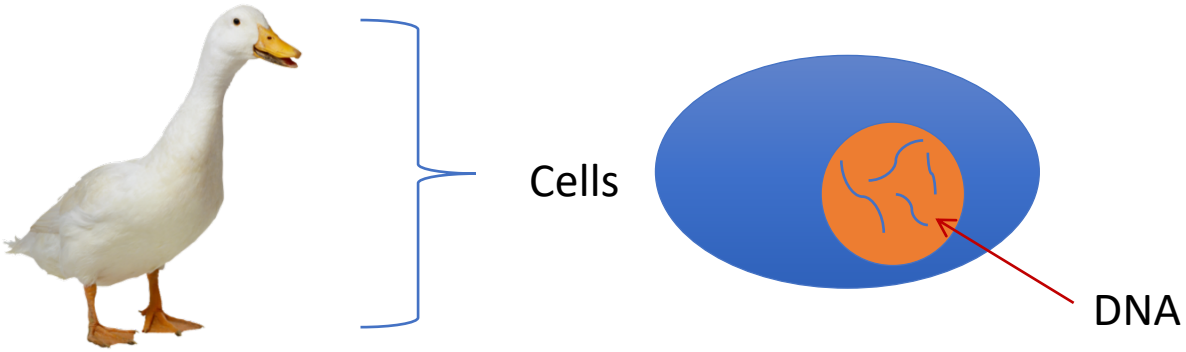
COMP90016 – Computational Genomics
Genomic I

