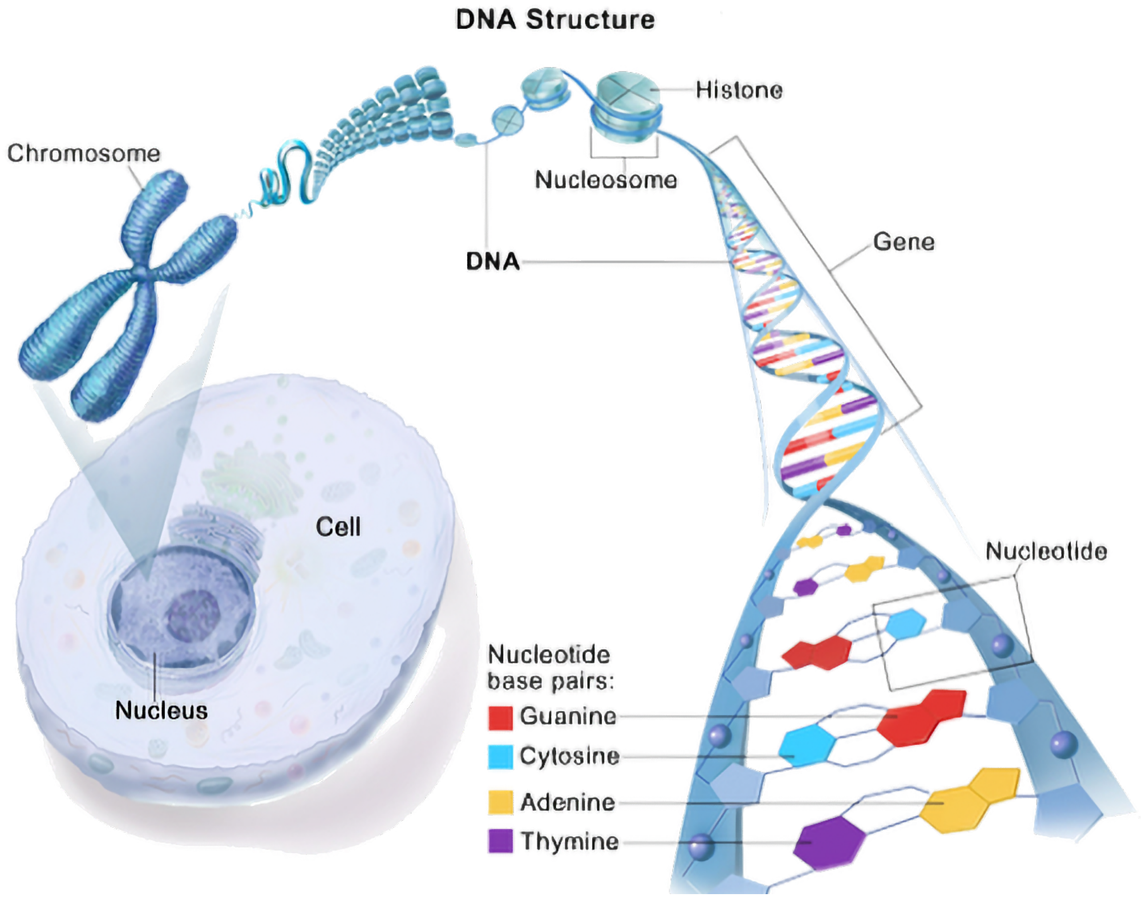

Human cell
Plasma membrane
Actin filament
Rough E.R.
Smooth E.R.
Peroxisome
Polyribosome
Chromatin
Nuclear pore
Nuclear envelope
Nucleolus
Nucleus
Ribosome
Mitochondria
Cytoplasm
Centriole
Golgi apparatus
Microtubule

Genome

- **Genome**: all genetic information of an organism or cell
- Around **3.2 billion nucleotide bases** long in humans
- But length is not important - human genome is very similar to mice, apes, even some worms, but **we can utilise the genes more efficiently**
- **Genotype**: set of genes organism carries
- **Phenotype**: observable characteristics

