

DTU





**DTU Health Technology
Bioinformatics**

Introduction to NGS

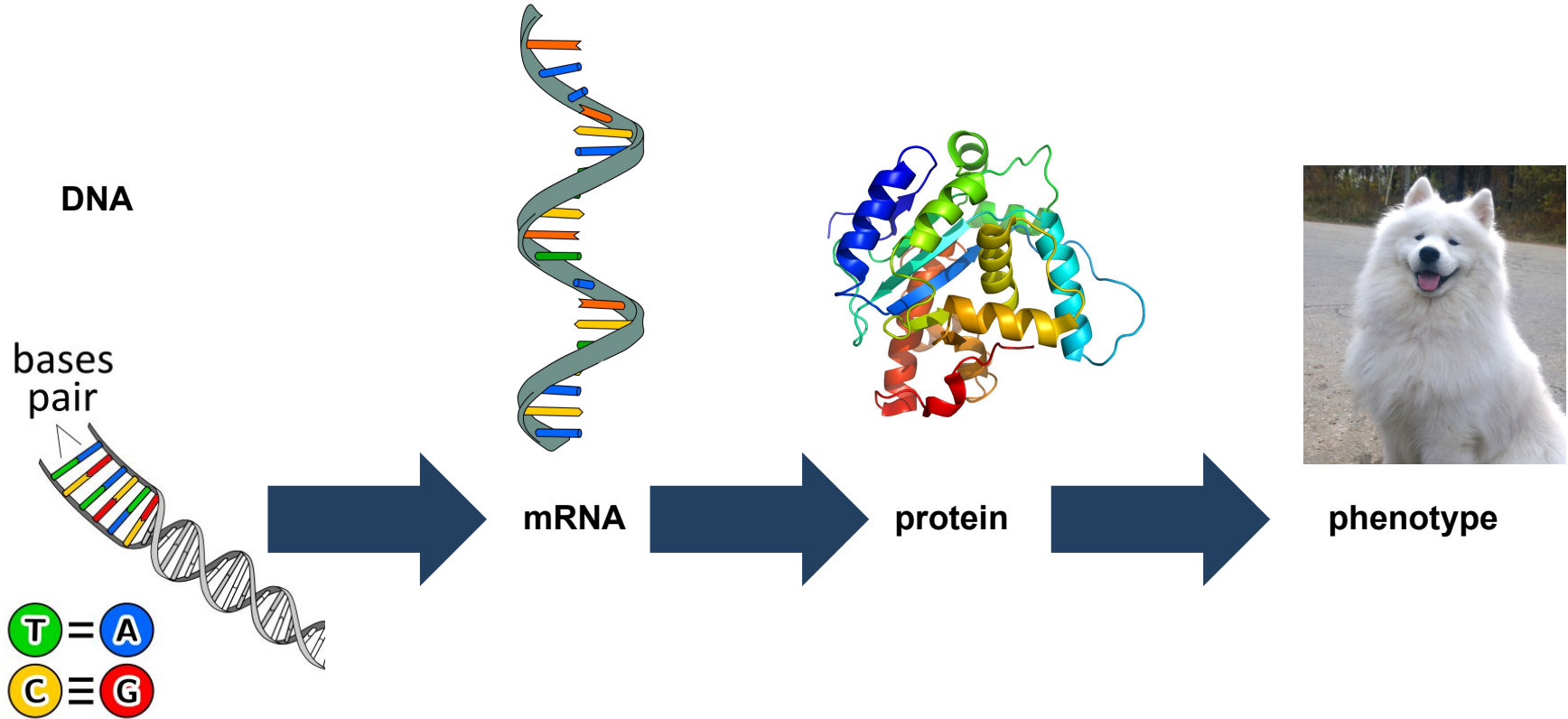
*Gabriel Renaud
Associate Professor
Section of Bioinformatics
Technical University of Denmark
gabriel.reno@gmail.com*

Menu

- What is sequencing? why?
- Basic nomenclature

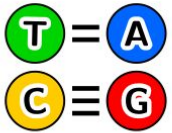
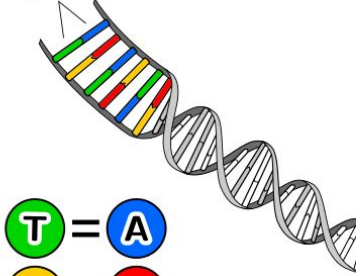
What is sequencing?

Remember high school?



