Bacterial and viral genomics

David Šmajs
Department of Biology
Faculty of Medicine
Masaryk University
Brno, Czech Republic

Bacterial and viral genomics

What is genomics?

DNA Sequencing approaches

Genome size and structure, mutation rate

Genomics of human pathogens: examples

- The Black Death in Europe in 1347
- Ebola, influenza

Changes in human genome selected by pathogens

What is genomics?

Genomics is the study of whole genomes of organisms, and incorporates elements from genetics. **Genomics** uses a combination of recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analyze the structure and function of genomes.

In 1952, Alfred Hershey and Martha Chase demonstrated with a series of experiments that DNA, not protein, is responsible for carrying genetic traits that may be inherited. James Watson and Francis Crick discovered the double helix structure of DNA in 1953.

Milestones

1952: Hershey and Chase – genetic information is encoded in DNA

1953: Watson and Crick – DNA structure

1972: Sanger started work on DNA sequencing

1977: first DNA virus sequenced (ΦX174 bacteriophage)

1995: first bacterium sequenced (*Haemophilus influenzae)*

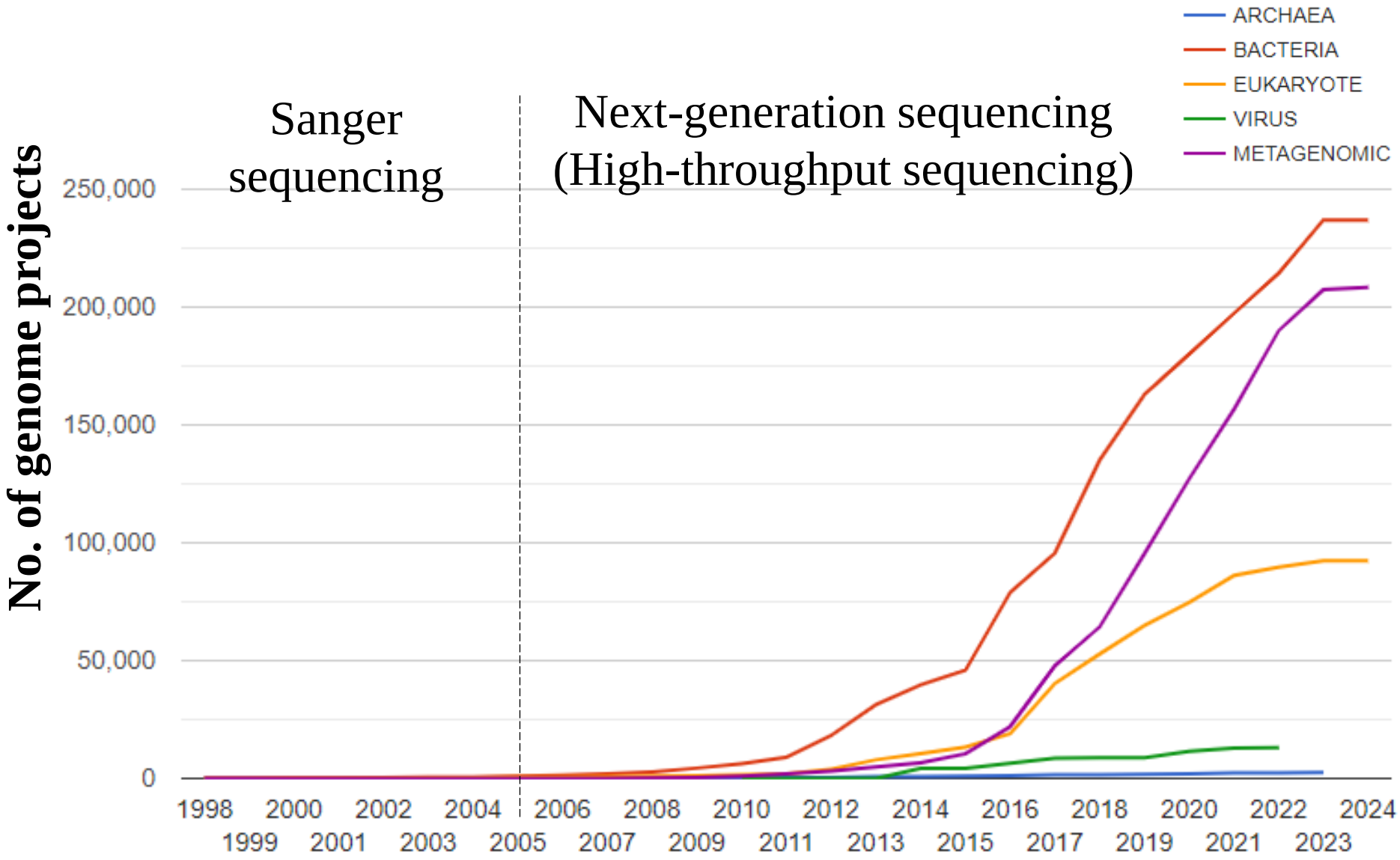
1996: first eukaryotic genome sequenced (*Saccharomyces cerevisiae*)

1998: first multicellular organism sequenced (*Caenorhabditis elegans*)

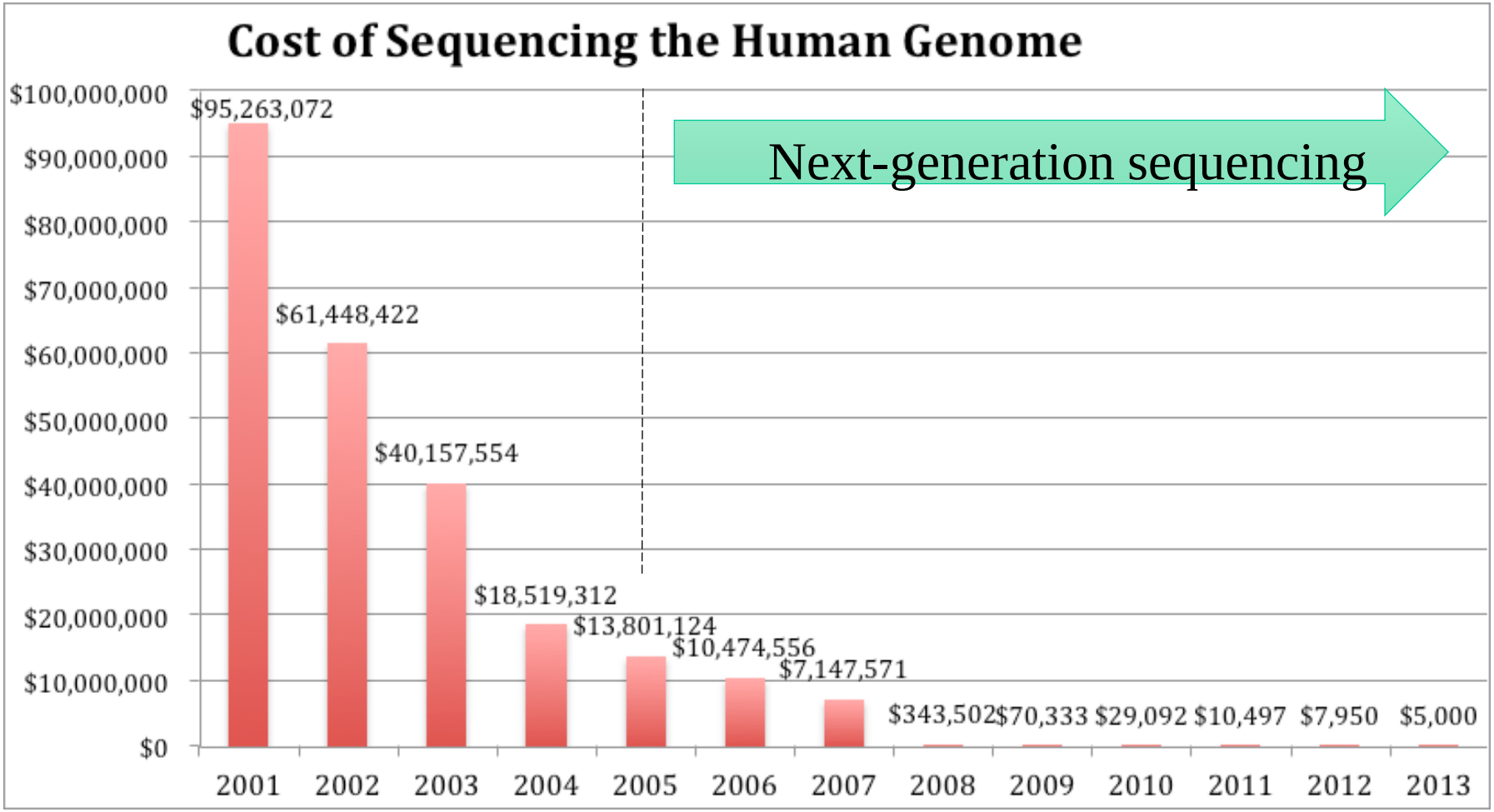
2003: human genome completed

2023: sequenced 506 618 genomes

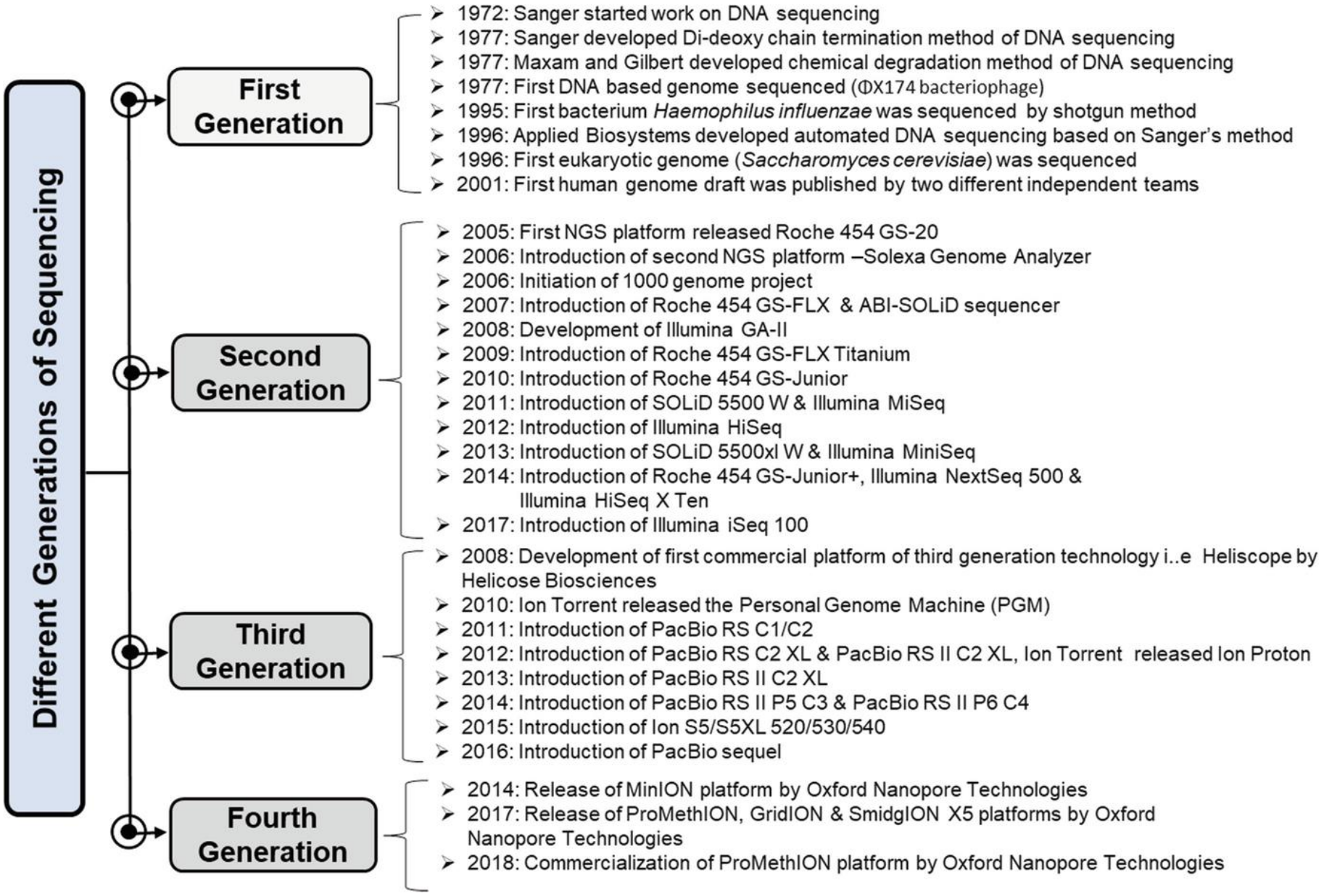
- Archaea: 5 371
- Bacteria: 433 319
- Eukarya: 49 843
- Viruses: 18 085



