
DTU

DTU Health Technology
Bioinformatics

Introduction to NGS

Gabriel Renaud
Associate Professor
Section of Bioinformatics
Technical University of Denmark
gabriel.reno@gmail.com

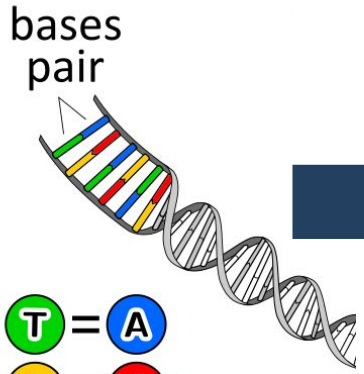
Menu

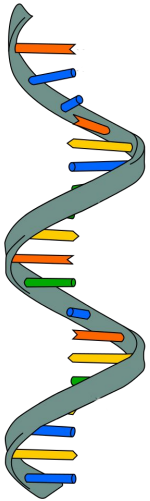
- What is sequencing? why?
- Basic nomenclature

What is sequencing?

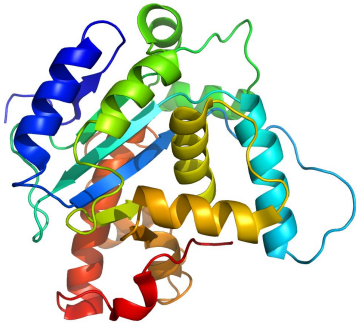
Remember high school?

