Introduction to bioinformatics

leo.planche@universite-paris-saclay.fr

Course Overview

i. Introduction to genetics (Today)

ii. Introduction to population genetics

iii. Data analysis (Python)

Today’s Overview

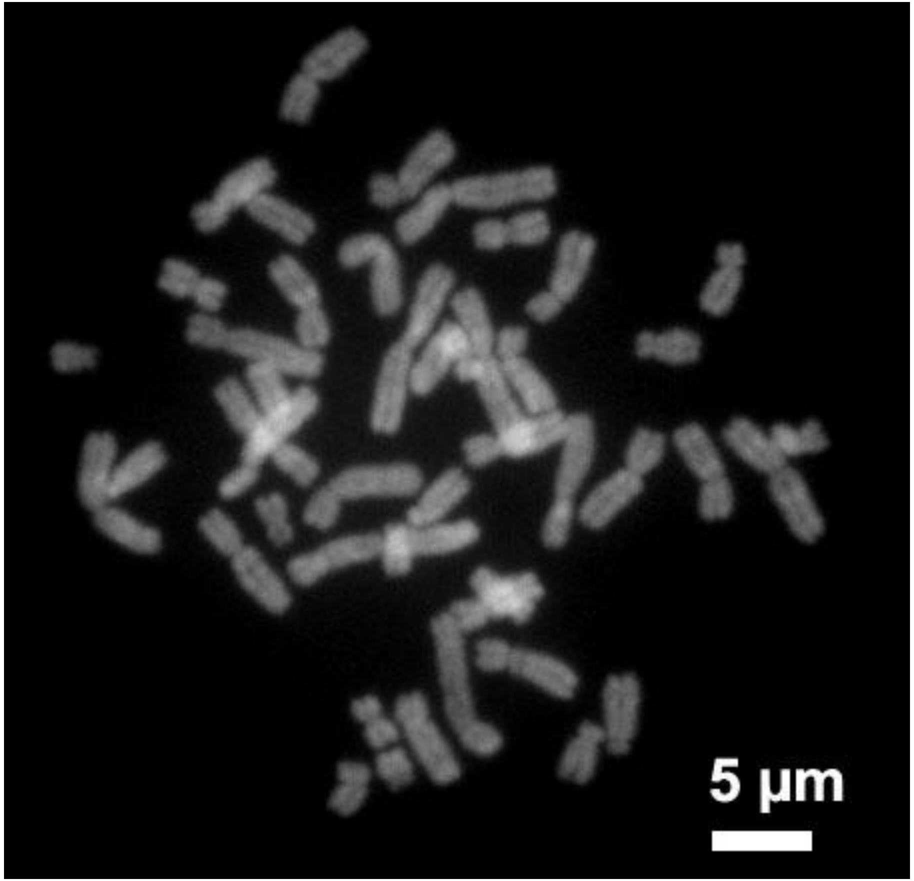
i. From DNA to Phenotypes

ii. A brief history of genetics

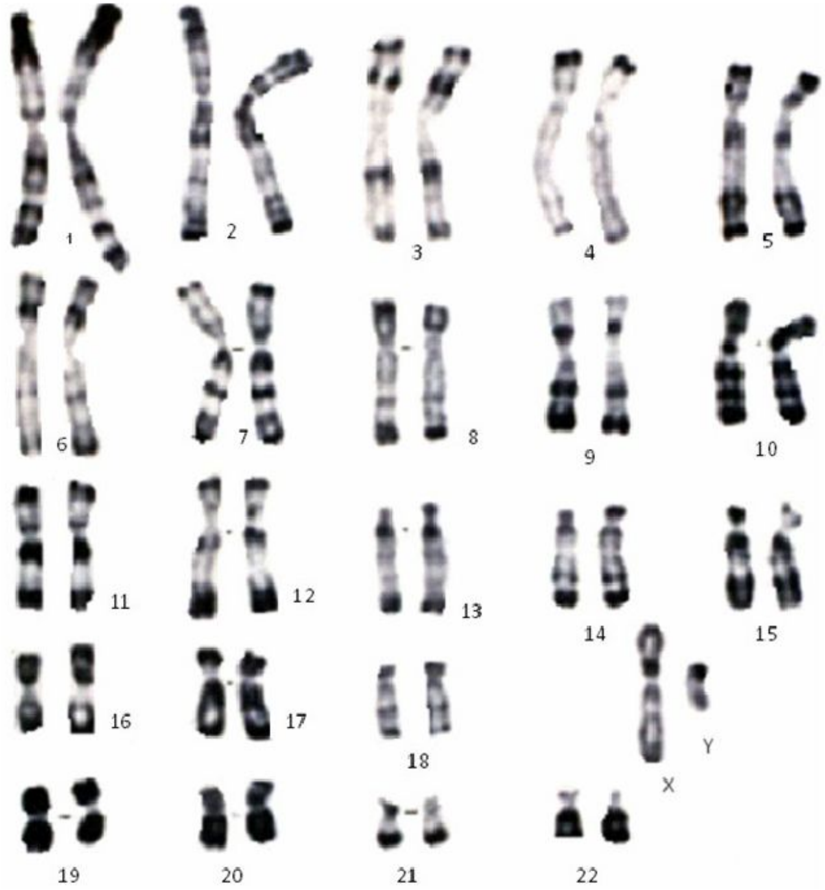
iii. Genetic analysis task: phylogeny

DNA

DNA

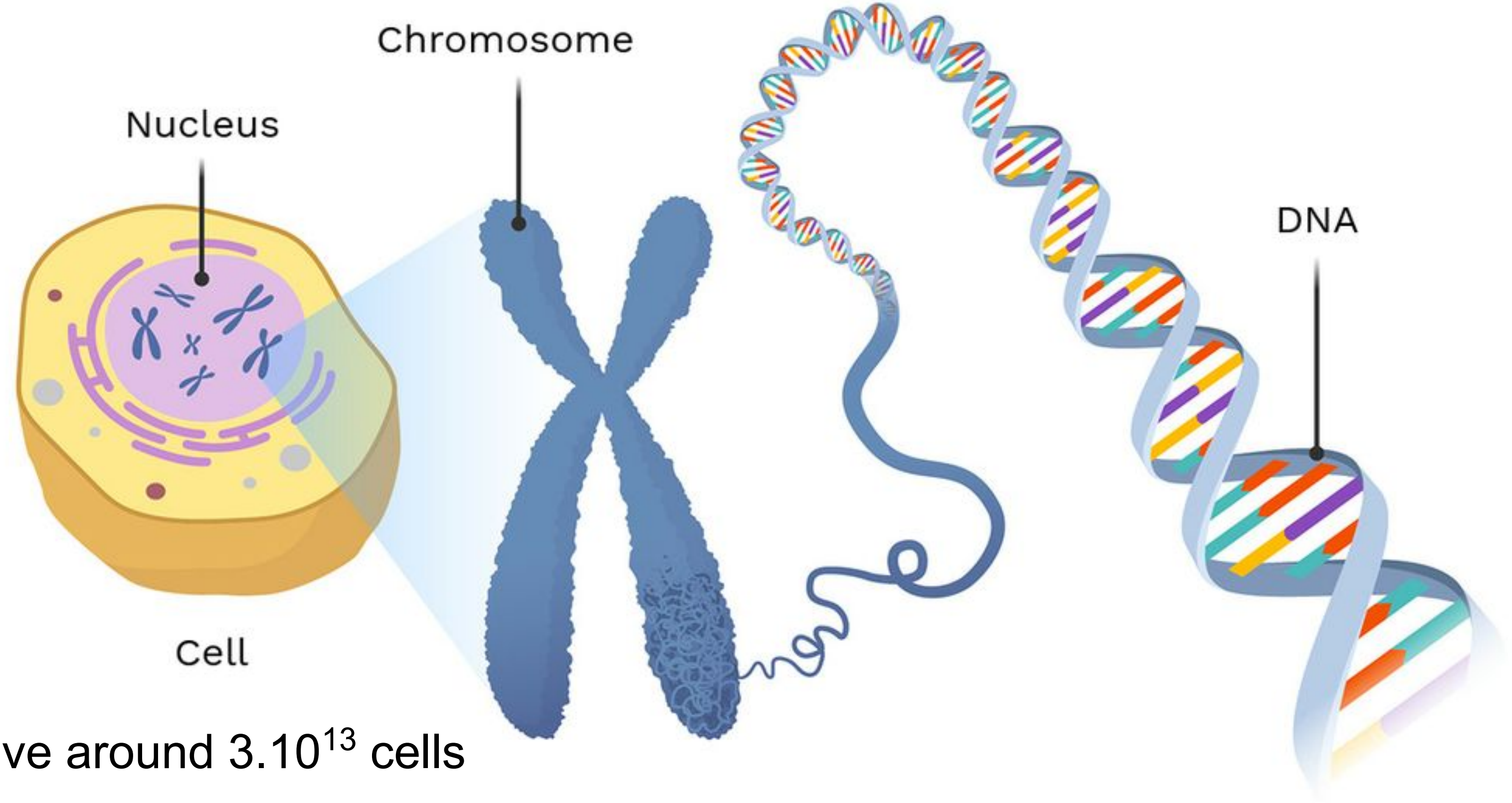


Human genome



- 46 chromosomes (sequences of DNA)
- Chromosomes come by pair, we say that humans are **diploids**
- For each pair, one chromosome comes from the father, one comes from the mother

Where is the DNA present?



Humans have around 3.10^{13} cells

What is the DNA made of?

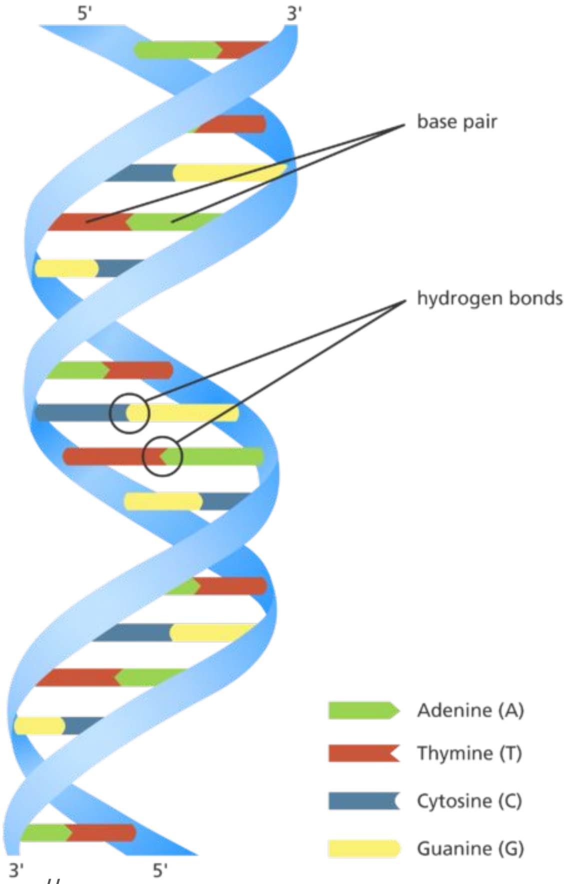


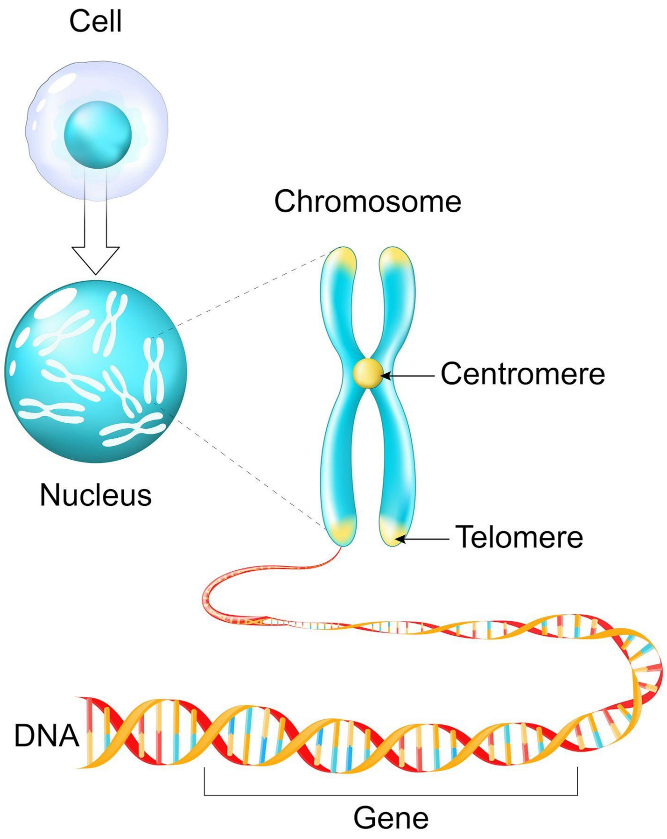
Image: https://www.yourgenome.org

- DNA stands for deoxyribonucleic acid
- DNA is a sequence of **nucleotides** (A, C, G, T) whose bases pair across two strands: A with T, and C with G.
- Human genome has a length of around 3 billions nucleotides (!)