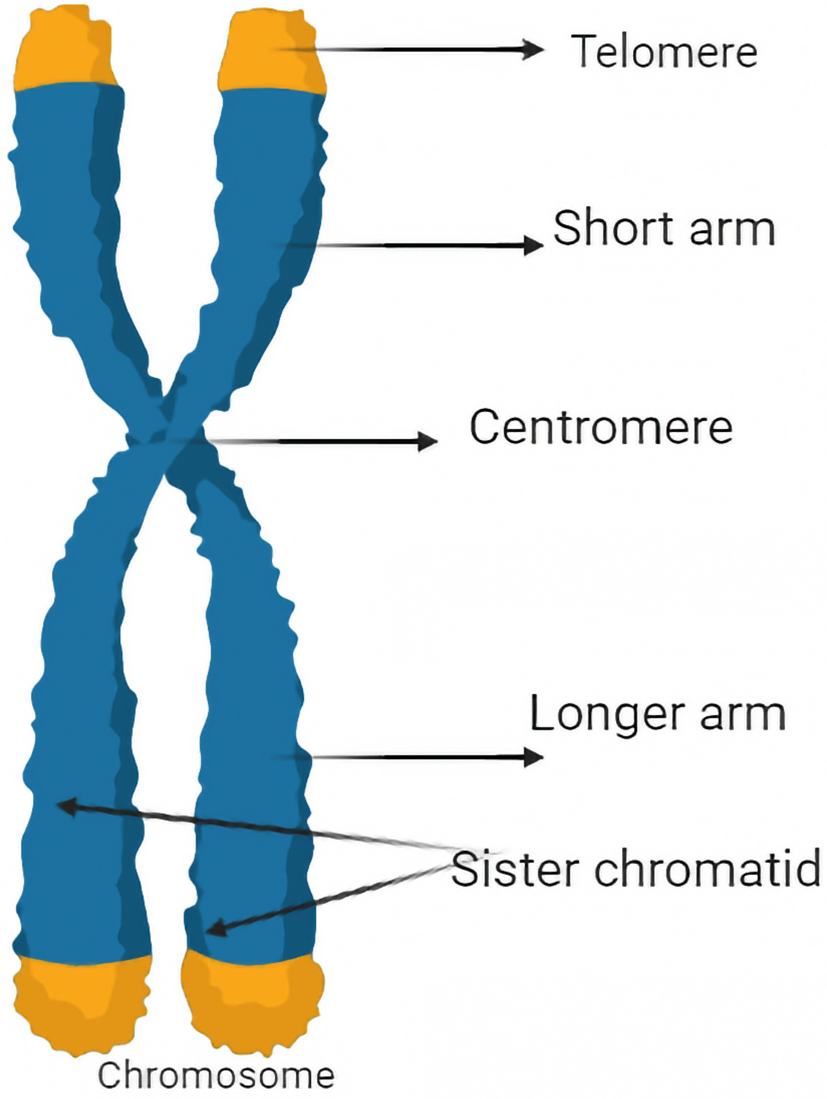
Organization of Human Genome

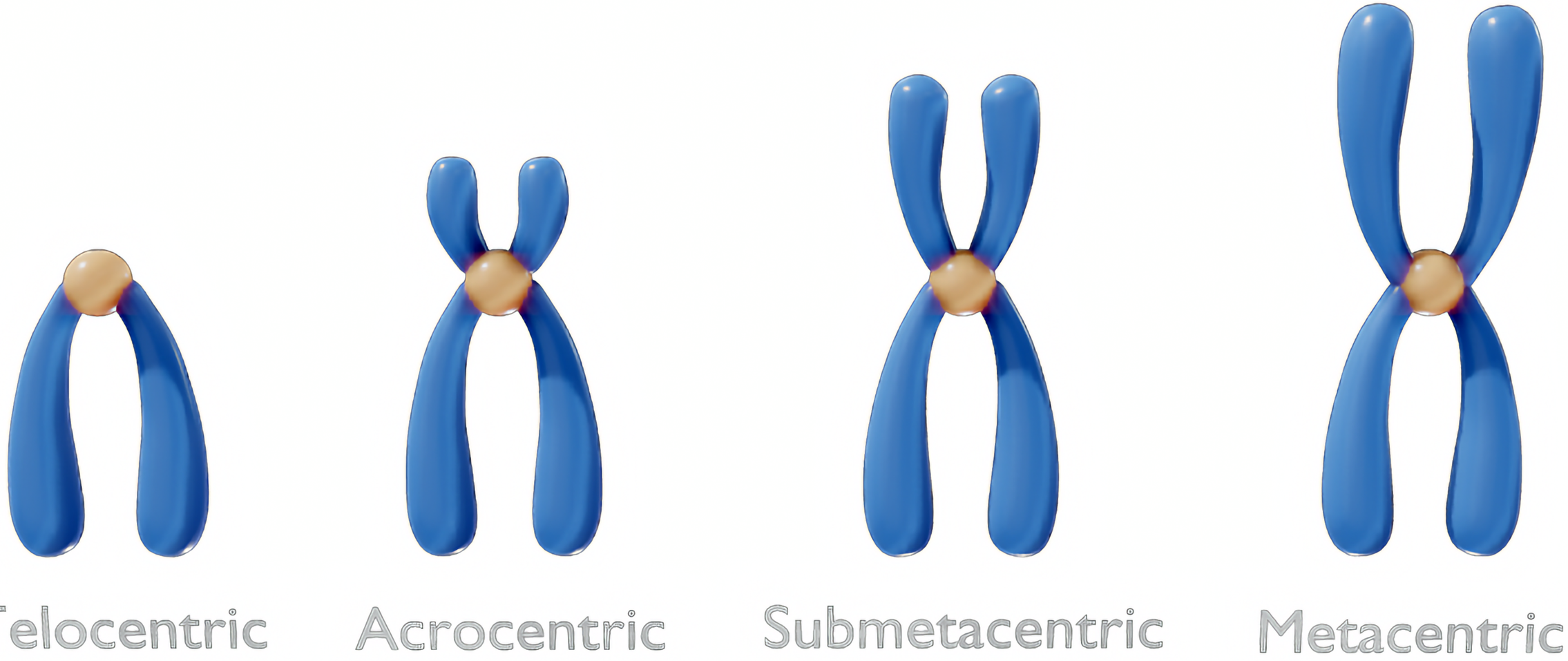
- DNA condensates from a single strand to a supercoiled structure: **Chromosome**
  - **DNA —> Nucleosome —> Chromatin**
- **Human nucleus genome**:
  - 3200Mb distributed between 24 linear molecules
  - Coding part represents only 1,5%
  - Repetitive regions - 45%

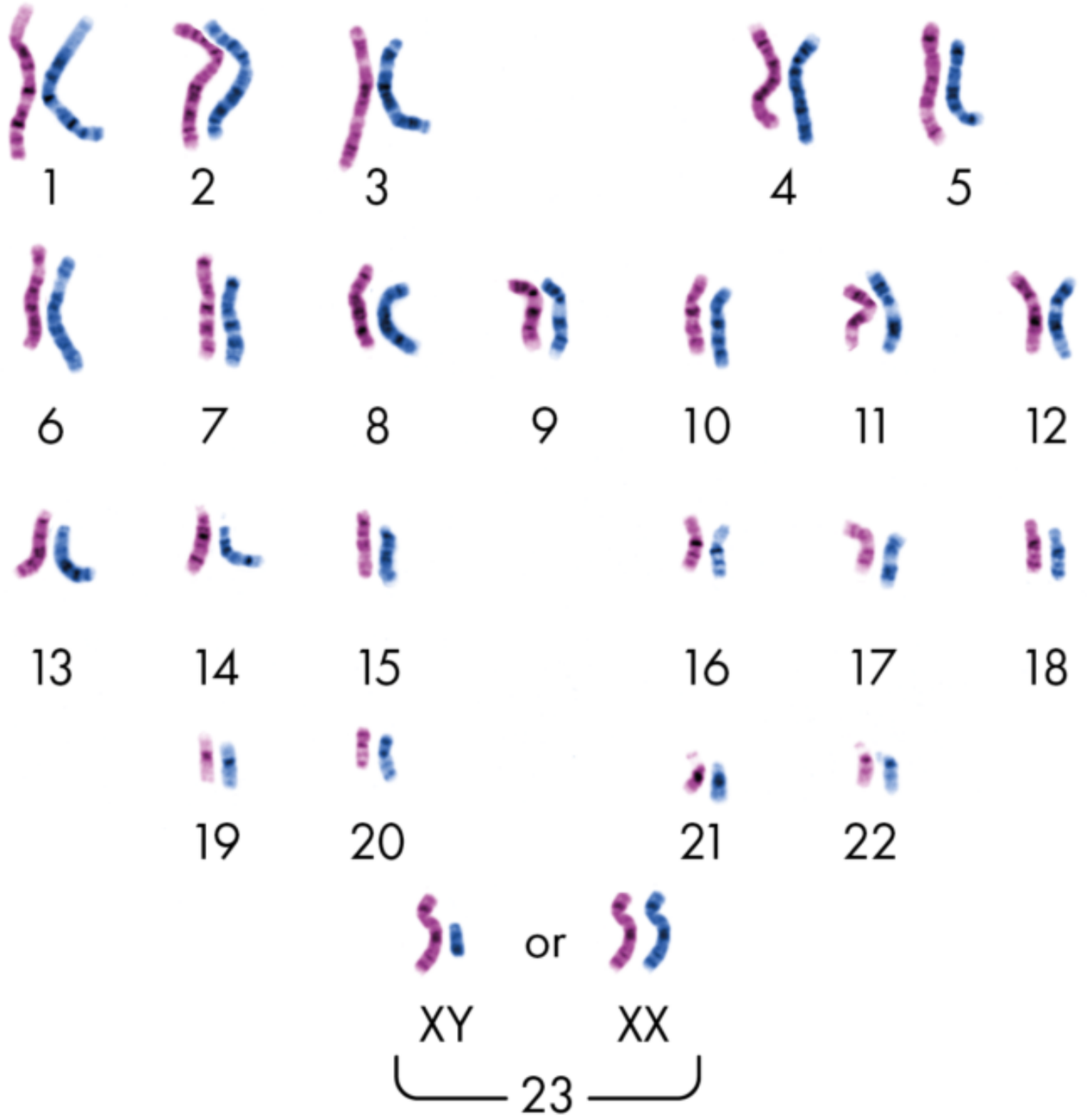


Chromosome

- **Human DNA is divided into 22 autosomal + 1 sex chromosome**
- **Each human has 2 chromosomal pairs**: 23 chromosomes from 1st parent, 23 chromosomes from 2nd parent + mitochondrial chromosome
- **Telomeres**: repetitive sequence protecting coding DNA
- **Centromere**: links sister chromatids (halves of the chromosome) during cell division