DNA





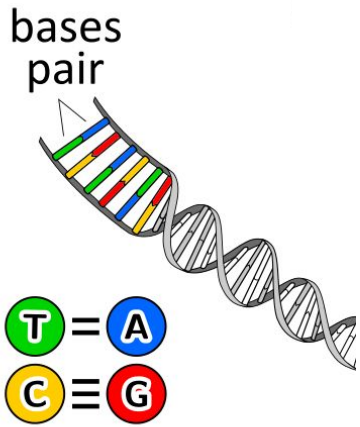
mRNA



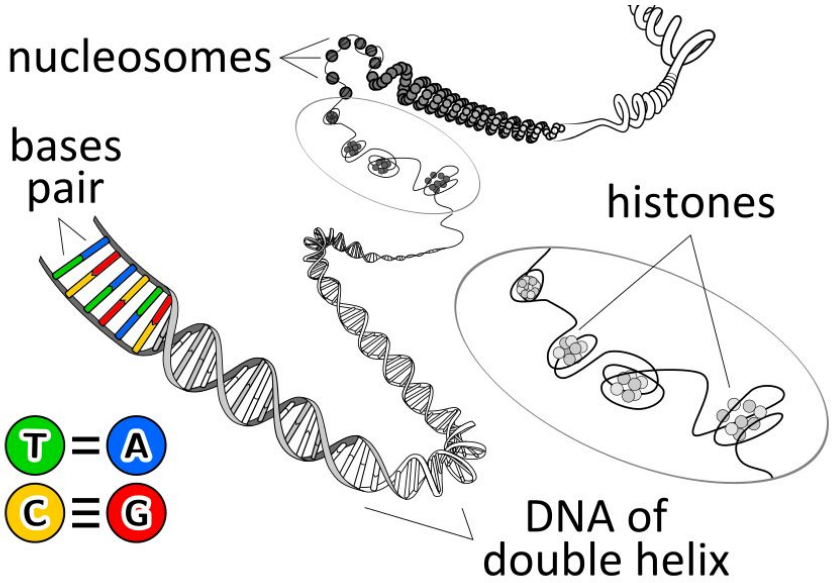
protein

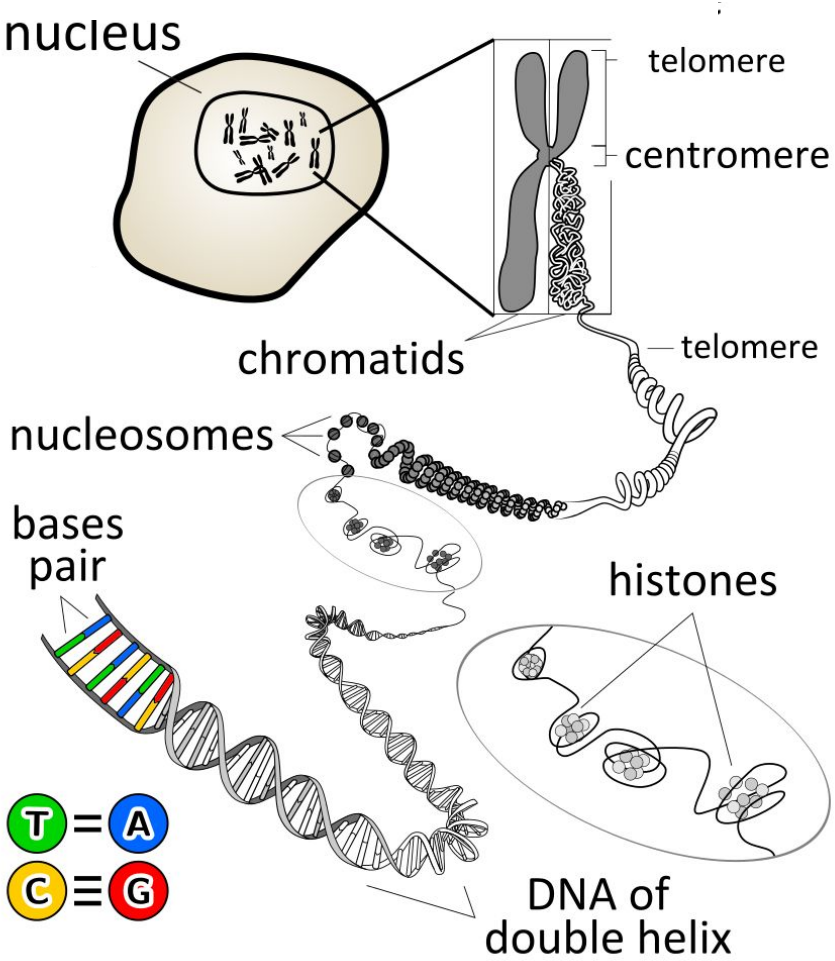
phenotype

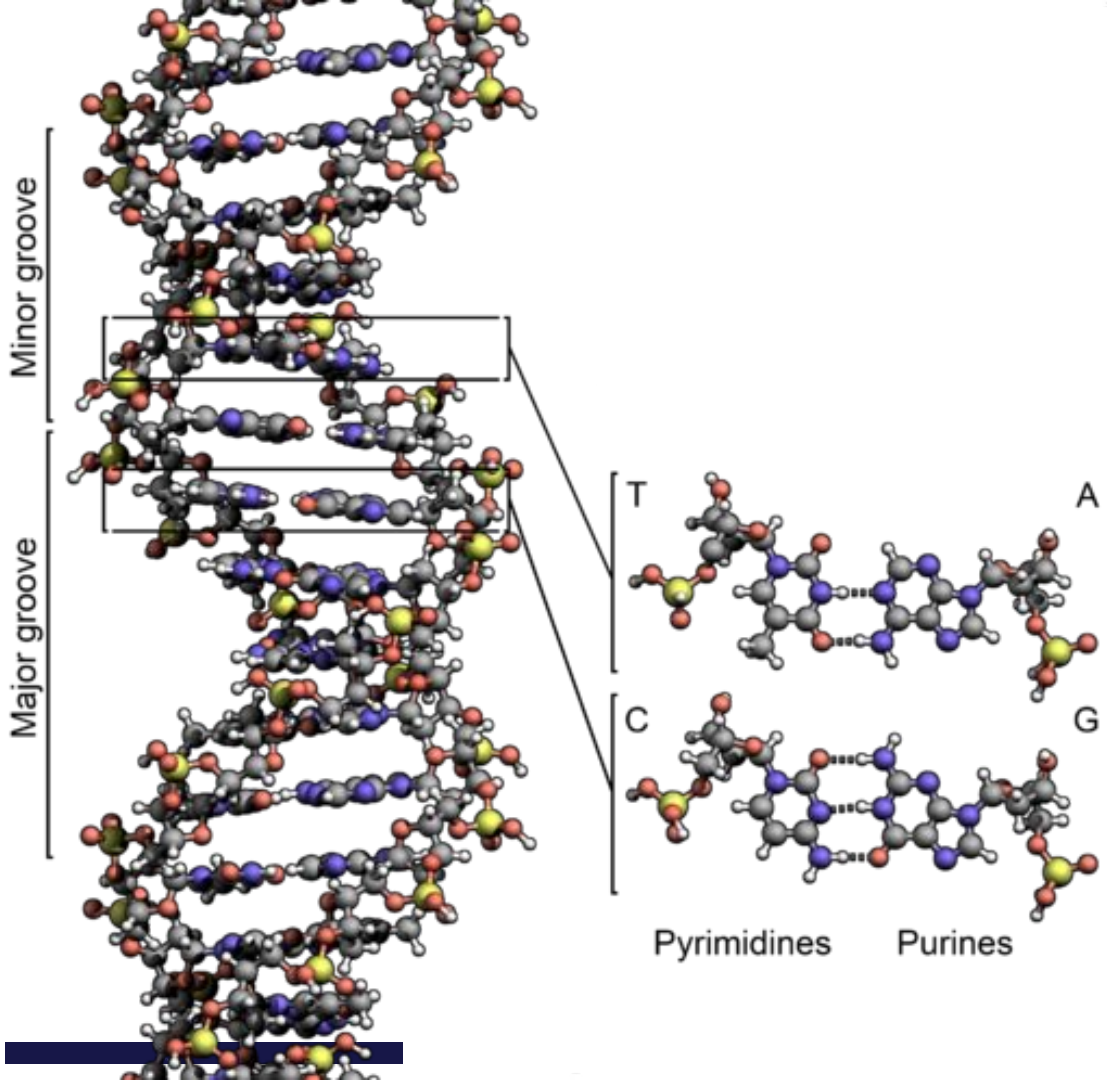
DNA



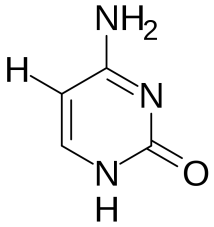
A few reminders about DNA …


nucleosomes
bases pair
histones
T = A
C ≡ G
DNA of double helix

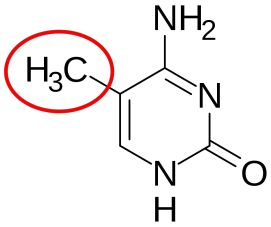

nucleus
telomere
centromere
chromatids
telomere
nucleosomes
bases
pair
histones
T = A
C ≡ G
DNA of
double helix



The “4” bases of DNA



cytosine

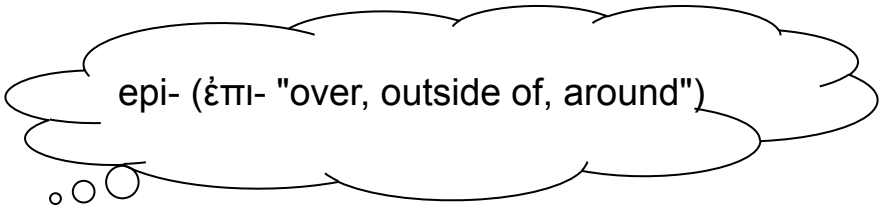


methylated cytosine

methylation

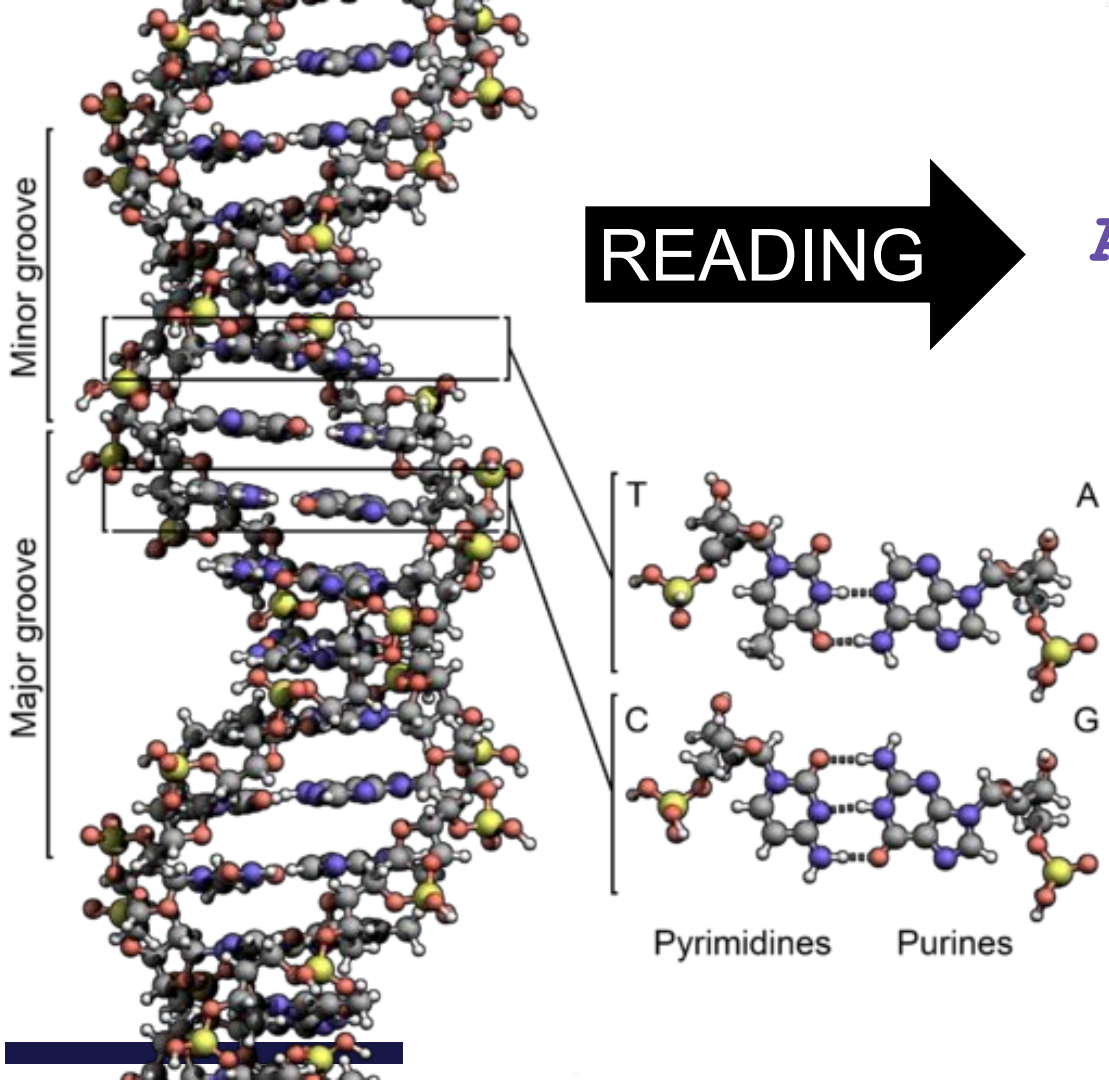
Genetics:

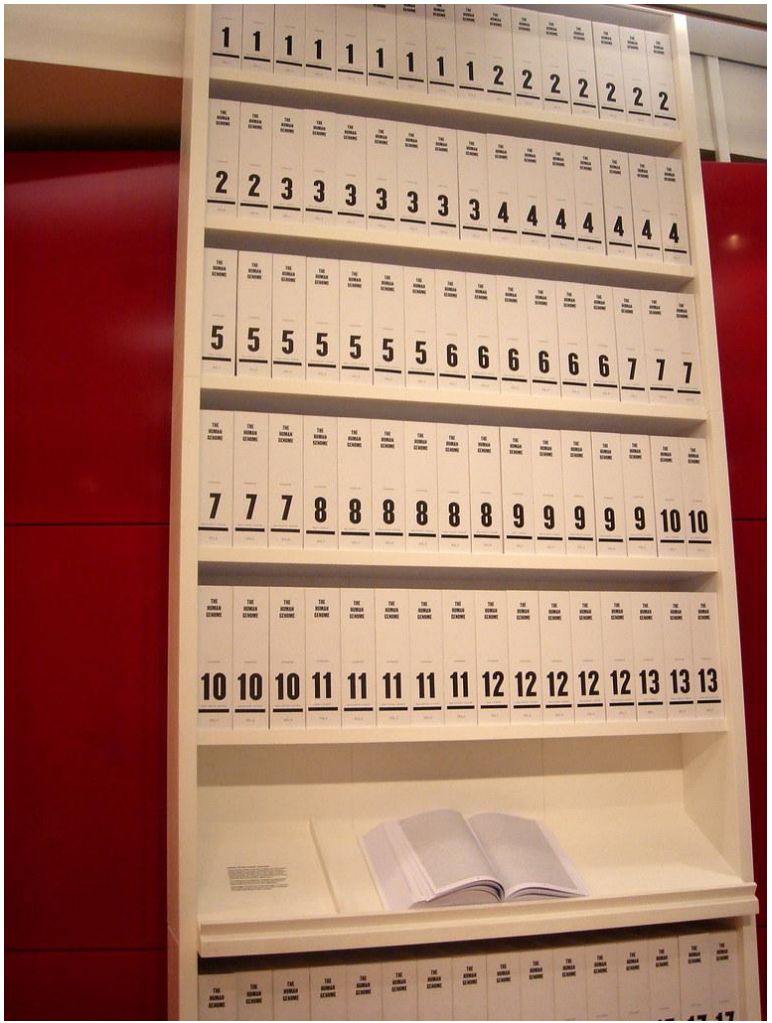
- A, C, G, T



Epigenetics:

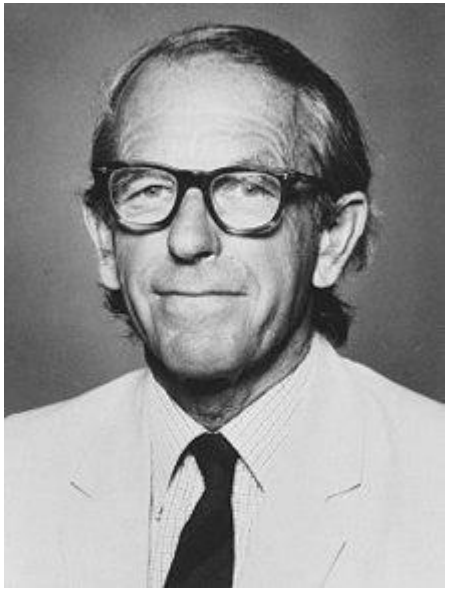
- Methylation
- Nucleosome positions


Minor groove
Major groove
READING
AGCAATCTCAATTACA
T
A
C
G
Pyrimidines
Purines



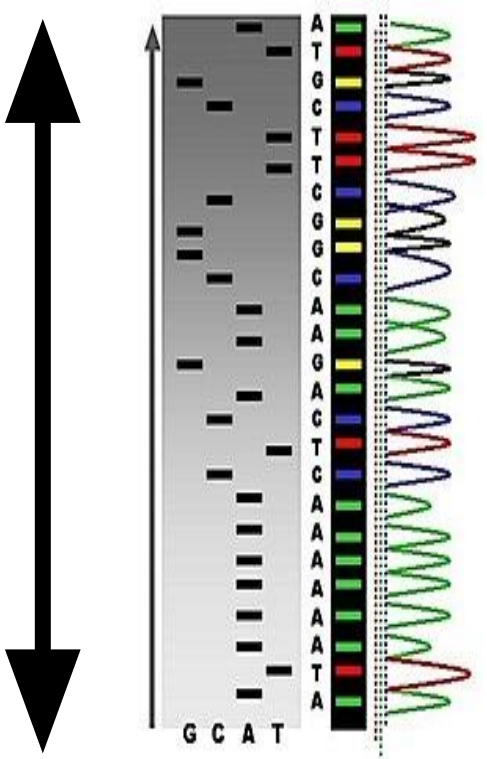
Human genome
3 billion letters

If we study Next-Generation Sequencing (NGS), why “next”? What was before?