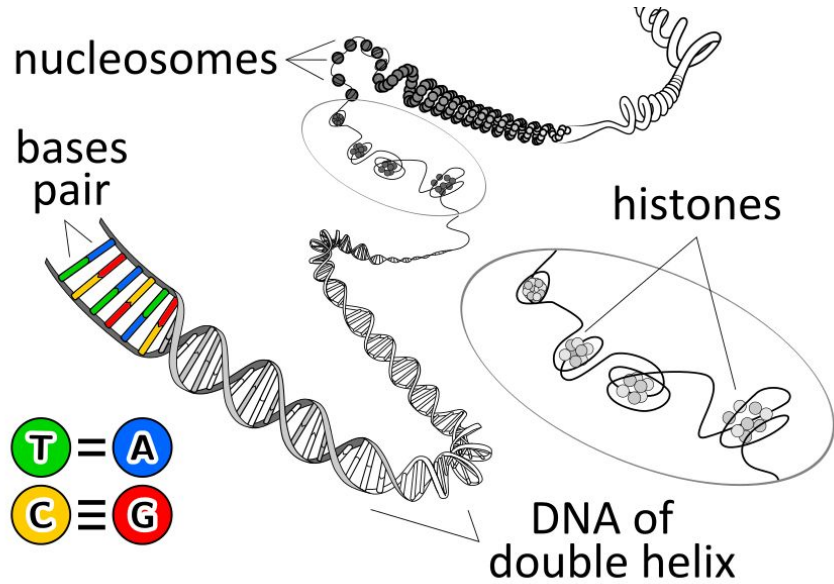
DNA

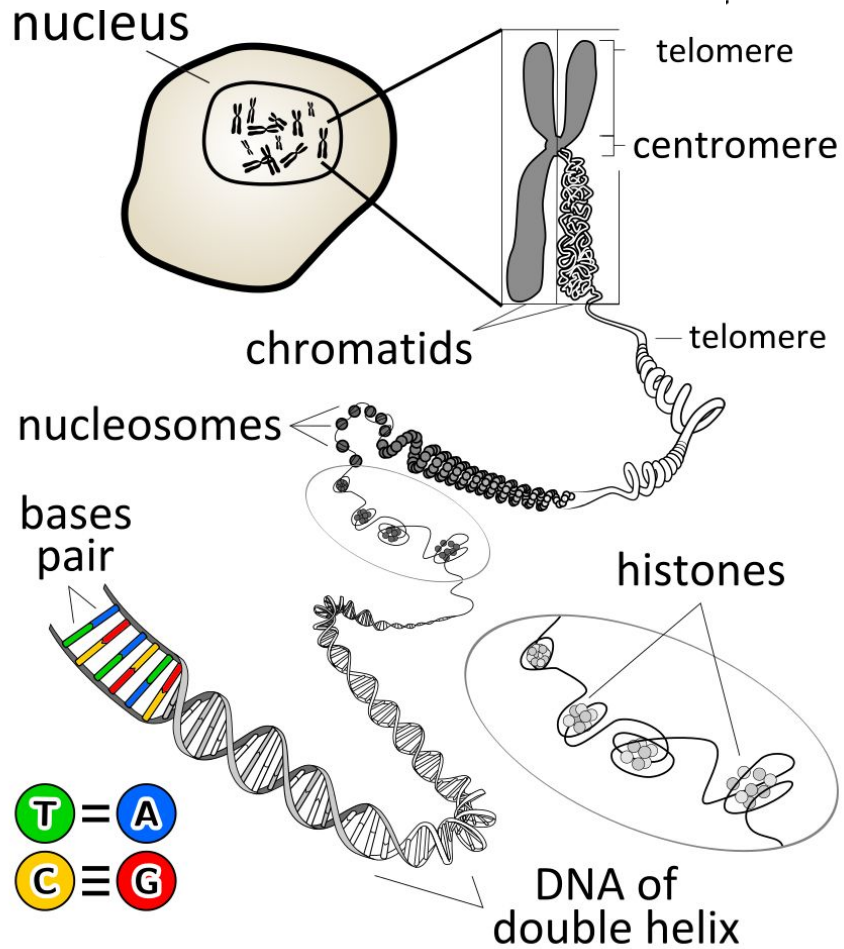
bases
pair



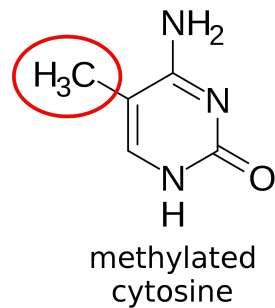
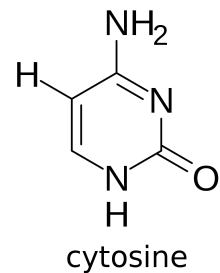
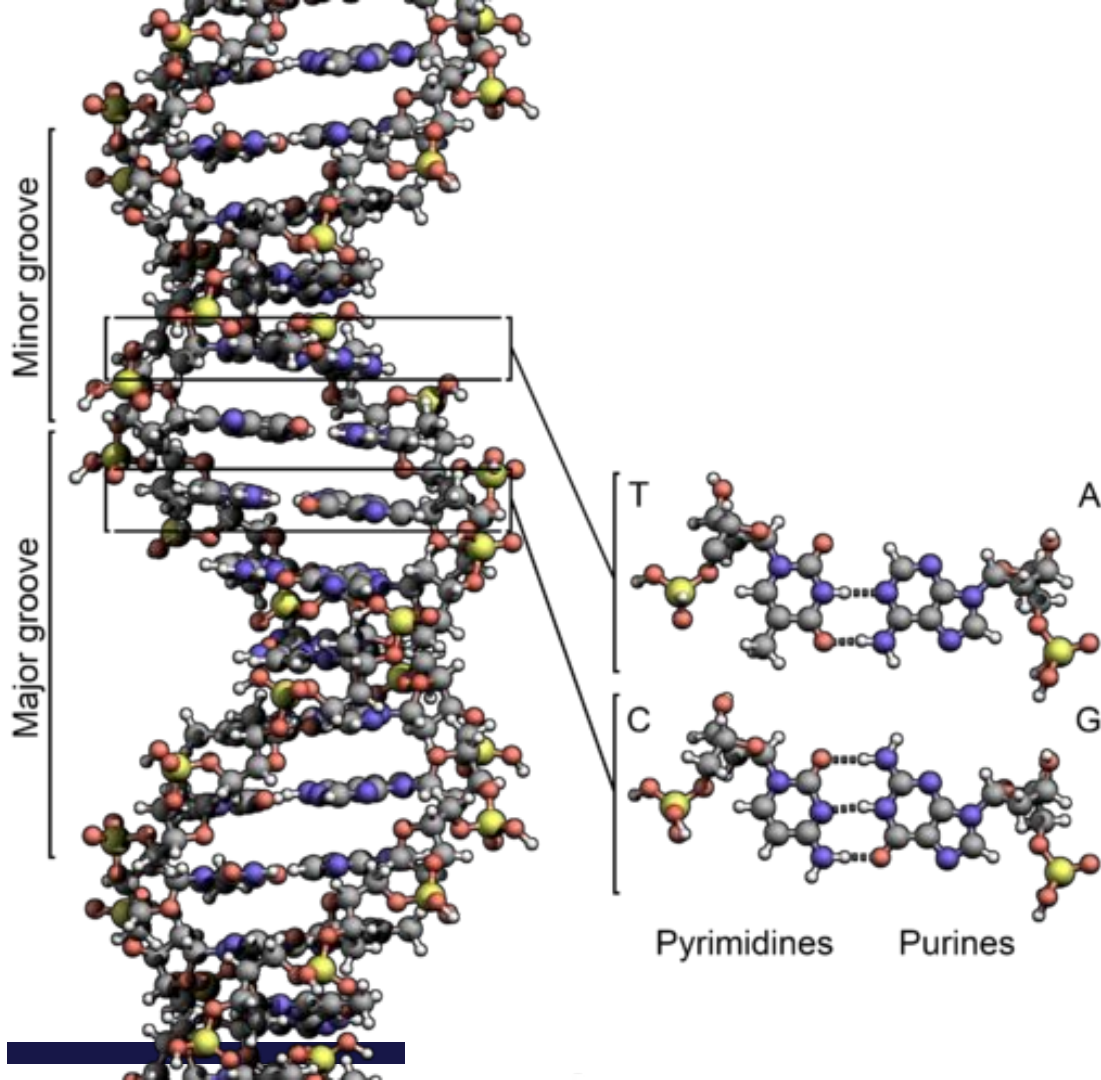
A few reminders about DNA ...







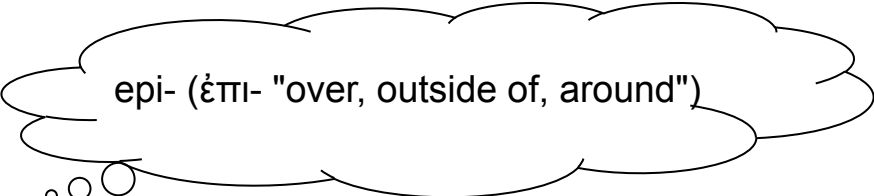
The "4" bases of DNA



methylation

Genetics:

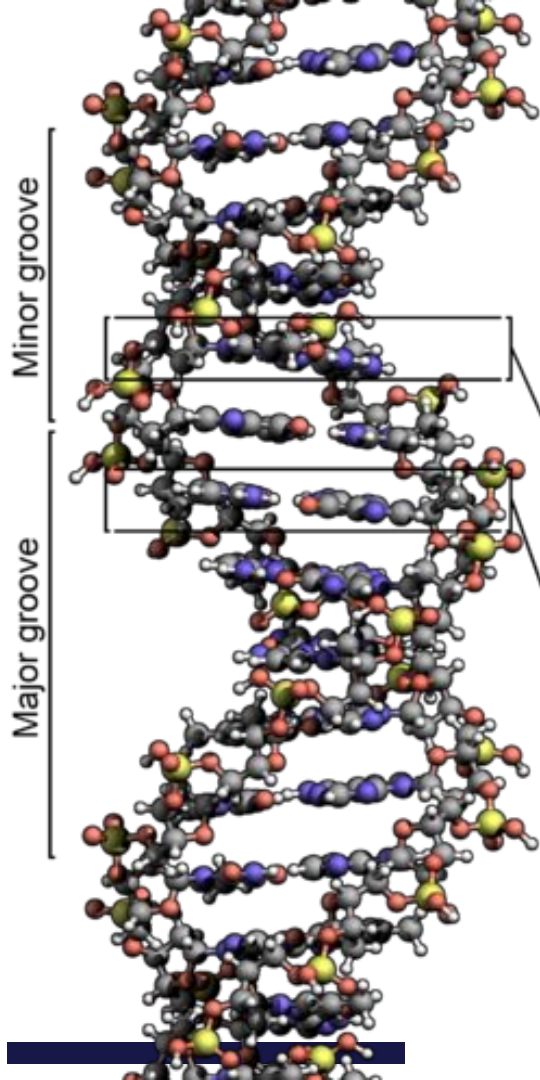
- A, C, G, T



epi- (ἐπι- "over, outside of, around")

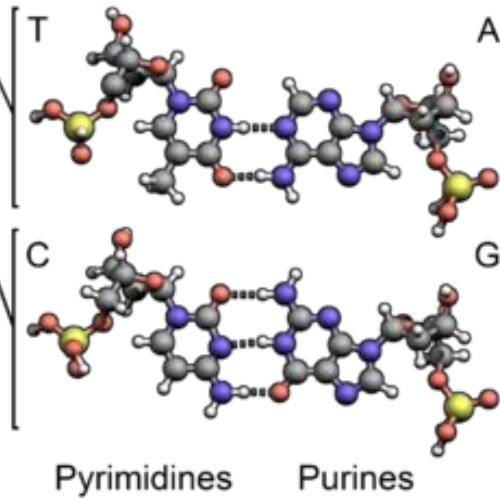
Epigenetics:

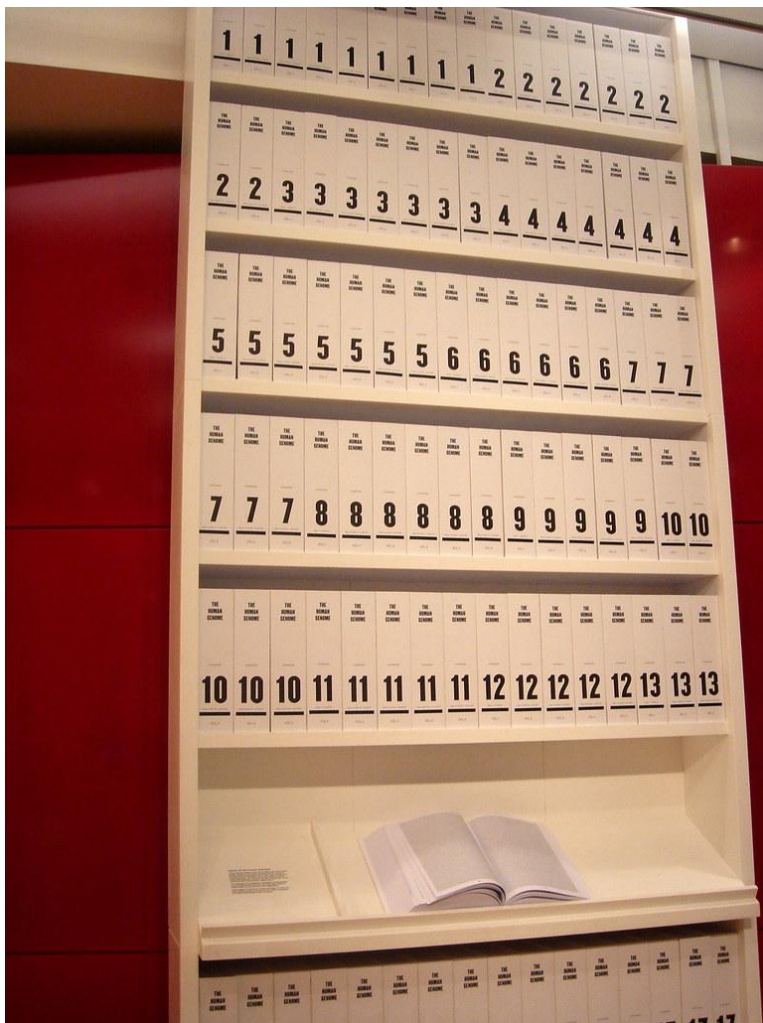
- Methylation
- Nucleosome positions



READING

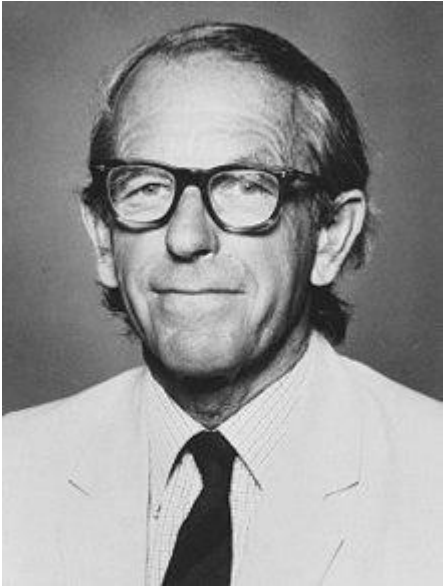
AGCAATCTCAATTACA





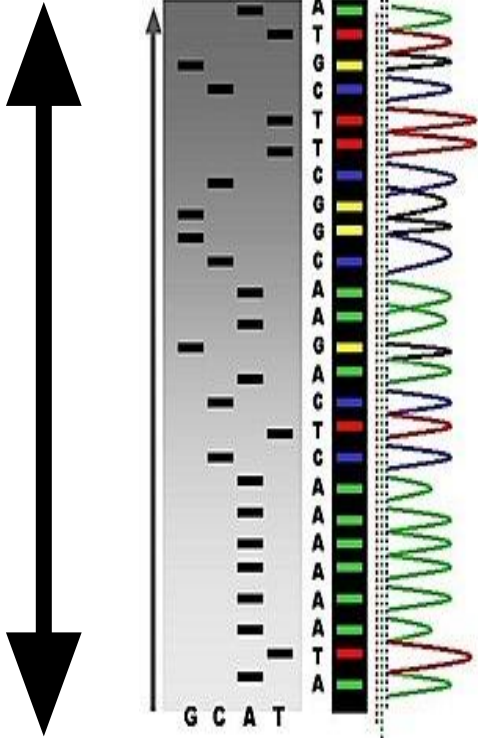
Human genome
3 billion letters

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



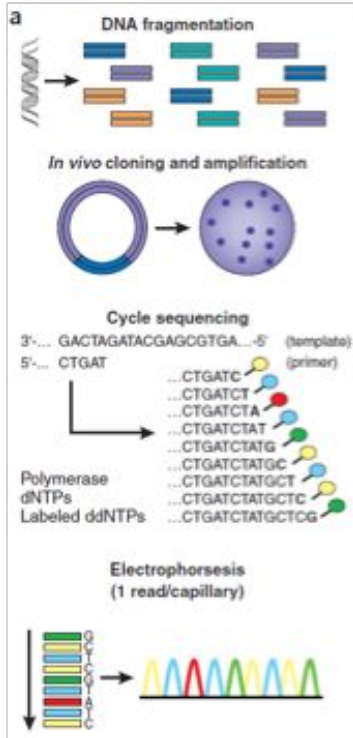
Frederick Sanger 1918 - 2013

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



AGG**T**TTATTGGTACT



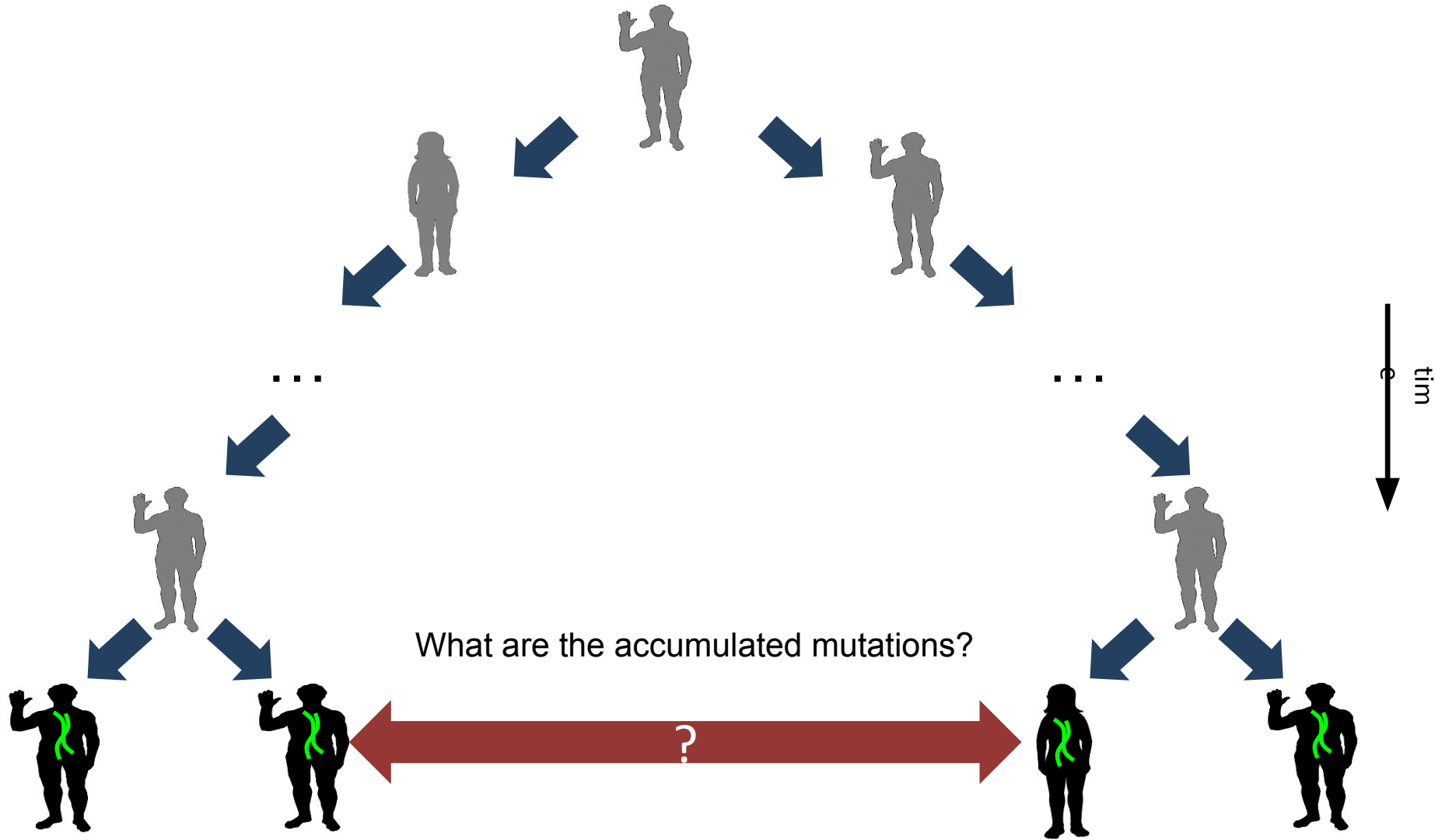
AGG**T**TTATTGGT**A**GT



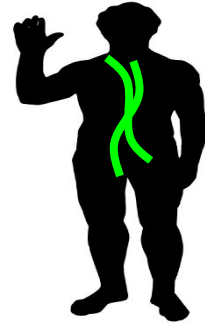
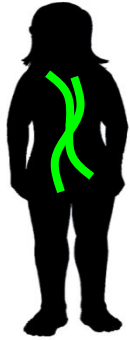
AGGATTAT**C**GGTACT



AAGATTAT**C**GGTACT

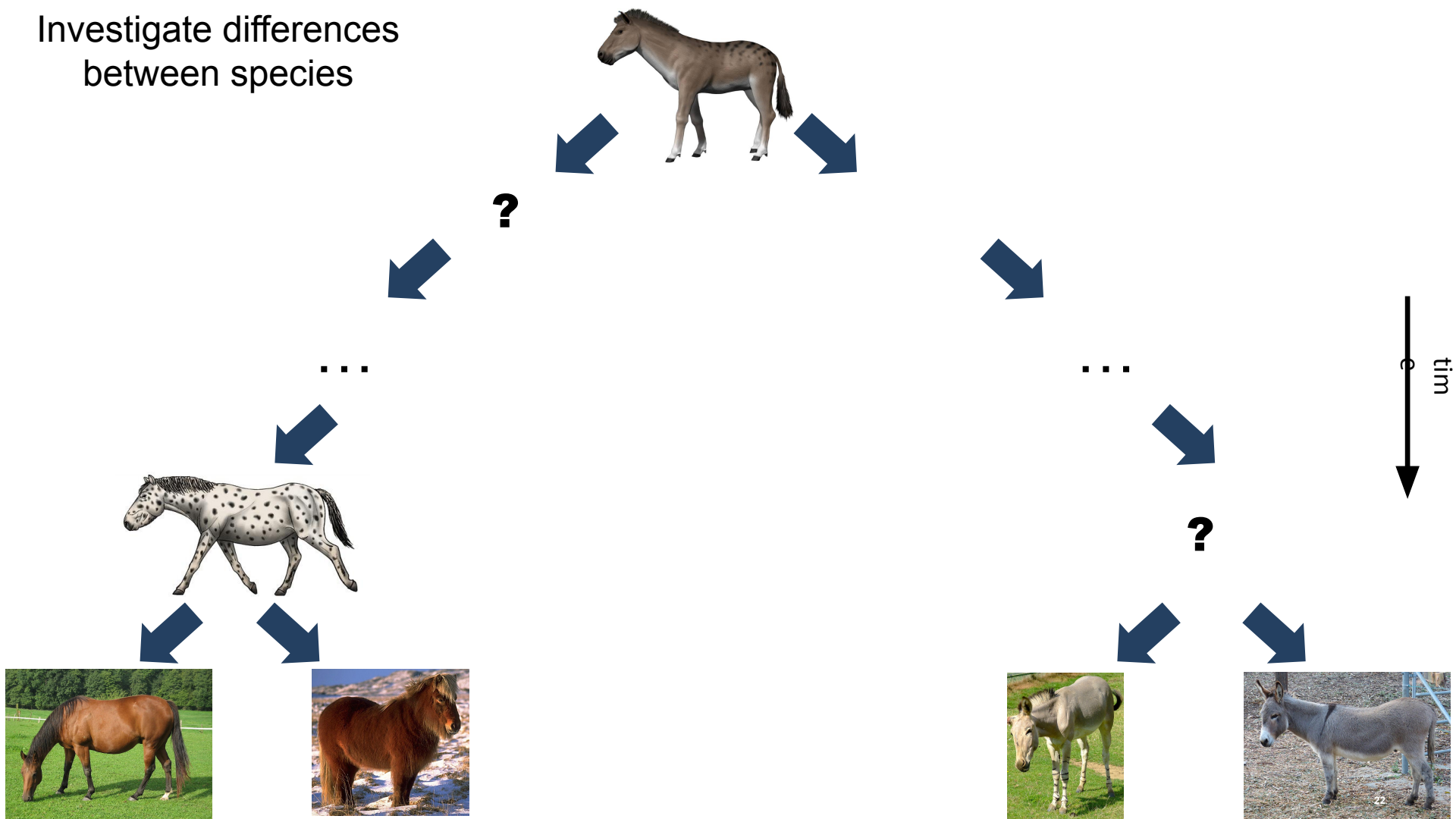


Investigate differences within a species



AGGTTATTGGTAGT
AAGATTATCGGTACT

Investigate differences between species