https://gold.jgi.doe.gov/



NGS cost in 2024: $100 per human genome


Different Generations of Sequencing
First Generation
1972: Sanger started work on DNA sequencing
1977: Sanger developed Di-deoxy chain termination method of DNA sequencing
1977: Maxam and Gilbert developed chemical degradation method of DNA sequencing
1977: First DNA based genome sequenced (ΦX174 bacteriophage)
1995: First bacterium Haemophilus influenzae was sequenced by shotgun method
1996: Applied Biosystems developed automated DNA sequencing based on Sanger's method
1996: First eukaryotic genome (Saccharomyces cerevisiae) was sequenced
2001: First human genome draft was published by two different independent teams
Second Generation
2005: First NGS platform released Roche 454 GS-20
2006: Introduction of second NGS platform –Solexa Genome Analyzer
2006: Initiation of 1000 genome project
2007: Introduction of Roche 454 GS-FLX & ABI-SOLiD sequencer
2008: Development of Illumina GA-II
2009: Introduction of Roche 454 GS-FLX Titanium
2010: Introduction of Roche 454 GS-Junior
2011: Introduction of SOLiD 5500 W & Illumina MiSeq
2012: Introduction of Illumina HiSeq
2013: Introduction of SOLiD 5500xl W & Illumina MiniSeq
2014: Introduction of Roche 454 GS-Junior+, Illumina NextSeq 500 & Illumina HiSeq X Ten
2017: Introduction of Illumina iSeq 100
Third Generation
2008: Development of first commercial platform of third generation technology i..e Heliscope by Helicose Biosciences
2010: Ion Torrent released the Personal Genome Machine (PGM)
2011: Introduction of PacBio RS C1/C2
2012: Introduction of PacBio RS C2 XL & PacBio RS II C2 XL, Ion Torrent released Ion Proton
2013: Introduction of PacBio RS II C2 XL
2014: Introduction of PacBio RS II P5 C3 & PacBio RS II P6 C4
2015: Introduction of Ion S5/S5XL 520/530/540
2016: Introduction of PacBio sequel
Fourth Generation
2014: Release of MinION platform by Oxford Nanopore Technologies
2017: Release of ProMethION, GridION & SmidgION X5 platforms by Oxford Nanopore Technologies
2018: Commercialization of ProMethION platform by Oxford Nanopore Technologies

Main sequencing technologies and their characteristics

Technology (manufacturer)	Sequencing chemistry	Platform	Read length (bp)	Throughput (Gb/h run)	Best suited for:
Sanger	Dye terminator	ABI 3730xl	700–900		*De novo* and metagenomics
454 (Roche)	Pyrosequencing	GS FLX	400–700	0.04	*De novo* and metagenomics
		GS Junior	400	0.004	*De novo* and metagenomics
Solexa (Illumina)	Sequencing by synthesis with reversible terminators	GaIIx	36–150	0.3	Resequencing
		HiSeq2000	36–100	**2.9**	Resequencing
		MiSeq	36–250	0.2	Resequencing
SOLiD (ABI)	Sequencing by ligation	5500xl	35–75	1	Resequencing
Heliscope (Helicos)	Sequencing by synthesis with virtual terminators	tSMS	25–55	1	Resequencing
Ion Torrent (Life Technologies)	Semiconductor sequencing	Ion torrent PGM	100–200	0.2	Resequencing
		Ion proton sequencer	100–200	**2.5**	Resequencing
PacBio (Pacific Bioscience)	SMRT technology	PacBio*RS*	**250–10 000**	0.1	Genome structure and metagenomics
Nanopore (Oxford Nanopore Technologies)	Ionic current sensing	GridION and MinION	**10 000–50 000**	*	*De novo* and genome structure