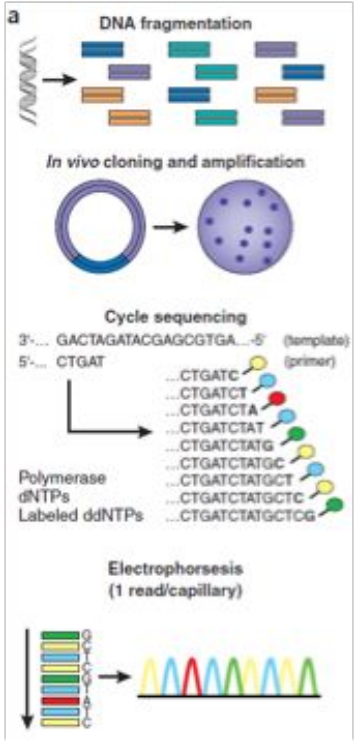
Frederick Sanger 1918 - 2013

1000 bases
x
96



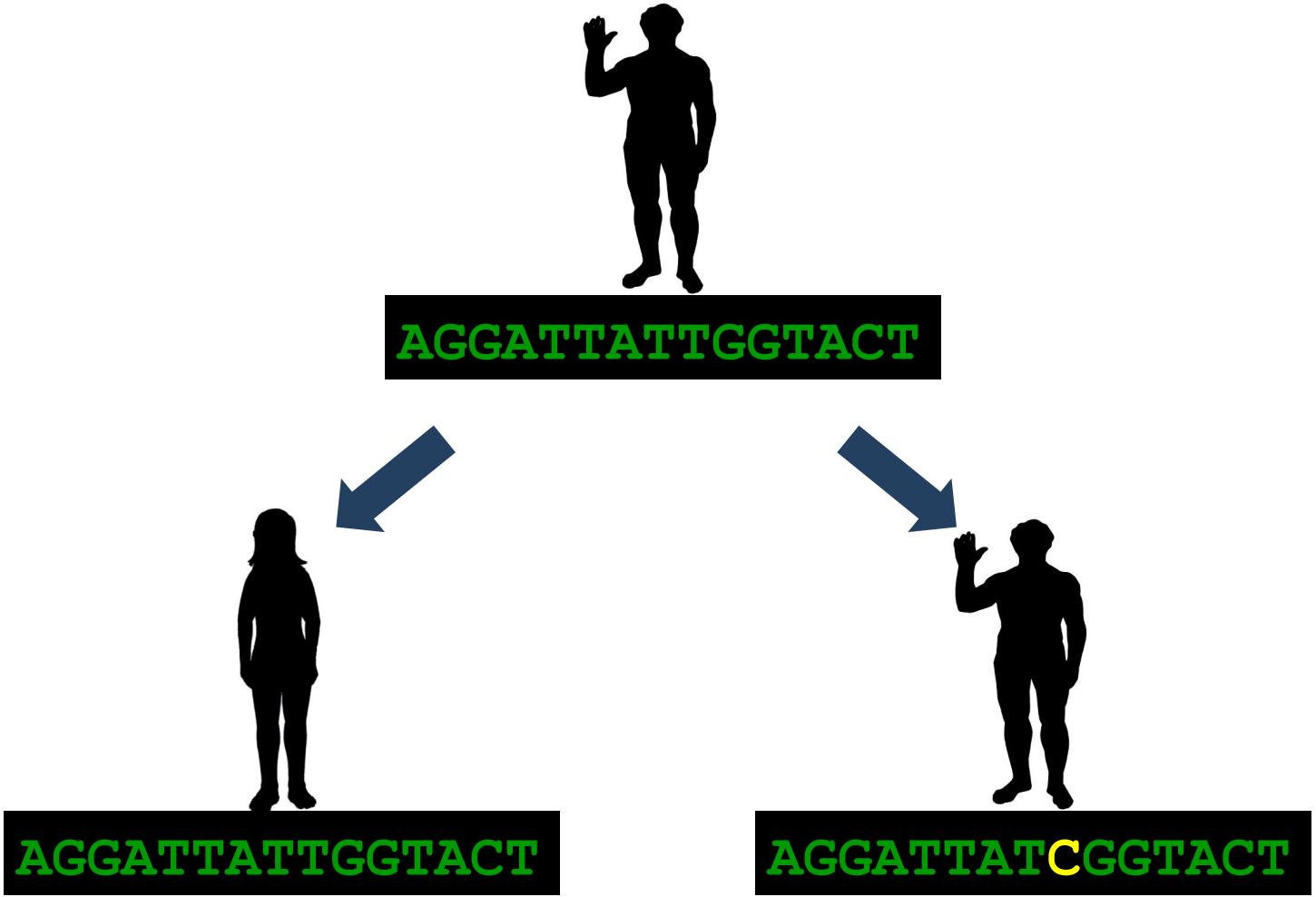
1977

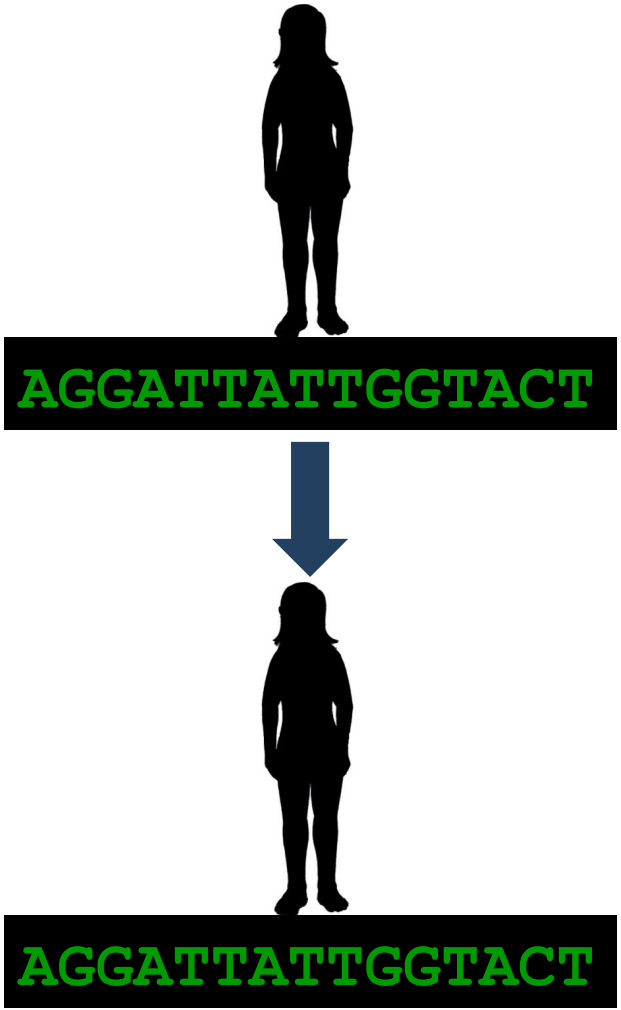
First generation: Sanger

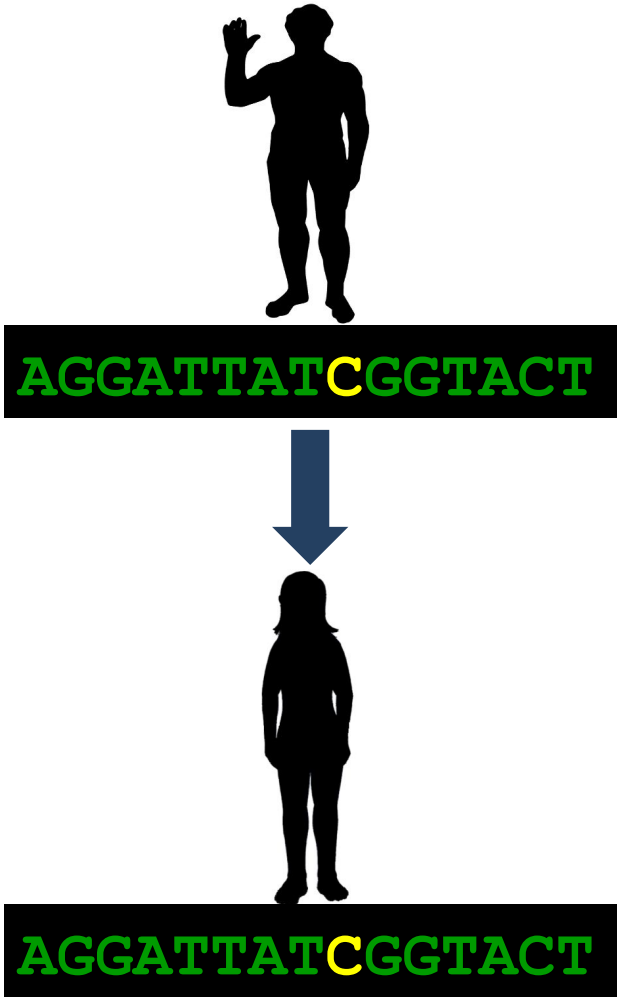


- Fragment DNA
- Clone into plasmid and amplify
- DNA polymerase and only 1 primer
- Sequence using labeled dinucleotides which cap seqs.
- Run capillary electrophoresis/gel and “read” DNA code
- Low output, long reads (~800-1200 nt), high quality
- Produces 96 reads / run

Why sequence?


AGGATTATTGGTACT
AGGATTATTGGTACT
AGGATTATCGGTACT


AGGATTATTGGTACT
AGGATTATTGGTACT


AGGATTATCGGTACT
AGGATTATCGGTACT

AGGATTATTGGTACT

AGGTTTATTGGTACT

AGGATTATCGGTACT

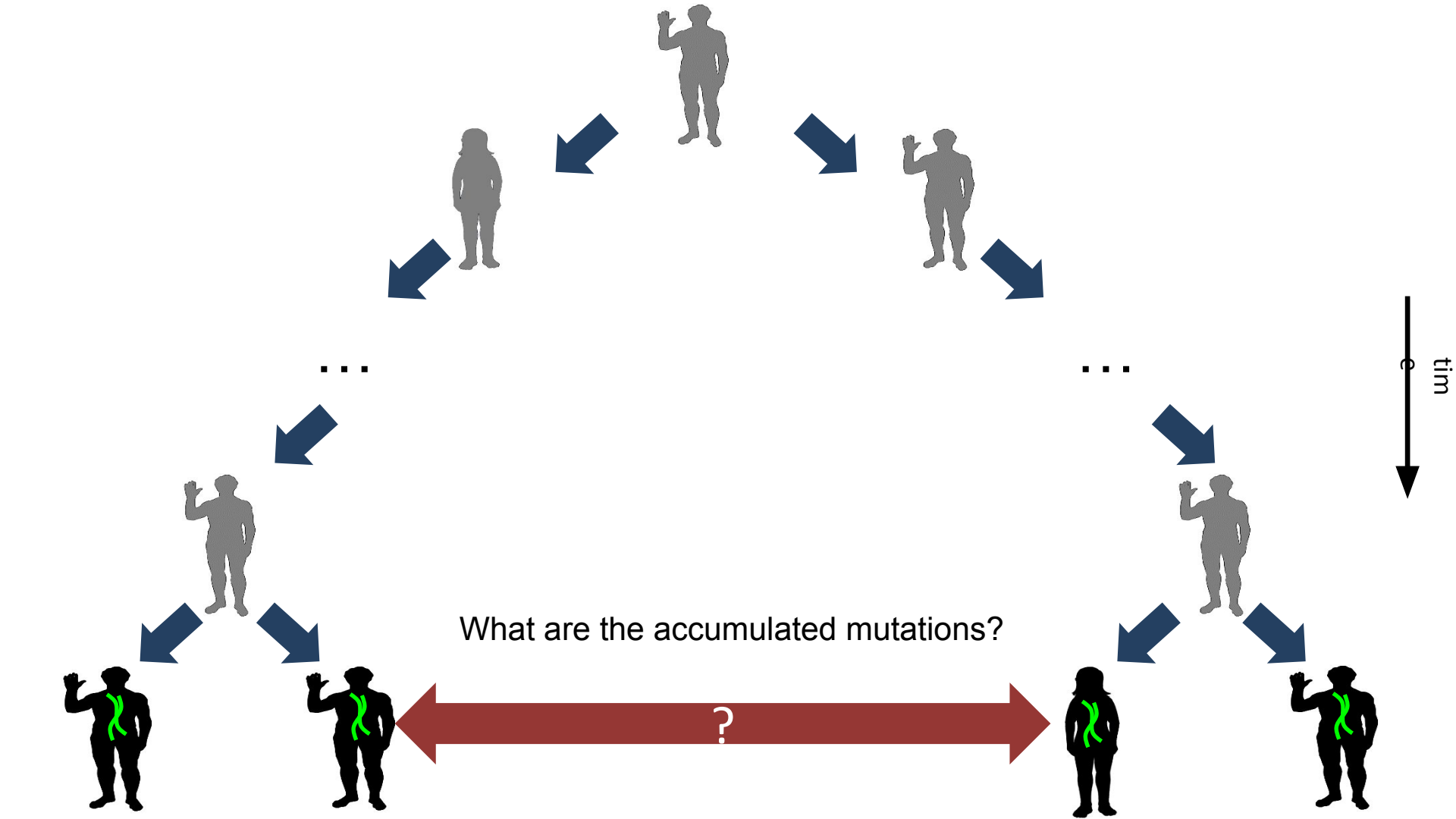
AGGATTATCGGTACT

AGGTTTATTGGTACT

AGGTTTATTGGTAGT

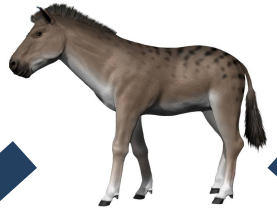
AGGATTATCGGTACT

AAGATTATCGGTACT


What are the accumulated mutations?
?
time

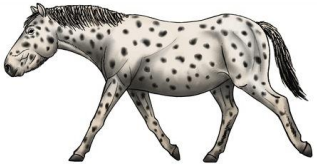
Investigate differences within a species

Investigate differences between species



?