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

Theodosius Dobzhansky, 1973

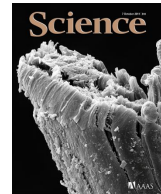
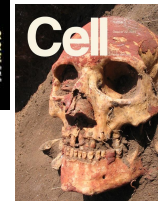
"Nothing in ~~Biology~~ Makes Sense Except in the Light of Evolution"

NGS

me, I made that up just now

What can we use it for?

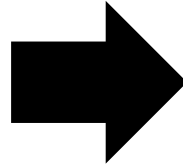
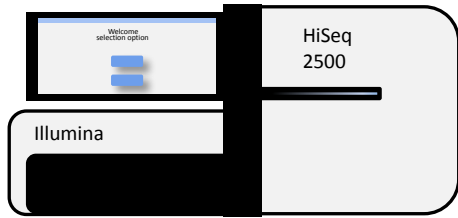
- Whole genome re-sequencing
- Population genomics
- Diagnostics
- Cancer genomics
- Ancient genomes
- Metagenomics
- RNA sequencing
- Single cell sequencing
- Genomic Epidemiology
- anything with DNA



Basic concepts

3 key concepts

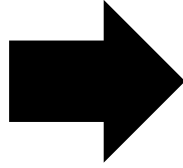
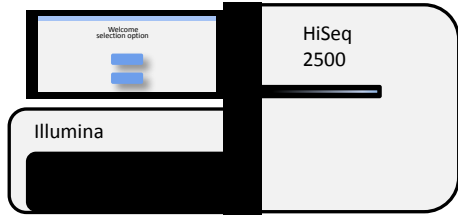
- Read length
- Throughput
- Types of errors



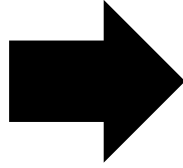
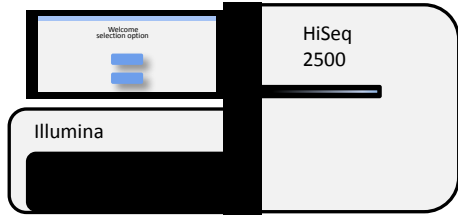
read length