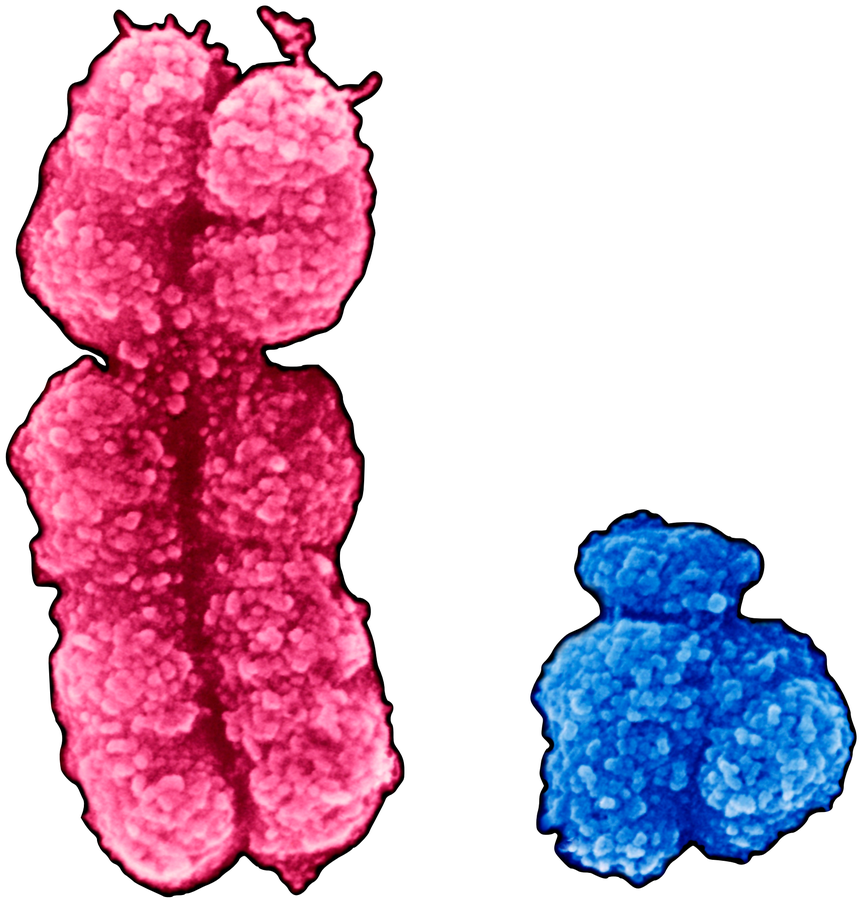



Telocentric
Acrocentric
Submetacentric
Metacentric


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
or
XY
XX
23

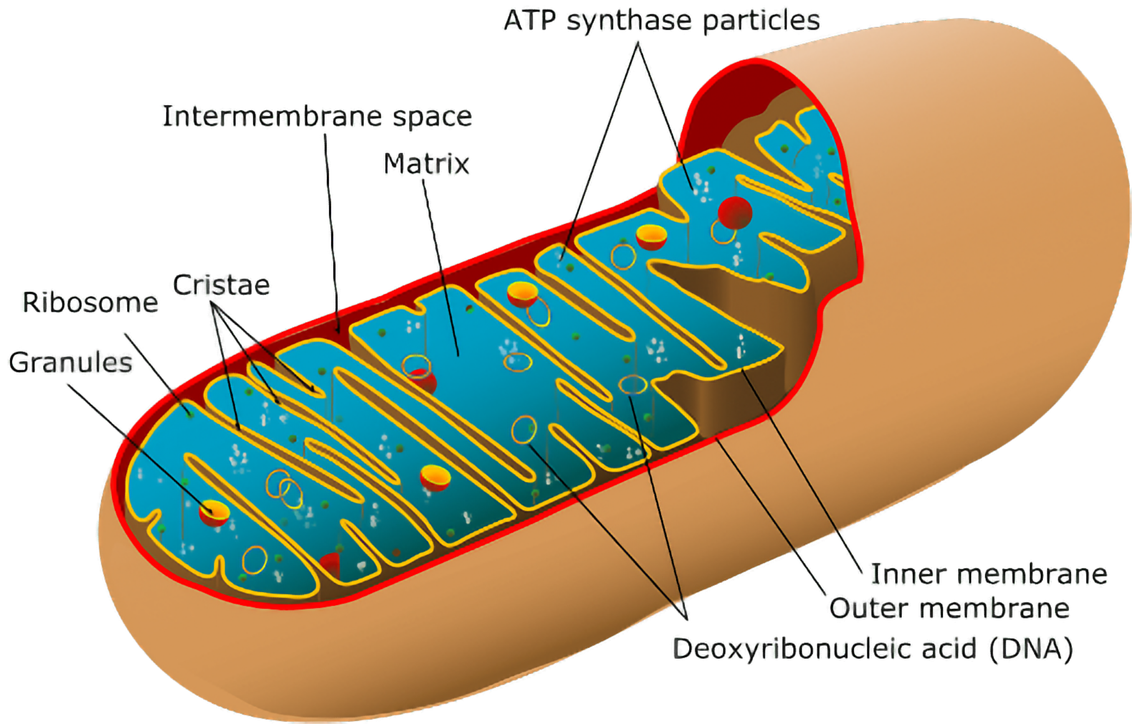
Gonosomes

- **Humans** and most other mammals have 2 sex chromosomes: **X** and **Y**, that determine the sex of an individual
- The recombination repair for Y is **missing**:
  - **Degeneration** of chromosome Y
- Issues with gene dose:
  - **Inactivation** of chromosome X