Department of Computing and Information Systems

The University of Melbourne

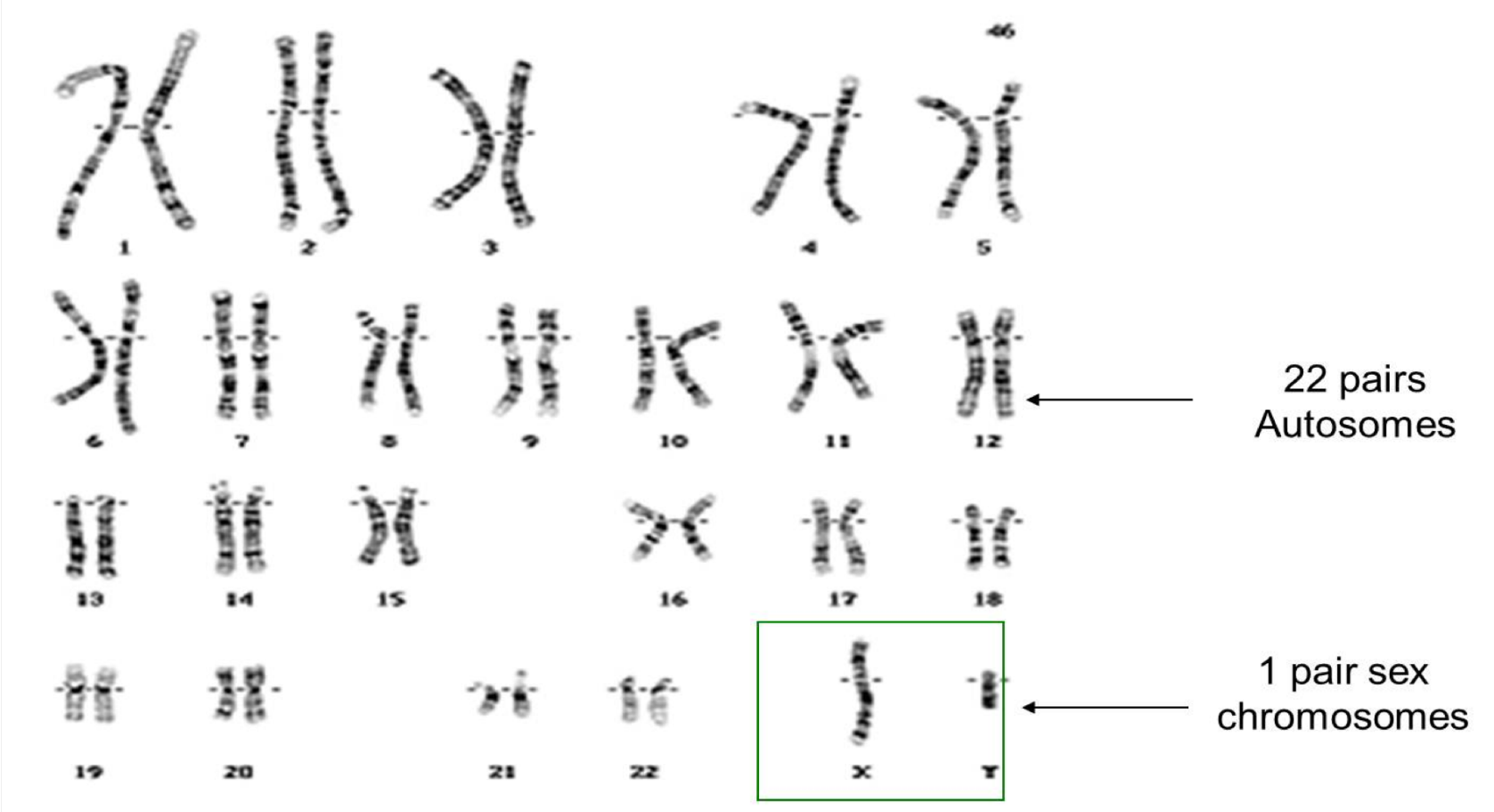
The Genome



- “An adult human being has approximately 100,000 billion cells, all originating from a single cell, the fertilized egg cell.”*
- The collection of DNA molecules in each cell is (mostly) identical.
- The collection of DNA in cells is called the organism’s genome.

*Nobel prize website

Organisation of DNA in Human Cells

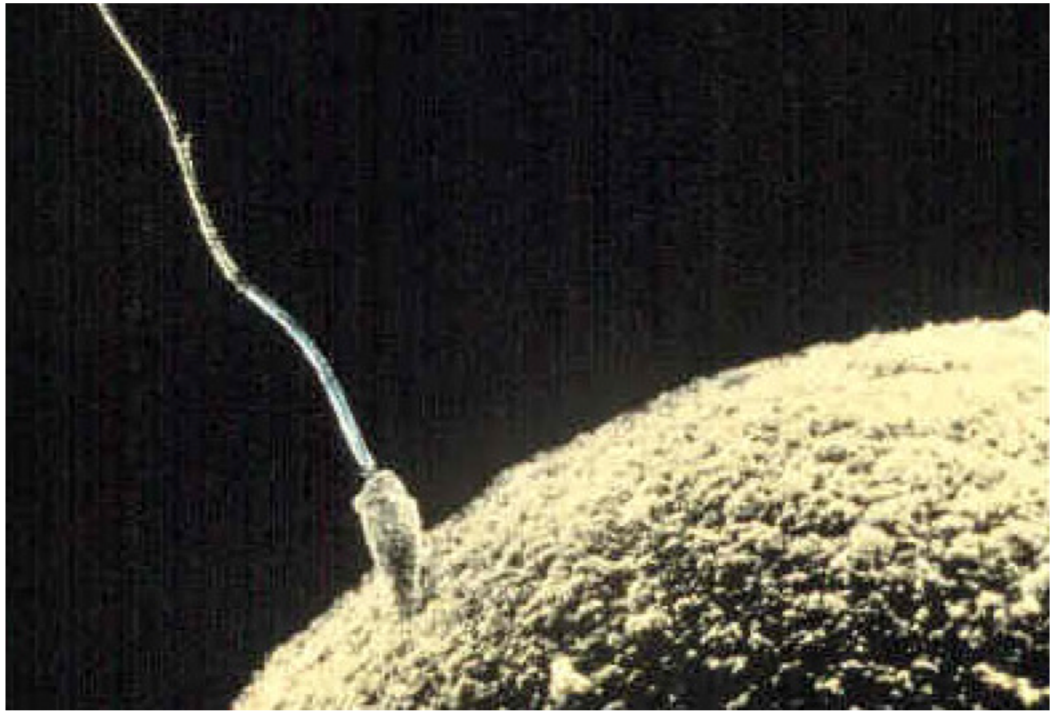


Note that not every organism organizes DNA like this – multiple diploid chromosomes. Some have only one molecule per cell (like viruses or bacteria). Some have more than two copies of each chromosome (seedless watermelon is triploid - it has three copies of each chromosome)

