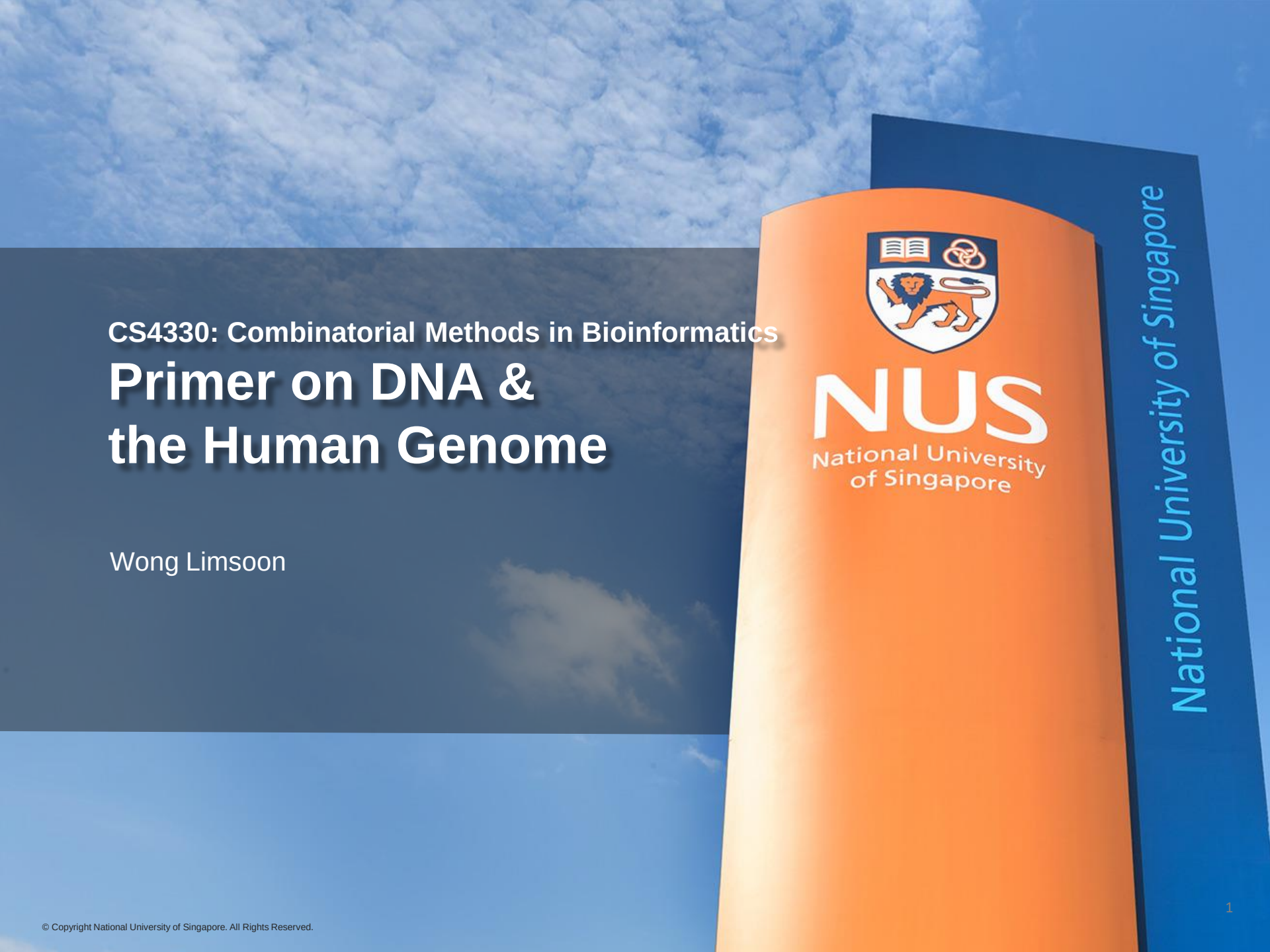


CS4330: Combinatorial Methods in Bioinformatics

Primer on DNA & the Human Genome

Wong Limsoon

Introduction to DNA

DNA (Deoxyribonucleic Acid)

- DNA, short for Deoxyribonucleic Acid, is a molecule found in all living organisms, holding the genetic instructions for development, functioning, growth, and reproduction.
- It serves as the fundamental building block of life, carrying the genetic information that determines the characteristics of an organism.
- Composed of long chains of nucleotides, DNA is responsible for passing hereditary traits from one generation to the next.