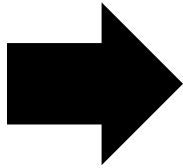
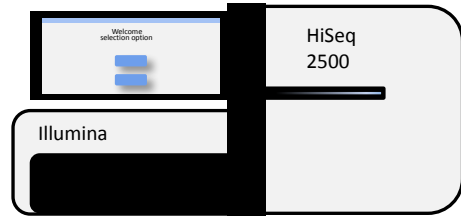
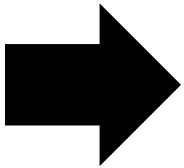
throughput def. 1



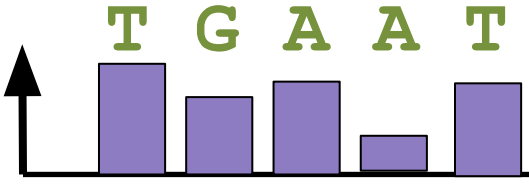
throughput def. 2



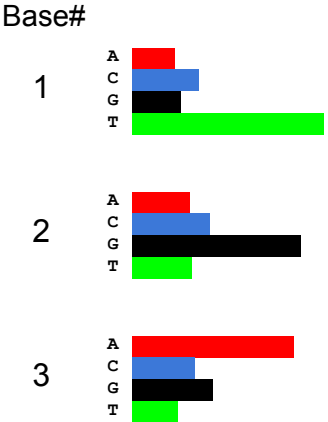
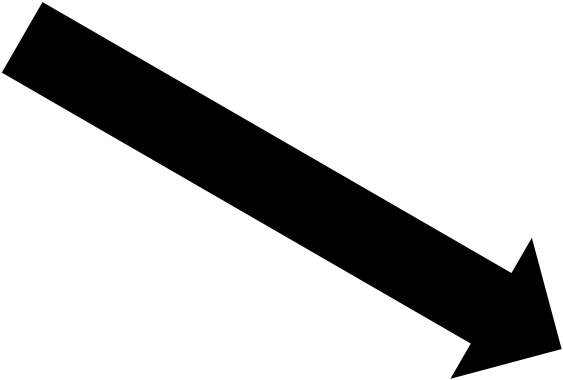
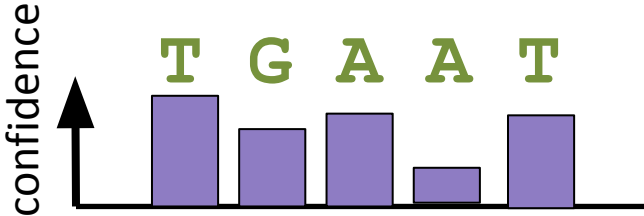
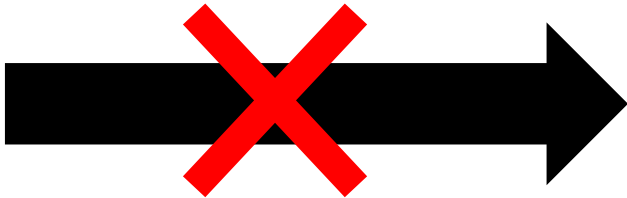
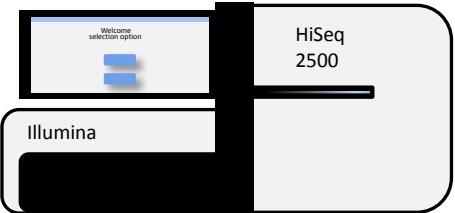
template



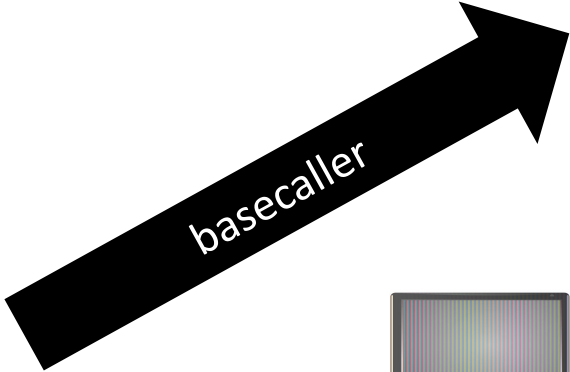
confidence



Key concept: basecalling




Raw Data



mismatch

template

AGCAATCTCAATTACAAAATATACACCAACAAA
AGCAATCTCAATTACAGATATACACCAACAAA

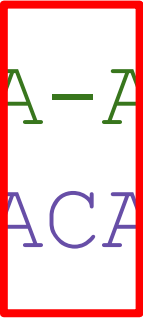


read

insert

template

AGCAATCTCAATTACA-~~A~~ATATACACCAACAA
AGCAATCTCAATTACACAATATACACCAACAA



read

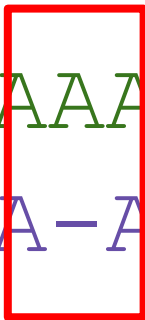
deletion

template

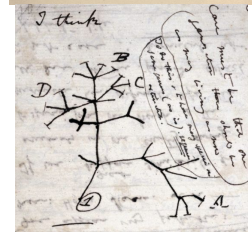
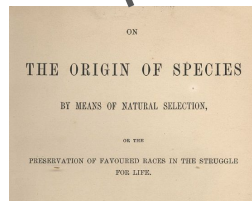
AGCAATCTCAATTACAAATATACACCAACAA

AGCAATCTCAATTACA-ATATACACCAACAA

read



1859



1900

Discovery of chromosomes
1880s

Versuche über Pflanzenhybriden.
Von
Gregor Mendel.1)

Einleitende Bemerkungen.
Befruchtungen welche an Ziernpflanz

1866

1941

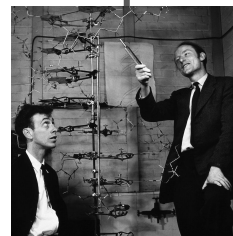
DNA is the molecule of heredity
1928-30s