. . .



. . .

?

time

"Nothing in Biology Makes Sense Except in the Light of Evolution"

Theodosius Dobzhansky, 1973

"**Nothing in Biology Makes Sense Except in the Light of Evolution**"

NGS

me, I made that up just now

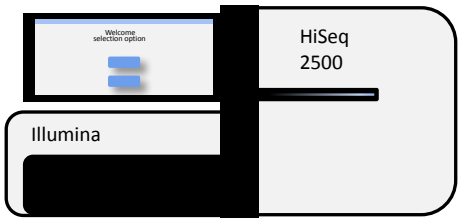
What can we use it for?

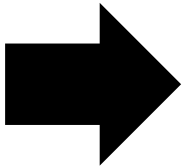
- Whole genome re-sequencing
- Population genomics
- Diagnostics
- Cancer genomics
- Ancient genomes
- Metagenomics
- RNA sequencing
- Single cell sequencing
- Genomic Epidemiology
- anything with DNA

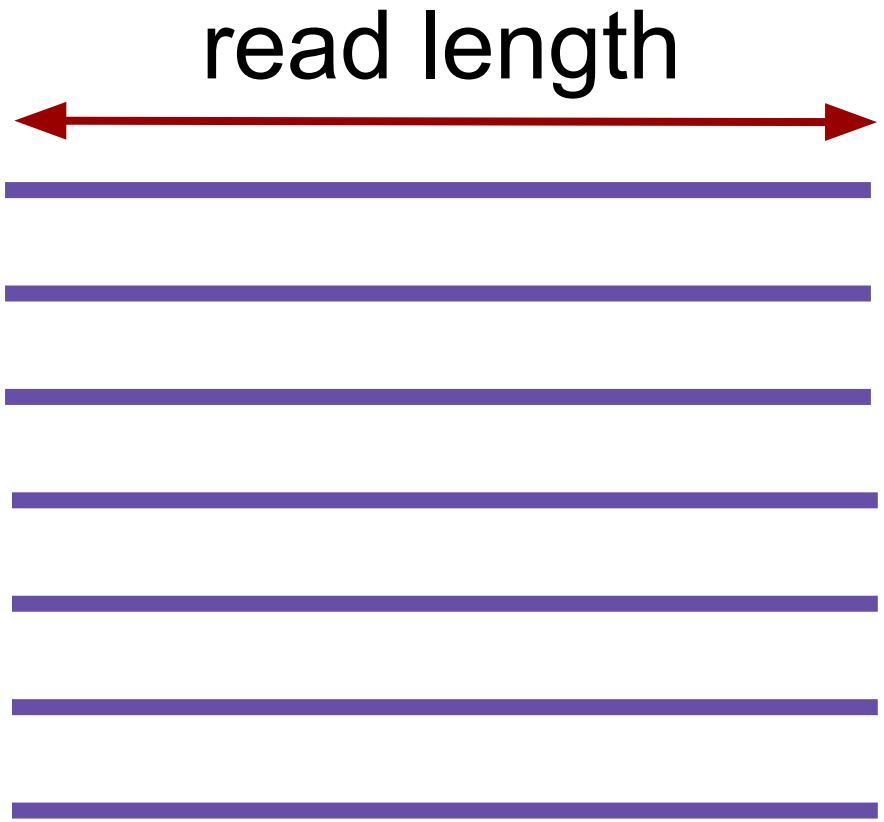
Basic concepts

3 key concepts

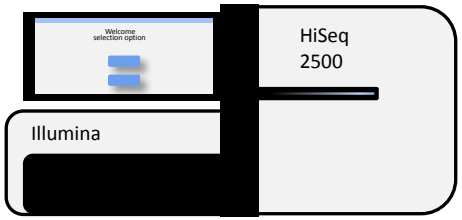
- Read length
- Throughput
- Types of errors