Image from dna-explained.com

Fertilization

- Upon fertilization, the sperm and egg fuse to initiate embryogenesis (the development of the embryo)
- A subset of the available DNA gets used of make up the genome of the offspring:
  - In humans, of the the 22*2 copies of the mother, 22 get selected. Another 22 out of 44 for the father.
  - This process is random.
  - There are $(2^{22})^2$ = 17.6×10^{12} combinations!



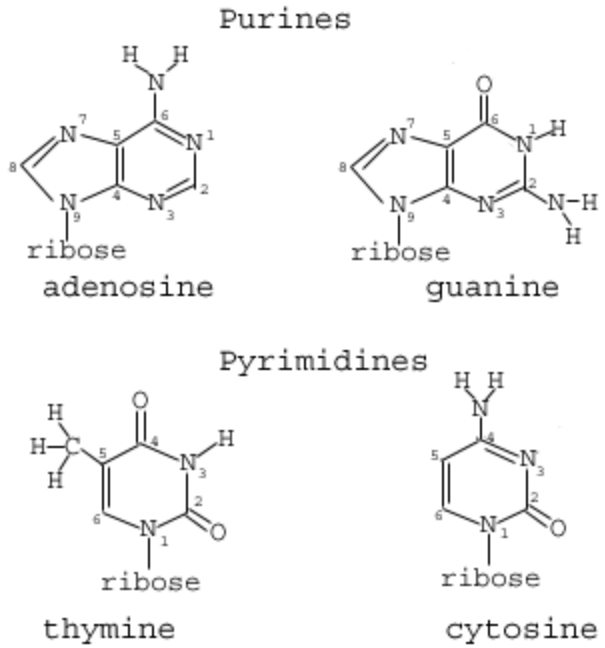
A sperm fusing with an egg – image from Wikipedia

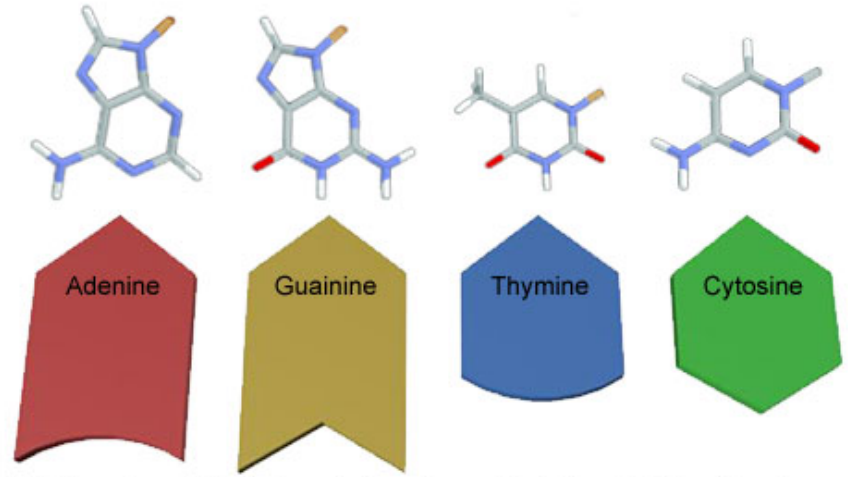
DNA

- DNA stands for *deoxyribonucleic acid*
- DNA is a macromolecule:
  - A big molecule
  - A linear arrangement of repeating units
  - Units are very similar, but the differences are important.
- Represent units by symbols (characters):
  - Basic alphabet of 4 symbols {TCAG}
  - Typical sequence: ACTGGTCAA…

DNA Bases: TCAG

- DNA primary structure (sequence)
  - 4 bases (nucleotides/nucleosides)





DNA Double Helix and Bases

- Salient features:
  - Size:
    - Two small bases (pyrimidines)
      T and C
    - Two larger bases (purines)
      A and G
  - Bases “fit” together:
    - A-T (weak, 2 hydrogen bonds)
    - G-C (stronger, 3 hydrogen bonds)
  - Therefore: Bases on one strand fully determine those on the other strand.