Genome Sequencing

Genome: 3 Gb

Cut genome into large pieces

Clone into BACs: 100 kb

Order based on sequence features (*markers*) = mapping

Cut again

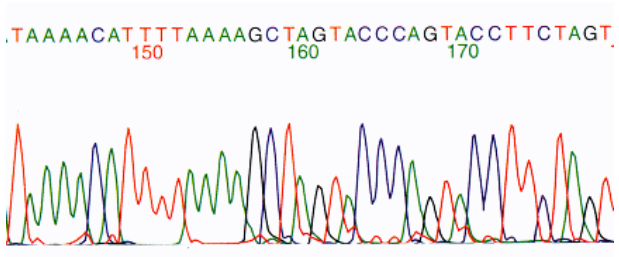
Assemble entire sequence

…**TTGTAAGTGAGAACAGGACGTATGTGGTTTTCTACTCCTGTGTT**…

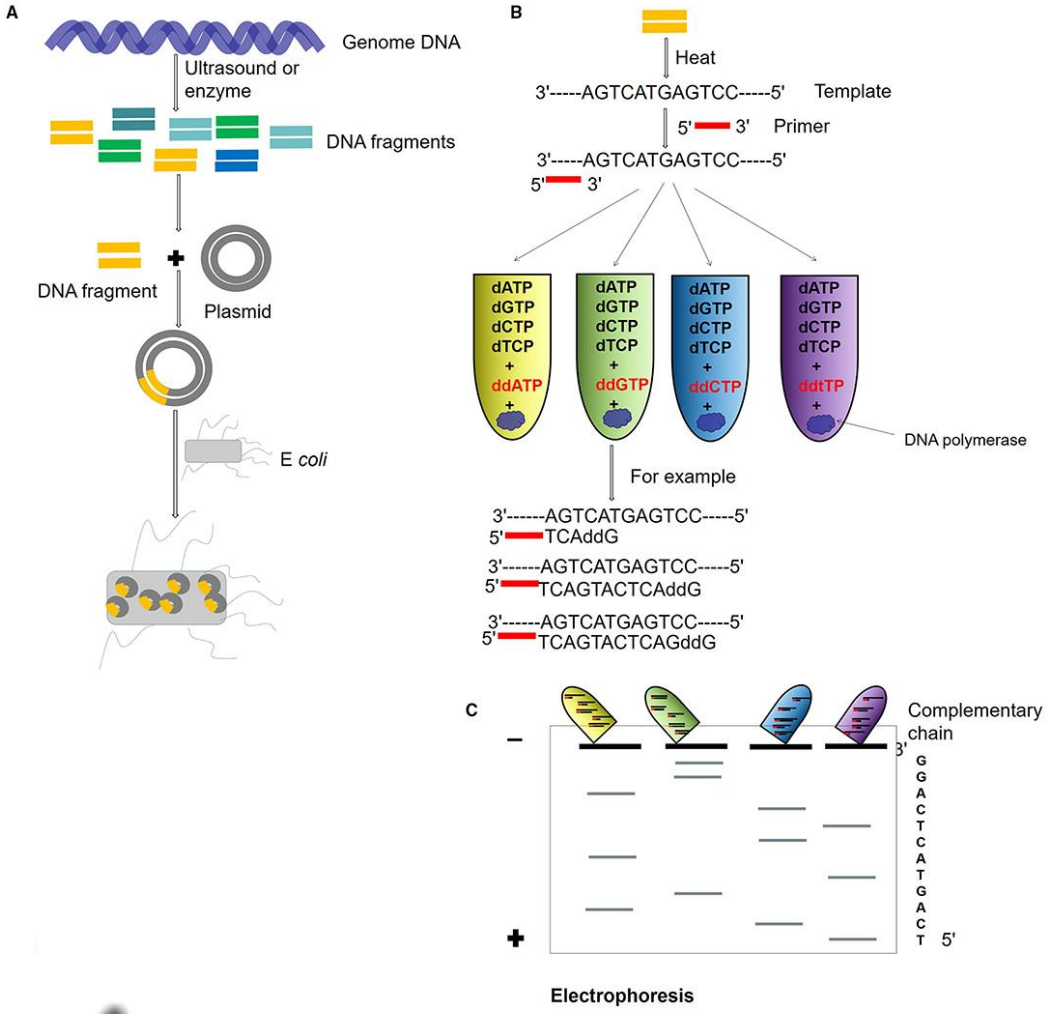
Assemble each BAC

TTGTAAGTGAGAACA
AGAACAGGACGTATGTGGT
TGTGGTTTTCTACTCC
CTACTCCTGTGTT

Sequence



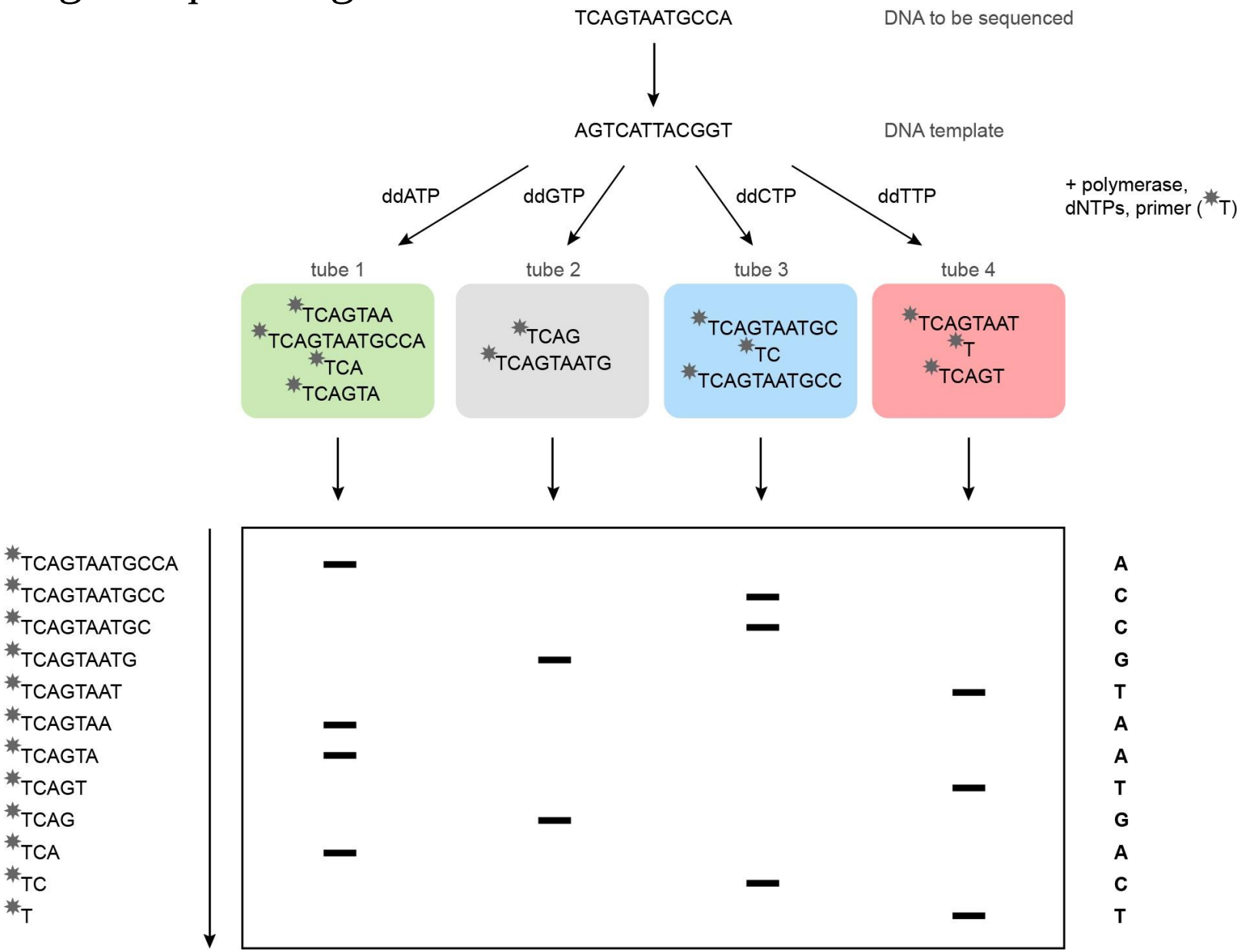
Sanger sequencing process (shotgun)

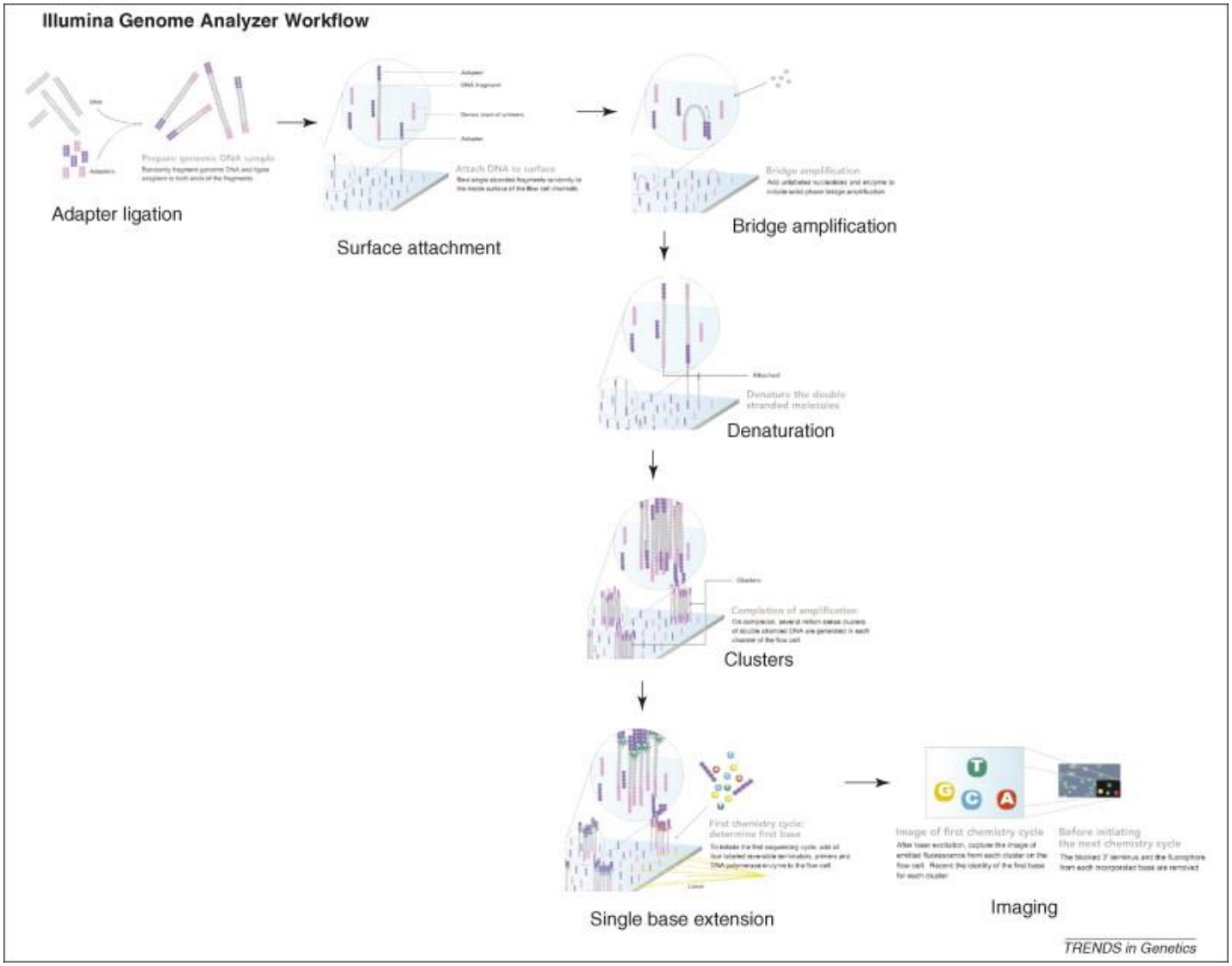


(A) Genomic DNA fragmented by ultrasound or enzymes, cloned to vector (plasmid), and transformed into *E. coli*. **(B)** The dideoxy chain termination method is used for sequencing. The PCR reaction mixture contains template, primers, DNA polymerase, and dNTPs. Moreover, four types of ddNTP (ddATP, ddTTP, ddCTP, and ddGTP) with fluorescent marks are added separately to four reactions. Since ddNTP does not form phosphodiester bond (does not contain hydroxyl group), incorporation of ddNTP into synthetized DNA strand results in termination of DNA amplification. **(C)** After PCR amplification, four PCR reaction are loaded to electrophoresis and DNA sequence is determined from positions of PCR products.