  War and Peace has around 500,000 words or 2 millions letters.
  An HD image has 1920×1080 = 2,073,600 pixels.

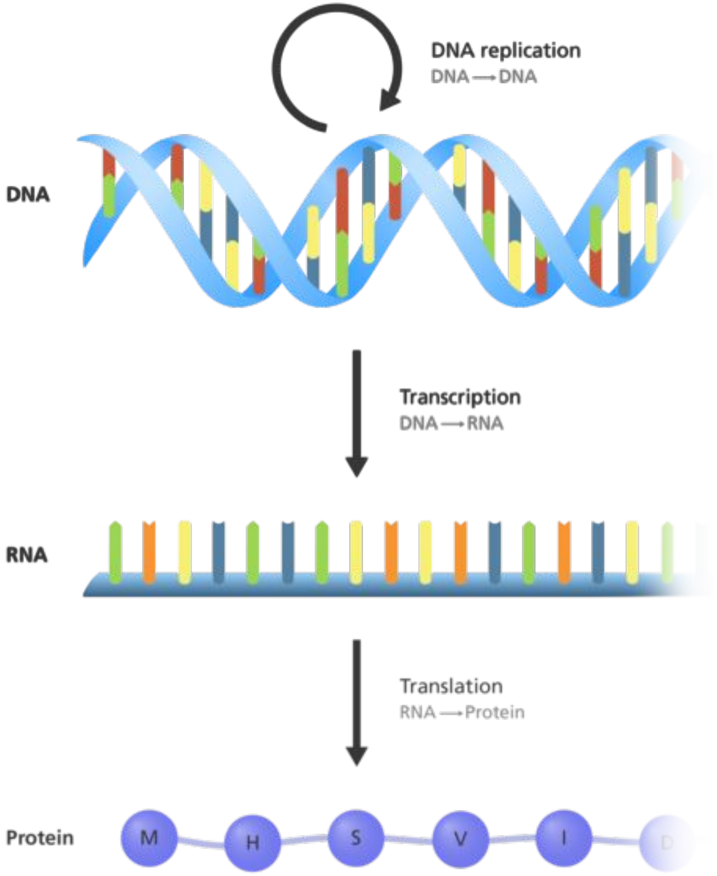
READING THE ENTIRE GENOME CODE...
Listen to this part: 'AGGCTAATC CGCATAACTG'
Wow!
VOL 9
S. Harris

What is the purpose of DNA? What is a gene?



- A gene is a **genomic sequence** that codes primarily for a protein
- Most of human DNA is non-coding (>98%) (!)
- A gene if a sequence of length around 1500 nucleotides

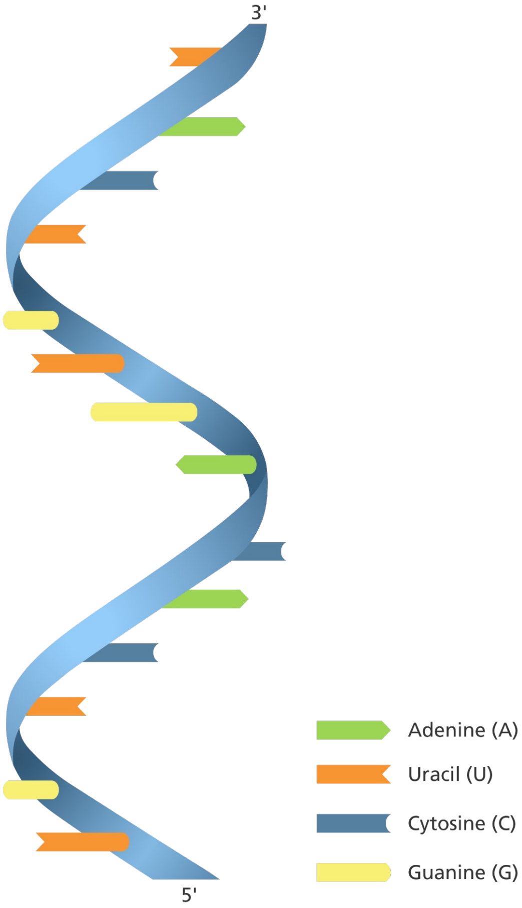
What is the purpose of DNA? How does it work?



Central dogma of molecular biology:

"DNA makes RNA, and RNA makes protein"

What is the RNA made of?



- RNA stands for ribonucleic acid
- RNA is a sequence of nucleotides (A, C, G, **U**)
- RNA molecules perform diverse roles, information transfer, catalysis, regulation, and forming the core of the protein-production machinery (mRNA).