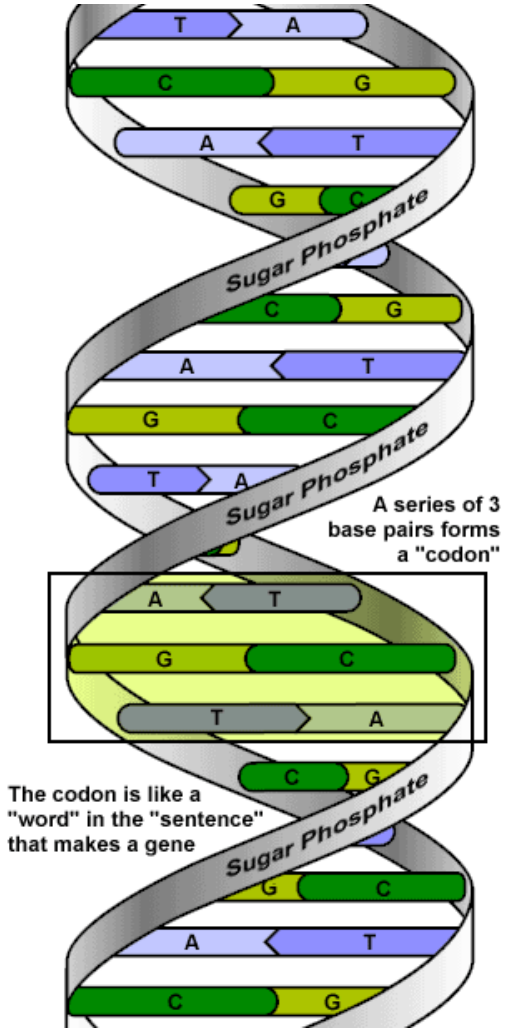


Image from David Chynoweth, Agricultural and Biological Engineering, University of Florida

Nomenclature

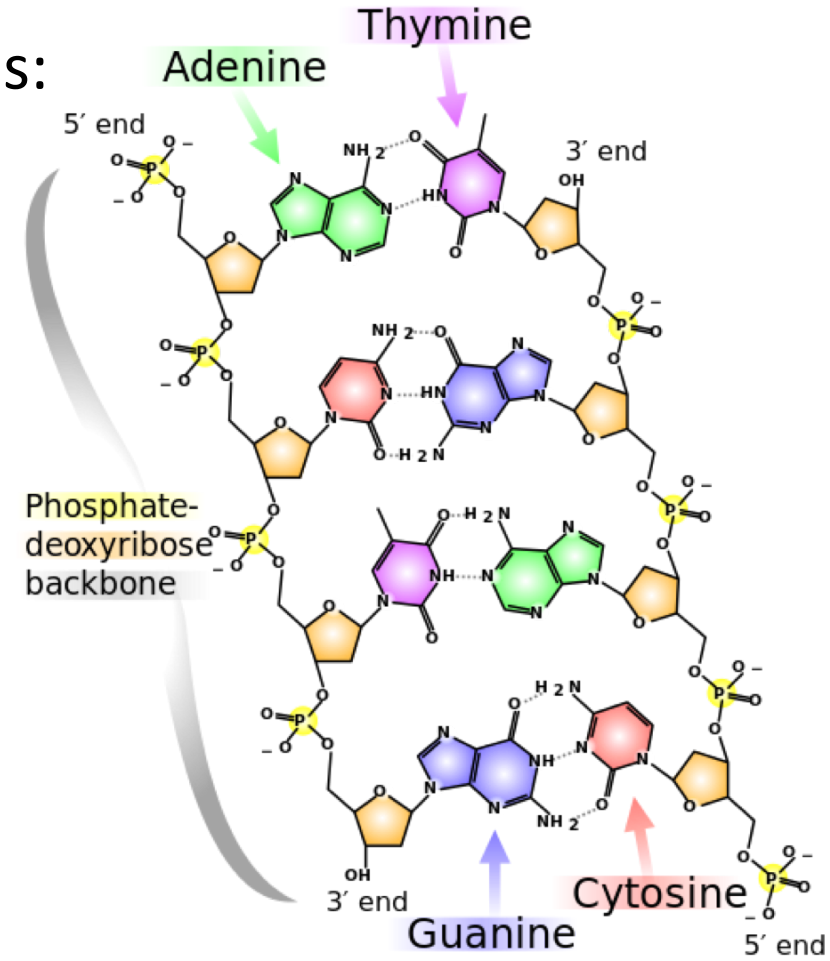
- The alphabet symbols T,C,A,G are connected with a backbone to form a DNA strand.
- The backbone has no information (but a direction).
- In this subject, we are not concerned with whether the symbol is connected to part or all of the backbone unit, so we can use the terms *base, nucleoside,* and *nucleotide* interchangeably.