Zhang et al. 2021; https://doi.org/10.3389/fmicb.2021.766364

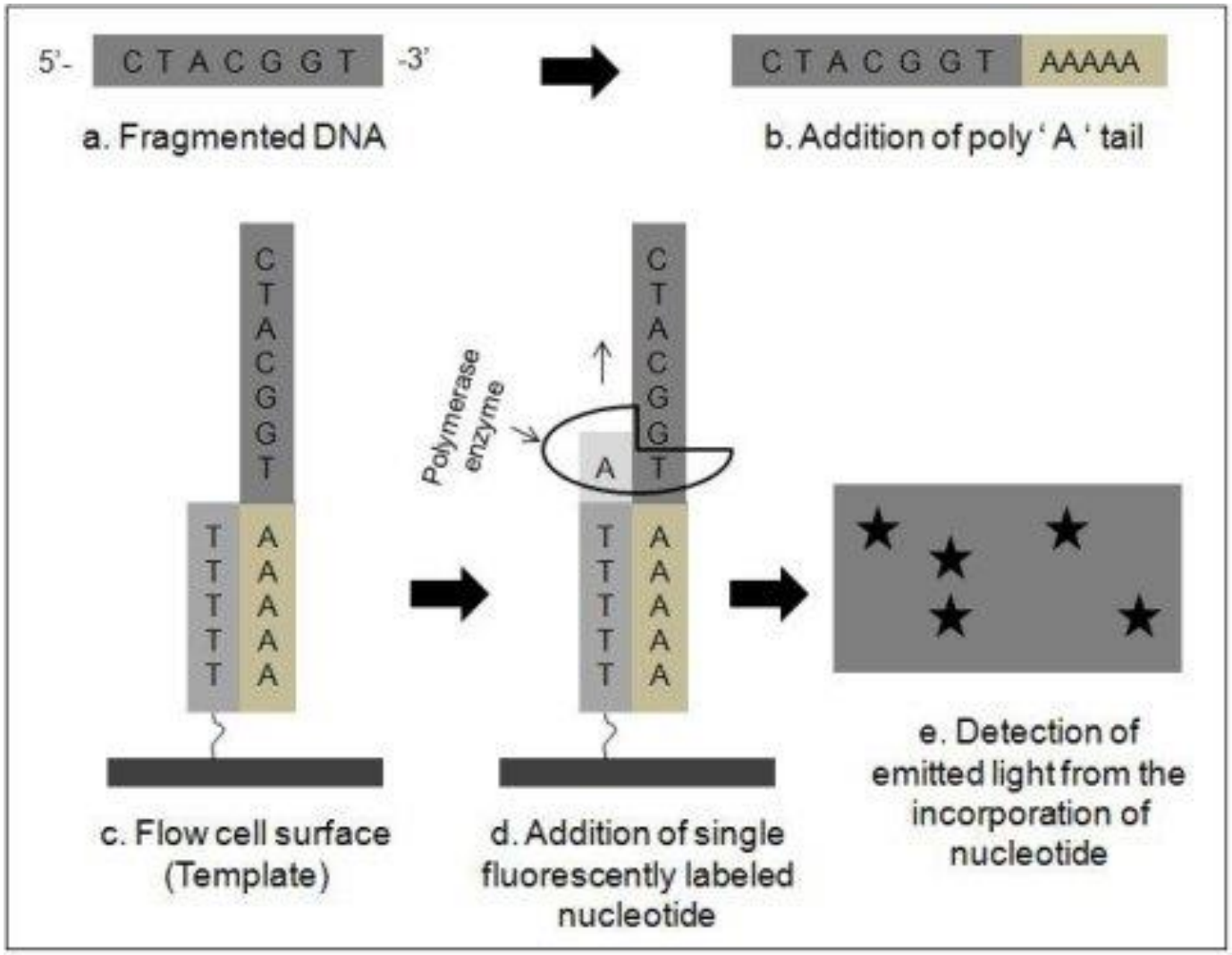
Sanger sequencing



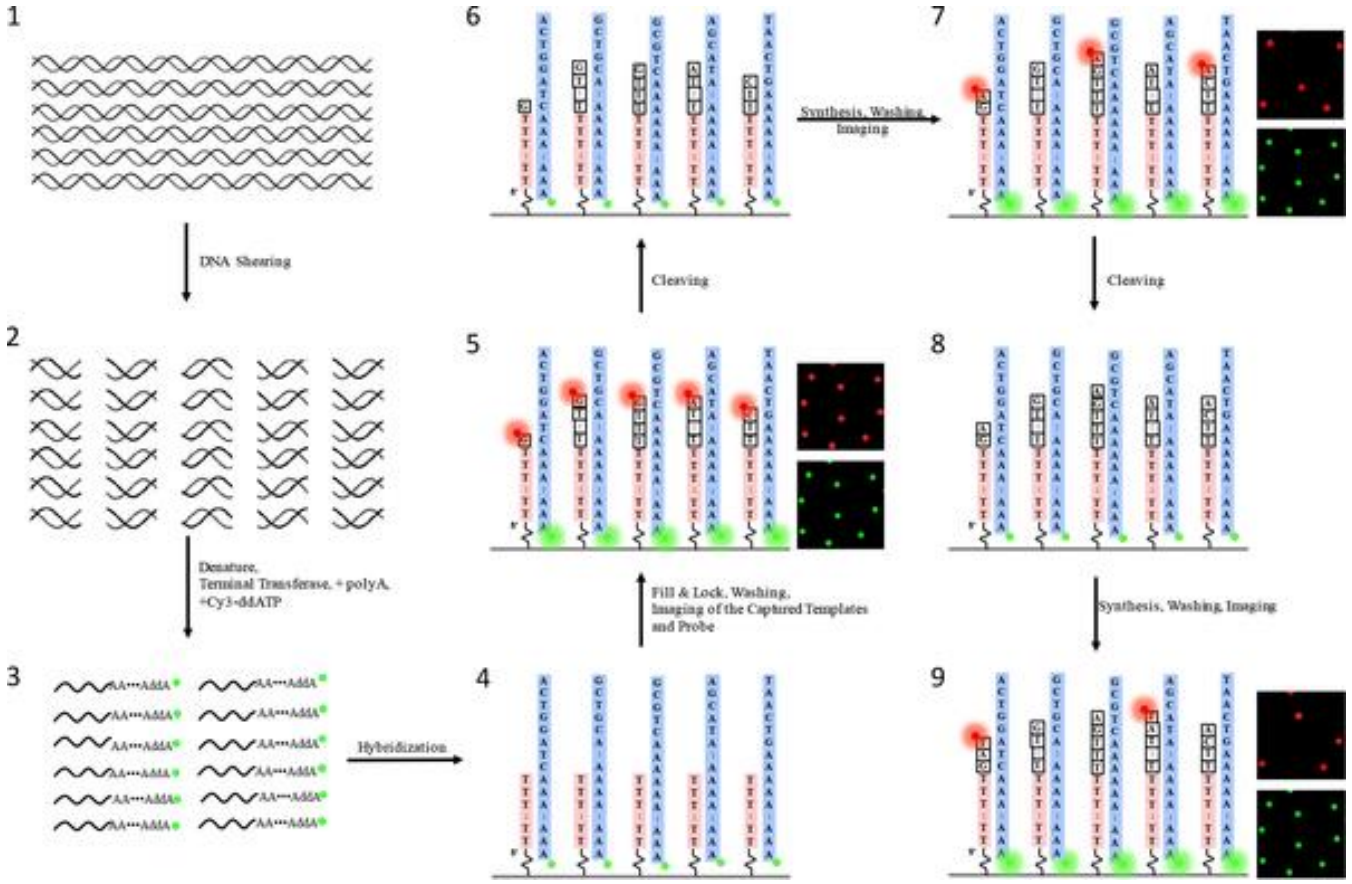


Illumina workflow. Starting from similar fragmentation and adapter ligation steps, the library is added to a flow cell for bridge amplification (an isothermal process that amplifies each fragment into a cluster). The cluster fragments are denatured, annealed with a sequencing primer and subjected to sequencing by synthesis using 3′ blocked labeled nucleotides.

Helicos single molecule sequencing method



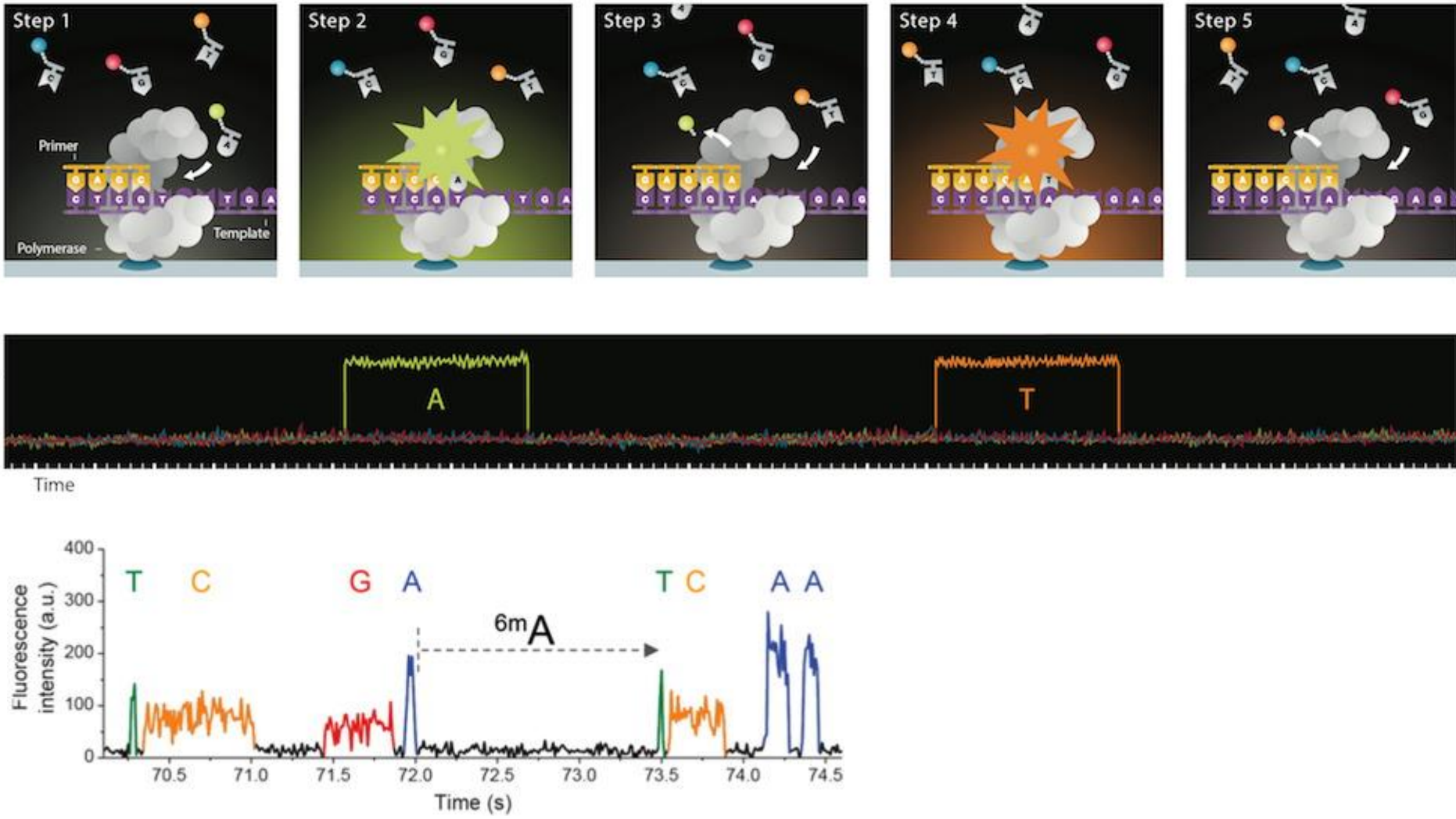
Sample preparation and sequencing process for single molecule sequencing of biological samples.



Zhao L, Deng L, Li G, Jin H, Cai J, et al. (2017) Single molecule sequencing of the M13 virus genome without amplification. PLOS ONE 12(12): e0188181. https://doi.org/10.1371/journal.pone.0188181
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0188181

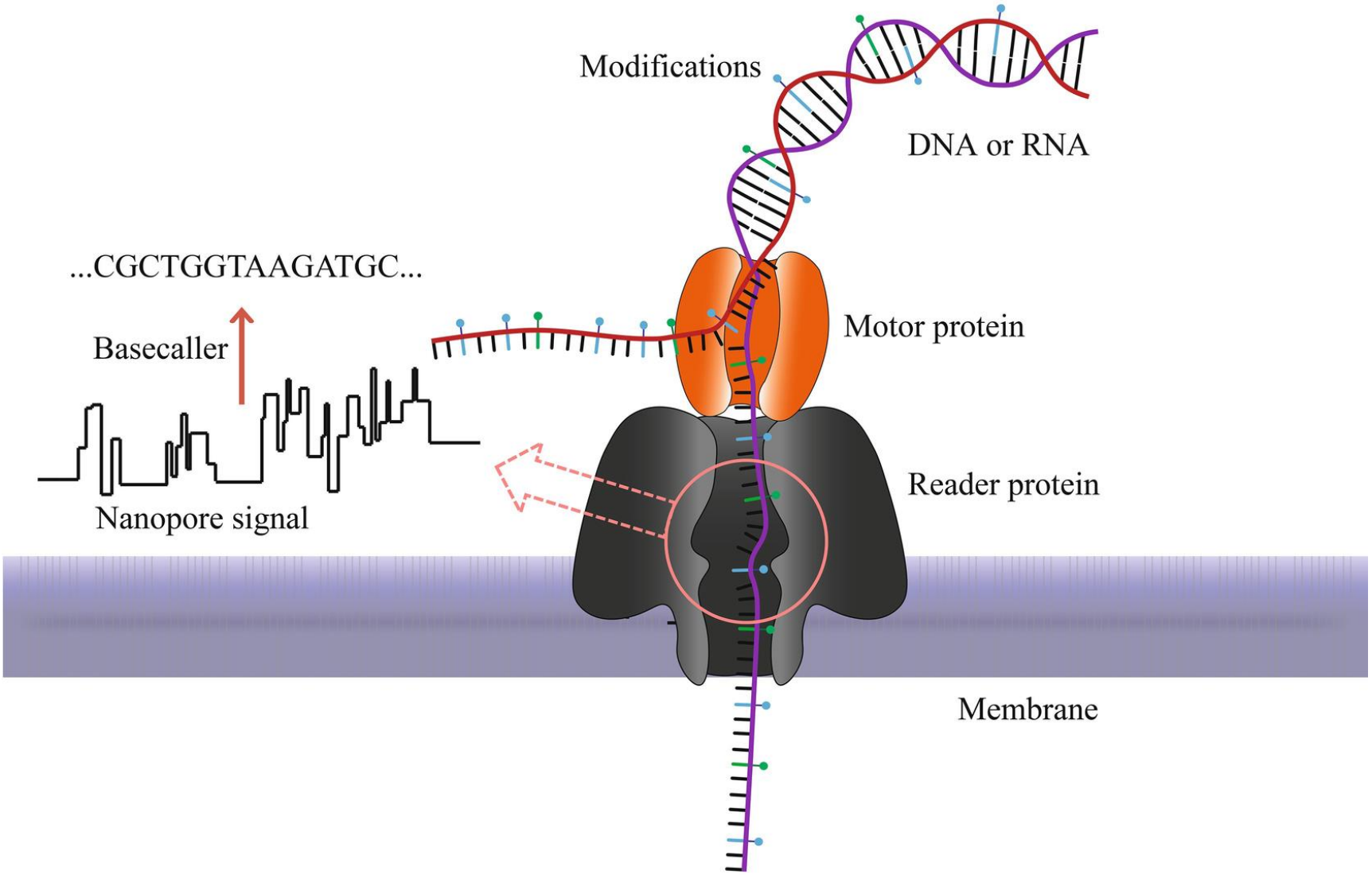
A new class of third-generation sequencing platforms capable of directly measuring DNA and RNA sequences at the single-molecule level without amplification. Here, we use the new GenoCare single-molecule sequencing platform from Direct Genomics to sequence the genome of the M13 virus. Our platform detects single-molecule fluorescence by total internal reflection microscopy, with **sequencing-by-synthesis chemistry.** We sequenced the genome of M13 to a depth of 316x, with 100% coverage. We determined a consensus sequence accuracy of 100%. In contrast to GC bias inherent to NGS results, we demonstrated that our single-molecule sequencing method yields minimal GC bias.