Welcome
selection option
HiSeq
2500
Illumina

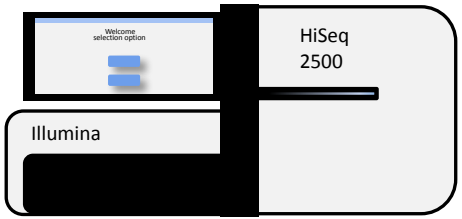



read length

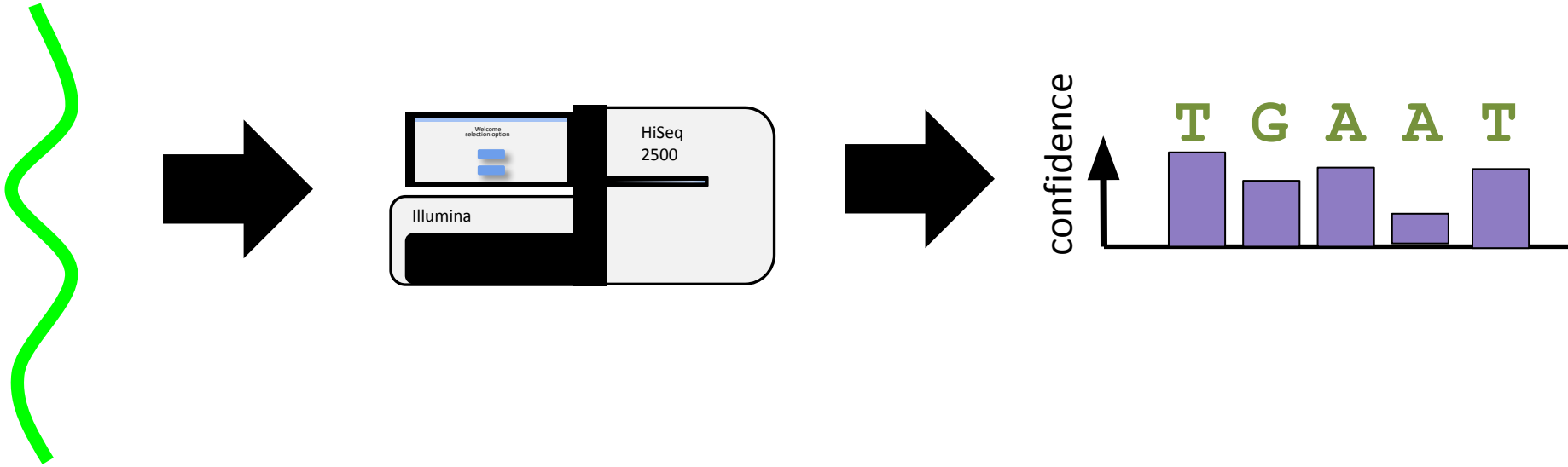
throughput def. 1



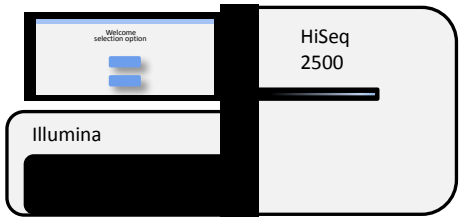
throughput def. 2

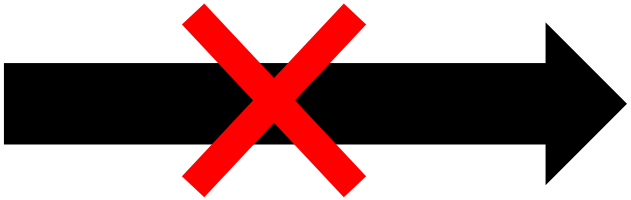


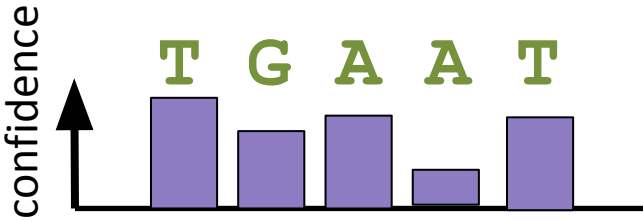
template

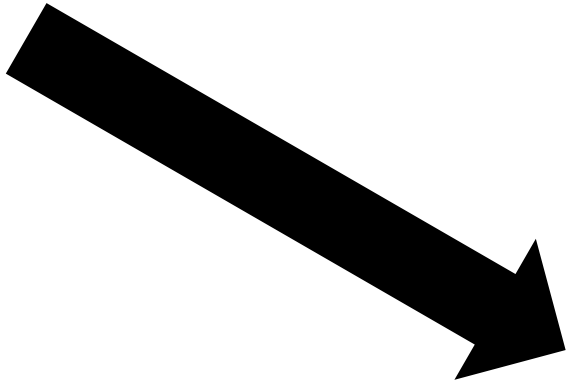

Welcome
selection option
HiSeq
2500
Illumina
confidence
T G A A T

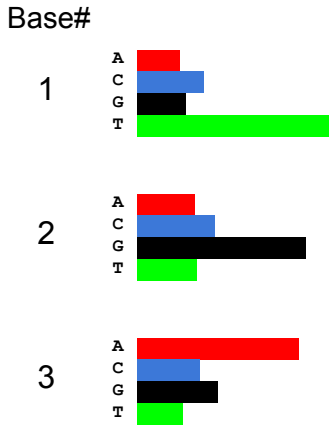
Key concept: basecalling



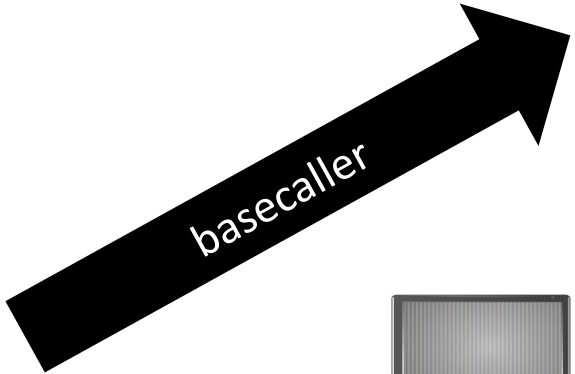






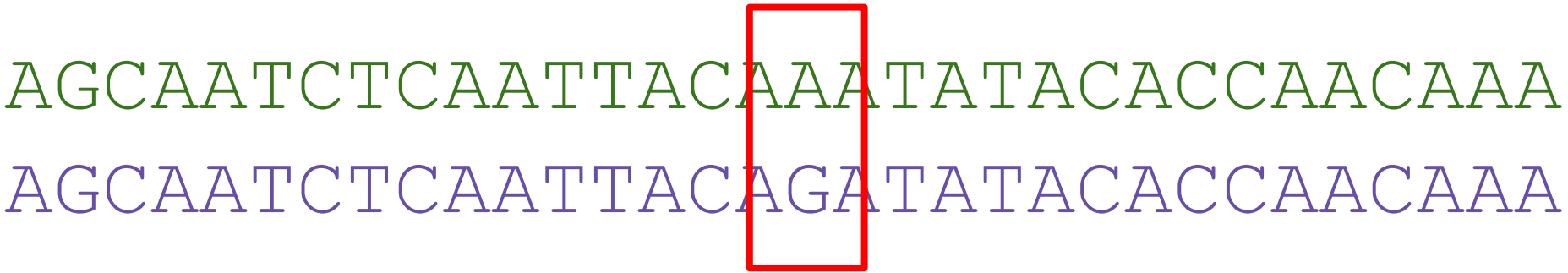


Raw Data



mismatch

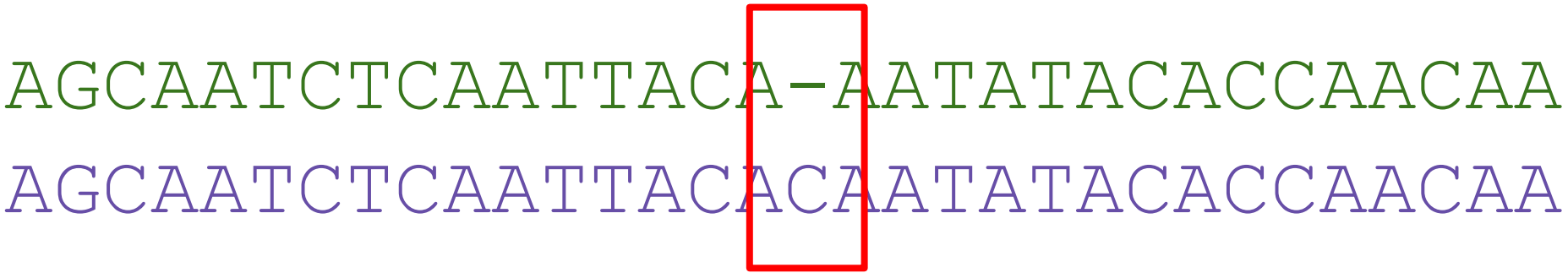
template



read

insert

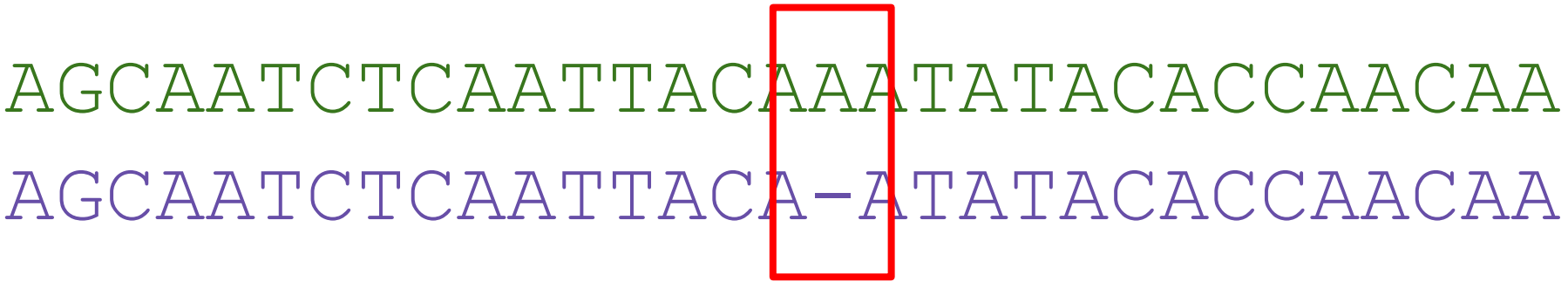
template



read

deletion

template

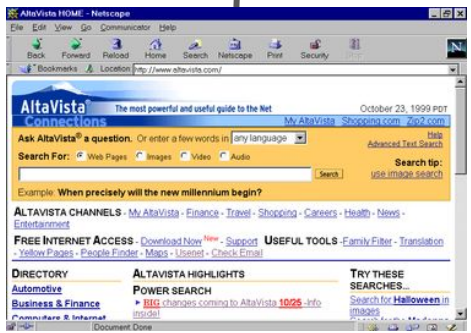


read

1977 1983 1989 1995 2001 2006 2012 2018 2024

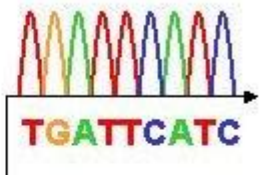
1977 1983 1989 1995 2001 2006 2012 2018 2024