Transcription: DNA→RNA

DNA: A T G G A T A G A C C G T G A C G T A A C

↓ Transcription ↓

RNA: A U G G A U A G A C C G U G A C G U A A C

Translation: RNA→Proteins

RNA: A U G G A U A G A C C G U G A C G U A A C

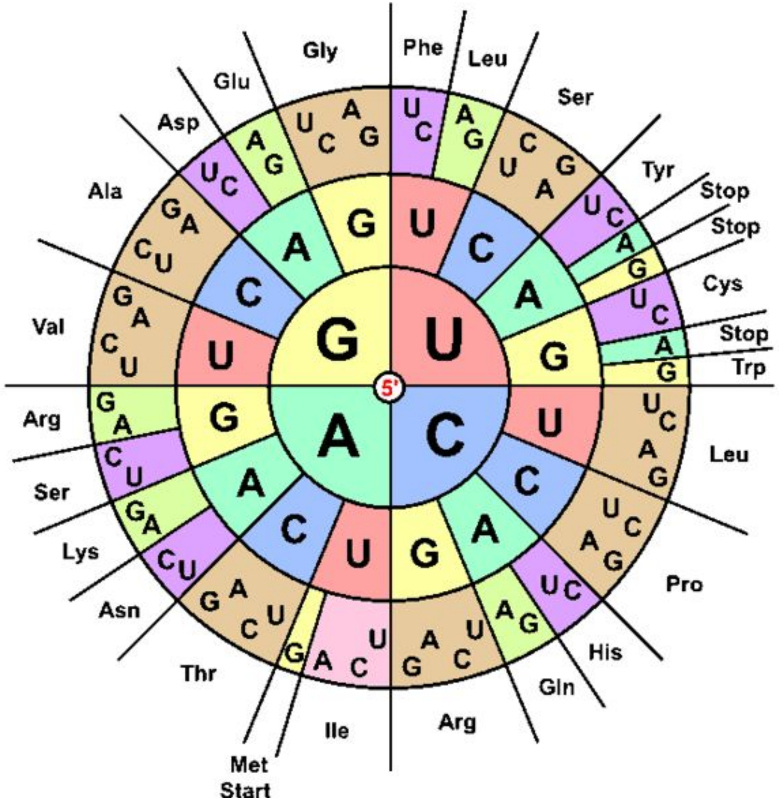
Translation: RNA→Proteins

RNA: A U G G A U A G A C C G U G A C G U A A C

codon

↓ Translation ↓

Protein: Met



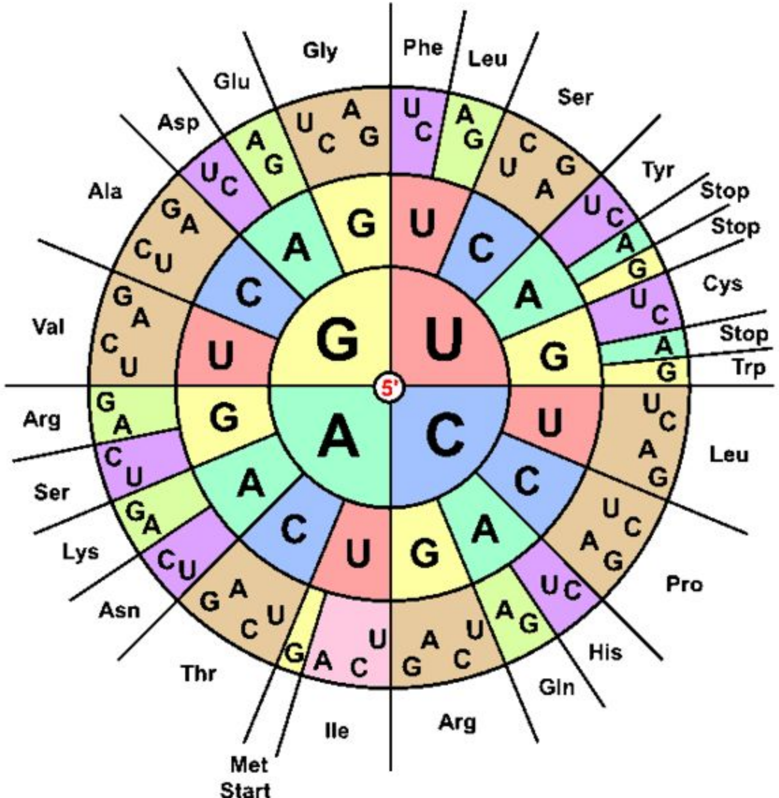
Translation: RNA→Proteins

RNA: A U G G A U A G A C C G U G A C G U A A C

codon

↓ Translation ↓

Protein: Met — Asp

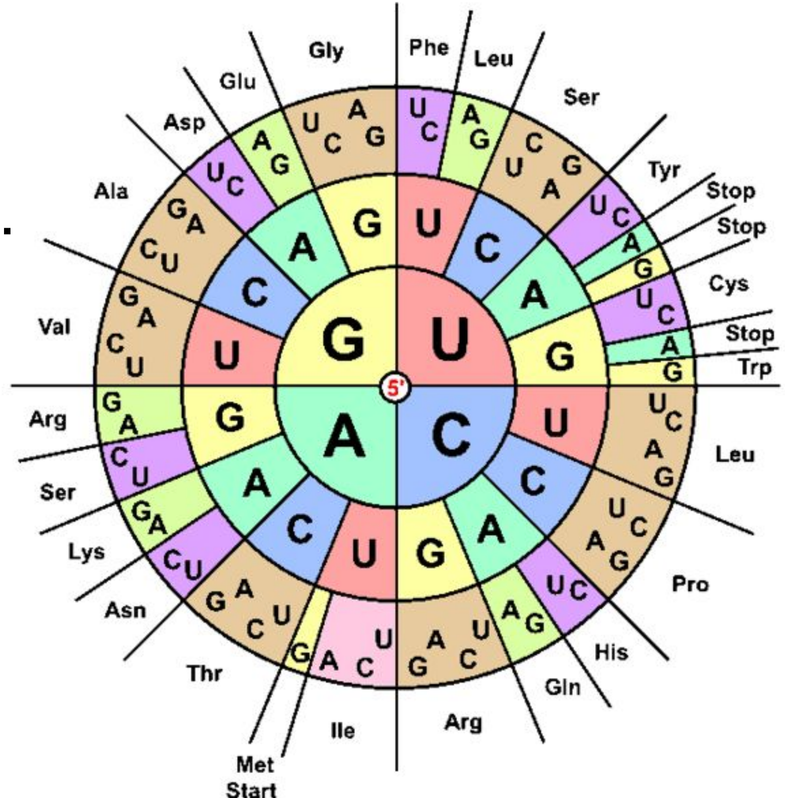


Translation: RNA→Proteins

RNA: A U G G A U A G A C C G U G A C G U A A C

↓ Translation ↓

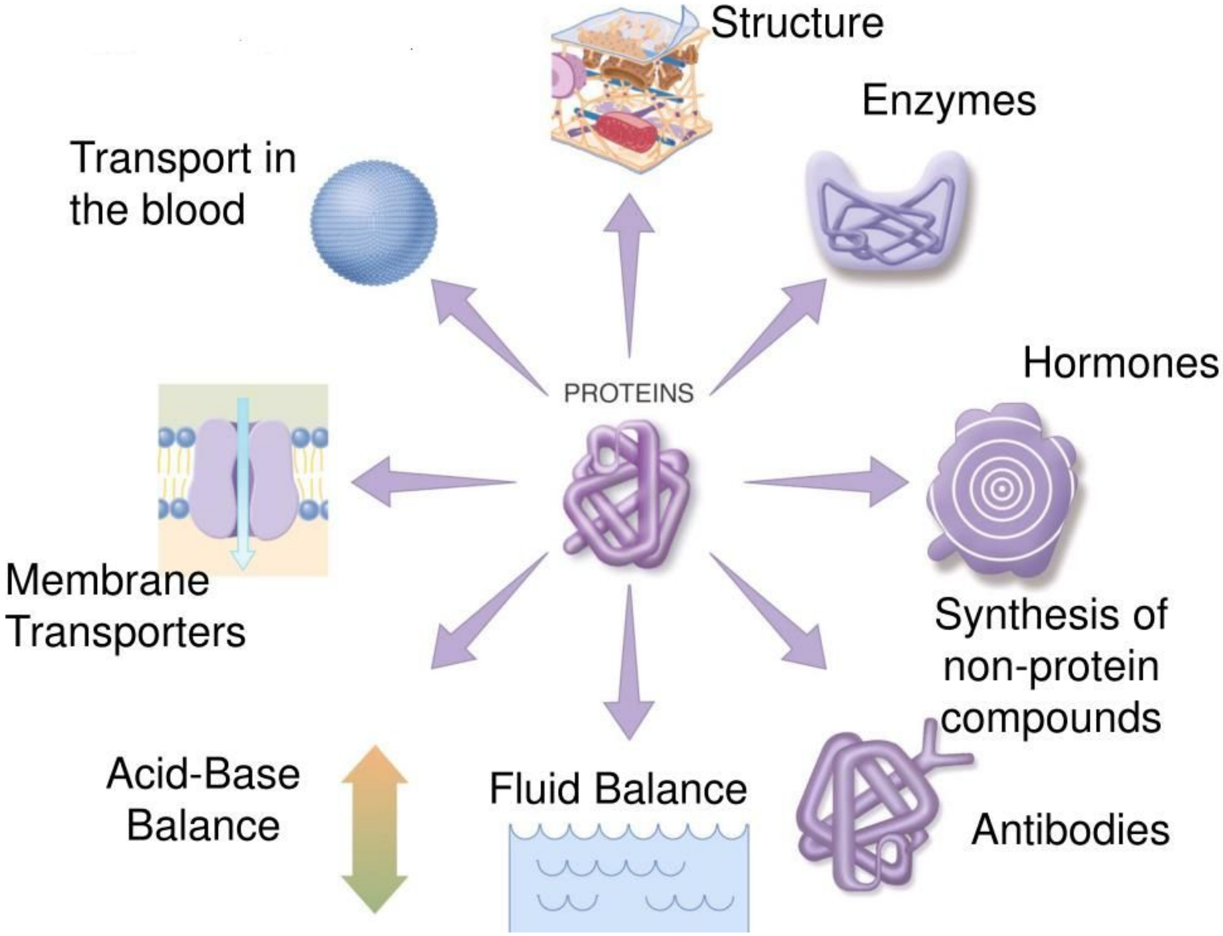
Protein: Met — Asp — Arg — …



A protein is a sequence of amino acids, it starts with the codon “Start” (Met) and it stops with the codon “Stop”

There are 24 amino acids that compose proteins

What do proteins do?

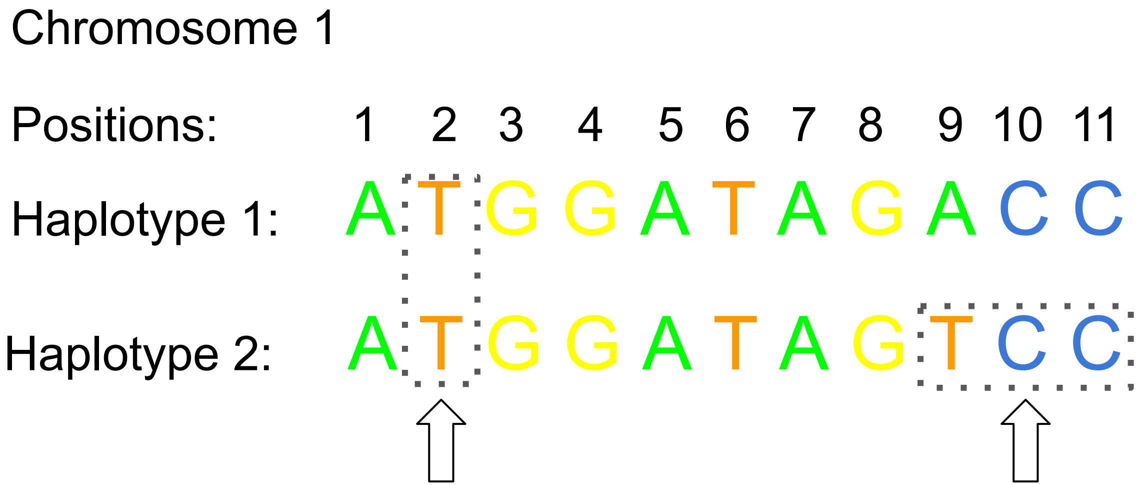


Summary

- DNA is present in (nearly) all cells.
- Human DNA is divided into 46 chromosomes (22 pairs +X/Y), each chromosome is one long sequence of DNA.
- A sequence of DNA can be simply seen as one potentially very long text over the alphabet (A,C,G,T).
- One important purpose of DNA is to code for proteins, a sequence of DNA which codes for a protein is called a gene.
- First DNA is translated into mRNA which is then translated into a chain of amino acids, which forms a protein.
- Proteins perform the majority of functional roles in cells and the body.

Genotype and phenotype

Genotype



Chromosome 1

Positions:	1	2	3	4	5	6	7	8	9	10	11
Haplotype 1:	A	T	G	G	A	T	A	G	A	C	C
Haplotype 2:	A	T	G	G	A	T	A	G	T	C	C

The genotype at position 2 on chromosome 1 is T/T

The haplotype 2 at position 9-11 on chromosome 1 is TCC

- The **genotype** of an individual is its complete set of genetic material.
- A **haplotype** is a group of **alleles** inherited together on the same chromosome.

Genotype vs Phenotype

Chromosome 1

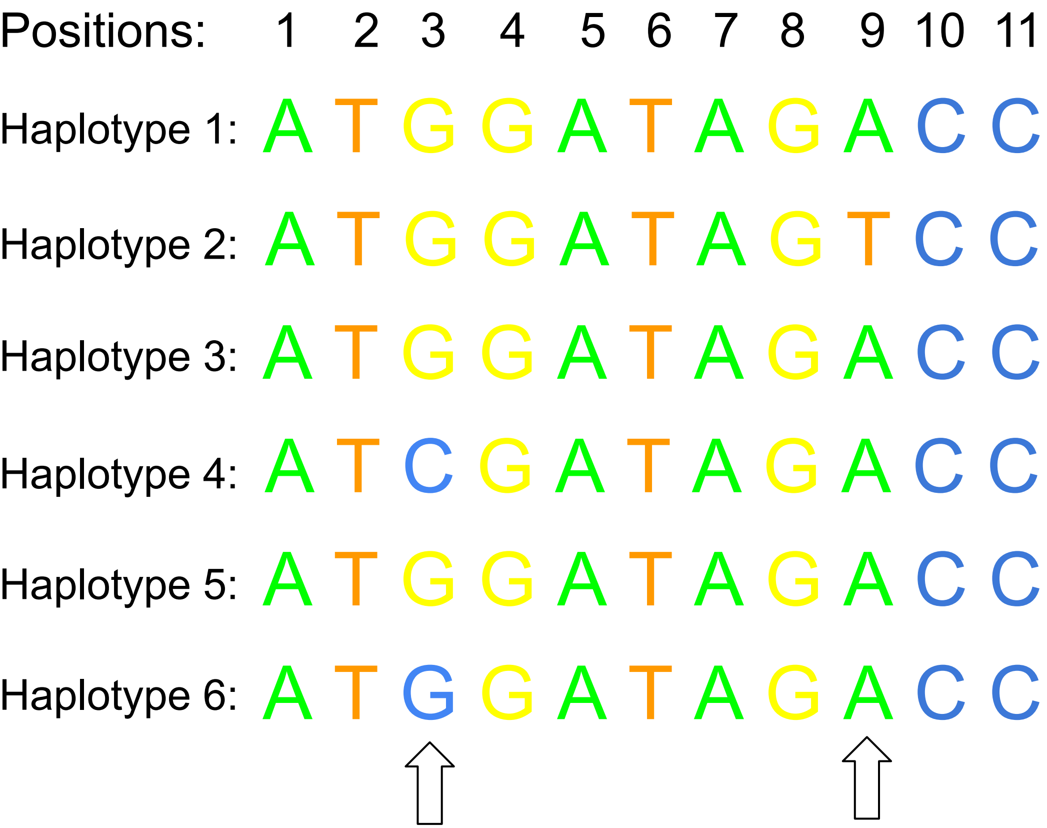
Positions:	1	2	3	4	5	6	7	8	9	10	11
Haplotype 1:	A	T	G	G	A	T	A	G	A	C	C
Haplotype 2:	A	T	G	G	A	T	A	G	T	C	C

- The **genotype** of an individual is its complete set of genetic material.
- The **phenotype** is its traits or characteristics resulting from genes and the environment.

⚠️ Phenotype is not only about visible physical traits, but about most of gene expression, biochemical traits or disease susceptibility.

Variants

Chromosome 1



Positions:	1	2	3	4	5	6	7	8	9	10	11
Haplotype 1:	A	T	G	G	A	T	A	G	A	C	C
Haplotype 2:	A	T	G	G	A	T	A	G	T	C	C
Haplotype 3:	A	T	G	G	A	T	A	G	A	C	C
Haplotype 4:	A	T	C	G	A	T	A	G	A	C	C
Haplotype 5:	A	T	G	G	A	T	A	G	A	C	C
Haplotype 6:	A	T	G	G	A	T	A	G	A	C	C

Single-nucleotide polymorphism (SNPs)

- Around 99.5% of the genome is identical for all humans. The positions that are not always identical are called **SNPs**.
- The genome of two humans will be around 99.9% identical.
- For a lot of research only SNPs matter, reducing the size of the genome to around 10-70 millions. (!)

Reference genome

Position	Ref	Alt	Hap1	Hap2	
1	A	.	A	A	Homozygous reference
2 3	T	.	T	T	
4 5	G	.	G	G	
6 7	G	C	G	G	
8 9	A	.	A	A	
10	T	.	T	T	
	G	A	A	A	Homozygous alternative
	G	.	G	G	
	T	A	A	T	Heterozygous
	C	.	C	C	

- The **human reference genome** is a representative DNA sequence for comparison and analysis.
- An allele is called **reference** if it matches the human reference genome at that position, and **alternative** if it differs.
- An individual is **homozygous** at a position if both its alleles are identical, **heterozygous** otherwise.

Summary

- The genotype of an individual is its complete set of genetic material.
- Through interaction with the environment, the genotype determines the phenotype i.e. the individual’s observable traits, including physical features, biochemical traits, and susceptibility to disease.
- Most positions (>99%) in the genomes have the same allele for all humans, the ones that do not are called SNPs and are responsible for the observable differences.
- The human reference genome is a representative DNA sequence for comparison and analysis.