PacBio sequencing



Detection of methylated bases using PacBio sequencing
PacBio sequencing can detect modified bases, including m^6A (also known as ^{6m}A), by analyzing variation in the time between base incorporations in the read strand. The figure is adapted with permission from Pacific Biosciences. a.u. stands for arbitrary unit.

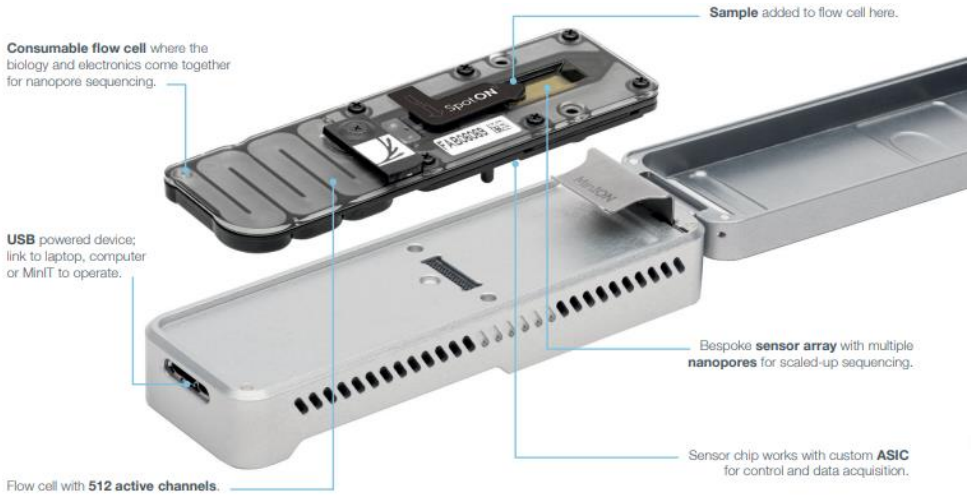
Nanopore sequencing

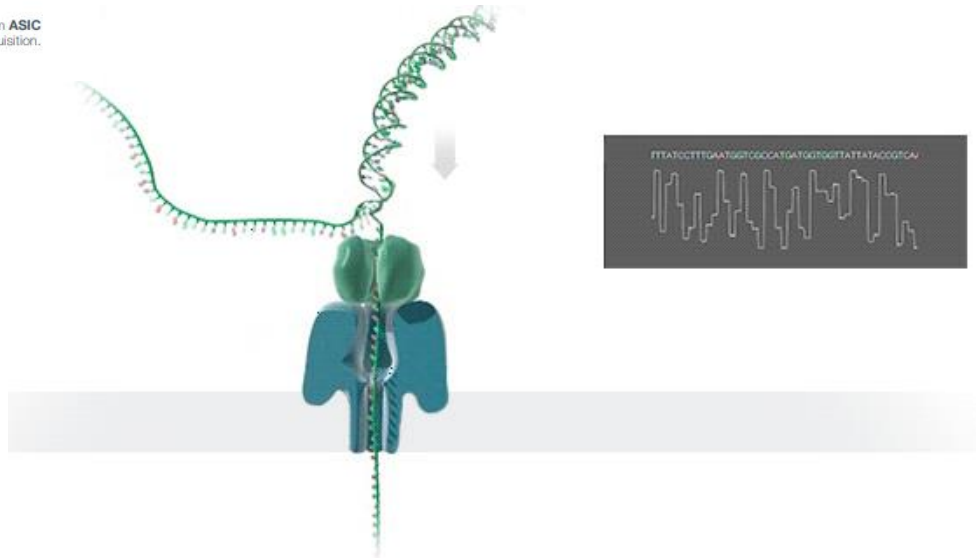


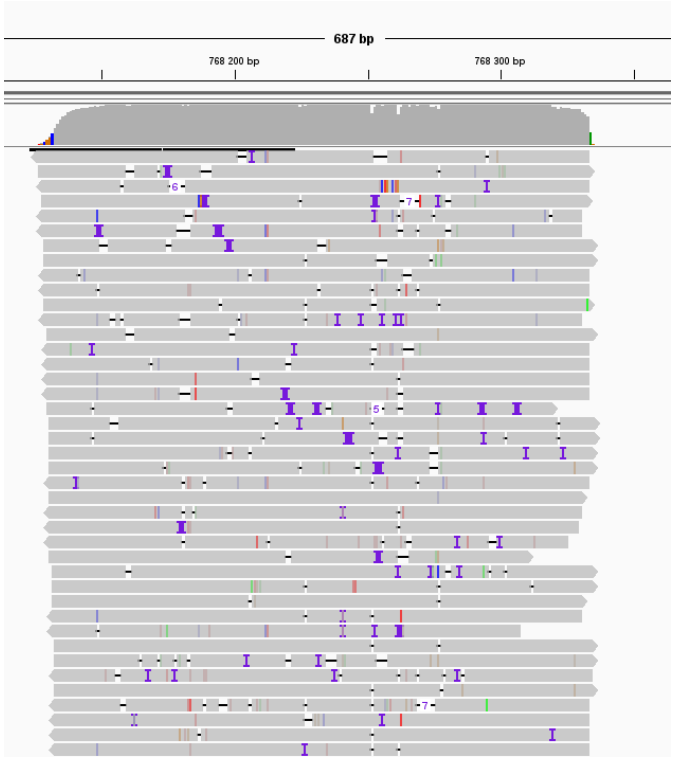
MinION sequencing

- Available since May 2015

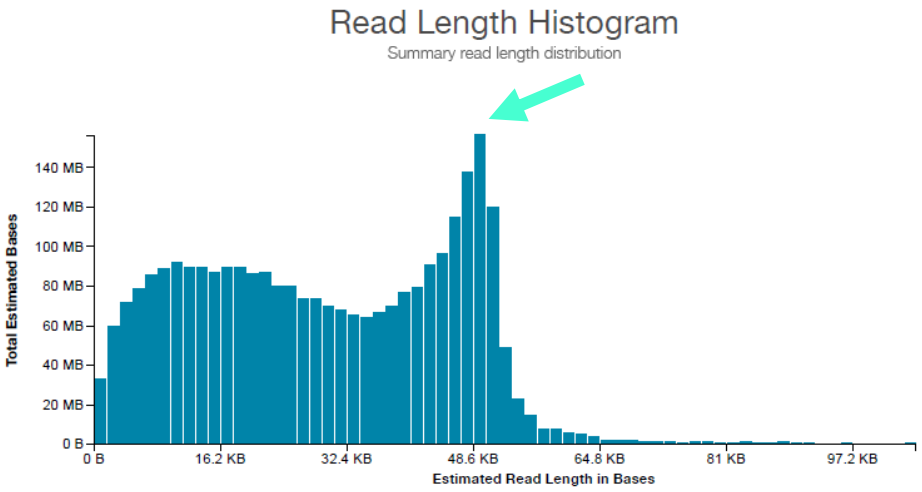
- 512 channels, 2 048 pores
- Theoretical output 50 Gb (run for 72 hours at 420 bases / second)
  - 30Gb-10x human genome
- Nanopores read the length of DNA presented to them, longest read so far: > 4 Mb
- Whole genome, targeted sequencing, whole transcriptome, metagenomics
- Newest flowcell and chemistry
  - Raw reads accuracies Q20 (1 in 100)- 99%
  - Consensus accuracies Q50 (1 in 1000 000)-99.999%
- Real-time analysis


Sample added to flow cell here.
Consumable flow cell where the biology and electronics come together for nanopore sequencing.
USB powered device; link to laptop, computer or MinIT to operate.
Bespoke **sensor array** with multiple **nanopores** for scaled-up sequencing.
Sensor chip works with custom **ASIC** for control and data acquisition.
Flow cell with **512 active channels**.




687 bp
768 200 bp
768 300 bp


Channel states panel
Run state: sequencing
280
Sequencing
130
Pore
36
Recovering
25
Inactive
41
Unclassified

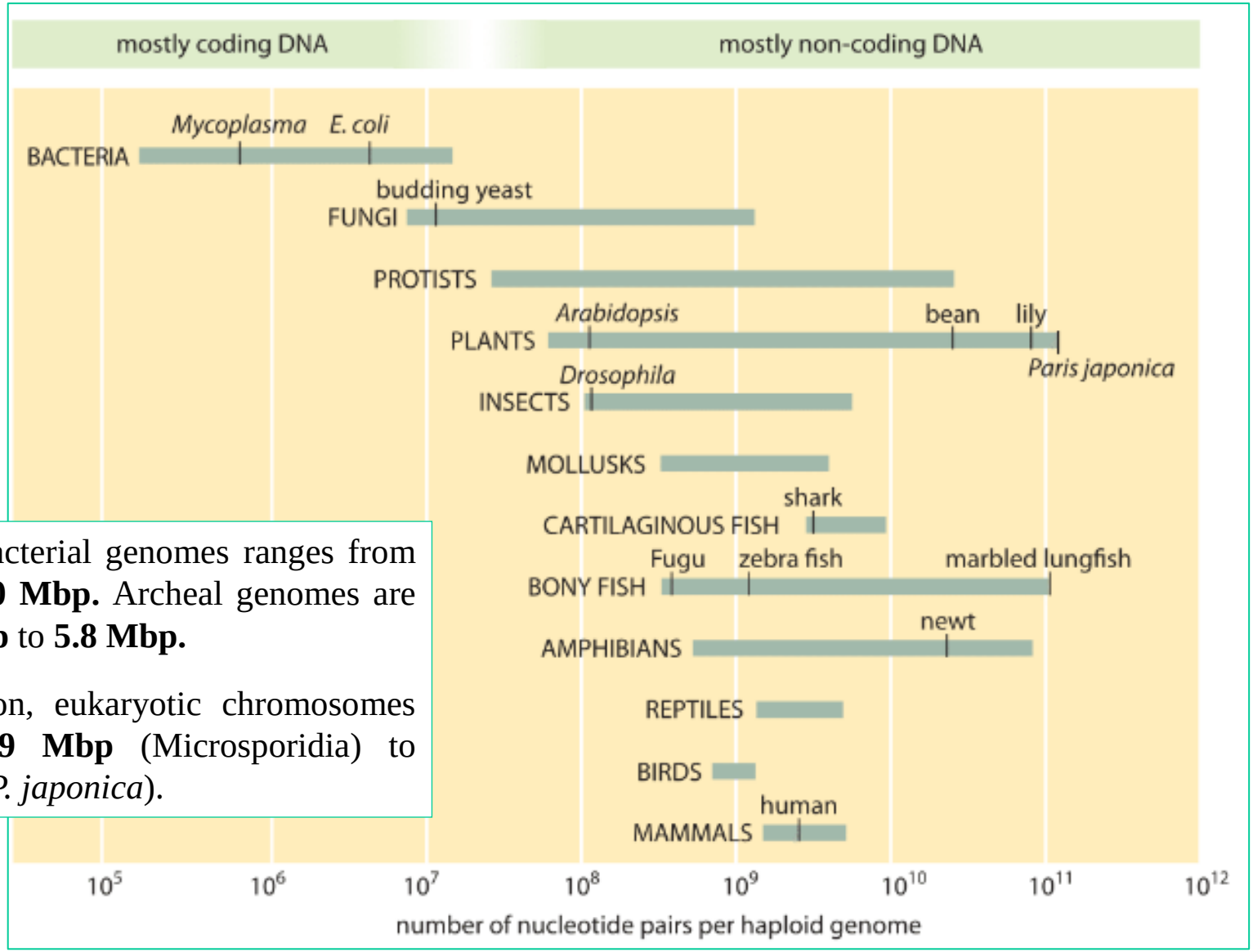

Read Length Histogram
Summary read length distribution
Total Estimated Bases
140 MB
120 MB
100 MB
80 MB
60 MB
40 MB
20 MB
0 B
0 B
16.2 KB
32.4 KB
48.6 KB
64.8 KB
81 KB
97.2 KB
Estimated Read Length in Bases