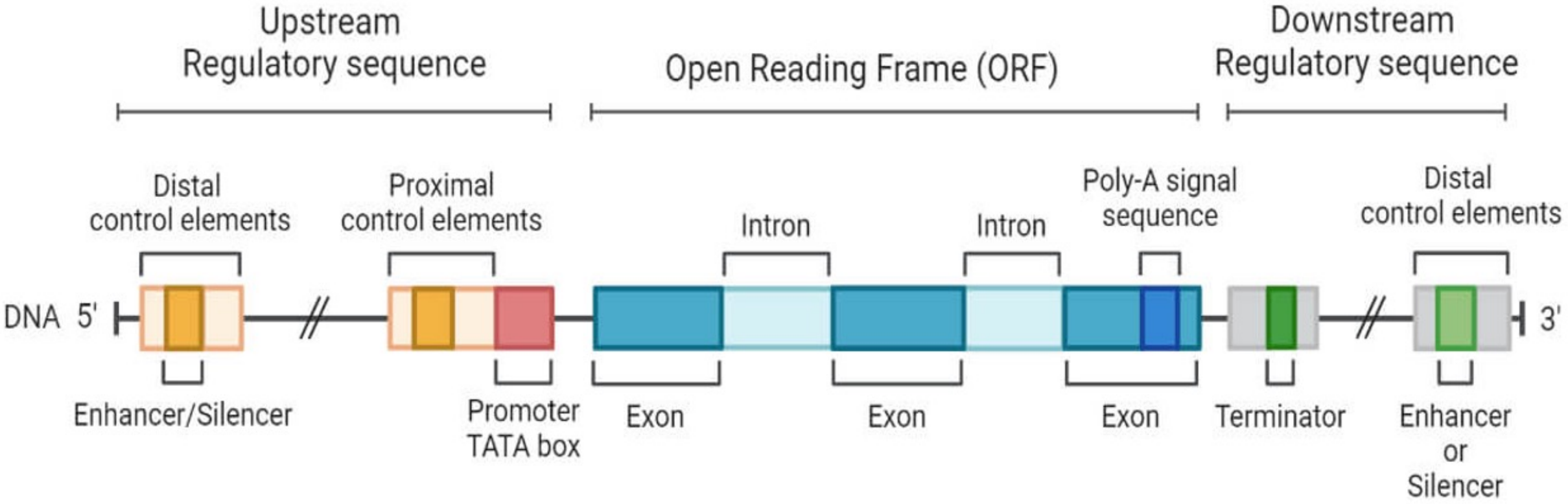
Mitochondrion

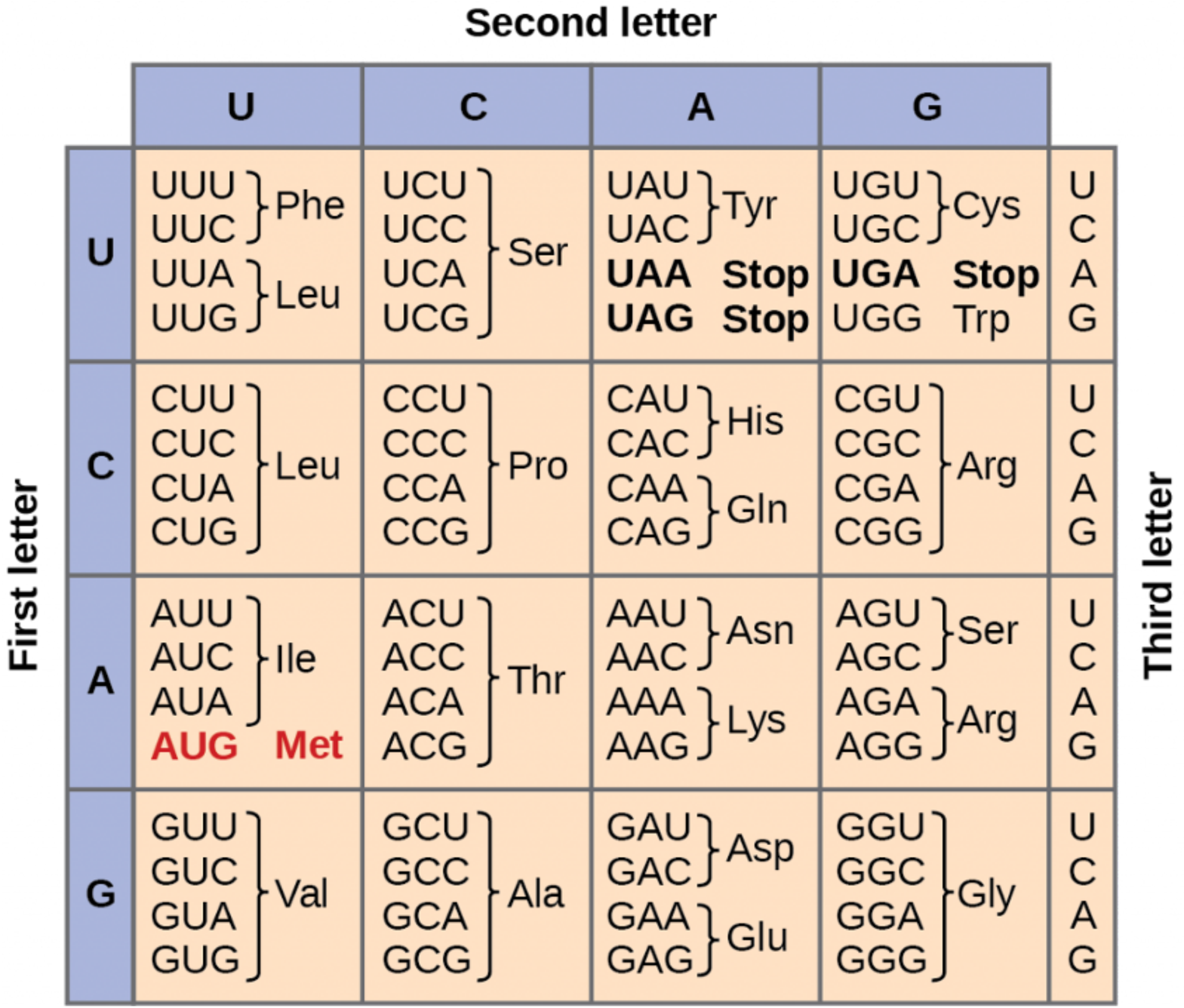
- Respiration center of the cell
- **Mitochondrial genome:**
  - Length: 16 569bp
  - Number of copies: 100 - 100 000
  - Extremely compact - 93% coding
- Inherited **only from mother**



Genes

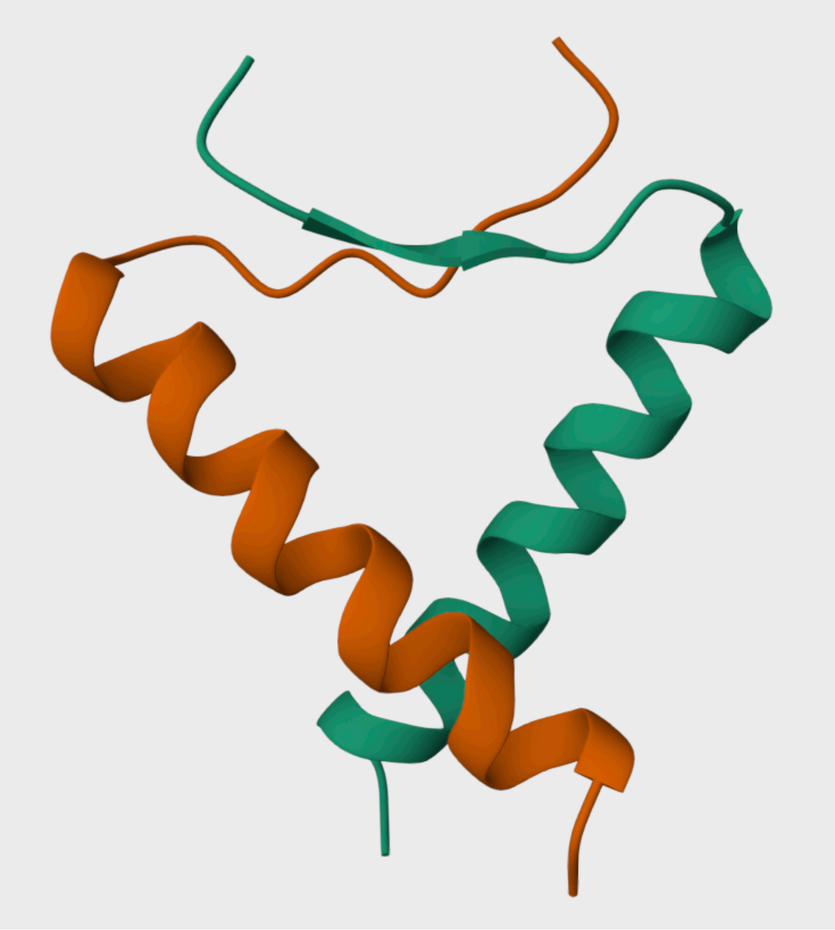
- Basic unit of heredity
- Segment of DNA containing instructions for building specific proteins
- Human genome consists of ~ 25 000 genes
- Gene regions:
  - **Exon** - coding sequence, only regions that are passed to mRNA
  - **Intron** - non-coding sequence, contains regulatory sequences
  - **Upstream & downstream regulatory sequences** - switch gene "on" & "off"