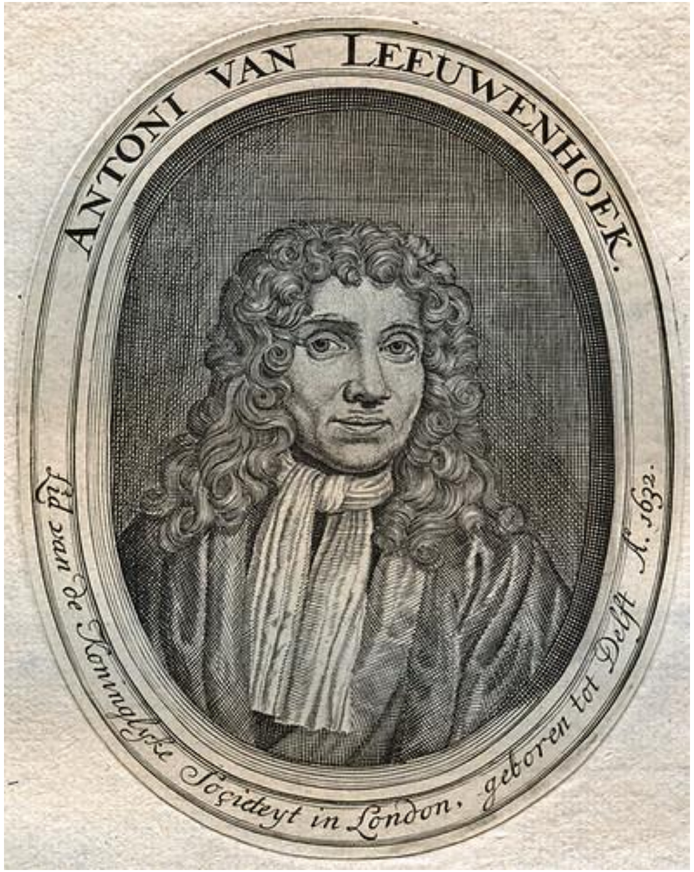
A brief history of genetics

Timeline

- 1683 - discovery of bacteria
- 1859 - Darwin’s *On the Origin of Species*, natural selection
- 1865 - Mendel’s laws
- 1953 - double helix suggested by Watson-Crick *
- 1955 - discovery of DNA and RNA polymerase by Arthur Kornberg *
- 1978 - first method to sequence DNA (Frederick Sanger **) and sequencing of first genome (5kb virus)
- 1983 - invention of PCR. Polymerase chain reaction (PCR) is a method that amplifies a specific DNA sequence, generating millions of copies for analysis. **
- 2000 - draft of the first human genome
- 2012 - CRISPR gene editing *

* Nobel prizes

1683 - discovery of bacteria

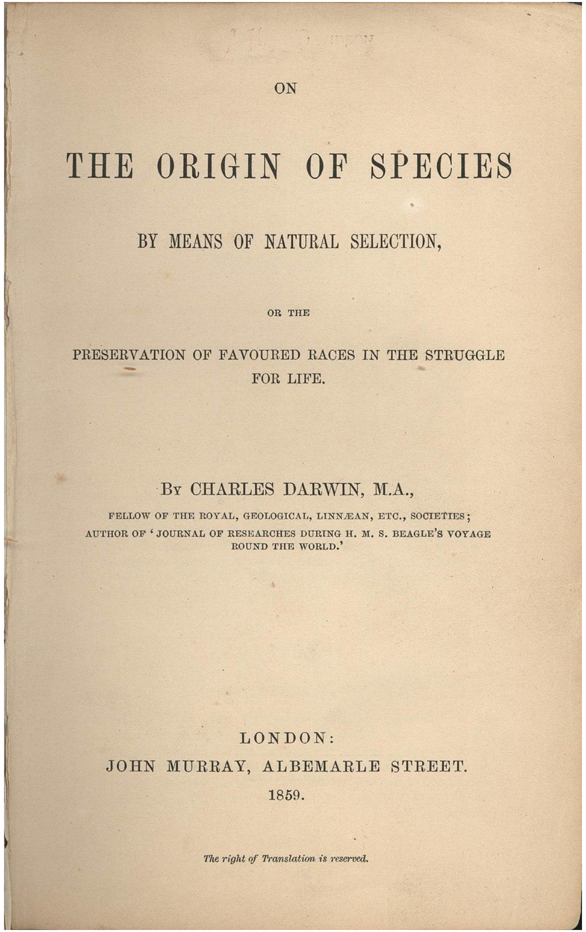

ANTONI VAN LEEUWENHOEK.
Lid van de Konninglyke Societeyt in London, geboren tot Delft A. 1632.

Antonie van Leeuwenhoek (1632-1723)

- Improved and created lenses for microscopes
- First to observe and describe bacteria and other microscopic organisms.
- For this reasons considered “the Father of Microbiology"

Note: Genomics is classified as “Molecular Biology”

1859 - Darwin’s *On the Origin of Species*, natural selection



ON

THE ORIGIN OF SPECIES

BY MEANS OF NATURAL SELECTION,

OR THE

PRESERVATION OF FAVOURED RACES IN THE STRUGGLE
FOR LIFE.

BY CHARLES DARWIN, M.A.,

FELLOW OF THE ROYAL, GEOLOGICAL, LINNÆAN, ETC., SOCIETIES;
AUTHOR OF ‘JOURNAL OF RESEARCHES DURING H. M. S. BEAGLE’S VOYAGE
ROUND THE WORLD.’

LONDON:
JOHN MURRAY, ALBEMARLE STREET.
1859.

The right of Translation is reserved.

- Species evolve and share common ancestors.
- Natural selection drives adaptation.
- Individuals in a population vary significantly from one another, these **variations accumulate over time and are heritable**

Lamarck vs Darwin (simplified):
Individual adaptations to the environment drive evolution vs random variation filtered by natural selection drives evolution.

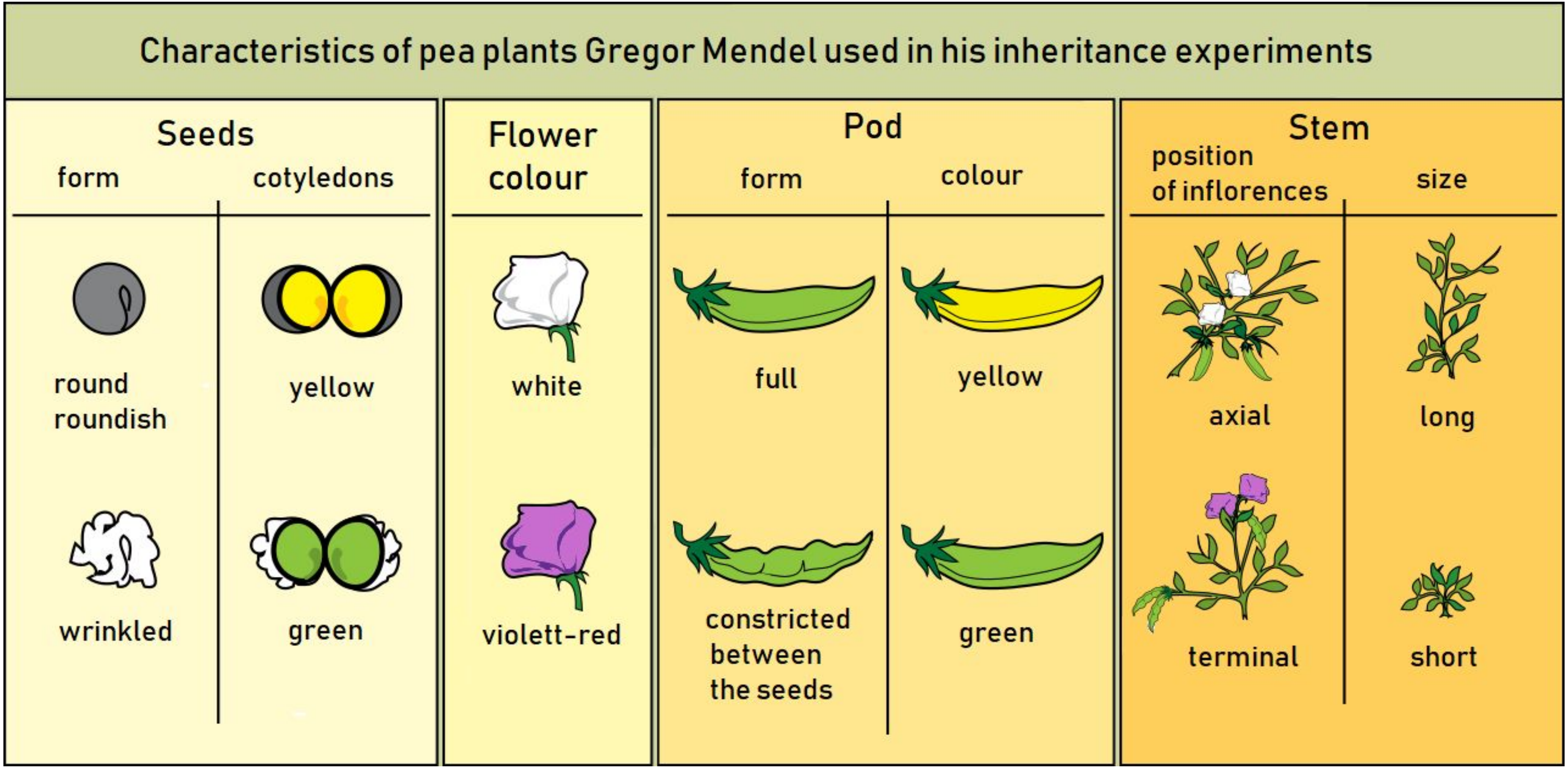
Genetics mostly confirm Darwin’s theory.

1865 - Mendel’s laws

“Gregor Mendel, the Moravian Augustinian friar who founded the modern science of genetics” - Wikipedia

- Made a large number of experiments with plants to better understand heredity in 1860s
- Reported his work in 1865
- Work remained unknown for 35 years
- His results are now known as “**Mendel's laws of inheritance**”

1865 - Mendel’s laws - Experiments



Characteristics of pea plants Gregor Mendel used in his inheritance experiments

Seeds		Flower colour	Pod		Stem	
form	cotyledons		form	colour	position of inflorences	size
round roundish	yellow	white	full	yellow	axial	long
wrinkled	green	violett-red	constricted between the seeds	green	terminal	short

1865 - Mendel’s laws - Experiments

size
long
short

- When tall plant crossed with short plant, he always got a tall plant. He used the term “**dominant**” for the tall character and “**recessive**” for the short character.
- This was true regardless of which parent (male or female) was tall.
- This confirmed earlier observations that both parents contribute equally.
- He then allowed hybrids to self pollinate.
- He ended up with 787 tall plants and 277 short plants, a proportion of 2.84:1 which is roughly equal to 3:1.

1865 - Mendel’s First Law: Law of Segregation

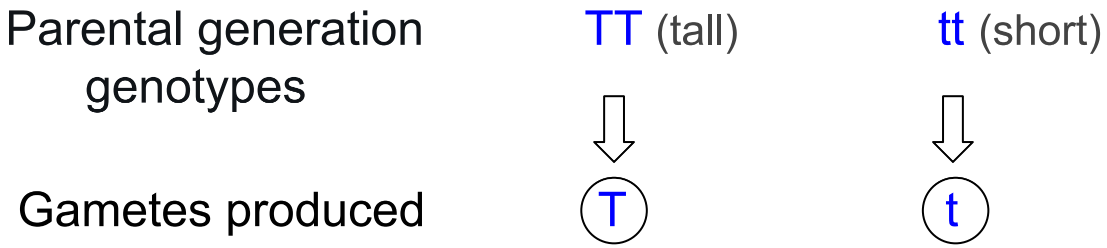
Mendel’s Law of Segregation:
During gamete formation, the alleles for each gene (i.e. tallness or shortness) segregate from each other so that each gamete carries only one allele for each gene.

Note : Mendel wouldn’t have used the term “gene”, in german he used “elementen” which is usually translated in that context as “characteristics”, “traits” etc.

Note 2: His observation went against the popular idea at that time of “blending”, i.e. that a short and a tall plant should give a medium sized offspring

1865 - Mendel’s First Law: Law of Segregation

Mendel’s Law of Segregation:
During gamete formation, the alleles for each gene (i.e. tallness or shortness) segregate from each other so that each gamete carries only one allele for each gene.



Parental generation genotypes

TT (tall)

tt (short)

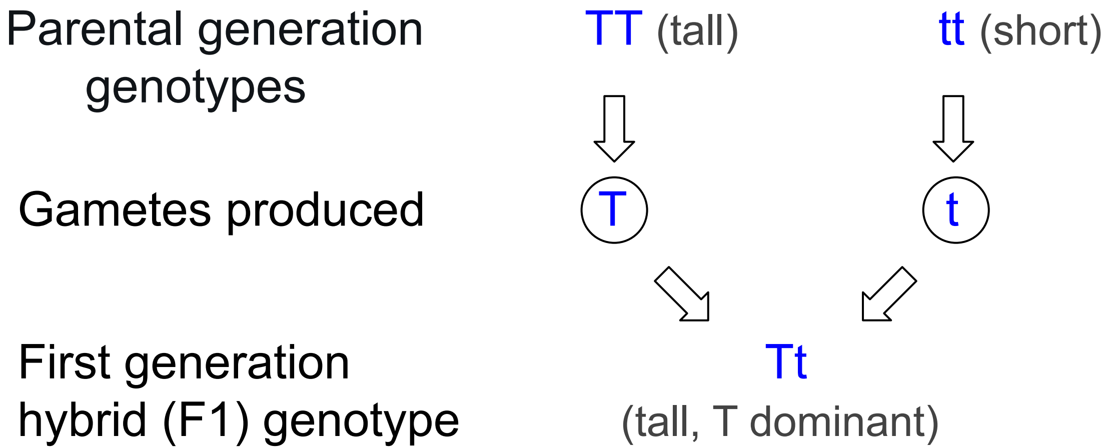
Gametes produced

T

t

1865 - Mendel's First Law: Law of Segregation

Mendel's Law of Segregation:
During gamete formation, the alleles for each gene (i.e. tallness or shortness) segregate from each other so that each gamete carries only one allele for each gene.