Early statistical genetics
1920s



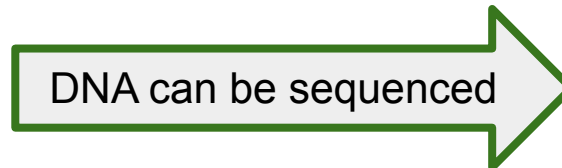
Wright Fischer Haldane

1953



DNA structure published
1953

1977



2024

1977 1983 1989 1995 2001 2006 2012 2018 2024



DNA can be sequenced

1977

1983

1989

1995

2001

2006

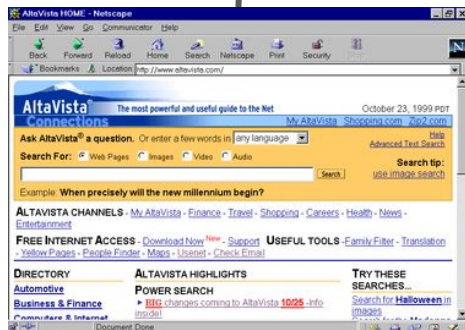
2012

2018

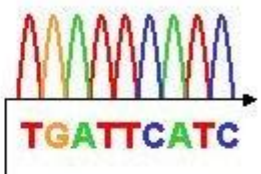
2024

STAR WARS

THE LORD OF THE RINGS



1977 1985 1989 1995 2001 2006 2012 2018 2024



Sanger



Illumina



SOLID



Ion Torrent



Oxford Nanopore



PacBio



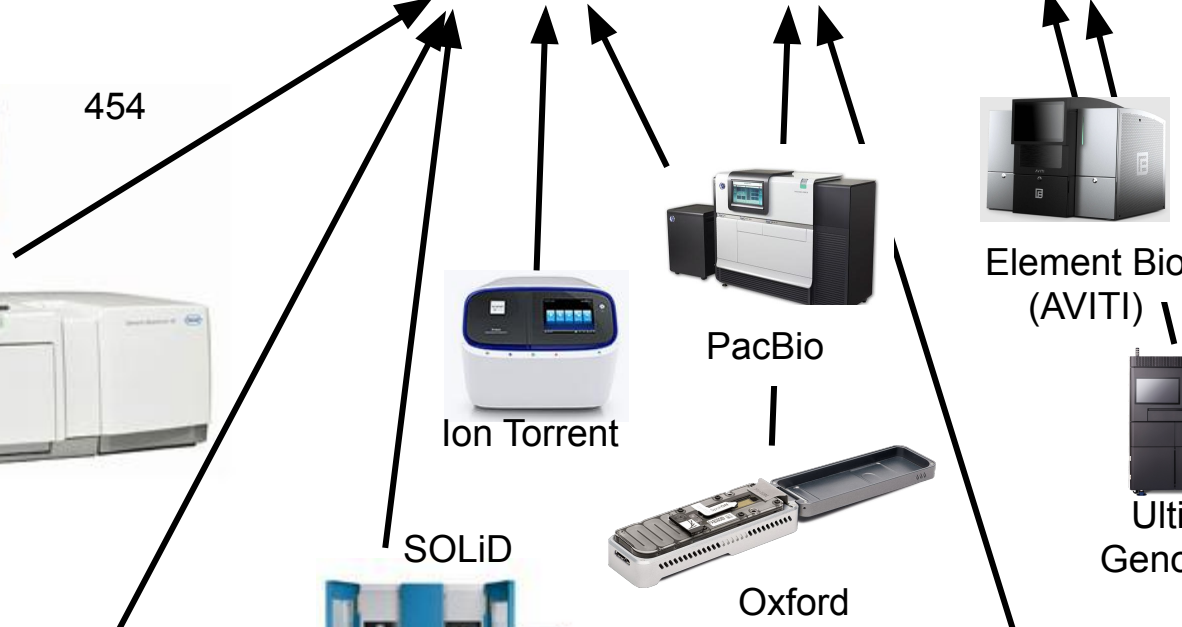
BGI



Element Bio (AVITI)

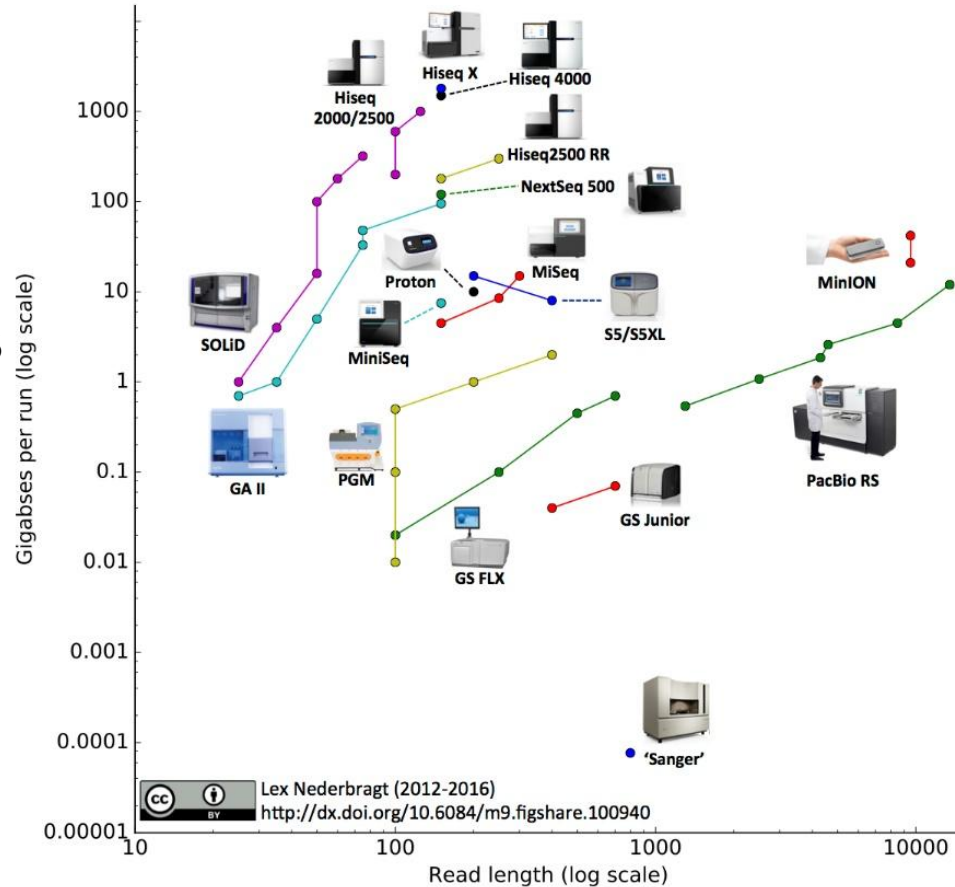


Ultima Genomics

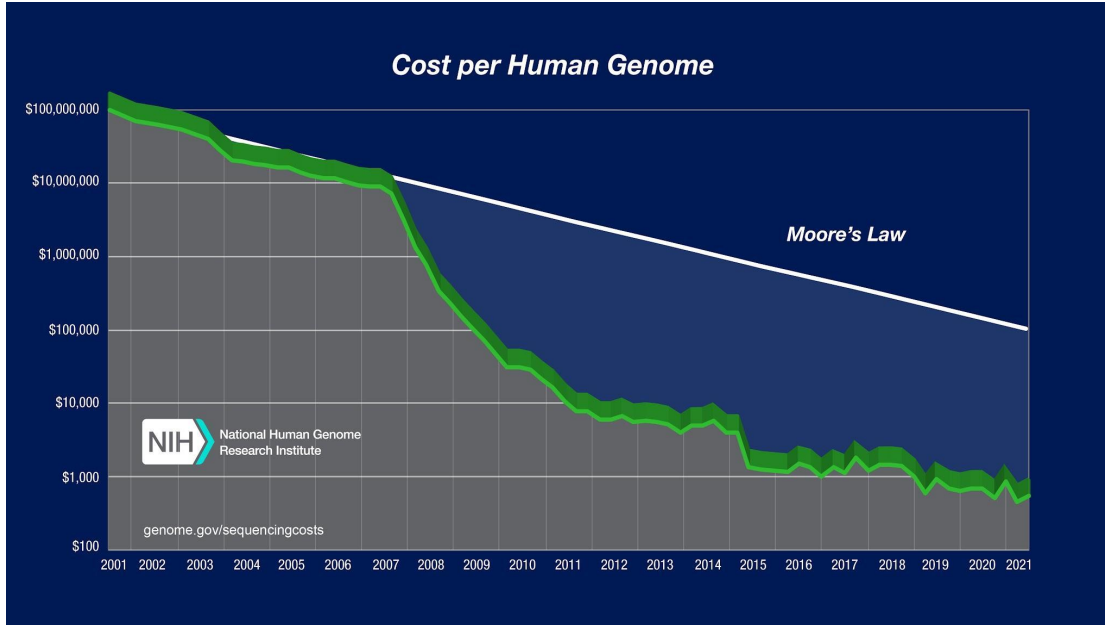


1st generation of NGS

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



Sequencing costs



NHGRI - Sept 2022

- Computer speed and storage capacity is **doubling every 18 months** and this rate is steady
- DNA sequence data **is doubling faster than computer speeds!**