Strands

- The two sides of the double helix are called strands.
- The strands of the DNA are referred to as…
  - *positive/negative*
  - *+/-*
  - *forward/reverse.*
  - Also: *sense/anti-sense* (but this is only meaningful in the context of *genes* – later lectures)
- There is usually no biological significance as to which one is + and which one -.
- However, it is important to note that strands have a distinct *reading direction,* which is dictated by the chemical properties of the helix. Strands can only be read in the so-called 5’ to 3’ direction.
  The direction are opposite on the two strands (hence the forward/reverse nomenclature)
- Since one strand defines the other, they are referred to as *complementary.*

DNA Example and Reverse Complement

- A (single-stranded) sequence of DNA could look like this:
  - `AAAAAAAAAAAA`
    Reading direction: left to right.
- Then, the opposing (complementary) strand would be defined as:
  - `TTTTTTTTTTTT`
    Reading direction: right to left
  - Each A pairing up with a T as prescribed.
- Another example:
  - `ACCGTAG` (read as `ACCGTAG`)
    `TGGCATC` (read as `CTACGGT`)
  - `ACCGTAG` is the reverse complement of `CTACGGT`

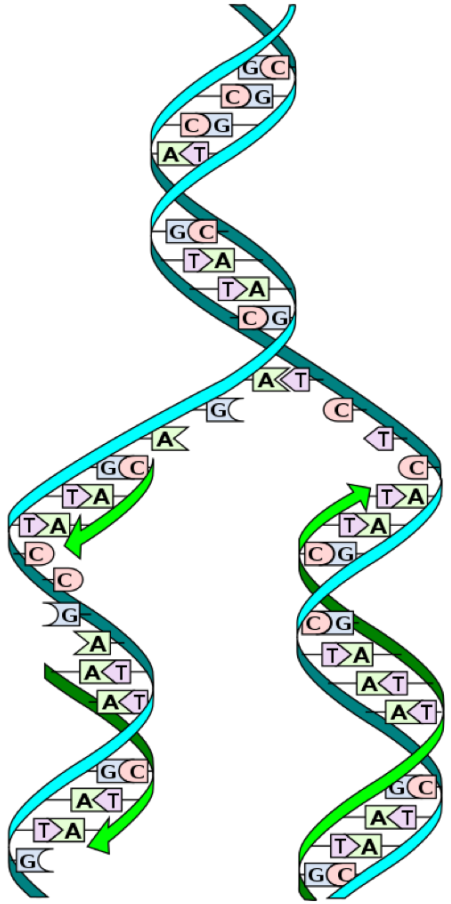


Double Helix

- We have learned that DNA is arranged as a double helix.
- Seemingly, there are two completely redundant copies of the same information.
- What advantage or purpose might this redundancy offer?