Parental generation genotypes

TT (tall)

tt (short)

Gametes produced

T

t

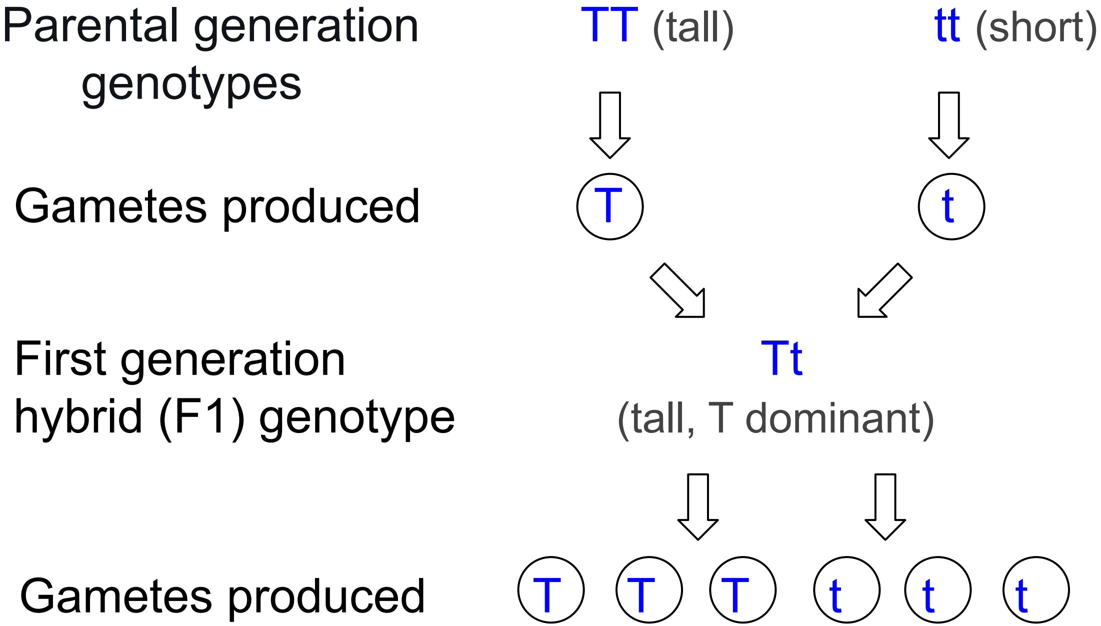
First generation hybrid (F1) genotype

Tt

(tall, T dominant)

1865 - Mendel’s First Law: Law of Segregation

Mendel’s Law of Segregation:
During gamete formation, the alleles for each gene (i.e. tallness or shortness) segregate from each other so that each gamete carries only one allele for each gene.



Parental generation genotypes

TT (tall)

tt (short)

Gametes produced

T

t

First generation hybrid (F1) genotype

Tt

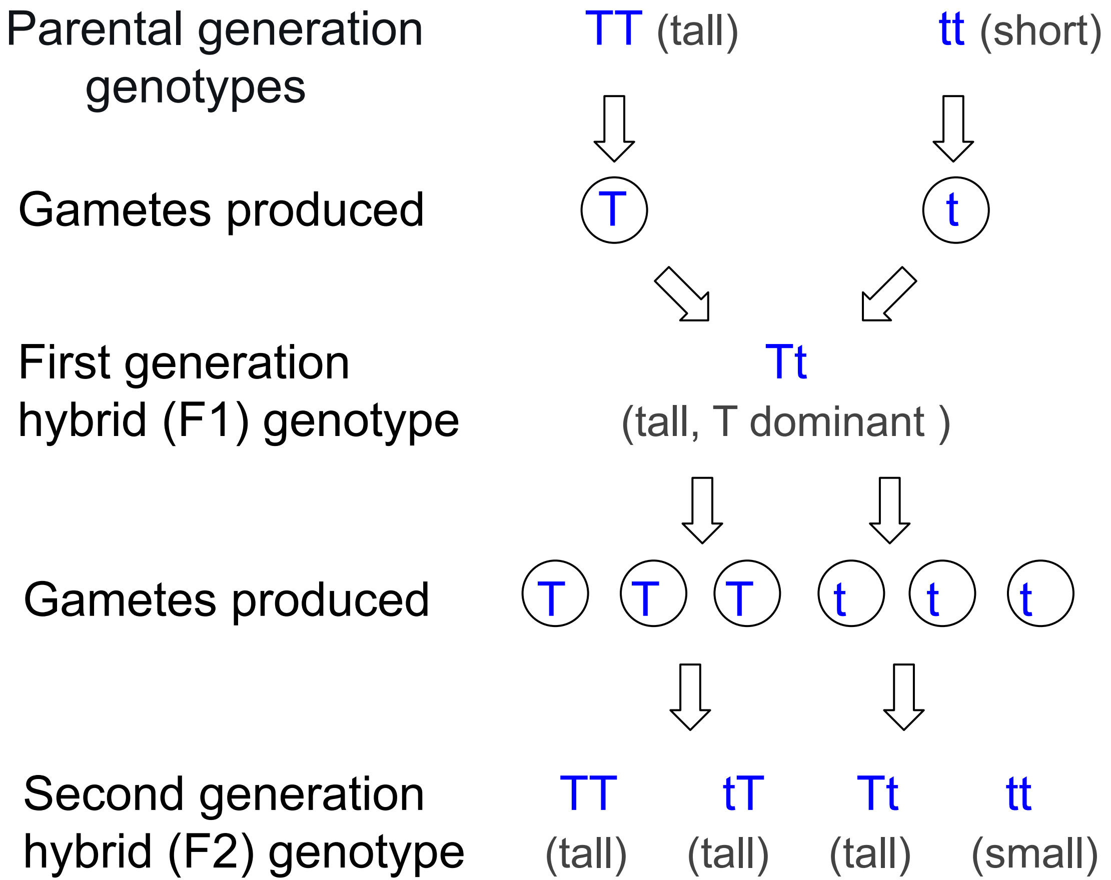
(tall, T dominant)

Gametes produced

T T T t t t

1865 - Mendel’s First Law: Law of Segregation

Mendel’s Law of Segregation:
During gamete formation, the alleles for each gene (i.e. tallness or shortness) segregate from each other so that each gamete carries only one allele for each gene.



Parental generation genotypes — TT (tall) tt (short)

Gametes produced — T t

First generation hybrid (F1) genotype — Tt (tall, T dominant)

Gametes produced — T T T t t t

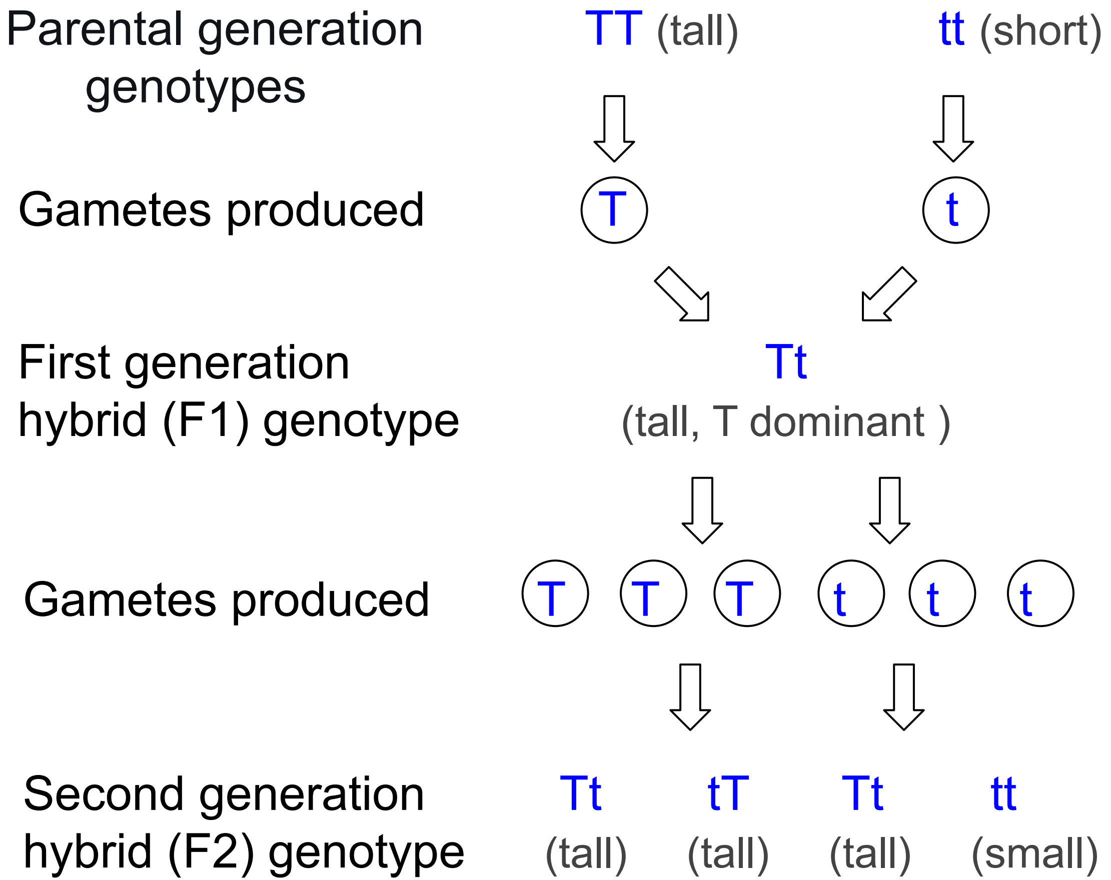
Second generation hybrid (F2) genotype — TT (tall) tT (tall) Tt (tall) tt (small)

Results: ¾ tall and ¼ small

1865 - Mendel’s First Law: Law of Segregation

Vocabulary recap!

- TT, Tt or tt are **genotypes**
- T and t are **alleles**
- TT and tt are **homozygous** genotypes
- Tt and tT are **heterozygous** genotypes
- Tall and small are **phenotypes**



Parental generation genotypes: TT (tall) tt (short)

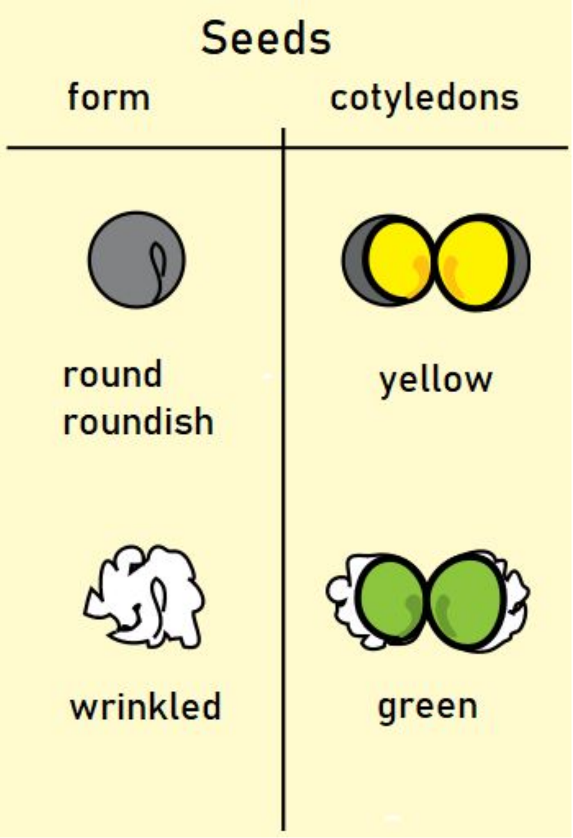
Gametes produced: T t

First generation hybrid (F1) genotype: Tt (tall, T dominant)

Gametes produced: T T T t t t

Second generation hybrid (F2) genotype: Tt (tall) tT (tall) Tt (tall) tt (small)

1865 - Mendel’s laws - Experiments



- He already know from experiments that round and yellow were dominant traits over wrinkled and green.

- When a round-yellow seeds plant crossed with wrinkled-green seeds plant, he always got round-yellow seeds plant (F1 hybrid). As expected.

- When self fertilizing F1 hybrid plants he got:
  - 9/16 round-yellow seeds plant
  - 3/16 round-green
  - 3/16 wrinkled-yellow
  - 1/16 wrinkled-green.