STAR WARS



THE LORD OF THE RINGS

1977 1985 1989 1995 2001 2006 2012 2018 2024



Sanger

454

Ion Torrent

PacBio

Element Bio

Ultima Genomi

Oxford Nanopore

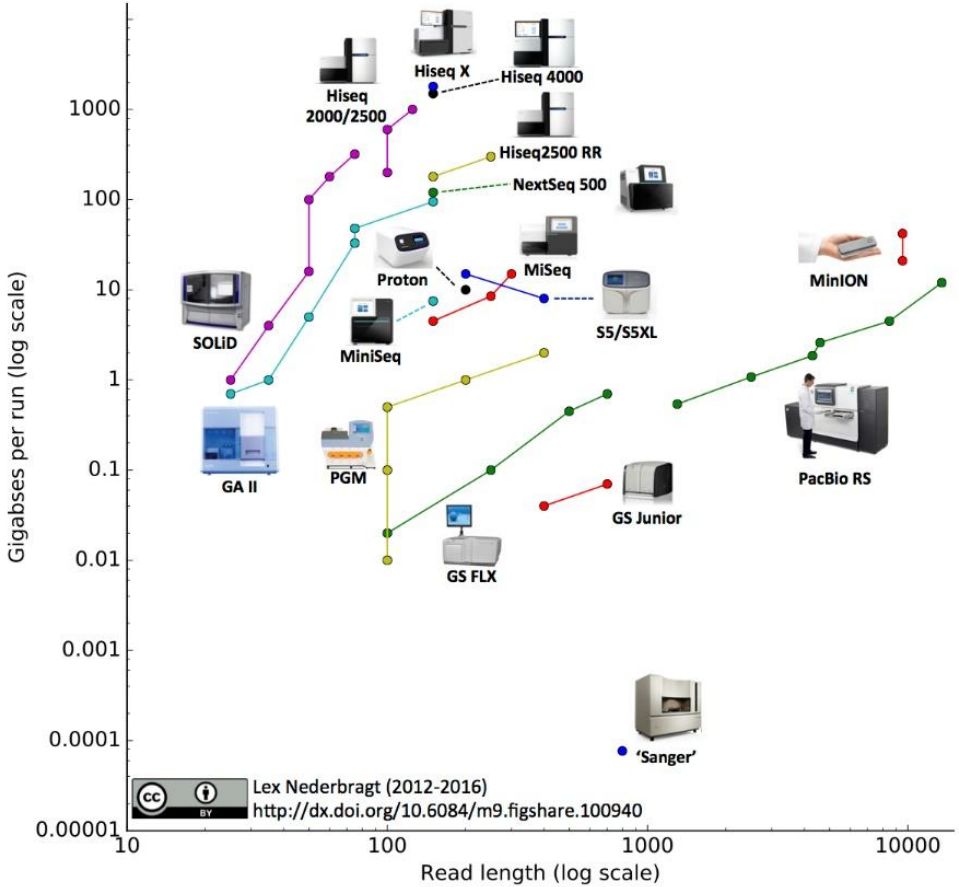
BGI

SOLiD

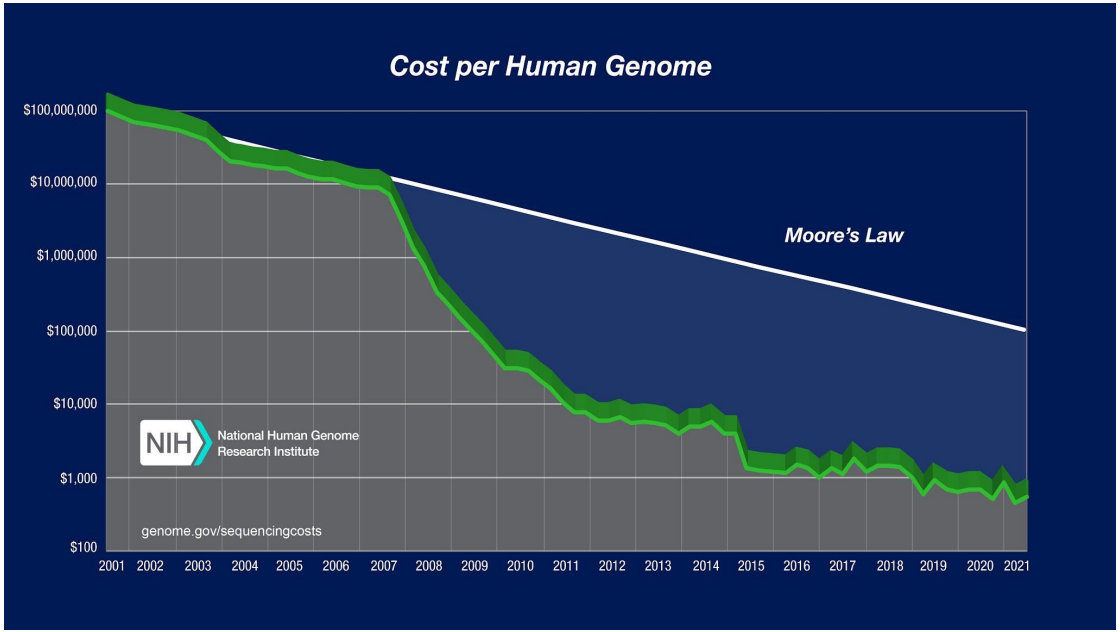
Illumina

1st generation of NGS

- 454 Life Sciences.
  – Bought by Roche 2007.
- Illumina/BGI is currently cheapest per GB
- Long-read sequencing is revolutionizing assembly



Sequencing costs



- Computer speed and storage capacity is **doubling every 18 months** and this rate is steady

- DNA sequence data **is doubling faster than computer speeds!**

NHGRI - Sept 2022

1990 - 2003

Picture: The Guardian

~5.54G USD (adj. for FY 2022)

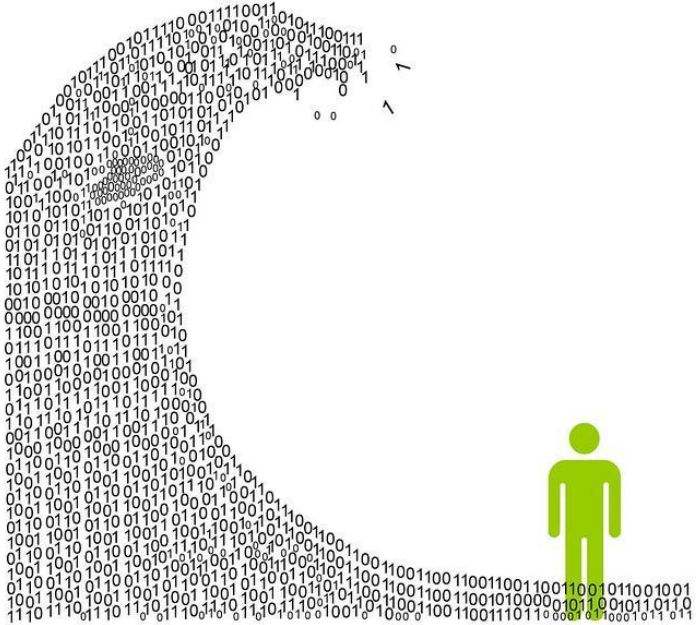
20 research centers, 6 countries

Human sequencing

- First draft genome of human in 2001, final 2004
- Estimated costs $5.54 billion USD, time 13 years

- Today:
  - 500-2000 USD for one genome
  - A couple of days
  - Will go down to 100 USD soon

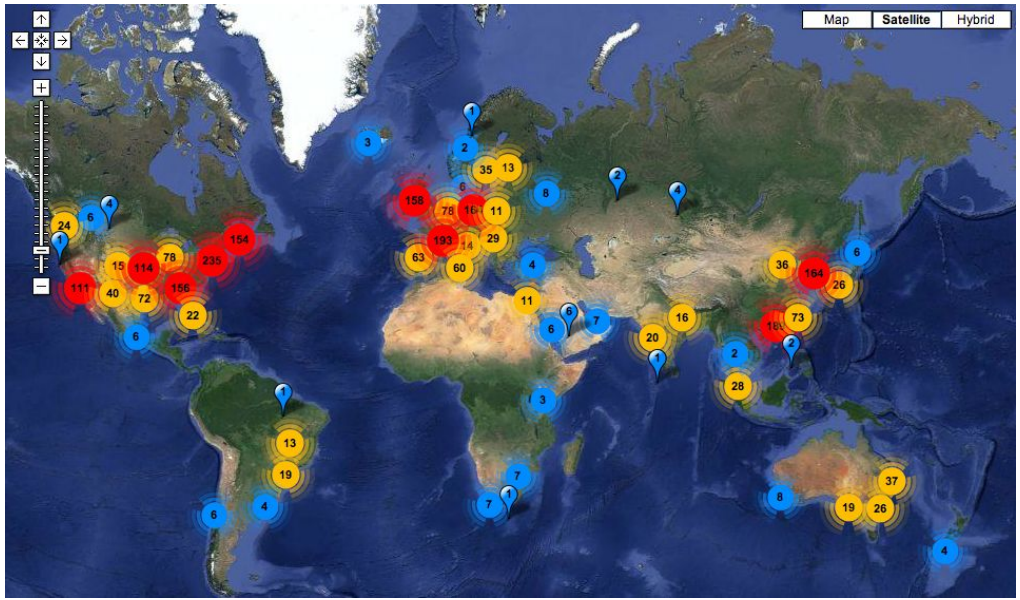
Storage and analysis



- Cost of sequencing is almost less than the cost of **storage and analysis**
- One Illumina NovaSeq system: almost 10,000 human genomes per year!
- A standard human (30-40x) whole-genome sequencing exp. would create 30-150 Gb of data