Upstream
Regulatory sequence
Open Reading Frame (ORF)
Downstream
Regulatory sequence
Distal
control elements
Proximal
control elements
Intron
Intron
Poly-A signal
sequence
Distal
control elements
DNA 5'
3'
Enhancer/Silencer
Promoter
TATA box
Exon
Exon
Exon
Terminator
Enhancer
or
Silencer


Second letter
First letter
Third letter
U
C
A
G
UUU Phe
UUC
UUA Leu
UUG
UCU Ser
UCC
UCA
UCG
UAU Tyr
UAC
UAA Stop
UAG Stop
UGU Cys
UGC
UGA Stop
UGG Trp
CUU Leu
CUC
CUA
CUG
CCU Pro
CCC
CCA
CCG
CAU His
CAC
CAA Gln
CAG
CGU Arg
CGC
CGA
CGG
AUU Ile
AUC
AUA
AUG Met
ACU Thr
ACC
ACA
ACG
AAU Asn
AAC
AAA Lys
AAG
AGU Ser
AGC
AGA Arg
AGG
GUU Val
GUC
GUA
GUG
GCU Ala
GCC
GCA
GCG
GAU Asp
GAC
GAA Glu
GAG
GGU Gly
GGC
GGA
GGG

Proteins

- Fundamental building blocks of life
- Large molecules of amino acids, whose three-dimensional structure is crucial for correct functionality
- Crucial for enzyme, hormone and antibody synthesis, function, growth, and repair
- Dietary protein is broken down into amino acids used to assemble required proteins or, if in excess, converted to energy



3 R’s of Gene Physiology

Regulation, Replication & Recombination

Regulation

- DNA contains recognition sites for molecular machinery
  - Binding sites, enhancers, silencers
- DNA regulates expression of RNA —> synthesis of proteins
  - **Transcription**: process during which **DNA is** used (**transcribed**) to synthesize RNA molecule
  - **Translation**: process during which **RNA is** used (**translated**) to assign a position of amino acid
- **Methylation**: Chemical modification of DNA that's silencing genes

Transcription

- Molecular machinery must bind to dsDNA
  - **Region**: Promoter
- Strands of dsDNA are split apart and RNA synthesis occurs
- Elongation and overall RNA synthesis proceeds until reaching a specific region
  - **Region**: Terminator
- Pre-mRNA is further modified and is transported from Nucleus to Cytoplasm
  - **Modification**: Addition of 3′ poly-A tail and 5′ cap to mature into mRNA

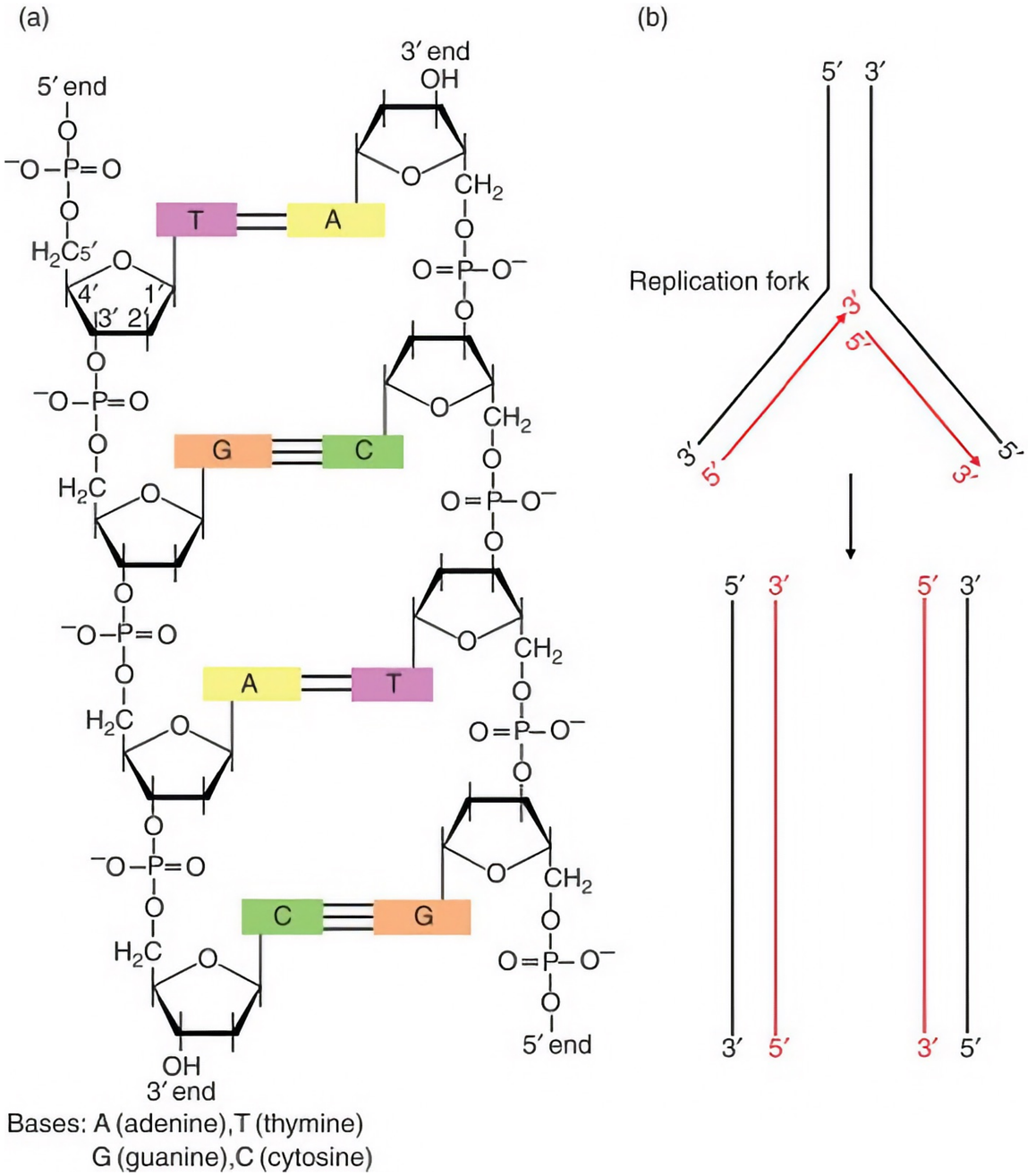
Translation

- mRNA left Nucleus and is now in cytoplasm, where it is found by Ribosome
- Molecular machinery searches for specific codon on mRNA - **START codon** - once found, **initiation** of Translation begins and first amino acid is assigned
- From this point on, amino acids are assigned to every codon, forming a long chain (Peptide chain) until **STOP codon** is reached
- After hitting STOP, the chain is released and begins maturing, eventually becoming a functional protein

Replication

- DNA replicates in every cell cycle with high precision and speed
- Multiple origins of replication spread across the genome
- Replication machinery splits dsDNA strands (Antiparallel):
  - **Coding**, + strand: 5′ —> 3′ (Sense strand)
  - **Template**, - strand: 3′ —> 5′ (Antisense strand)
- Both separated strands get their other half synthesized, creating another DNA molecule