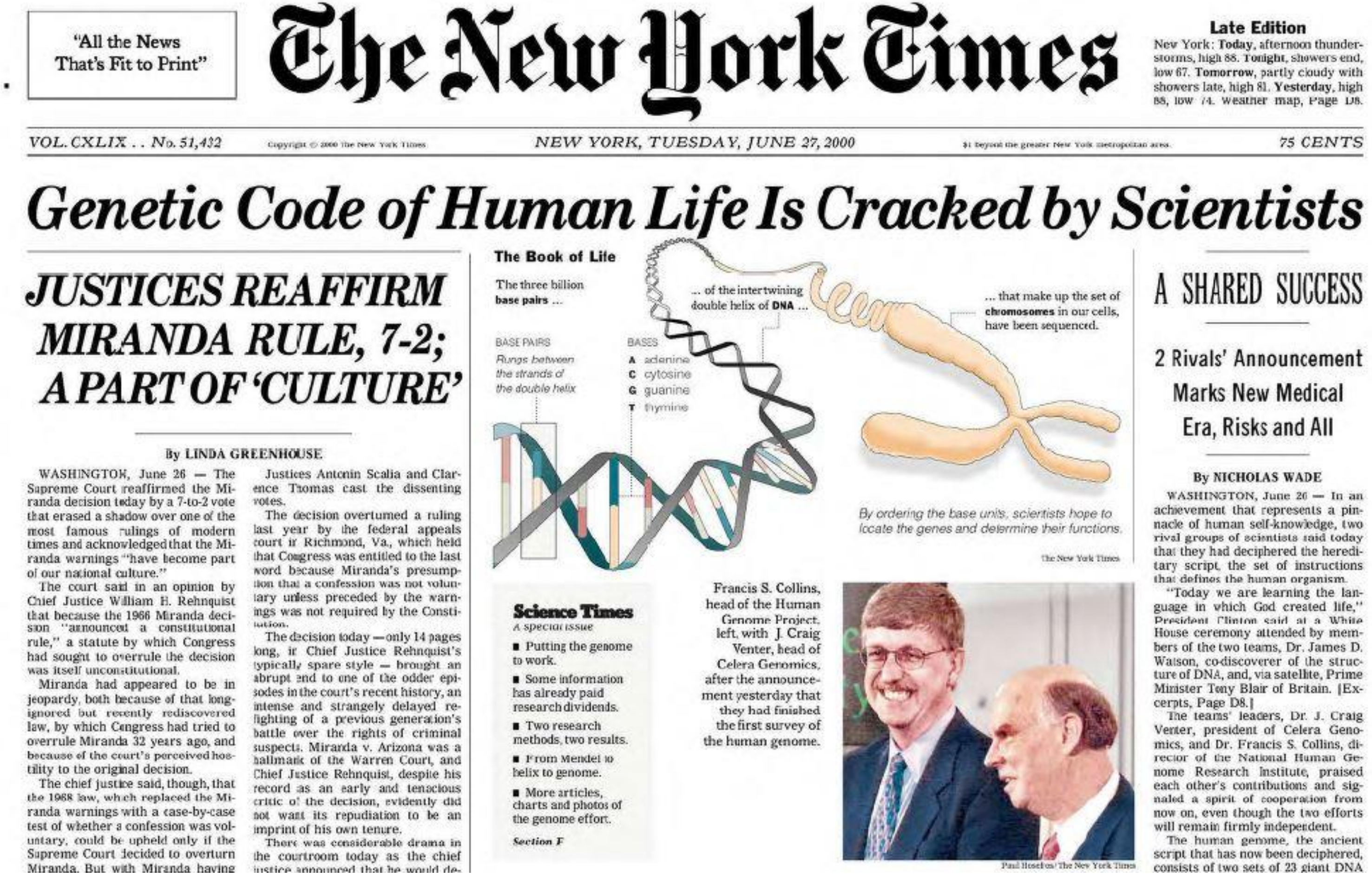
- What he **did not** get was: ¾ round-yellow ¼ wrinkled-green

1865 - Mendel’s Second Law: Law of Independence

Mendel’s Law of independence:
Genes of different traits segregate independently during the formation of gametes.

Note : This is actually untrue if the genes are on the same chromosome, this is called **linkage** and can be very important.

2000 - Draft of the first human genome



"All the News That's Fit to Print"

The New York Times

Late Edition

New York: **Today**, afternoon thunderstorms, high 88. **Tonight**, showers end, low 67. **Tomorrow**, partly cloudy with showers late, high 81. **Yesterday**, high 88, low 74. Weather map, Page D8.

VOL. CXLIX . . No. 51,432 Copyright © 2000 The New York Times NEW YORK, TUESDAY, JUNE 27, 2000 $1 beyond the greater New York metropolitan area. 75 CENTS

Genetic Code of Human Life Is Cracked by Scientists

JUSTICES REAFFIRM MIRANDA RULE, 7-2; A PART OF 'CULTURE'

By LINDA GREENHOUSE

WASHINGTON, June 26 — The Supreme Court reaffirmed the Miranda decision today by a 7-to-2 vote that erased a shadow over one of the most famous rulings of modern times and acknowledged that the Miranda warnings "have become part of our national culture."

The court said in an opinion by Chief Justice William H. Rehnquist that because the 1966 Miranda decision "announced a constitutional rule," a statute by which Congress had sought to overrule the decision was itself unconstitutional.

Miranda had appeared to be in jeopardy, both because of that long-ignored but recently rediscovered law, by which Congress had tried to overrule Miranda 32 years ago, and because of the court's perceived hostility to the original decision.

The chief justice said, though, that the 1968 law, which replaced the Miranda warnings with a case-by-case test of whether a confession was voluntary, could be upheld only if the Supreme Court decided to overturn Miranda. But with Miranda having

Justices Antonin Scalia and Clarence Thomas cast the dissenting votes.

The decision overturned a ruling last year by the federal appeals court in Richmond, Va., which held that Congress was entitled to the last word because Miranda's presumption that a confession was not voluntary unless preceded by the warnings was not required by the Constitution.

The decision today — only 14 pages long, in Chief Justice Rehnquist's typically spare style — brought an abrupt end to one of the odder episodes in the court's recent history, an intense and strangely delayed refighting of a previous generation's battle over the rights of criminal suspects. Miranda v. Arizona was a hallmark of the Warren Court, and Chief Justice Rehnquist, despite his record as an early and tenacious critic of the decision, evidently did not want its repudiation to be an imprint of his own tenure.

There was considerable drama in the courtroom today as the chief justice announced that he would de-

The Book of Life

The three billion **base pairs** ...

... of the intertwining double helix of **DNA** ...

... that make up the set of **chromosomes** in our cells, have been sequenced.

BASE PAIRS

Rungs between the strands of the double helix

BASES

A adenine
C cytosine
G guanine
T thymine

By ordering the base units, scientists hope to locate the genes and determine their functions.

The New York Times

Science Times

A special issue

- Putting the genome to work.
- Some information has already paid research dividends.
- Two research methods, two results.
- From Mendel to helix to genome.
- More articles, charts and photos of the genome effort.

Section F

Francis S. Collins, head of the Human Genome Project, left, with J. Craig Venter, head of Celera Genomics, after the announcement yesterday that they had finished the first survey of the human genome.

Paul Hosefros/The New York Times

A SHARED SUCCESS

2 Rivals' Announcement Marks New Medical Era, Risks and All

By NICHOLAS WADE

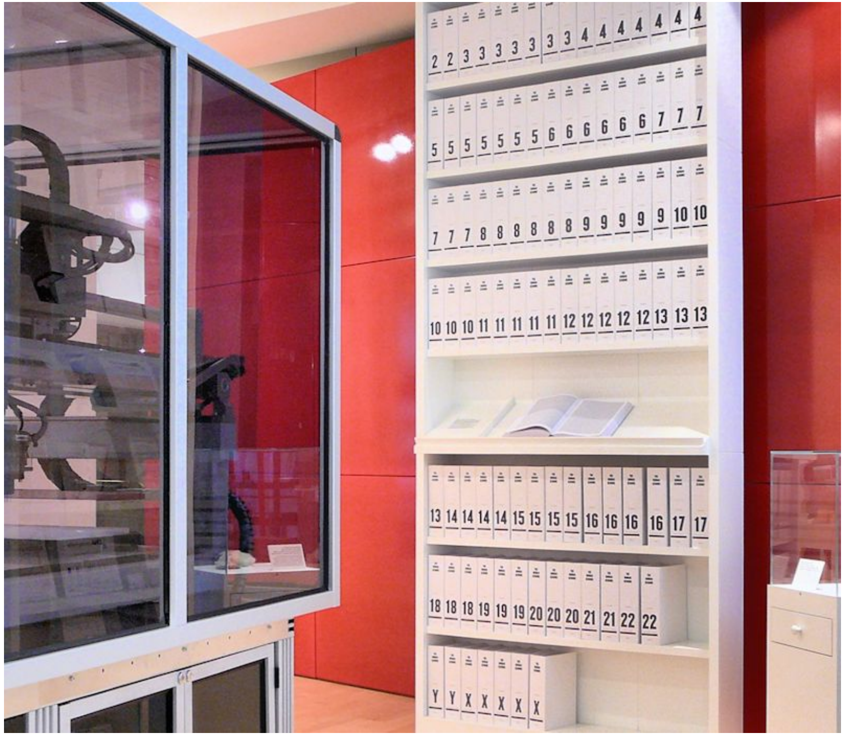
WASHINGTON, June 26 — In an achievement that represents a pinnacle of human self-knowledge, two rival groups of scientists said today that they had deciphered the hereditary script, the set of instructions that defines the human organism.

"Today we are learning the language in which God created life," President Clinton said at a White House ceremony attended by members of the two teams, Dr. James D. Watson, co-discoverer of the structure of DNA, and, via satellite, Prime Minister Tony Blair of Britain. [Excerpts, Page D8.]

The teams' leaders, Dr. J. Craig Venter, president of Celera Genomics, and Dr. Francis S. Collins, director of the National Human Genome Research Institute, praised each other's contributions and signaled a spirit of cooperation from now on, even though the two efforts will remain firmly independent.

The human genome, the ancient script that has now been deciphered, consists of two sets of 23 giant DNA

2000 - Draft of the first human genome

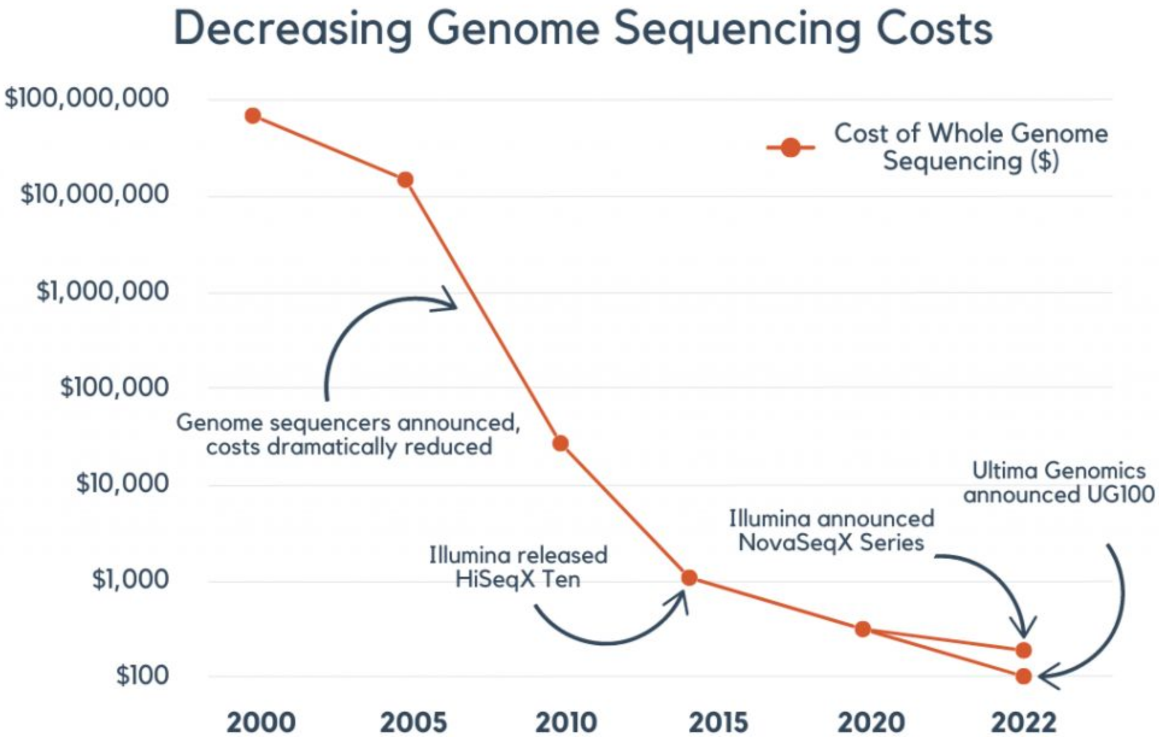


Wellcome Collection, London

A few numbers:

- Lasted for 13 years from 1990 to 2003
- Costed around $3 billion
- Thousands of researches across many countries
- Sequenced 92% of the human genome