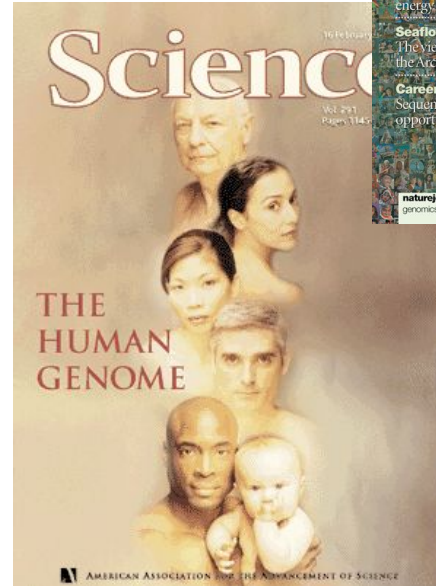
1990 - 2003

Picture: The Guardian

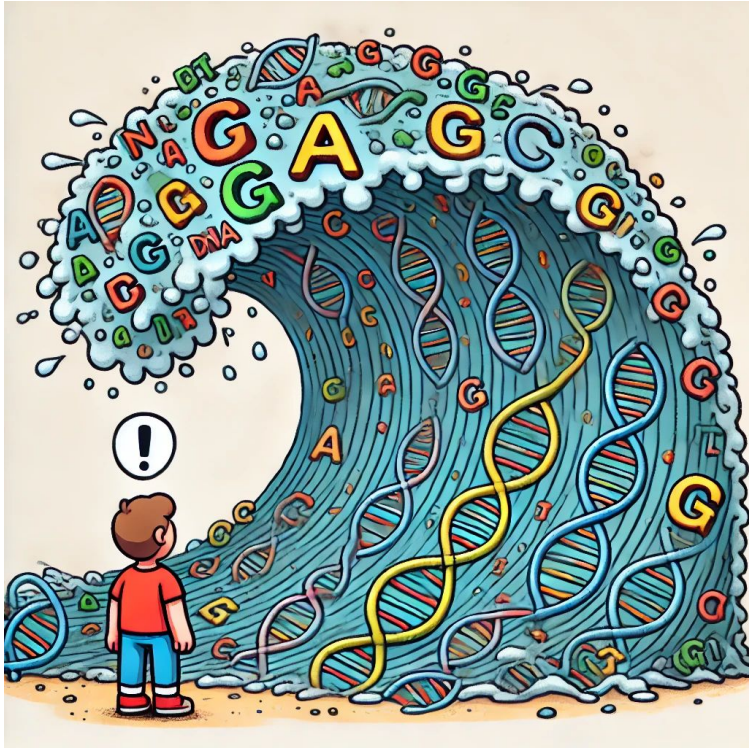
~5.54G USD (adj. for FY 2022)
20 research centers, 6 countries

Human sequencing

- First draft genome of human in 2001, final 2004
- Estimated costs \$5.54 billion USD, time 13 years
- Today:
 - 200-1000 USD for one genome
 - A couple of days
 - Will go down to 100 USD soon



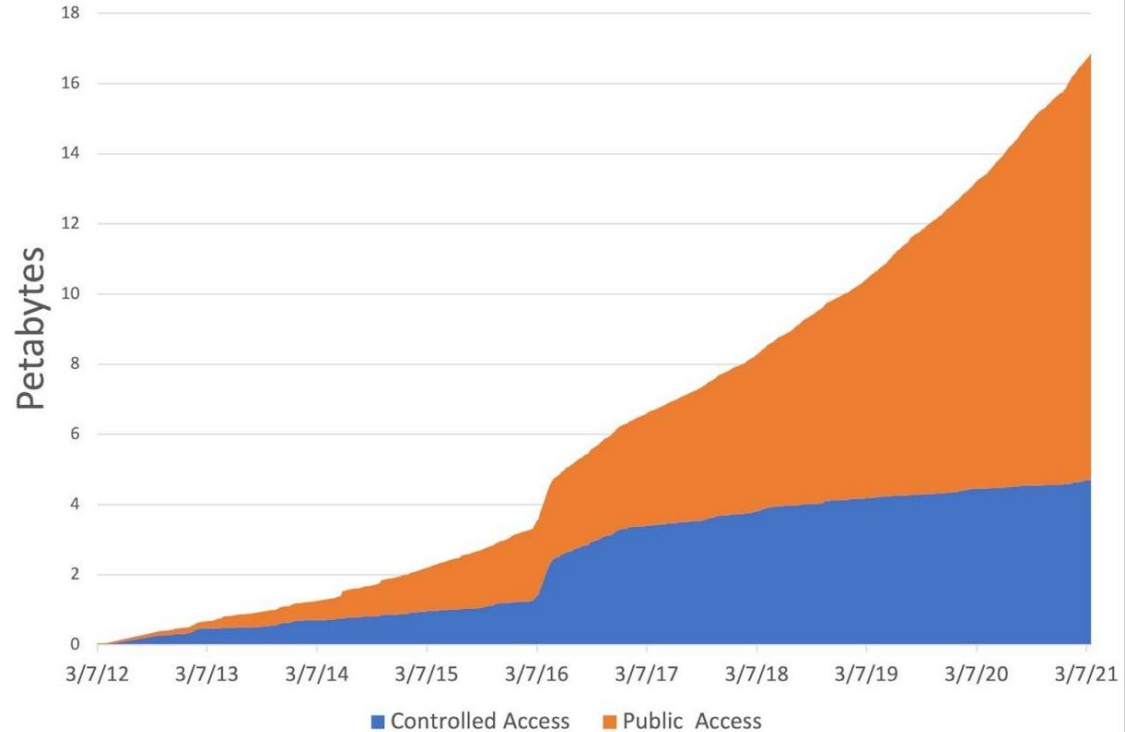
Storage and analysis



- Cost of sequencing is almost less than the cost of **storage and analysis**
- One Illumina NovaSeq system: almost 10k-20k human genomes per year! 1 year+300 NovaSeqs to sequence every person in Denmark...
- A standard human (30-40x) whole-genome sequencing exp. would create 30-150 Gb of data

Sequence Read Archive (SRA)

SRA Growth



What is 16Pb?

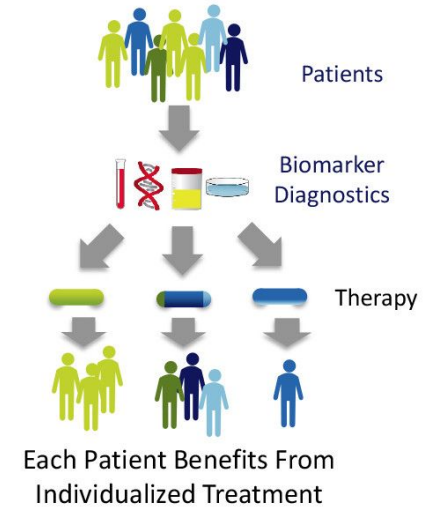
- At 4 GB per HD movie, 16 PB could store approximately 4 million HD movies.
- If you watched these back-to-back (24/7), it would take you about **912 years** to finish them all.

The X Genomes projects

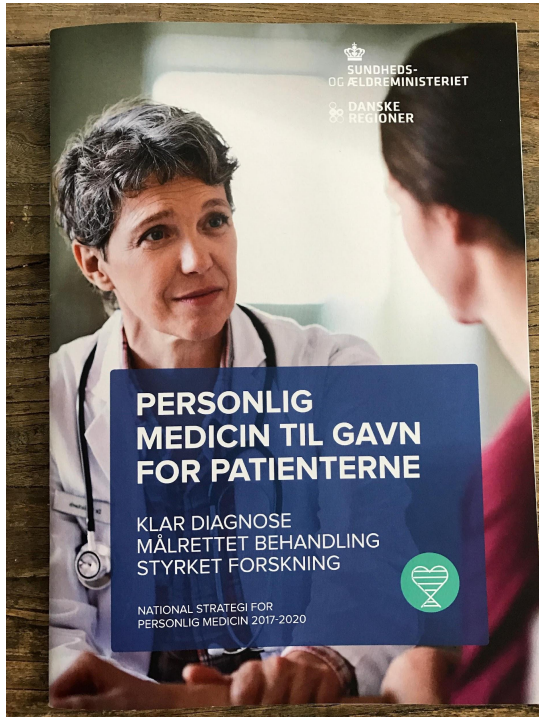
- Human population projects
 - 1000 genomes project (2500 individuals)
 - Genomics England (100k individuals)
 - US Precision Medicine (1 million individuals)
- 100K pathogens project, Earth Microbiome project, Cancer genome project, Plants and animals, insects,...

NGS in the clinic

- Diagnostics of patients (+ fetus)
- Focused treatment of cancer patients
- Sequencing of bacterial isolates
- More and more: mRNA vaccines
- Country-wide projects:
 - UK, UAE, Qatar, Finland, China, ...
 - DK: Danish regions want to sequence 100k individuals
 - USA: 245k individuals, “All of US” published 2024



Personalized medicine

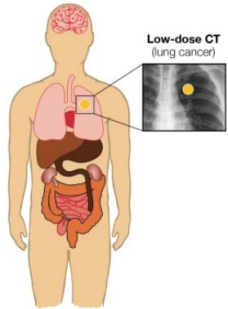


- Giving the same medication to all will not work
- Personalized medicine initiative in DK
- Sequence 100,000 patients on hospitals
- Use extensive registry data
- Current: 100M DKK (estimated 2G DKK)

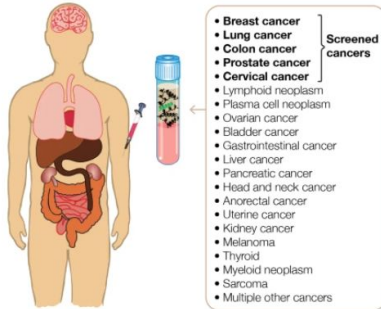
Preventive medicine/prenatal via cell-free DNA

GRAIL

A "One test-one cancer" approach



B "One test-many cancers" approach



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

<https://emea.illumina.com/clinical/reproductive-genetic-health/nip.t.html>

Cancer vaccines

What To Know About Personalized mRNA Cancer Vaccines After Promising Trials From Moderna And Merck

Robert Hart Forbes Staff

Robert Hart is a London-based Forbes senior reporter.