Importance of DNA in living organisms

- DNA carries the genetic code that determines the traits and characteristics of an organism, such as eye color, hair texture, susceptibility to diseases, etc.
- It plays a pivotal role in cellular processes, including protein synthesis, cell division, and the overall functioning of an organism's cells and systems.
- Understanding DNA helps scientists comprehend the underlying mechanisms of inheritance, evolution, and the development of various life forms on Earth.

Role as the genetic blueprint

- DNA acts as a comprehensive instruction manual or blueprint for the development, growth, and functioning of organisms.
- It contains the information necessary for the construction and regulation of proteins, which are essential for the structure and function of cells.

DNA structure

Double helix structure

Two long strands that coil around each other like a twisted ladder

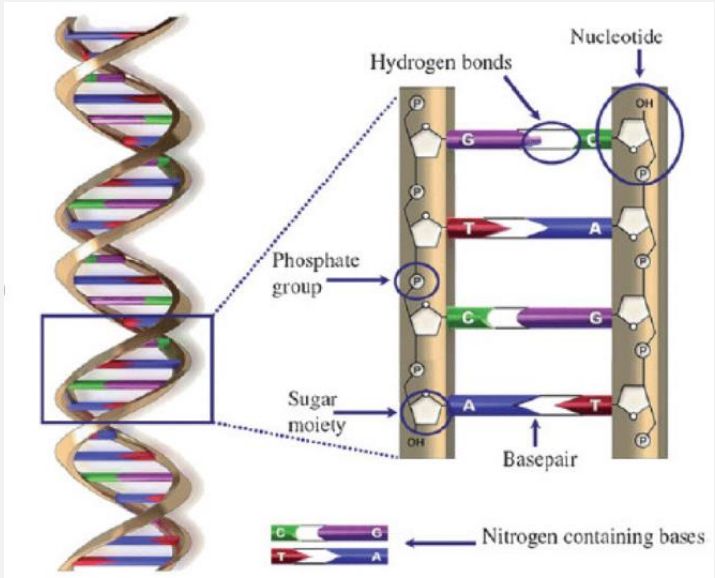


Image credit: Georgette Salieb-Beugelaar

Nucleotide components

Phosphate group – the "backbone"

Deoxyribose sugar – connects phosphate groups

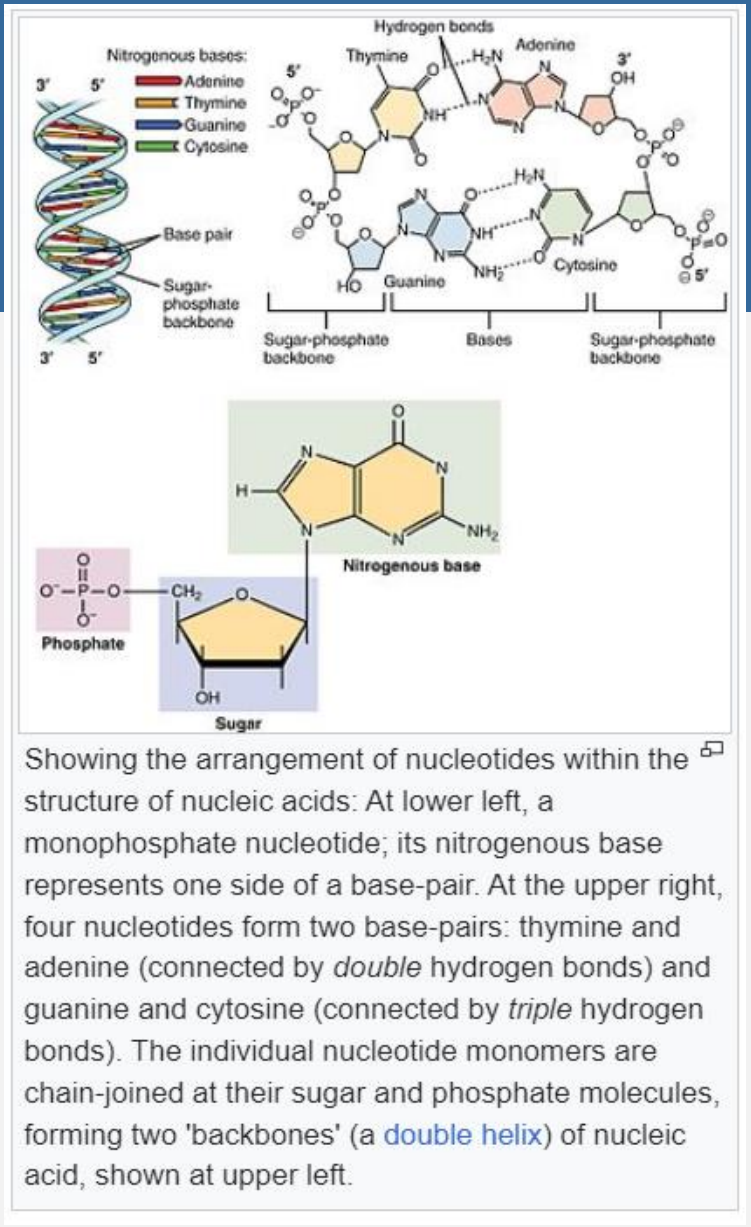
Nitrogenous bases – connect two DNA strands