Replication

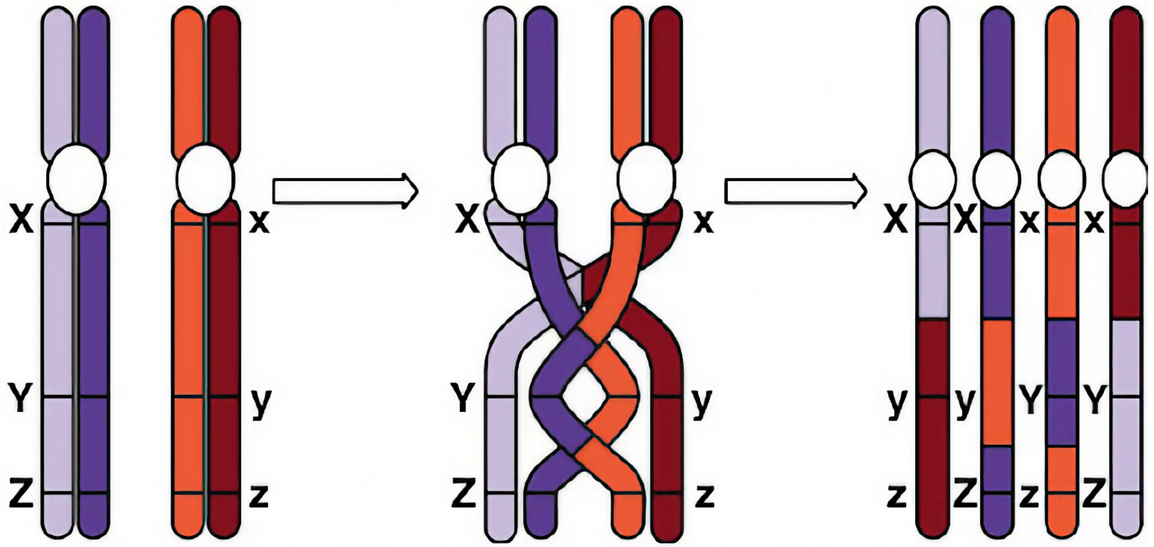


Telomeres

- Repetitive DNA sequences at the end of the chromosomes 5′ - TTAGGG - 3′
- Evolutionary patch to the linear chromosome problem
- They protect chromosome, but require special synthesis mechanism
  - This mechanism is typically reactivated in cancerous cells - *TERT* gene
- Shortening associated with worsening of health quality and aging

Recombination

- Exchange of genetic material
- Crossing-over: **Homologous Recombination**
  - Exchange of genetic information between 2 **non-sister chromatids**



DNA Repair

- Molecular machinery proofreads the DNA and repairs errors if necessary
  - Base Excision Repair, Nucleotide Excision Repair, Mismatch Mediated Repair
  - Double strand breaks are repaired by mechanisms like Homologous Recombination and Non-Homologous End-Joining (NHEJ)
- If the repair mechanisms are defective, the consequences are severe
  - Syndrome of spontaneous chromosomal instability
    - Cancer, immunodeficiency, neurodegeneration