GridION

- 5 flowcells to be run concurrently or individually
- 250Gb, flexible, Integrated compute

PromethION

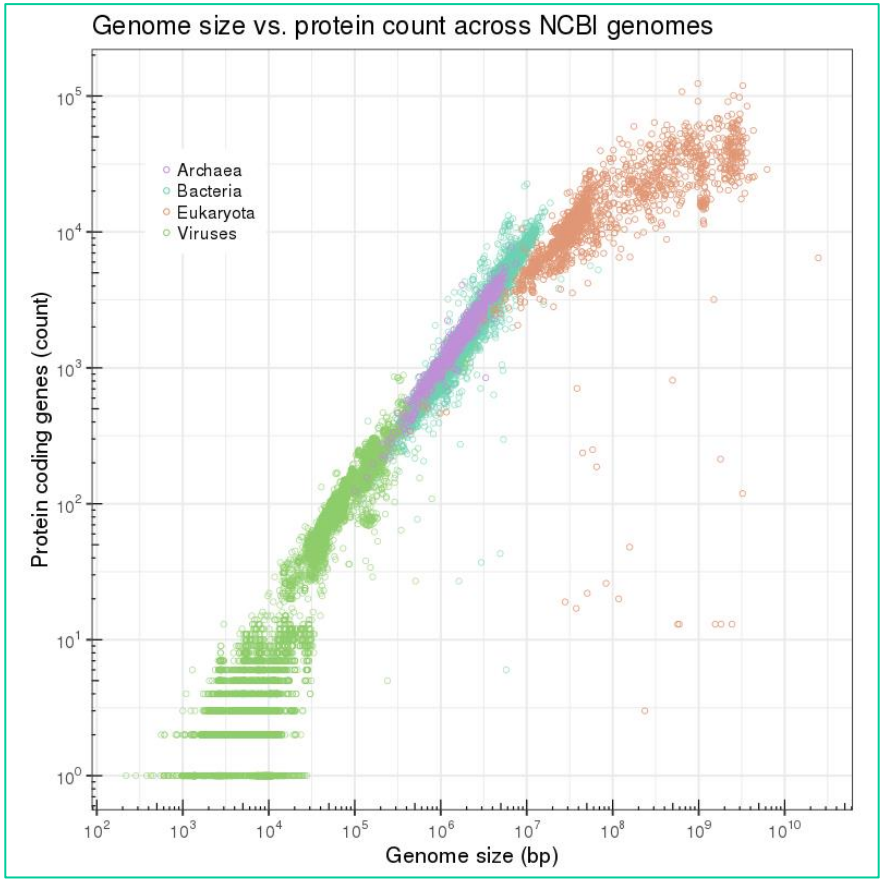
- 48 flowcells, 14Tb
- large scale, ultra-high throughput



- The size of bacterial genomes ranges from **0.6 Mbp** to **>10 Mbp.** Archeal genomes are smaller **0.5 Mbp** to **5.8 Mbp.**
- For comparison, eukaryotic chromosomes range from **2.9 Mbp** (Microsporidia) to **150,000 Mbp** (*P. japonica*).

http://book.bionumbers.org/how-big-are-genomes

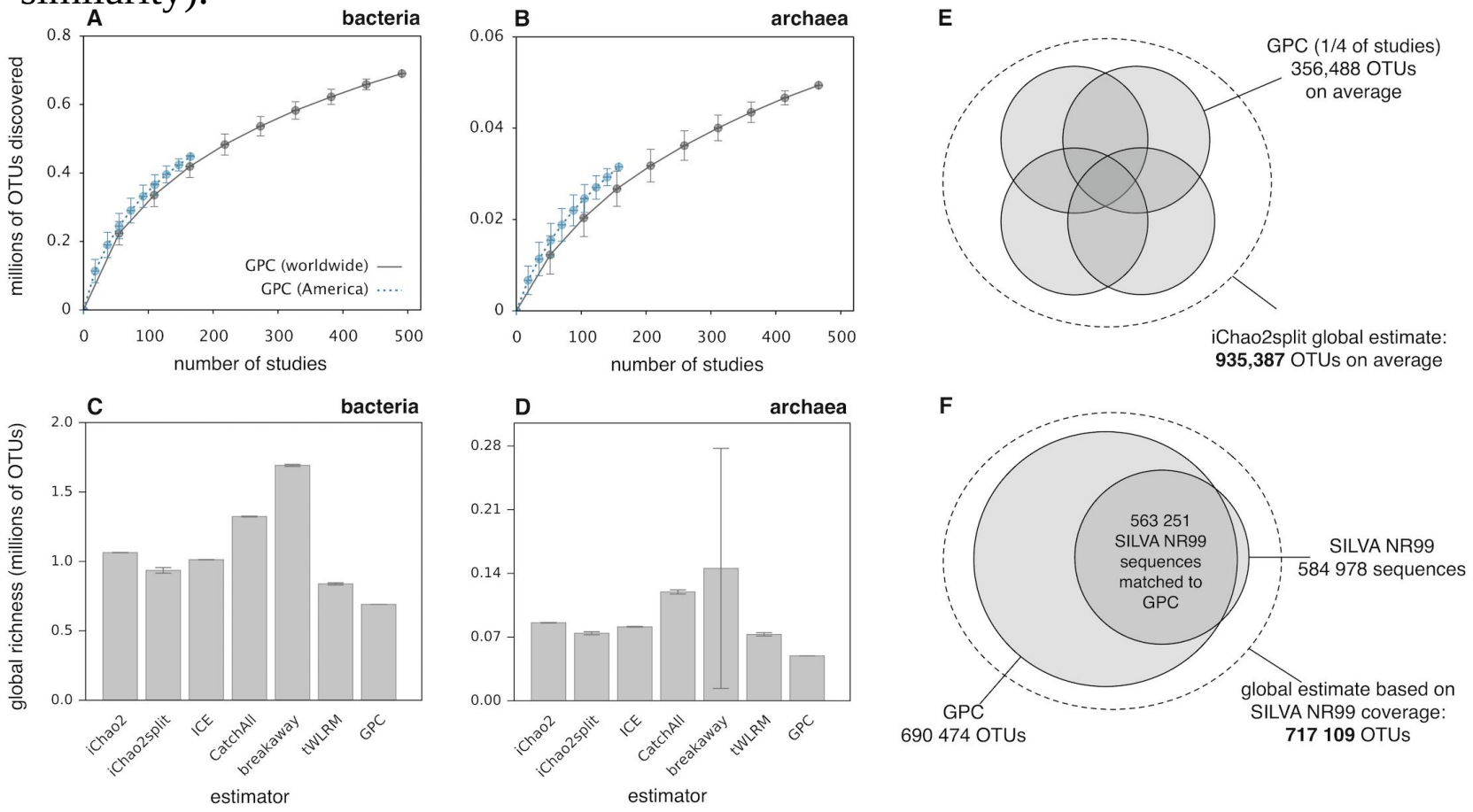
Genome size and structure, mutation rate



organism	genome size (base pairs)	protein coding genes
model organisms		
model bacteria *E. coli*	4.6 Mbp	4,300
budding yeast *S. cerevisiae*	12 Mbp	6,600
fission yeast *S. pombe*	13 Mbp	4,800
amoeba *D. discoideum*	34 Mbp	13,000
nematode *C. elegans*	100 Mbp	20,000
fruit fly *D. melanogaster*	140 Mbp	14,000
model plant *A. thaliana*	140 Mbp	27,000
moss *P. patens*	510 Mbp	28,000
mouse *M. musculus*	2.8 Gbp	20,000
human *H. sapiens*	3.2 Gbp	21,000
viruses		
hepatitis D virus (smallest known animal RNA virus)	1.7 Kb	1
HIV-1	9.7 kbp	9
influenza A	14 kbp	11
bacteriophage λ	49 kbp	66
Pandoravirus salinus (largest known viral genome)	2.8 Mbp	2500
bacteria		
C. ruddii (smallest genome of an endosymbiont bacteria)	160 kbp	182
M. genitalium (smallest genome of a free living bacteria)	580 kbp	470
H. pylori	1.7 Mbp	1,600
Cyanobacteria S. elongatus	2.7 Mbp	3,000
methicillin-resistant *S. aureus* (MRSA)	2.9 Mbp	2,700
B. subtilis	4.3 Mbp	4,100
S. cellulosum (largest known bacterial genome)	13 Mbp	9,400
eukaryotes - multicellular		
pufferfish *Fugu rubripes* (smallest known vertebrate genome)	400 Mbp	19,000
poplar *P. trichocarpa* (first tree genome sequenced)	500 Mbp	46,000
corn *Z. mays*	2.3 Gbp	33,000
dog *C. familiaris*	2.4 Gbp	19,000
chimpanzee *P. troglodytes*	3.3 Gbp	19,000
wheat *T. aestivum* (hexaploid)	16.8 Gbp	95,000
marbled lungfish *P. aethiopicus* (largest known animal genome)	130 Gbp	unknown
herb plant *Paris japonica* (largest known genome)	150 Gbp	unknown

http://book.bionumbers.org/how-big-are-genomes

There exist globally between **0.8** and **1.6** million prokaryotic OTUs (16S gene at 97% similarity).



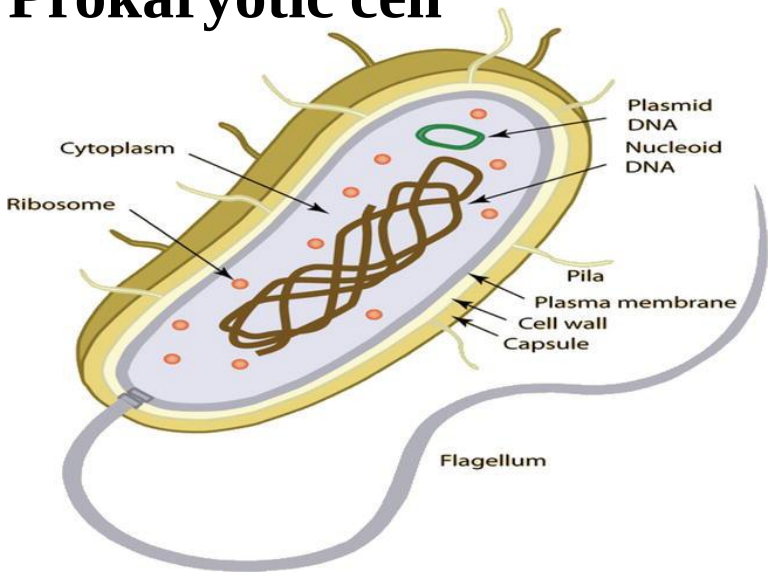
Estimating global prokaryotic OTU richness.
(A, B) Accumulation curves showing the number of bacterial (A) and archaeal (B) OTUs discovered, depending on the number of distinct studies included. Curves are averaged over 100 random subsamplings, and whiskers show corresponding standard deviations. Continuous curves were calculated using all studies (worldwide), while blue dashed curves were calculated using solely studies performed in the Americas or near American coasts. (C, D) Global OTU richness of Bacteria (C) and Archaea (D), estimated using the iChao2, iChao2split, ICE, CatchAll, breakaway, and tWLRM estimators. The number of OTUs discovered by the GPC is included for comparison (last bar).