DNA bases & pairing

Adenine (A) pairs with **Thymine (T)**, held by 2 hydrogen bonds

Cytosine (C) pairs with **Guanine (G)**, held by 3 hydrogen bonds

"Crick-Watson rule"



Showing the arrangement of nucleotides within the structure of nucleic acids: At lower left, a monophosphate nucleotide; its nitrogenous base represents one side of a base-pair. At the upper right, four nucleotides form two base-pairs: thymine and adenine (connected by *double* hydrogen bonds) and guanine and cytosine (connected by *triple* hydrogen bonds). The individual nucleotide monomers are chain-joined at their sugar and phosphate molecules, forming two 'backbones' (a double helix) of nucleic acid, shown at upper left.

Image credit: Wikipedia

DNA replication

DNA replication is the process by which a cell makes a copy of its DNA before cell division

It is fundamental for passing on genetic traits to offsprings

Steps: Initiation, elongation, termination

1. **Initiation**: The process starts at specific points on the DNA molecule called origins of replication. Enzymes unwind and separate the double helix, creating replication forks.
2. **Elongation**: DNA polymerase, along with other enzymes, adds complementary nucleotides to each of the separated DNA strands. The enzyme "reads" the existing DNA strand and synthesizes a new strand following base-pairing rules (A-T, C-G).
3. **Termination**: Once the entire DNA molecule is replicated, termination occurs, and two identical DNA molecules are formed.

Enzymes: DNA polymerase, helicase, ligase, etc.

- DNA polymerase is the primary enzyme responsible for adding new nucleotides to the growing DNA strand.
- Other enzymes, such as helicase, topoisomerase, and ligase, aid in unwinding, stabilizing, and joining DNA fragments during replication.

Central dogma of molecular biology

DNA is transcribed to RNA

RNA is translated to proteins

- Proteins are essential molecules that perform various functions in cells, including structural support, enzymatic reactions, and cellular signaling.
- The genetic code and protein synthesis are crucial for determining an organism's traits and functionalities.