Distributed data production

- Worldwide >900 centers
- >60 Pb pr year (2014)
- 20,000 Pb pr year (2025)
- Data transfer and storage become difficult



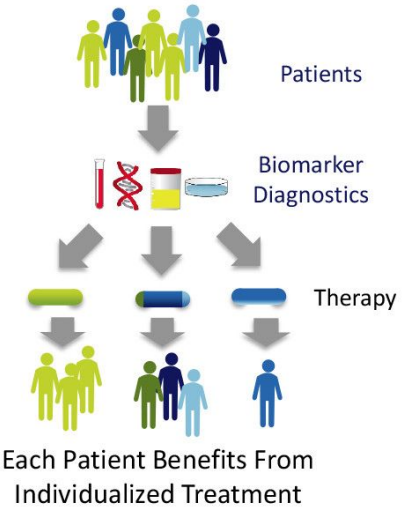
The X Genomes projects

- Human population projects
  - 1000 genomes project (2500 individuals)
  - Genomics England (100k individuals)
  - US Precision Medicine (1 million individuals)

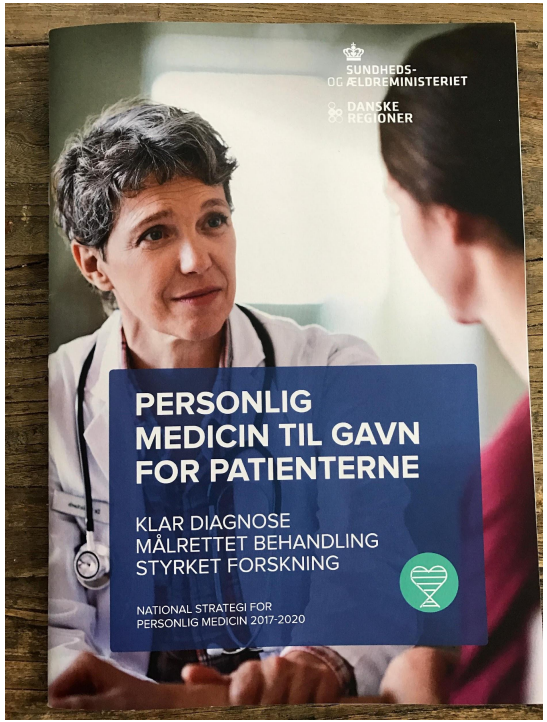
- 100K pathogens project, Earth Microbiome project, Cancer genome project, Plants and animals, Insects,…

NGS in the clinic

- Diagnostics of patients (+ fetus)
- Focused treatment of cancer patients
- Sequencing of bacterial isolates
- Country-wide projects:
  - UK, US, UAE, Qatar, Finland, China, …
  - DK: Danish regions want to sequence 100k individuals



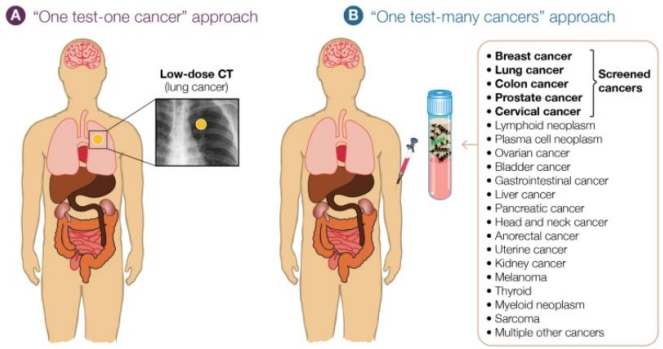
Personalized medicine



- Giving the same medication to all will not work
- Personalized medicine initiative in DK
- Sequence 100,000 patients on hospitals
- Use extensive registry data
- Current: 100M DKK (estimated 2G DKK)

Preventive medicine/prenatal

GRAIL



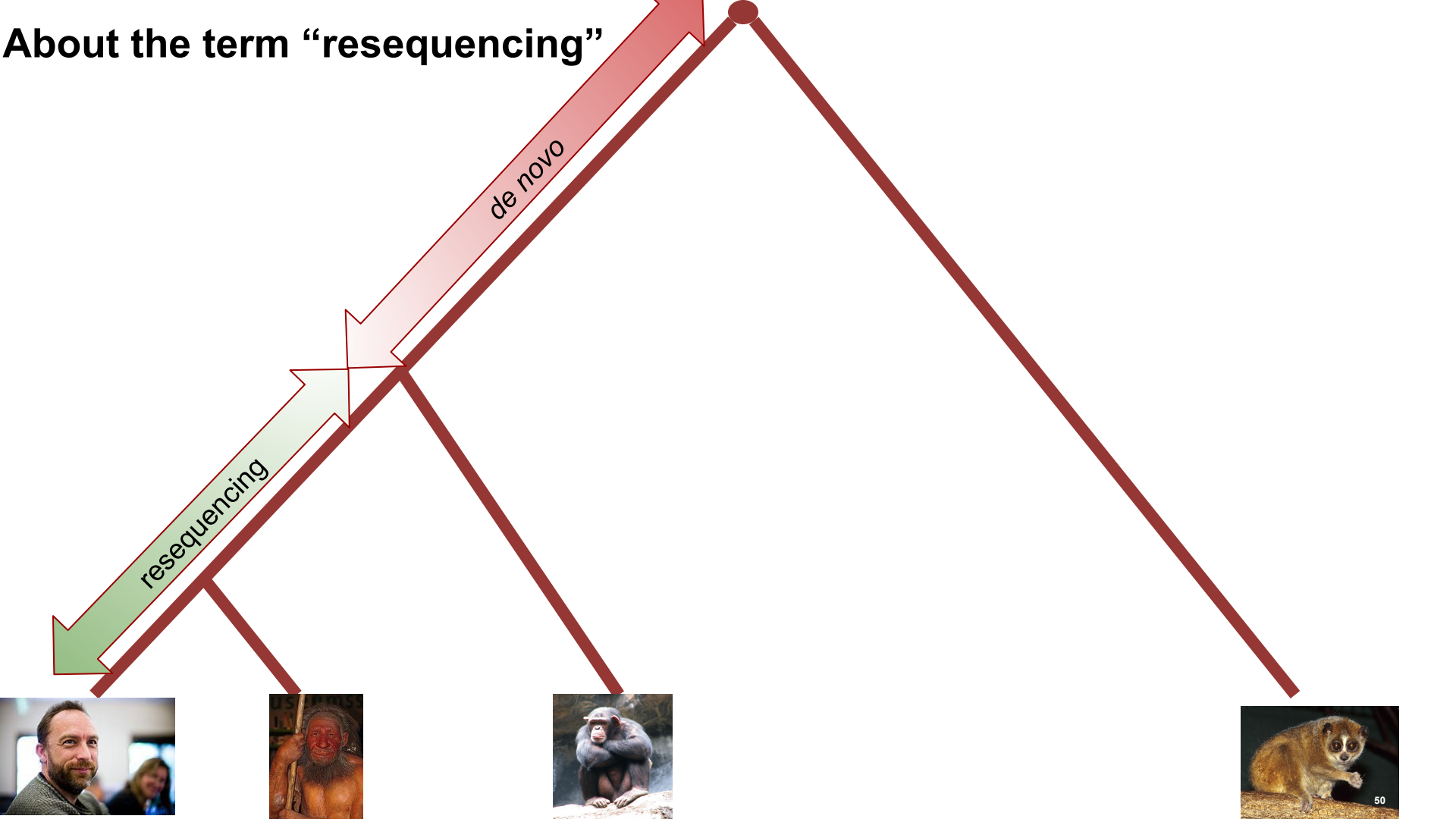
- cell free DNA
- prenatal screening

Ofman, Joshua J., et al. "GRAIL and the quest for earlier multi-cancer detection." Nature (2018).

NGS & bioinformatics

- Extreme data size causes problems
- Just transferring and storing the data
- Need computer clusters, large storage
- Think in fast and parallel programs
- Cloud computing increasingly used

About the term “resequencing”



Whole genome sequencing