Bacterial genomics – *Escherichia coli*

„*E. coli* is a single species with numerous recognized roles, from lab workhorse to beneficial intestinal commensal or deadly pathogen. The extant strains have disparate lifestyles as a result of differential niche expansion since their divergence 25–40 million years ago, ten times longer than the estimated divergence between chimpanzees and humans.“

E. coli is a Gram-negative, nonsporulating, facultative anaerobe. Cells are typically rod-shaped (2.0 x 0.5 μm).

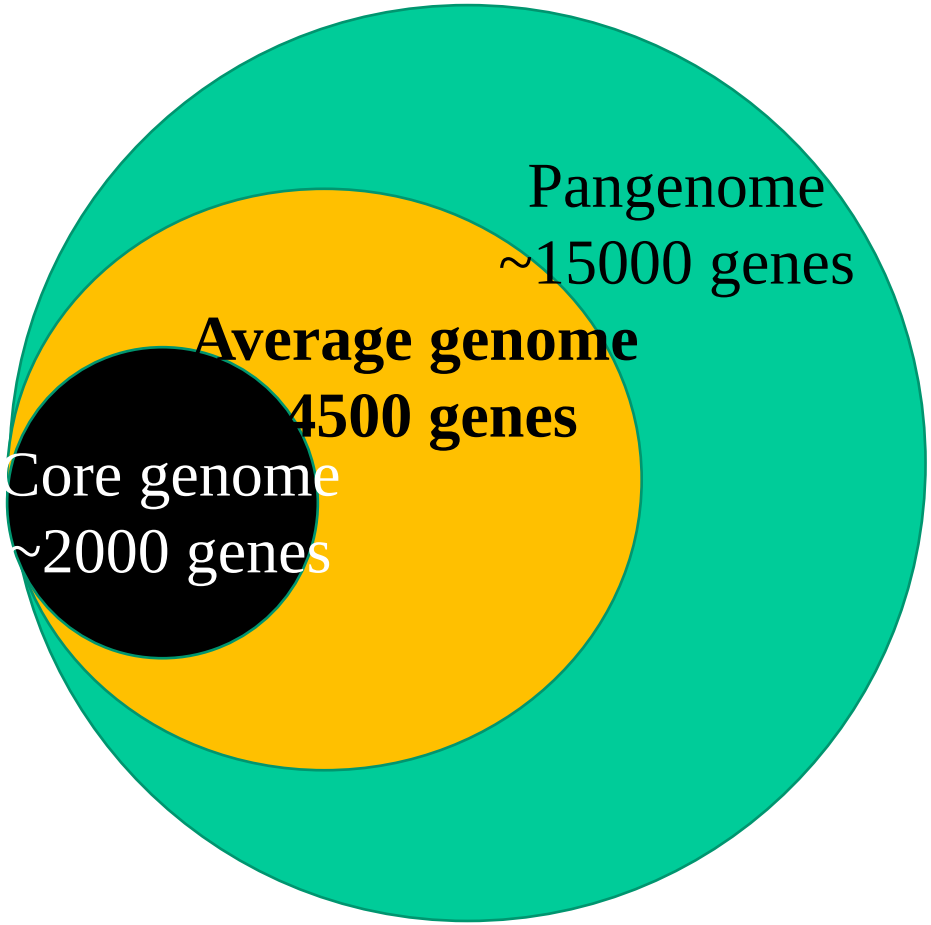
Prokaryotic cell



https://www.intechopen.com/chapters/84764

Hendrickson 2009 ; DOI: 10.1371/journal.pgen.1000335

E. coli genome types



Genome is the set of genetic information (genes) in one cell (strain/organism).

Core genome is the set of genes found across all strains.

Pangenome is the set of all genes from all strains.

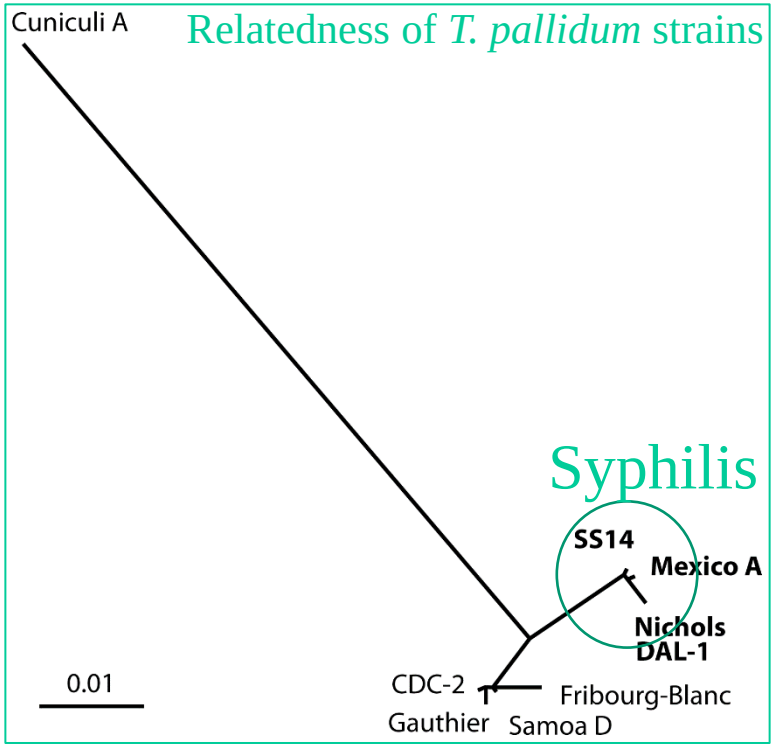
Accessory (cloud) genome is the set shared within only one or some strains.

>50% of *E. coli* genetic information is variable

Genetically monomorphic pathogens

T. pallidum genome (1.14 x 10^6 bp) contains minimal sequence diversity among syphilis strains (**99.99% identity).**

- there are two genomic groups



Other examples of genetically monomorphic pathogens are *Bacillus anthracis* (**anthrax**), *Yersinia pestis* (**plague**), *Burkholderia mallei* (**glanders**), *Escherichia coli* O157:H7 (**HUS**), *Mycobacterium tuberculosis* (**tuberculosis**), *Mycobacterium leprae* (**leprae**), and *Salmonella enterica* serovar Typhi (**typhoid fever**).

Mikalová et al.2010, https://doi.org/10.1371/journal.pone.0015713

Minimal bacterial genome

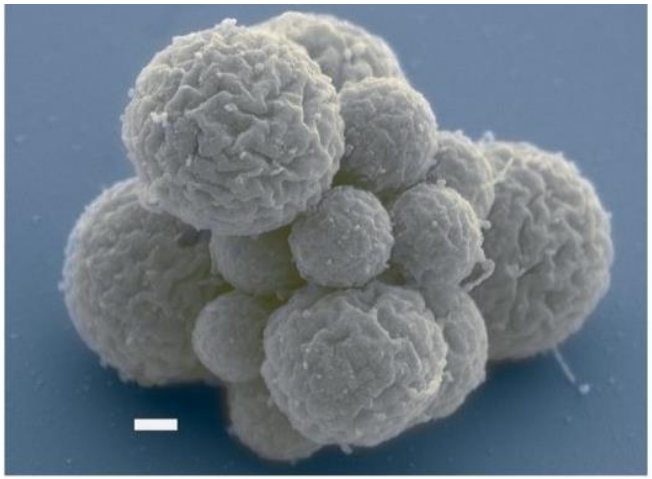
Mycoplasma genitalium (a human urogenital pathogen) has **the smallest genome** of all solitaire organisms**.** Its genome size is 580 kb and contains only **470 genes**.

It has a minimal metabolism and little genomic redundancy. One third of the proteins have unknown function.

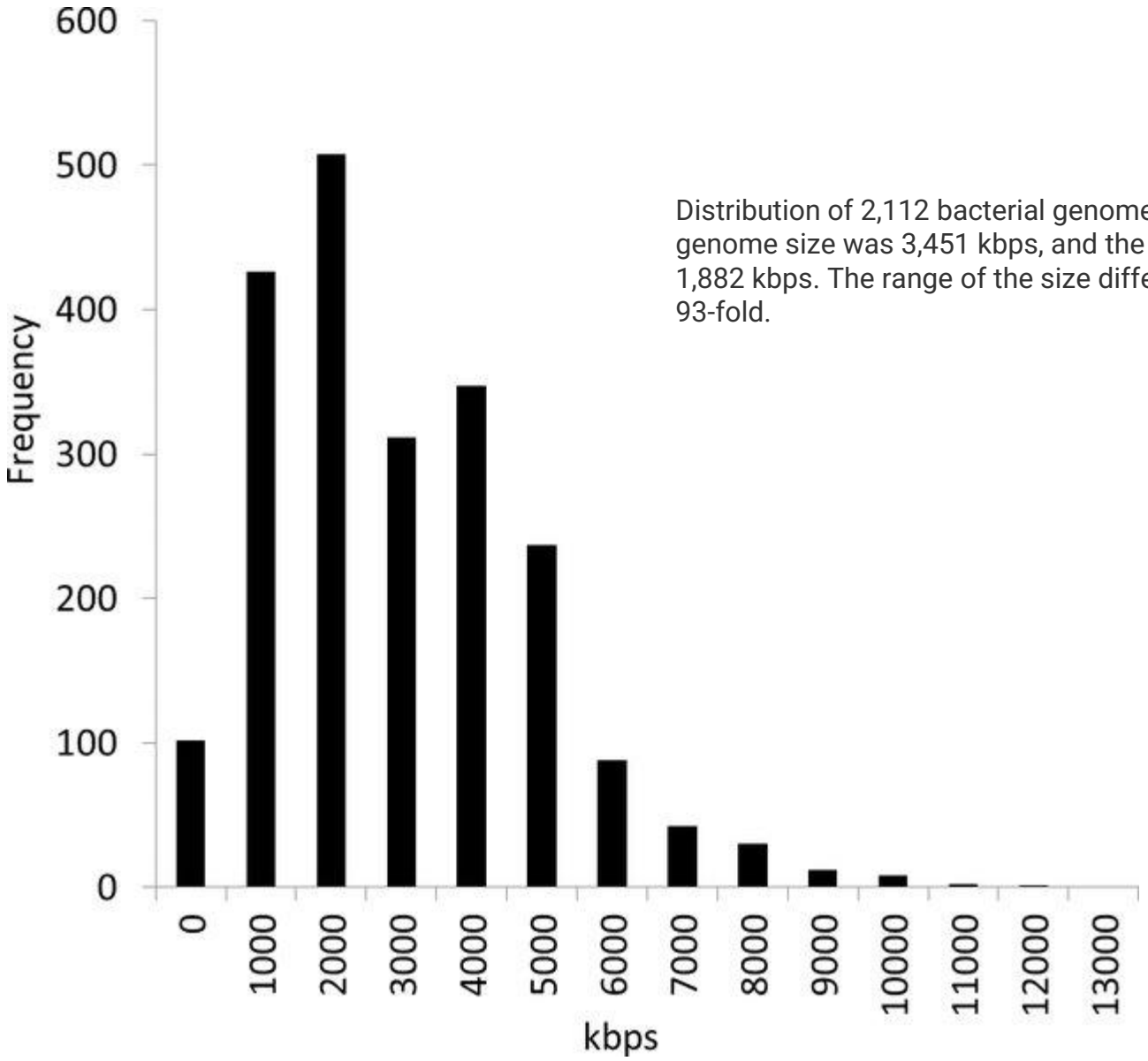
Synthetic genomes

JCVI-syn3.0, synthetic bacterium, encodes only 473 genes (genome 531.56 kbp).