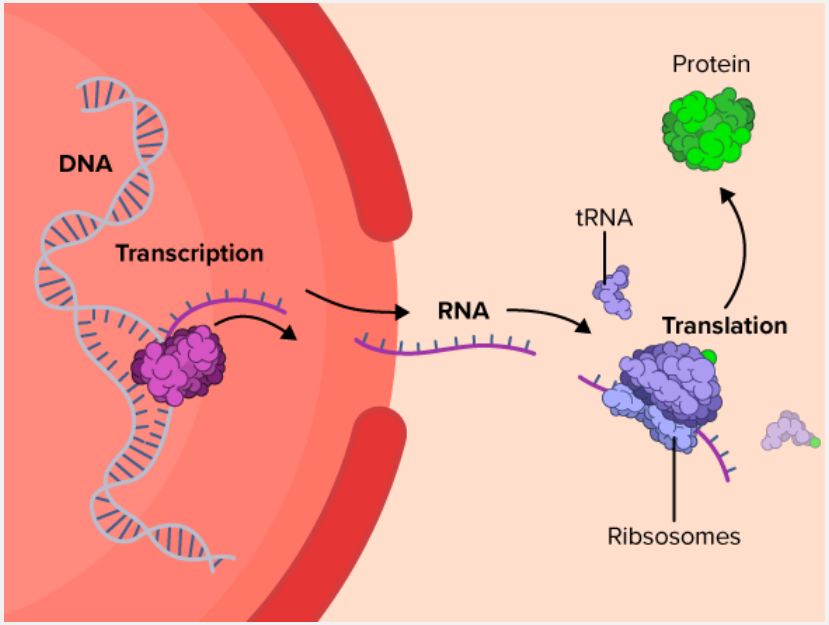


Image credit: cK-12

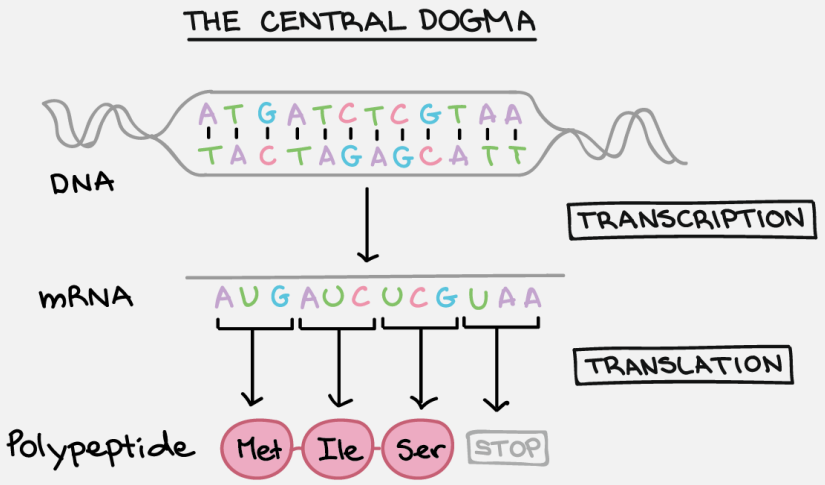


Image credit: Khan Academy

Transcription

Let’s watch this video together

https://www.youtube.com/watch?v=WsofH466lqk

And watch this video yourself

https://youtu.be/JQIwwJqF5D0

Translation

Let’s watch this video together

https://www.youtube.com/watch?v=5bLEDd-PSTQ

And watch this video yourself

https://youtu.be/ocAAkB32Hqs

Genetic code

A triplet code of nucleotides (codons) to encode specific amino acids

Start codon

ATG

Stop codons

TAA, TAG, TGA

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

DNA mutations

Mutations are changes in DNA seq that can alter the genetic info carried by an organism

- **Causes**: Mutations can be caused by exposure to radiation, certain chemicals, errors during DNA replication, or spontaneous changes.
- **Consequences**: Mutations can have varying effects, from no noticeable impact to severe alterations in protein structure or function. Some mutations can lead to genetic disorders or increased susceptibility to diseases.

Point mutations

- **Substitution**: One base pair is replaced by another (e.g., A to T, C to G).
- **Insertion**: An extra nucleotide is inserted into the DNA sequence.
- **Deletion**: A nucleotide is removed from the DNA sequence.

Chromosomal mutations

- **Deletion**: A part of a chromosome is missing.
- **Duplication**: A section of a chromosome is repeated.
- **Inversion**: A segment of a chromosome is reversed.
- **Translocation**: A segment of a chromosome breaks off and attaches to another chromosome.

Applications of DNA science

Forensics

- **DNA Fingerprinting**: It's a technique used to identify individuals by analyzing unique patterns in their DNA. It's crucial in forensic investigations to match suspects with biological evidence from crime scenes.
- **Criminal Justice**: DNA evidence plays a pivotal role in solving crimes, exonerating innocent individuals, and providing accurate identification in legal cases.

Medicine

- **Personalized Medicine**: DNA analysis allows tailoring medical treatments to individual genetic profiles, enabling more effective and precise therapies.
- **Genetic Testing**: It helps in identifying genetic predispositions to diseases, allowing for preventive measures or early interventions.