Follow



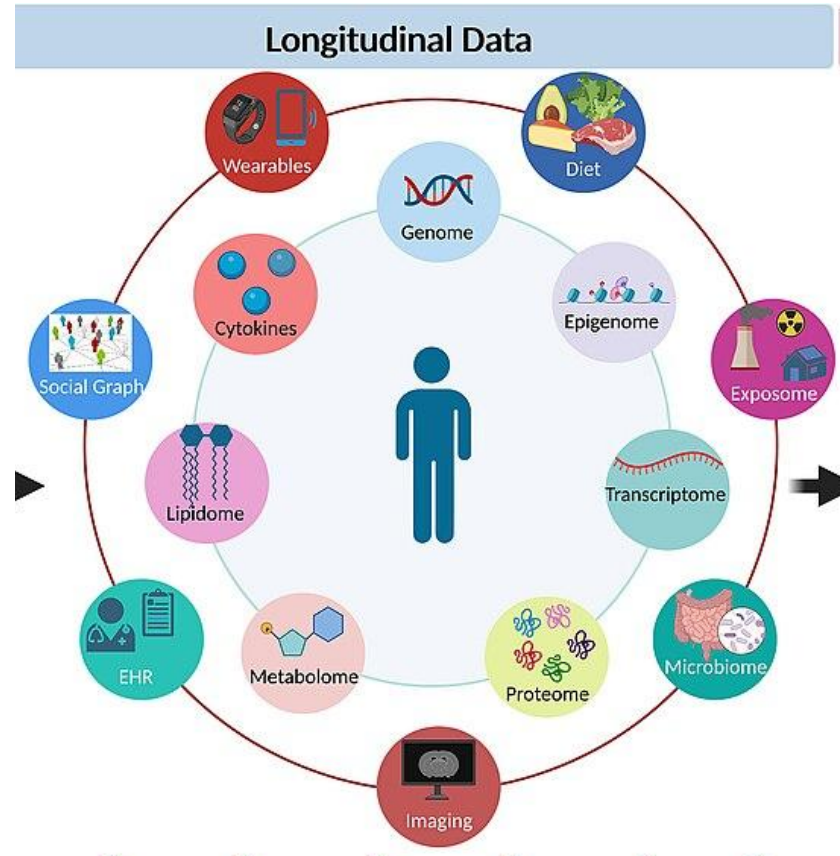
Jun 4, 2024, 11:12am EDT

TOPLINE Cancer vaccines are finally showing promise as Moderna and Merck touted promising data on an experimental skin cancer vaccine and the U.K. announced plans for a “landmark” scheme to test the technology across the country this week, after decades of research that could bring a new era of personalized medicine.



EVAXION

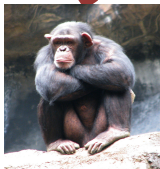
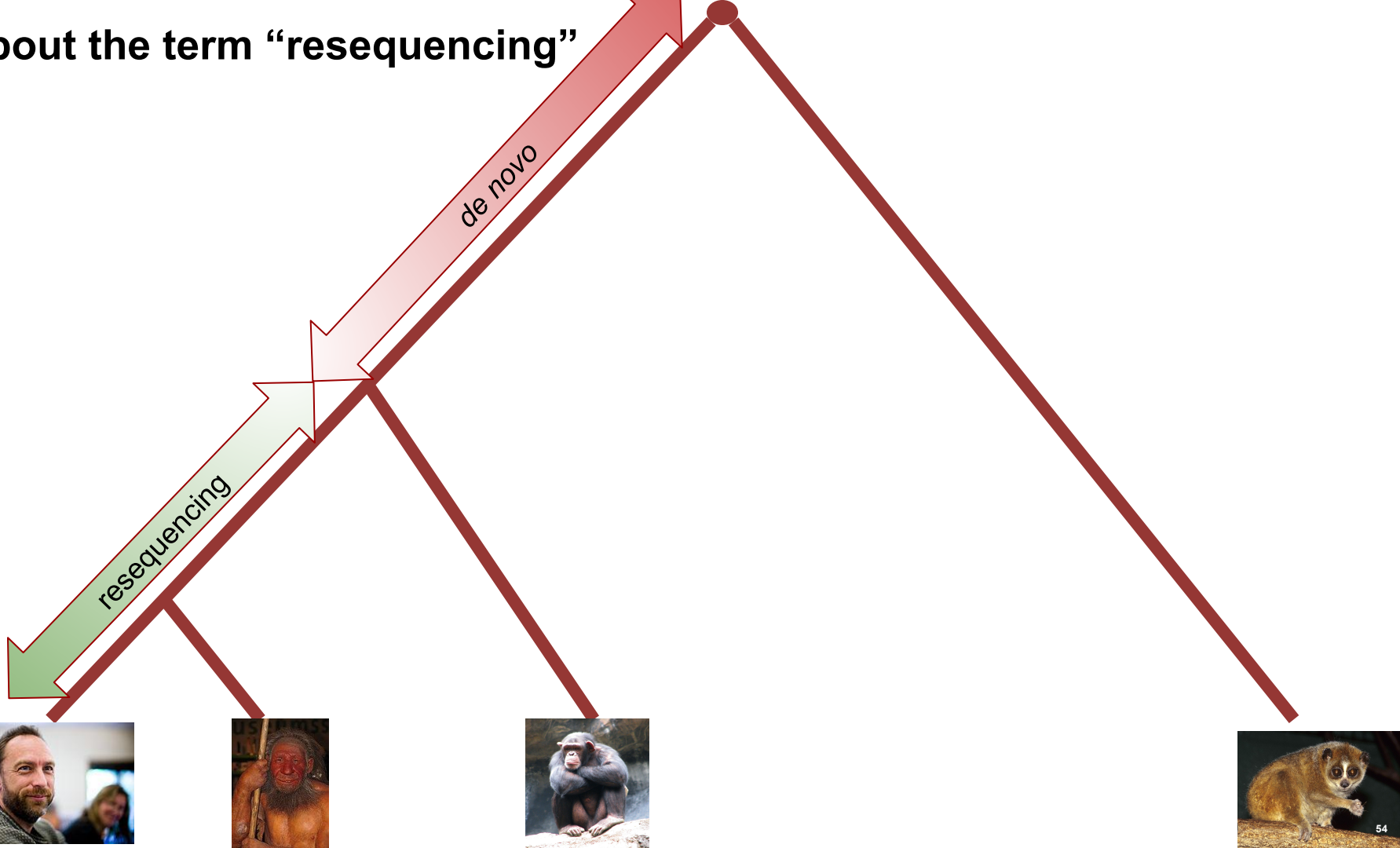
Multimomics: everything, everywhere, all at once



NGS & bioinformatics

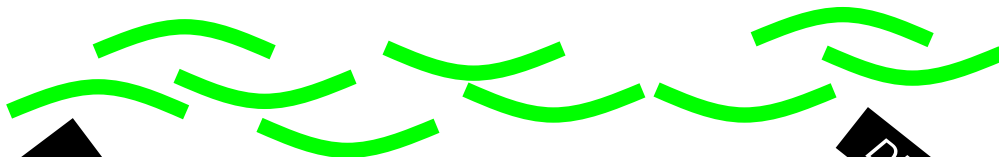
- Extreme data size causes problems
- Just transferring and storing the data
- Need computer clusters, large storage
- Think in fast and parallel programs
- Cloud computing increasingly used

About the term “resequencing”



Whole genome sequencing

Genome



We cover this on
Wednesday



reference



We cover this on Thursday



new reference