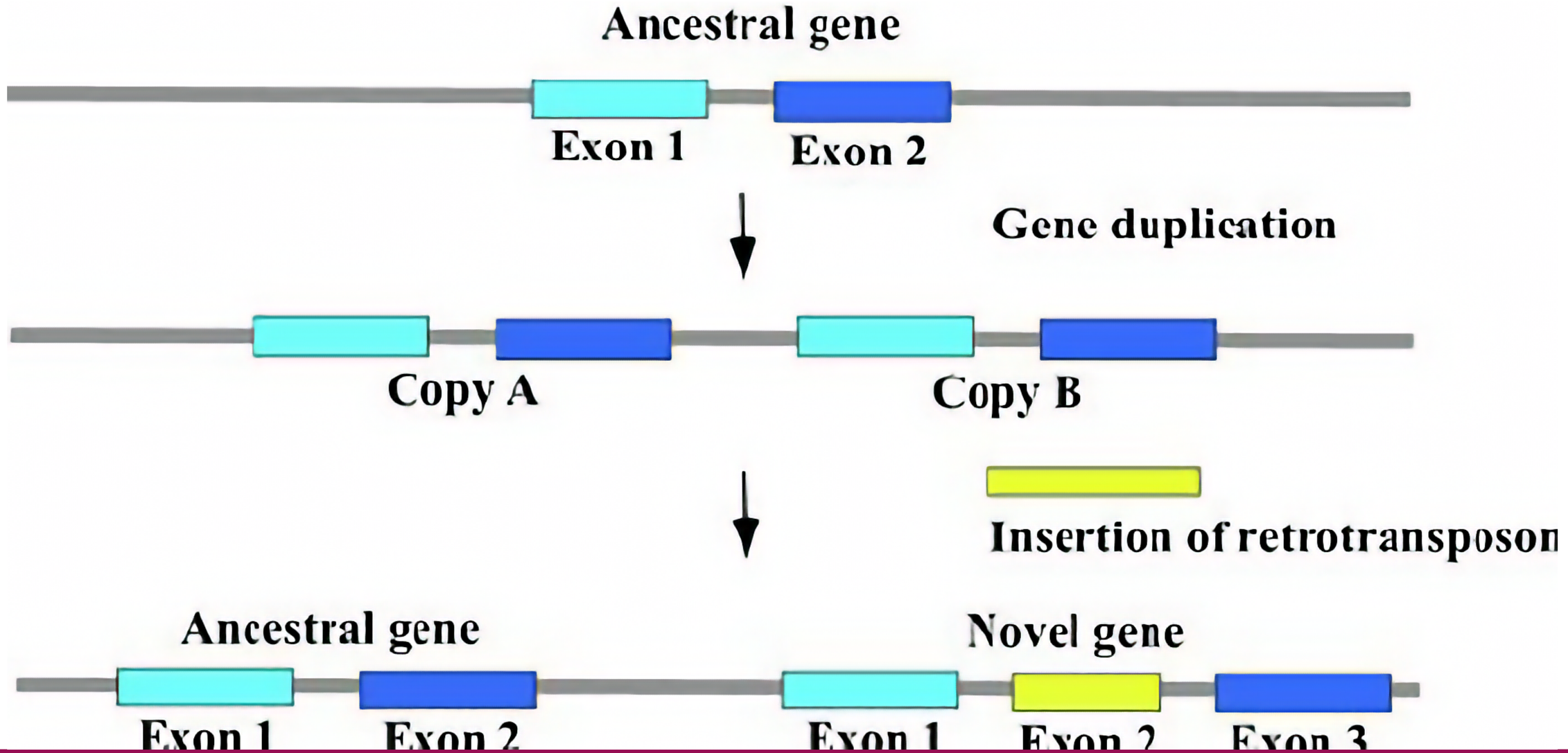
Architecture of Inheritance

Mutations & its types

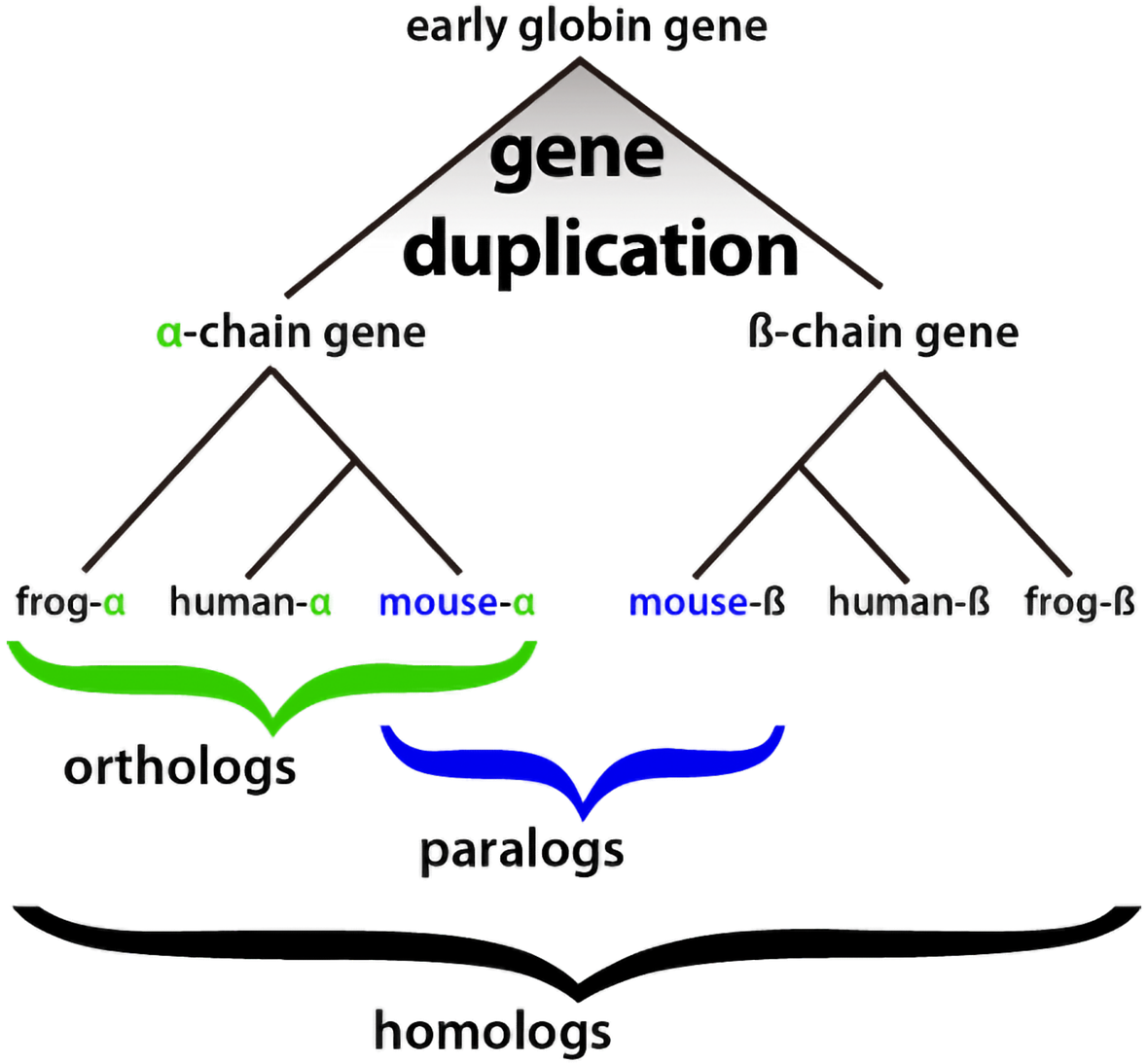
Organization of human genome

- Length of haploid human genome: 3200Mb
  - 3000Mb: euchromatin
  - 200Mb: heterochromatin
- High variability in the distribution of DNA content in individual chromosomes
- Types of sequences:
  - Unique: 55%
  - Repetitive: 45%

Gene duplication



Homology


early globin gene
gene duplication
α-chain gene
ß-chain gene
frog-α
human-α
mouse-α
mouse-ß
human-ß
frog-ß
orthologs
paralogs
homologs

Unique sequences

- **Genes coding proteins** ~ 25 000
  - 1.5% of human genome
  - High variability in number and size of exons, introns and distribution
- **Pseudogenes** ~ 20 000
  - Non-functional copies of genes
  - High sequence homology with active genes
  - Can be close to the original gene, but also on completely different chromosome

Repetitive sequences - Tandem repeats

- **Satellite DNA** (100kb+, repeat size 50-250bp)
  - Larger part of heterochromatin near centromere
- **Minisatellite DNA** (0,1-20kb)
  - Hypervariable: in proximity to telomeres
  - Telomeric
- **Microsatellite DNA** (Short Tandem Repeats)
  - Scattered across whole genome

Repetitive sequences - Interspersed repeats

- Blocks of repeats found across entire genome, but not in tandem arrangement
- Almost all of them are derived from DNA transposons
  - SINE (Short Interspaced Nuclear Elements): 280bp
    - Alu family - specific for primates
  - LINE (Long Interspaced Nuclear Elements): 6.1kb
  - LTR (Long Terminal Repeats)

Genetic diseases

- **Monogenic**: produced by a variant in single gene
- **Polygenic**: produced by joint contribution of multiple gene variants
- **Multifactorial**: result of multiple genetic and environmental factors interacting
- **Chromosomal**:
  - **Chromosomal variants**: Morphological variability
  - **Chromosomal aberrations**: Numerical or structural variability

Mutations

- Source of genetic variability
- Human cell gets 10s of thousands of new mutations **daily**, most of them are successfully repaired
  - Some of them could lead to cancer if not repaired
- Mutations are the driving force of evolution
- They are caused by physical, chemical and biological factors
  - UV light, Certain Chemicals (Ethidium Bromide), Viruses

Mutations - Substitutions

- A change of the base occurs
  - **Transition**: purine ↔ purine (A ↔ G); pyrimidine ↔ pyrimidine (C ↔ T)
  - **Transversion**: purine ↔ pyrimidine (A ↔ C) (G ↔ T)
- Impact depends on how the resulting strand was read
  - **Silent**: Substitution didn't change the amino acid
  - **Miss-sense**: Substitution did change the amino acid
  - **Non-sense**: Substitution created STOP codon

Mutations

- **Insertion**: DNA strand is interrupted by a different sequence
- **Deletion**: DNA strand loses 1 or more bp
- **Frame-Shift**: If the insertion/deletion wasn't a multiple of three, the reading frame shifts, and DNA is not read as intended
- **Duplications**: Segment of a gene is replicated
- **Inversions**: Part of the sequence breaks and flips 180°

Mutations

- **Expansions**: multiplication of a short sequence 10s to 100s of times
- **Fusion**: Two genes fuse together (NHEJ)
- **Gene Conversion**: Nonreciprocal transfer of sequence from donor to acceptor sequence
- **Loss-of-Function**
- **Gain-of-Function**