Agriculture

- **Genetically Modified Organisms (GMOs)**: DNA science enables the modification of plant and animal genomes to enhance traits like resistance to pests, improved nutritional content, or increased yield.

Biotech

- **Biopharmaceuticals**: DNA technologies are used in producing therapeutic proteins and vaccines, contributing to advancements in pharmaceuticals and biomedicine.
- **Bioremediation and Biofuels**: DNA science helps in engineering microorganisms for cleaning up pollutants and producing renewable energy sources like biofuels.

Some characteristics of human genome

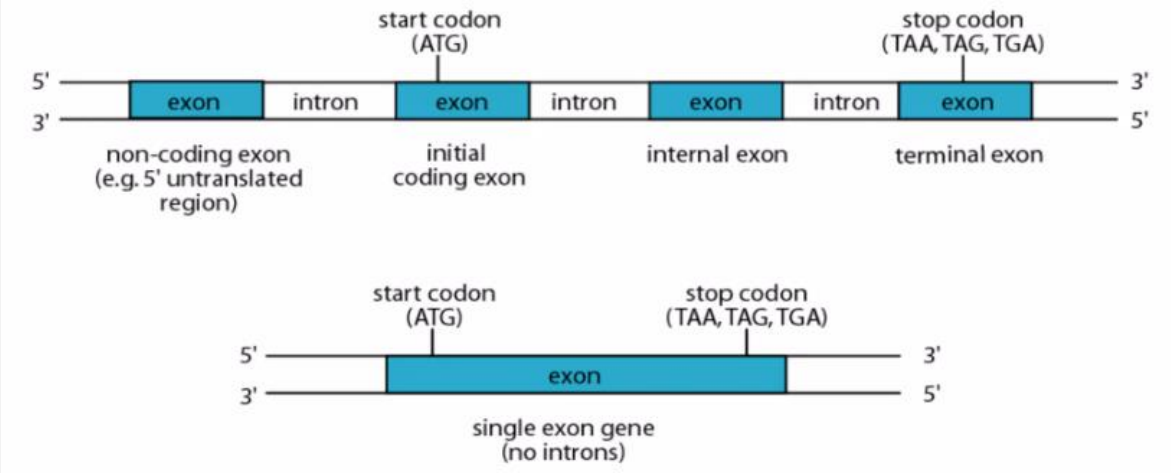
22 autosome pairs + 2 sex chromosomes

3 billion base pairs in the haploid genome

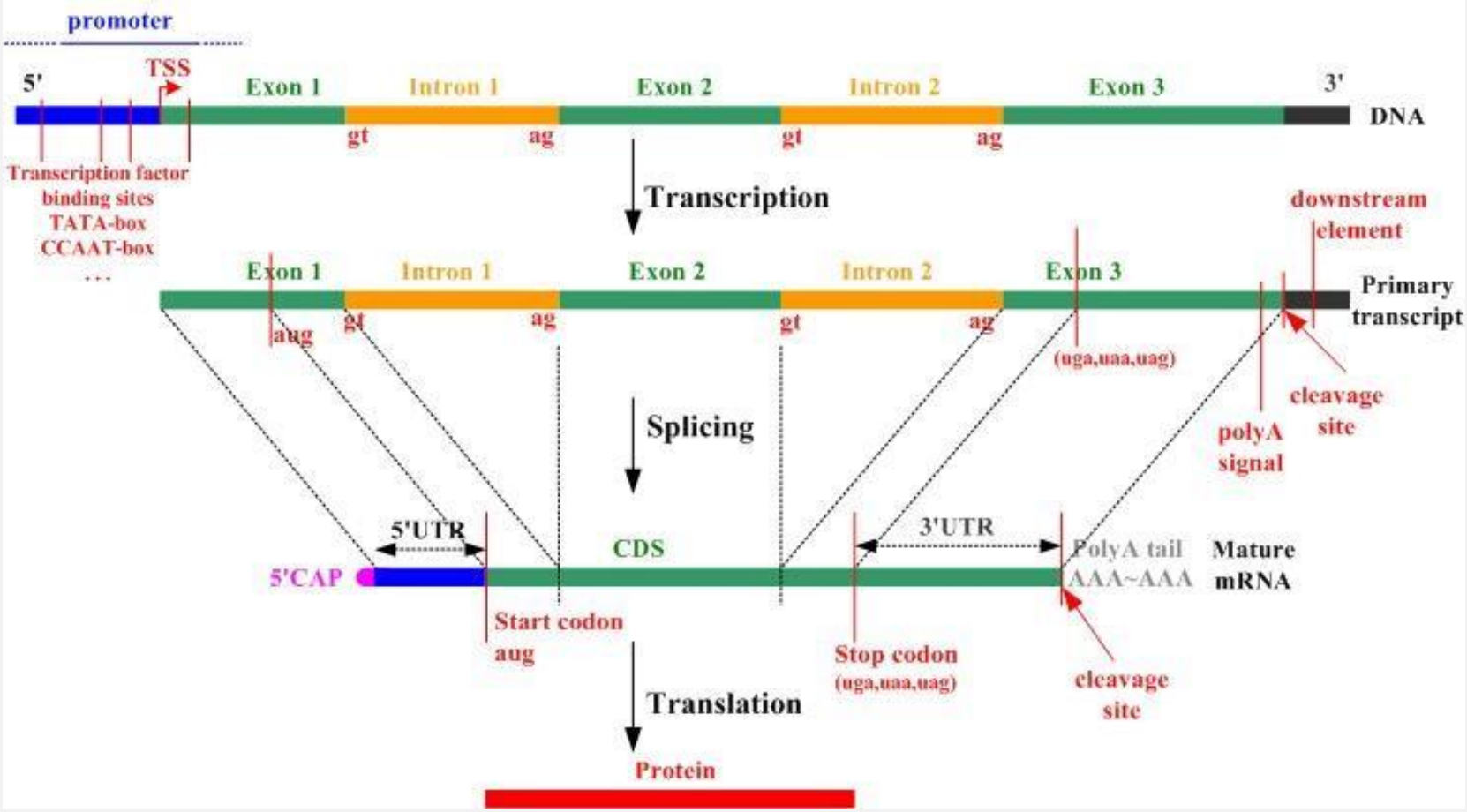
50% is repetitive regions

3% codes for proteins

Structure of protein-coding genes



Gene expression



Repeats in human genome

Interspersed repeats, ~45% of human genome

SINEs; e.g. ALU ~300bp @ ~1 million copies

LINEs; e.g. LINE-1 ~6000bp @ ~500k copies

Processed pseudo genes