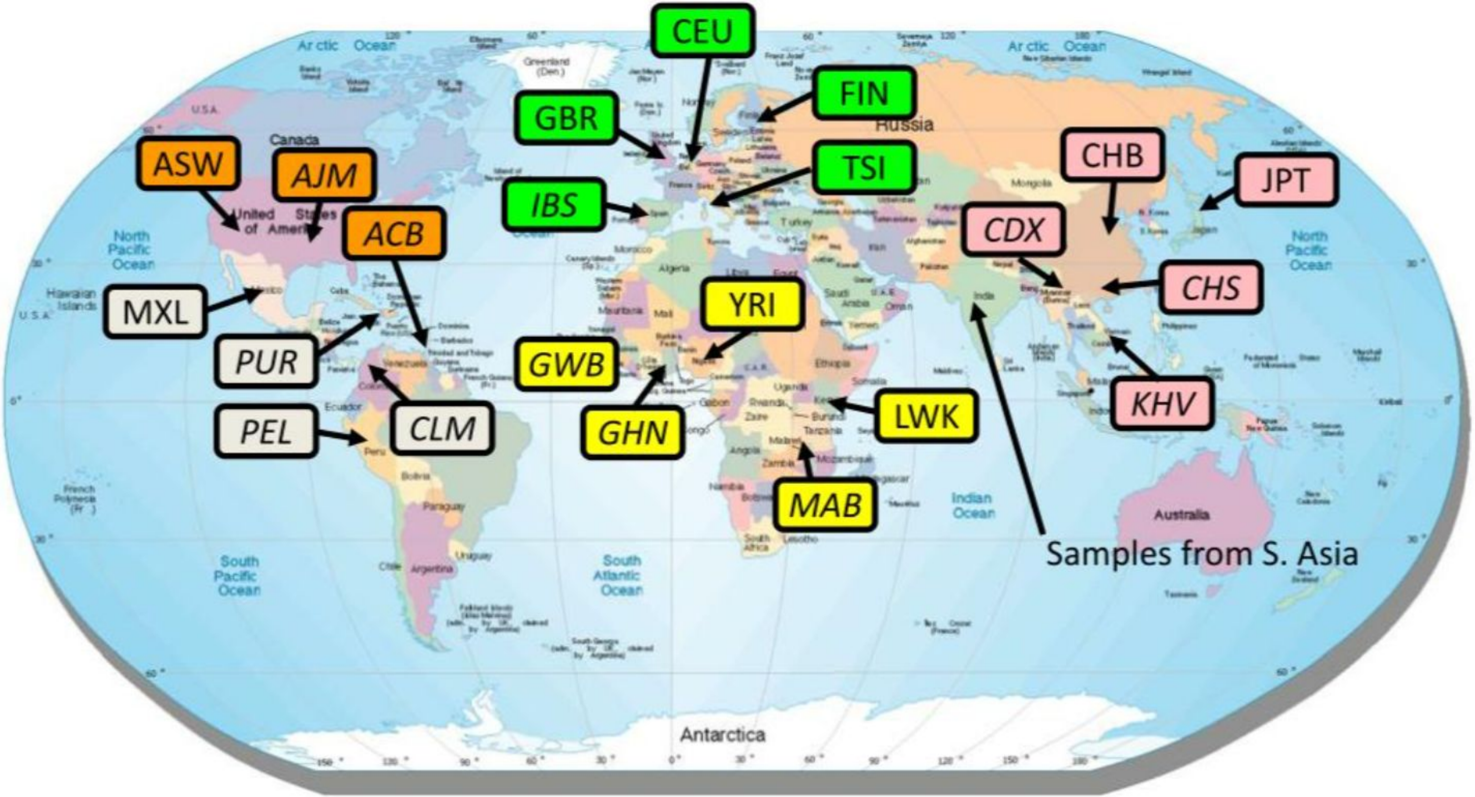
2000 and onward - Huge reduction in sequencing cost



Cost of whole genome sequencing has dramatically been reduced (faster than exponential decay) resulting in an fastly increasing amount of data.

Still a low amount of data compared to other fields, there are “only” thousands of available human genomes.

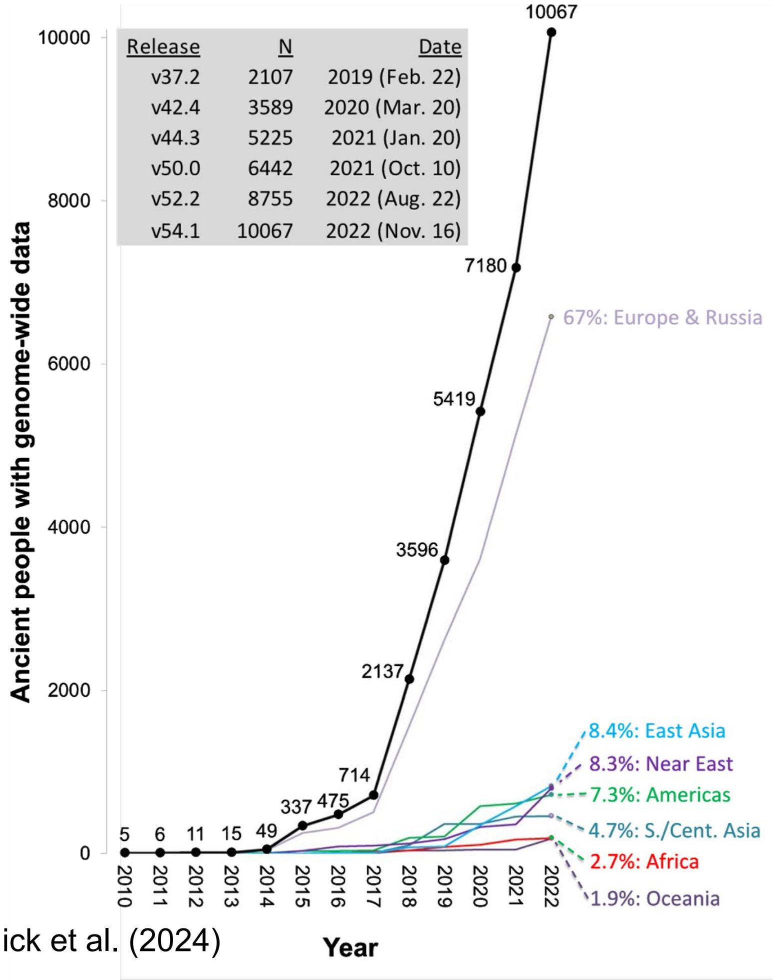
2006-2016 - 1000 Genome Project



Sequencing of around 100 genomes per different populations around the globe.

Data is publicly available.

2010-present - Ancient DNA revolution



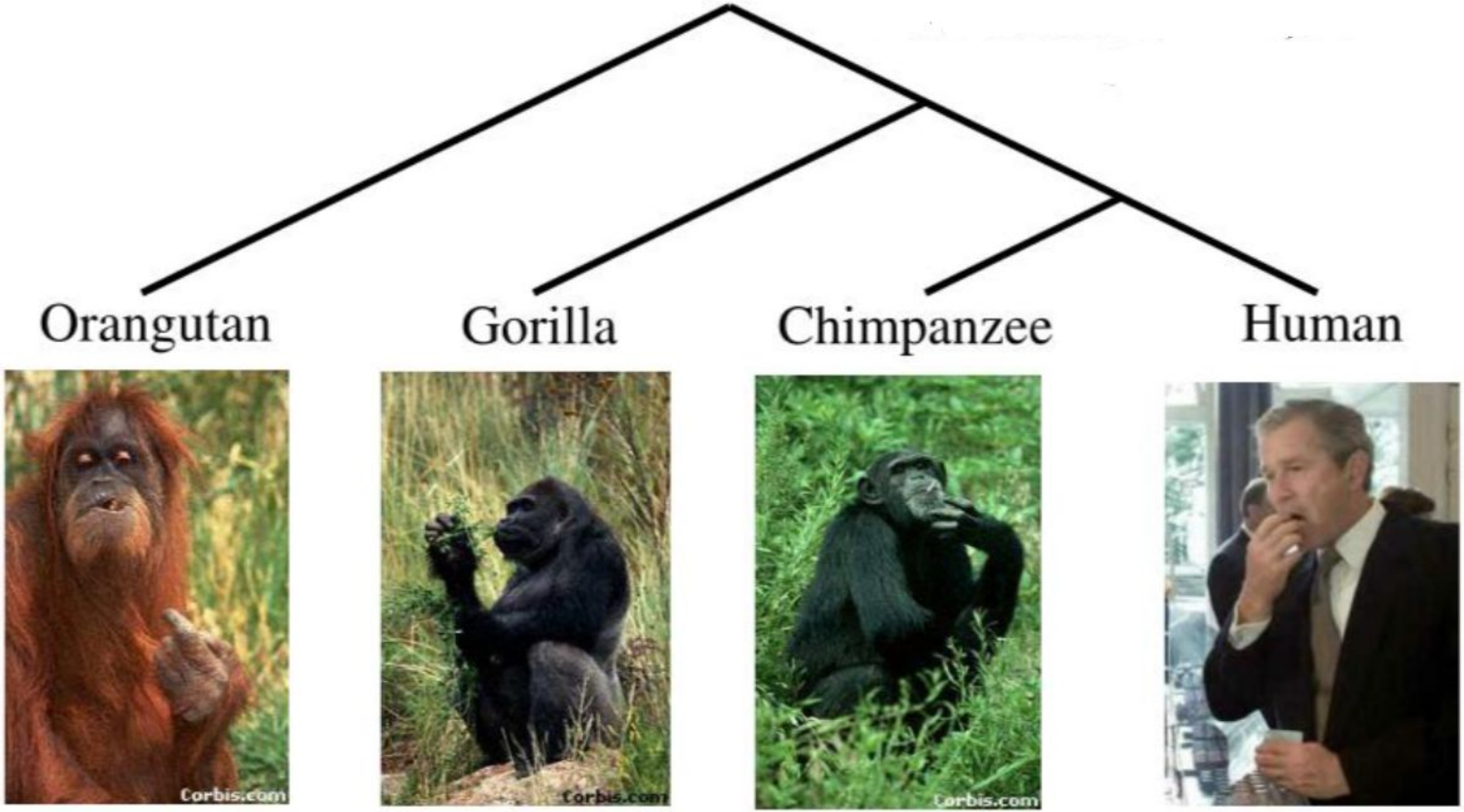
Release	N	Date
v37.2	2107	2019 (Feb. 22)
v42.4	3589	2020 (Mar. 20)
v44.3	5225	2021 (Jan. 20)
v50.0	6442	2021 (Oct. 10)
v52.2	8755	2022 (Aug. 22)
v54.1	10067	2022 (Nov. 16)

Mallick et al. (2024)

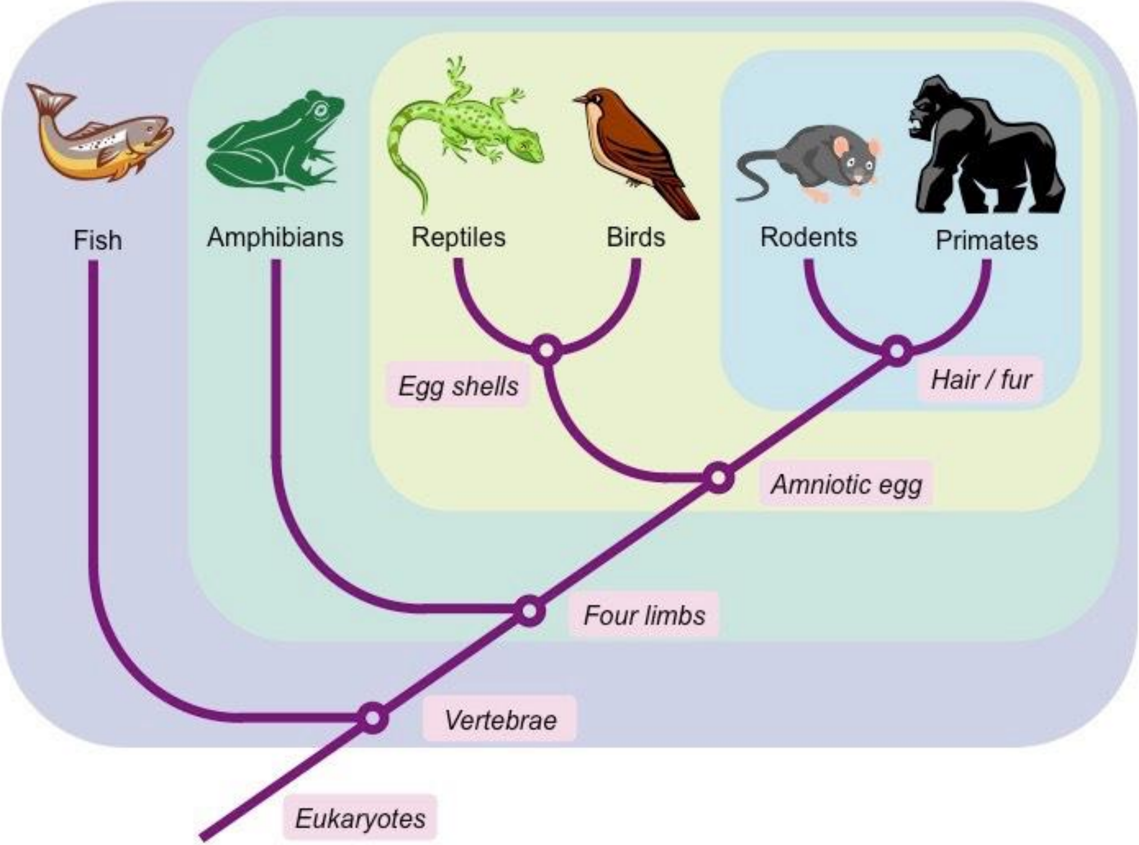
Ancient DNA is genetic material from non-contemporary individuals, ranging from a few hundred years old to tens of thousands of years old.