Mutations

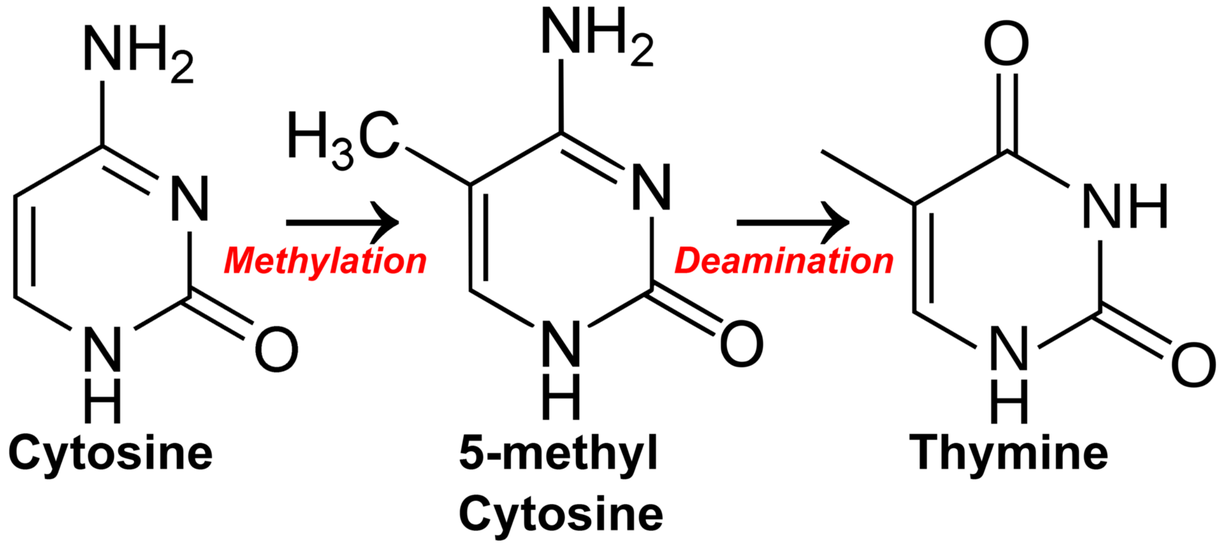
- **Driver Mutations**: Gives fitness advantage to cells that carry them
  - These can cause cells to become cancer cells
- **Passenger Mutations**: Neutral impact on phenotype
- Are mutations truly random or are they linked to the specific architecture of certain areas of genome, where they are more likely to occur?

Are mutations truly random?

- **Random Mutations**: Result of a unique mutation event
  - Population-genetic mechanisms
  - They usually originated a long time ago
  - Different frequency between populations
- **Mutations associated with architecture of human genome**:
  - Same type of DNA damage across all populations
  - Same frequency between populations

Mutation Hotspot - CpG

- CpG dinucleotides - mutation hotspot in all animals
  - Amount of SNP is 12x higher than in other dinucleotides
- The amount of CpG is noticeably repressed in human genome in comparison to other dinucleotides
  - Natural selection



Mutation Hotspot - Repetitive sequences

- **Hotspot of genomic instability**
  - Non-homologous crossing-over
  - Chromosomal breakage
  - Chromosomal aberrations
- **Mechanisms**:
  - Non Allelic Homologous Recombination
  - Non-Homologous End Joining
  - FoSTeS - Fork Stalling Template Switching

Chromosomal aberrations - Numerical

- **Polyploid**: Number of chromosomes is a multiple of the haploid (n) number
  - **Triploidy**: 69 chromosomes, Tetraploidy: 92 chromosomes...
- **Aneuploid**: Altered number of only some chromosomes
  - Monosomy: One of homologous chromosomes is missing
  - Trisomy: 3 instead of 2 homologous chromosomes
- **Uniparental diploidy**: Both sets of chromosomes from single parent
- **Uniparental disomy**: Both chromosomes of homologous pair from single parent

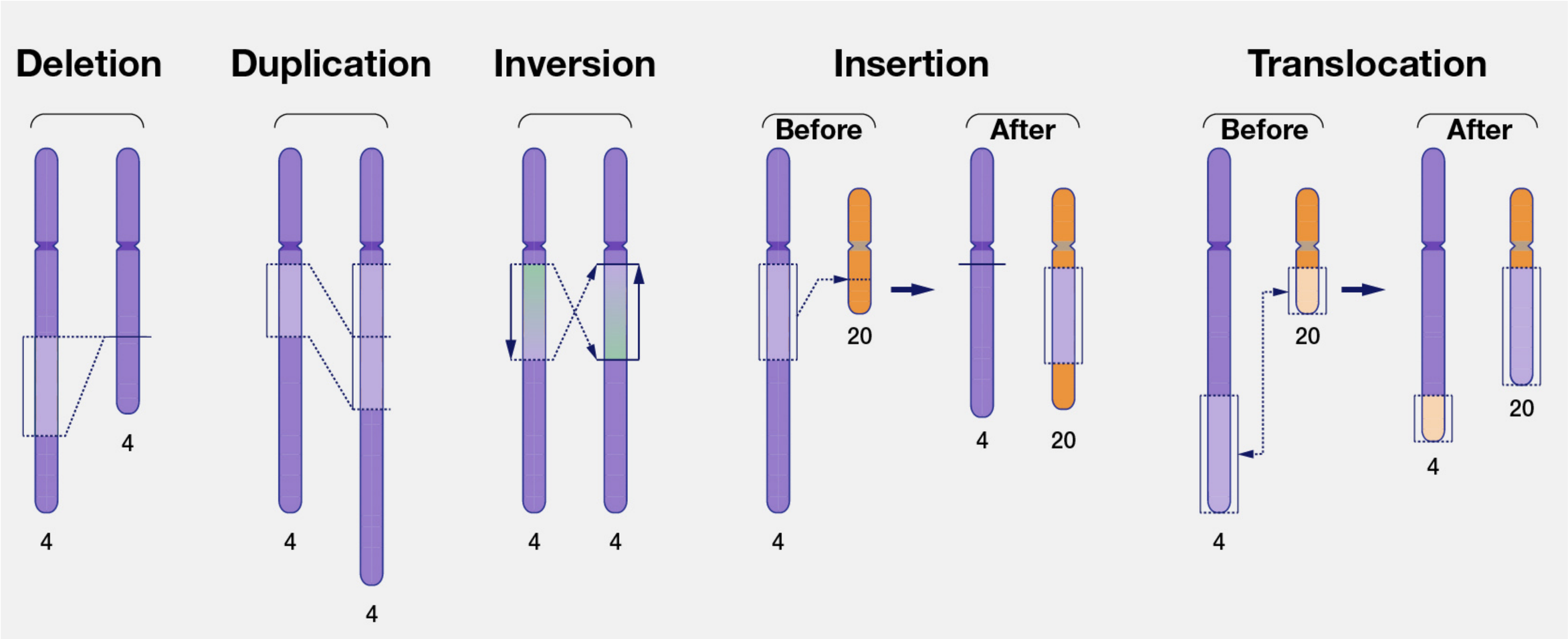
Chromosomal aberrations - Structural

- **Deletion**: loss of a region
  - Terminal: loss of the end region
  - Intersticial: Loss of the region between centromere and terminal region
- **Duplication**: Doubling of a region
- **Inversion**: Inversion of a region by 180°
  - Paracentric: Excludes centromere
  - Pericentric: Includes centromere
- **Ring Chromosome**: Both ends of the chromosome break off and the ends fuse

Chromosomal aberrations - Structural

- **Translocation**: transfer of certain region
  - **Intrachromosomal**: Transfer of a region within chromosome
  - **Reciprocal**: Transfer of regions between homologous or non-homologous chromosomes
  - **Centric fusion**: Fusion of two acrocentric chromosomes
- Based on the effect on phenotype:
  - **Balanced aberrations**: Little to no effect on phenotype
  - **Unbalanced aberrations**: Associated with severe conditions

Chromosomal aberrations - structural



Dictionary

- **SNP** (Single Nucleotide Polymorphism "Snip"): Substitution, Insertion or Deletion of a single nucleotide at specific position in the genome
- **Indel**: Insertion/Deletion of bases
- **CNV** (Copy Number Variation): Number of copies of a specific DNA segment. This varies among different individual genomes
- **SV** (Structural Variant): Differences between genomes involving tens of thousands of nucleotides, therefore larger segments of DNA
- **MSI** (Microsatellite Instability): Abnormal length of microsatellite repeats, due to defects in Mismatch repair

Questions?