They arise following gene duplication & loss

Simple sequence repeats

Microsatellites, ~10bp; e.g., $(A)_n$, $(CA)_n$, $(CGG)_n$

Minisatellites, ~500bp

Repeats in human genome

Segmental duplicates, ~5% of human genome

~1 kb to 300 kb copied intra- or inter-chromosomally

Duplicated regions share high sequence identity, ~99%

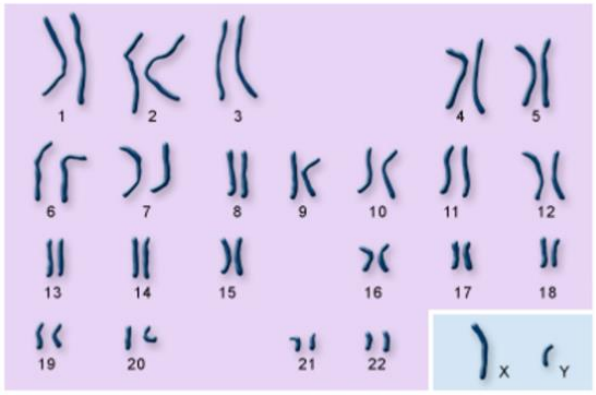
Tandem repeats

Repetitive DNA, sometimes spanning millions of bases

```
>gi|7407196|gb|AF236885.1|AF236885 Homo sapiens clone p10 chromosome 6,
telomeric repeat region
GGATCCCCCCCCAACTCATGACTGTCGGGCTATTTCCAGGCCGCATCGACAGTGAACAAAATCCTTTCTGT
TTGCAGCCCTGAATAATCAGGGTTAGGGTTAGGGTTAGGGGTTAGGGGTTGGGGTTGGGGTTAGGGTTAG
GGTTGGGTTAGGGTTAGGGTTAGGGGTTAGGGTTTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTAGGGTC
AGGGTCAGGGTCAGGGTCAGGGTCAGGGTTAAGGGTCAGGGTCAGGGTCAGGGTCAGGGTCAGGGGTTAG
GGTTAGGGTTAGGGTCAGGGTCAGGGTCAGGGTCAGGGTCAGGGGTTAGGGTCGGGGTCGGGGTCGGGGT
CAGGGGTCAAGGGTCAAGGGTCGGGGTCAGGGGTCAAGGGTCGGGGTCGGGGTCCGGGTCAGGGTGAGGG
TGAGGGTGAGGGTGAGGGTGGGGG
```