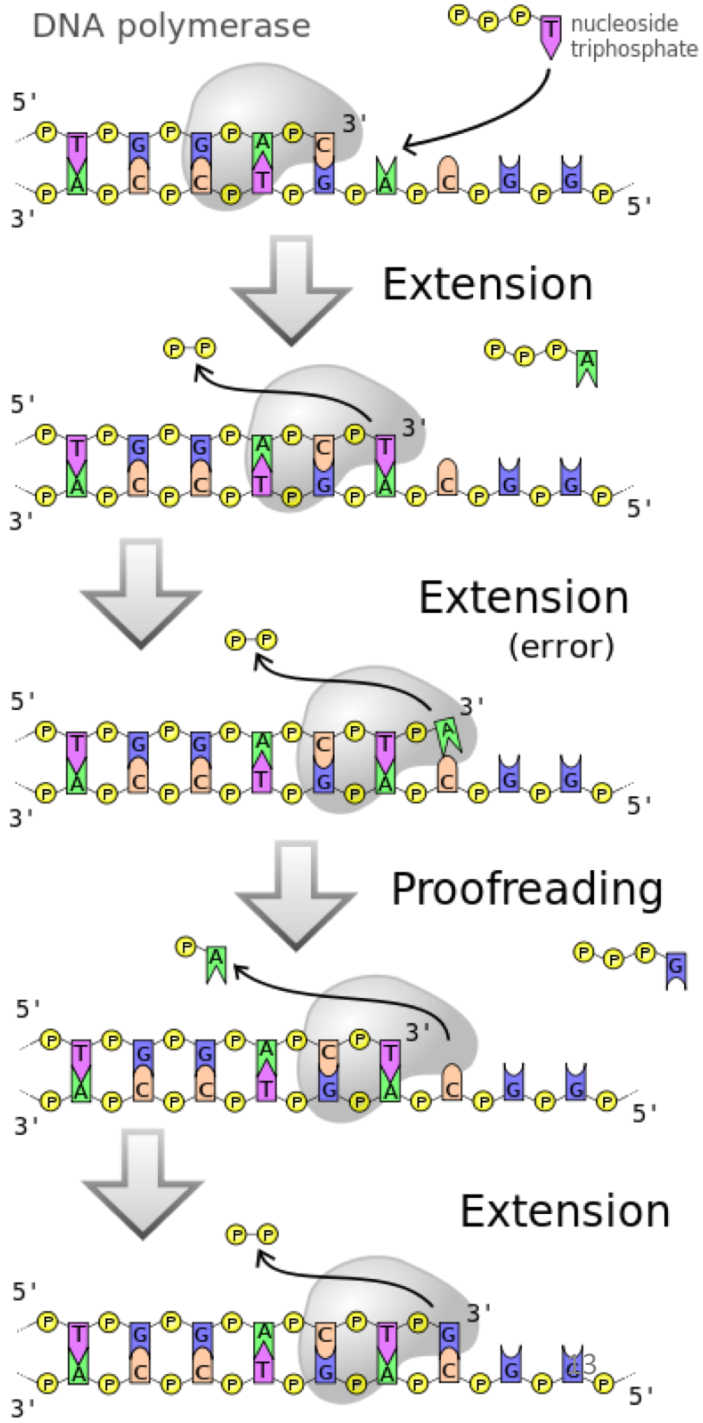
DNA Synthesis and Replication

Cells in our bodies are constantly dividing to form new copies.
This requires to duplicate the DNA, so that the daughter cell has its own genome.
DNA replication is achieved by synthesizing a new copy one base at a time:
The double helix gets unwound and complementary bases added to make two new helixes.
The enzyme that does this (complementing a single strand with matching bases) is called DNA polymerase.



DNA Polymerase

- Polymerase is an enzyme present in cells.
  - Enzymes are a class of protein, which act as a catalyst
  - That means they change things in the cells without being changed themselves.
- Polymerase adds nucleotides to single-stranded DNA one at a time from the 3’ end.
- This is used when replicating DNA for cell division.
- The same functionality can be used for other purposes, however. See next lecture.



Summary and Outlook

- DNA is present in each cell of every organism. There are multiple separate molecules that occur as pairs (referred to as chromosome pairs).
- DNA does not do very much – it is merely the blueprint for other molecules, such as RNA and Proteins.
  - We are going to revisit these in a future lecture.
- DNA can be replicated via DNA polymerase. It turns single-stranded DNA into double stranded by adding the according complementary bases one at a time.
- For now, we have enough information to get started on genomics – the study of genomes (DNA).