Genetic analysis task: phylogeny

Phylogeny


Orangutan
Gorilla
Chimpanzee
Human
Corbis.com
Corbis.com
Corbis.com

Phylogeny based on observable traits



Phylogenetics

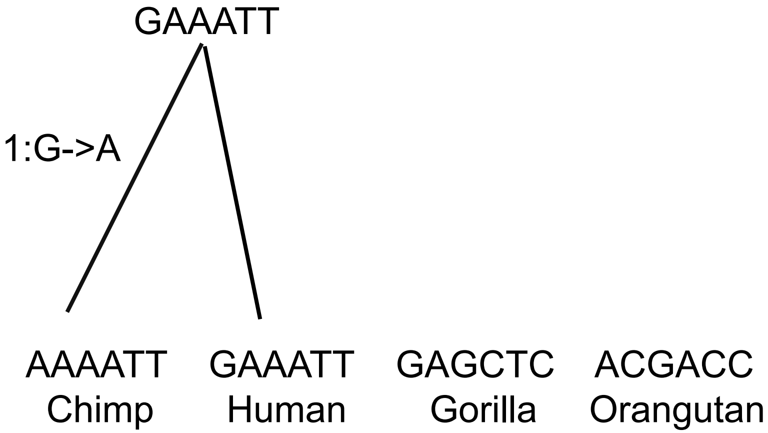
AAAATT
Chimp

GAAATT
Human

GAGCTC
Gorilla

ACGACC
Orangutan

Phylogenetics



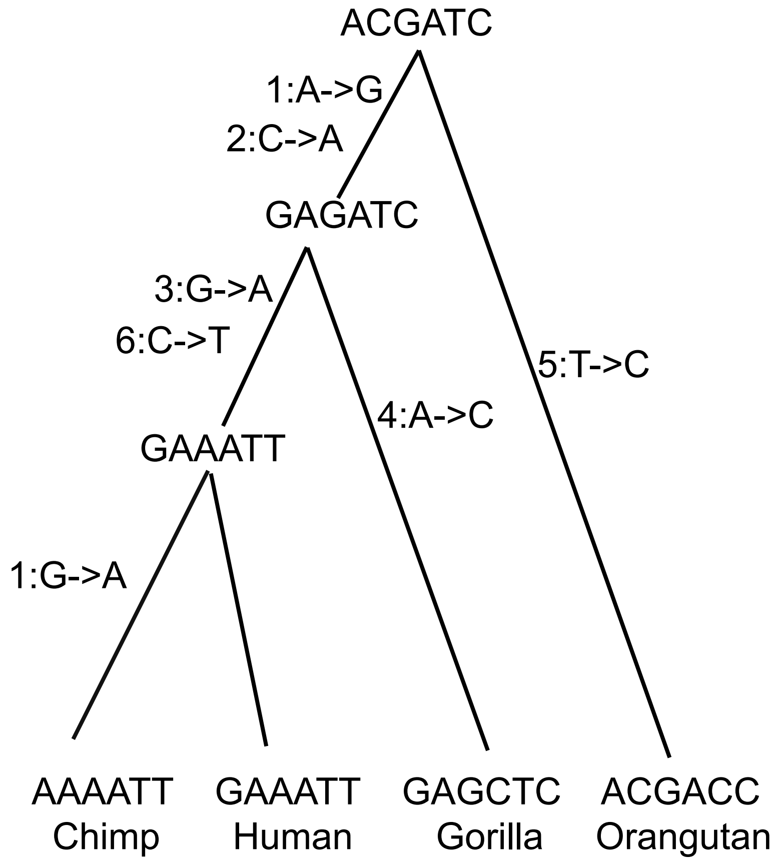
GAAATT

1:G->A

AAAATT GAAATT GAGCTC ACGACC

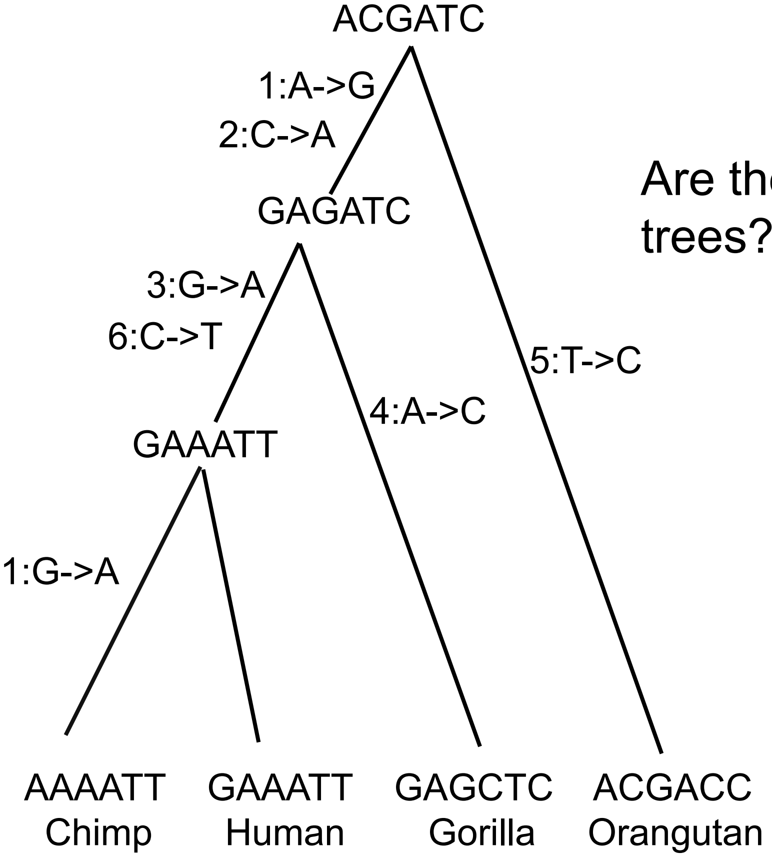
Chimp Human Gorilla Orangutan

Phylogenetics

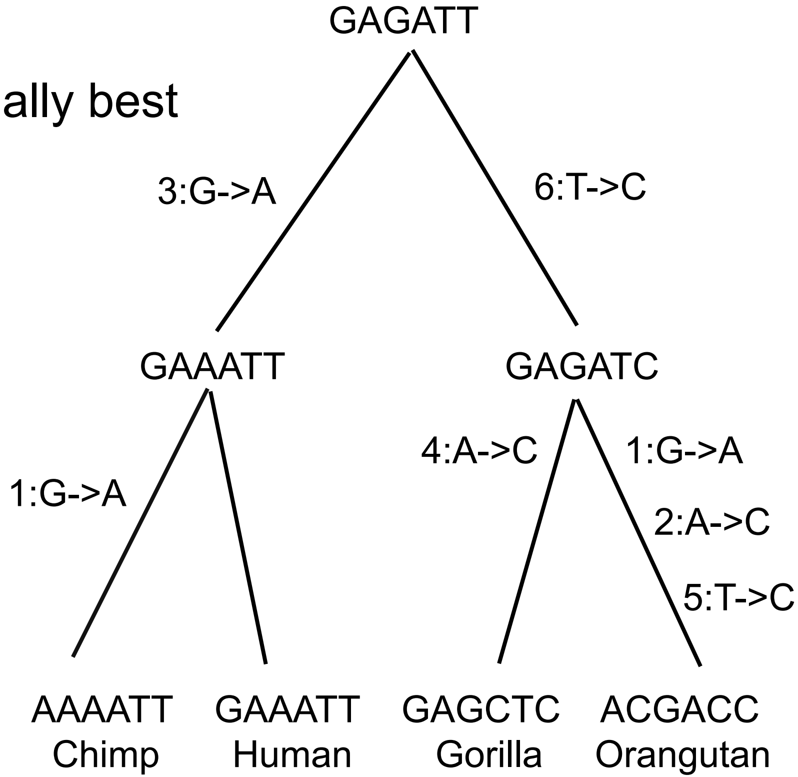


Is it the only tree?

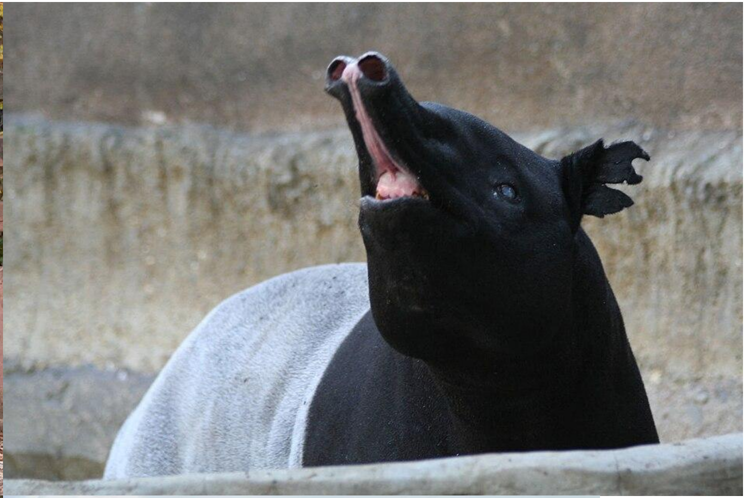
Phylogenetics



Are these the equally best trees?

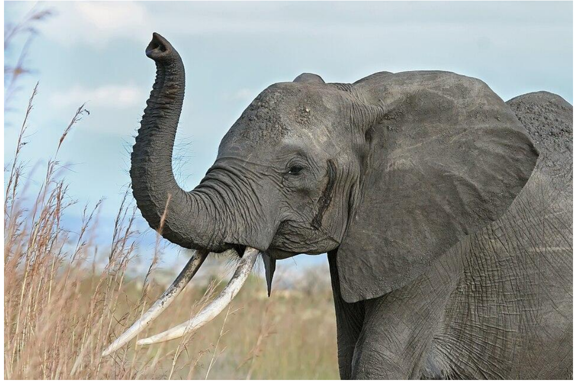


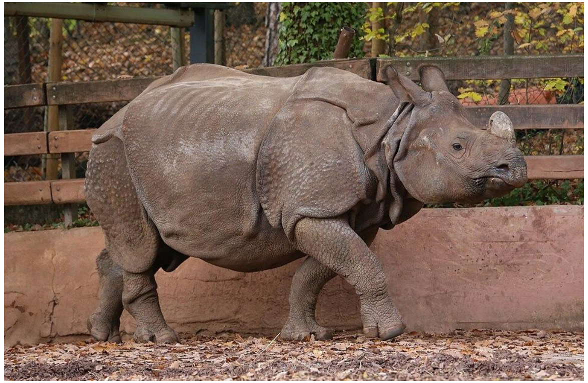
Surprises from phylogenetics: Pachydermata

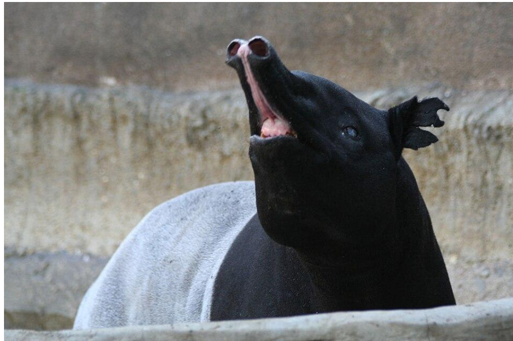


Surprises from phylogenetics: ~~Pachydermata~~

Whippomorpha :
cetaceans + hippopotamids







The end