Chromosomes on human genome

The 22 autosomes are numbered by size. The other two chromosomes, X and Y, are the sex chromosomes. This picture of the human chromosomes lined up in pairs is called a karyotype.



U.S. National Library of Medicine

Autosomes

22 pairs

Half of each pair is from father, half from mother

Sex chromosomes

Men have XY

Women have XX

Cell division & genetic recombination

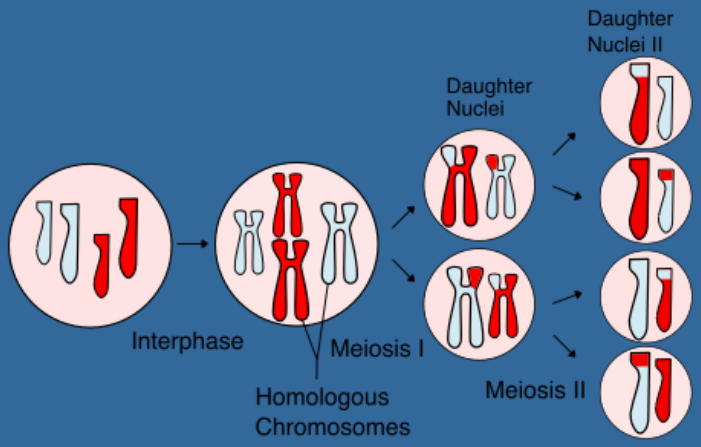


Image credit: Wikipedia

Let's watch this video together

https://youtu.be/BlnUNmfGn7I

At the end, one gamete from each chromosome goes into a different sperm or egg, forming haploid genomes

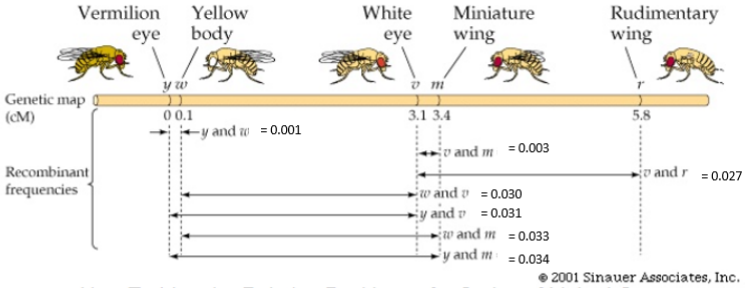
When a sperm fertilizes an egg, their haploid genomes combine to form a diploid genome

Genetic linkage

Let's watch this video together

https://youtu.be/ZeATszO-6e0

The closer two genetic loci are, the more likely they are passed to next generation together



How To Map the Relative Positions of a Series of Linked Genes:

- Generate a series of test crosses

Between all possible pair-wise combinations of available genetic markers

- Determine Recombination Frequencies
- % Recombination X100 = Map Distance in cM

e.g., for *y* & *w* 0.1% (.001) recombination **(NOT 0.010)** X 100 = 0.1cM

Exercise

~99.9% of the genome is identical between two persons

Is the similarity between the genomes of you and your father closer to 99.9% or 99.95%?

Does this matter whether you are a son or a daughter?

Does this matter whether you are compared to your father or mother?

These resources may help:
https://isogg.org/wiki/Autosomal_DNA_statistics
https://www.pfizer.com/news/articles/what%E2%80%99s-y-chromosome-handed-down-father-son

Good to read

S.K. Ng, “Molecular biology for the practical bioinformatician”, Chapter 1, *The Practical Bioinformatician*, WSPC 2004
https://www.comp.nus.edu.sg/~wongls/psZ/practical-bioinformatician/ch1-skintro/ch1-skintro.pdf