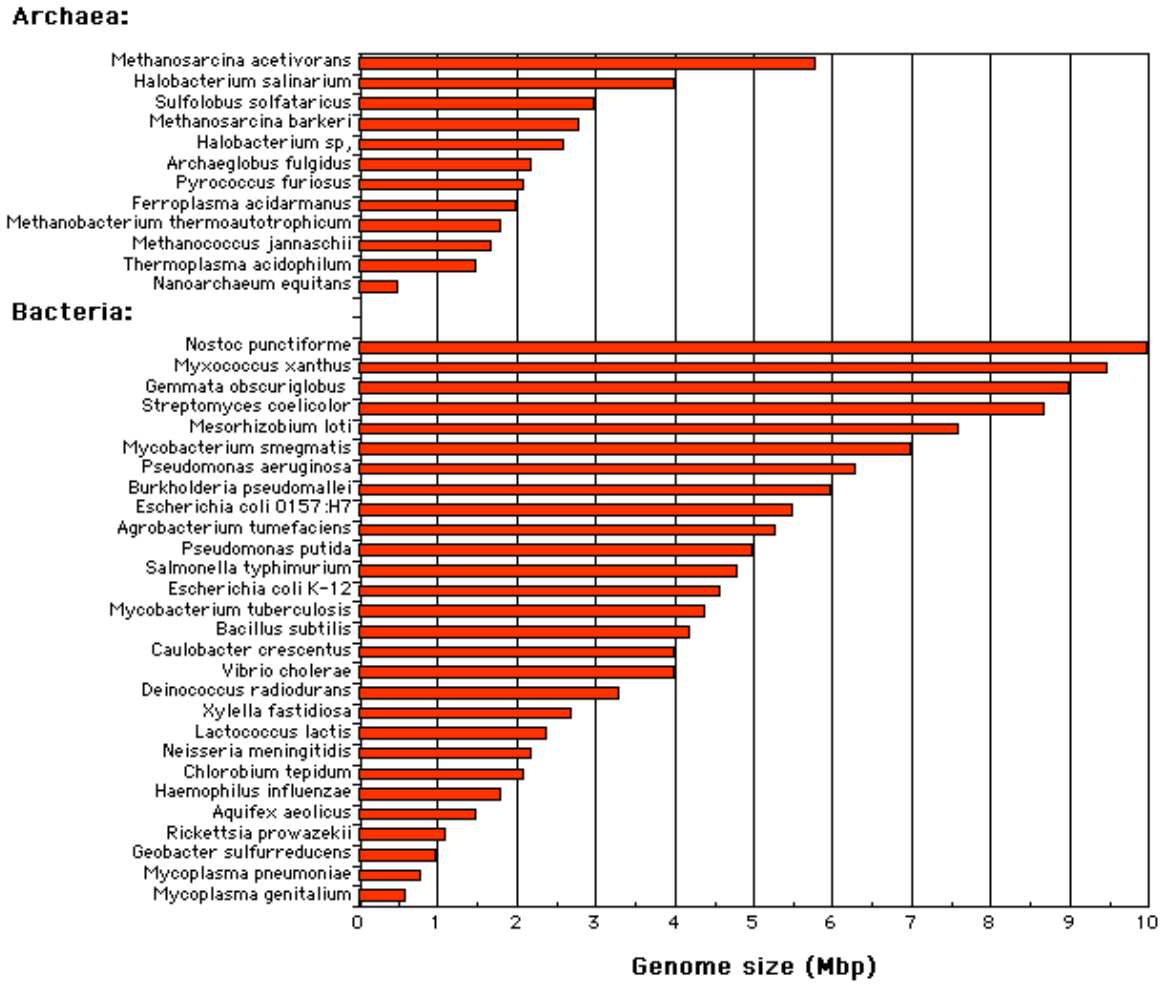
Proposed minimal set of 256 genes.



Hutchinson et al. 2016; https:// doi: 10.1126/science.aad6253.
Mushegian AR, Koonin EV 1996; https://doi.org/10.1073/pnas.93.19.10268



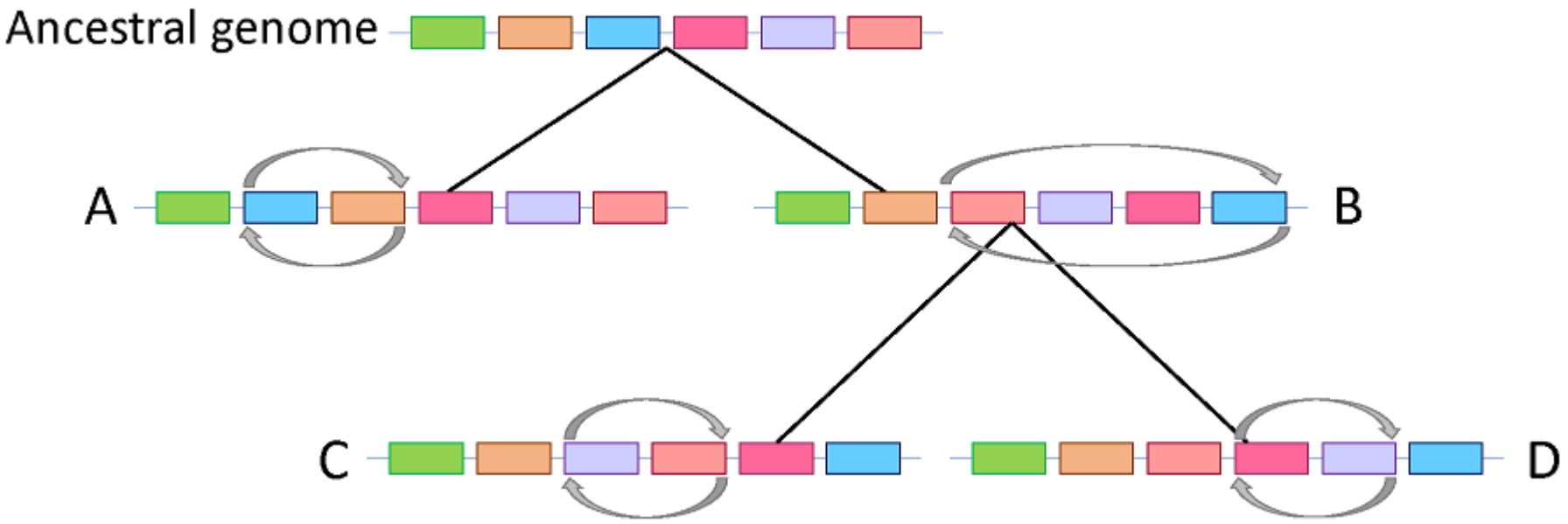
Distribution of 2,112 bacterial genome sizes. The average genome size was 3,451 kbps, and the standard deviation was 1,882 kbps. The range of the size difference was approximately 93-fold.



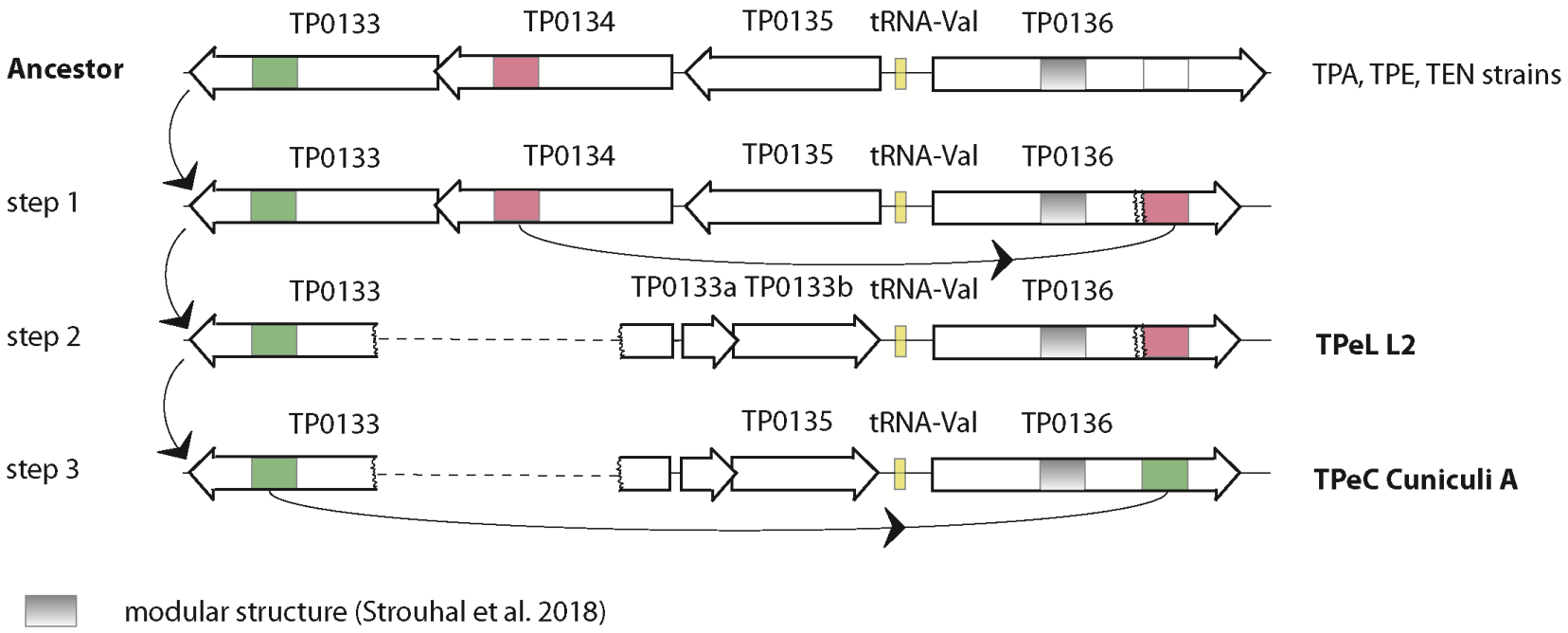
Not long ago it was thought that all prokaryotic genomes (both Bacteria and Archae) were much smaller than eukaryotic genomes. However, the application of new techniques for constructing physical maps and whole genome sequencing has demonstrated that there is tremendous diversity in the size and organization of prokaryotic genomes. The following figure shows some examples of genome sizes of Bacteria and Archae. The size of Bacterial chromosomes ranges from **0.6 Mbp to over 10 Mbp**, and the size of Archael chromosomes range from **0.5 Mbp to 5.8 Mbp.** (For comparison, Eukaryotic chromosomes range from **2.9 Mbp** (Microsporidia) to well over **4,000 Mbp**, although the largest genomes are littered with a tremendous amount of repetitive "junk" DNA.)

Genome rearrangements and evolution

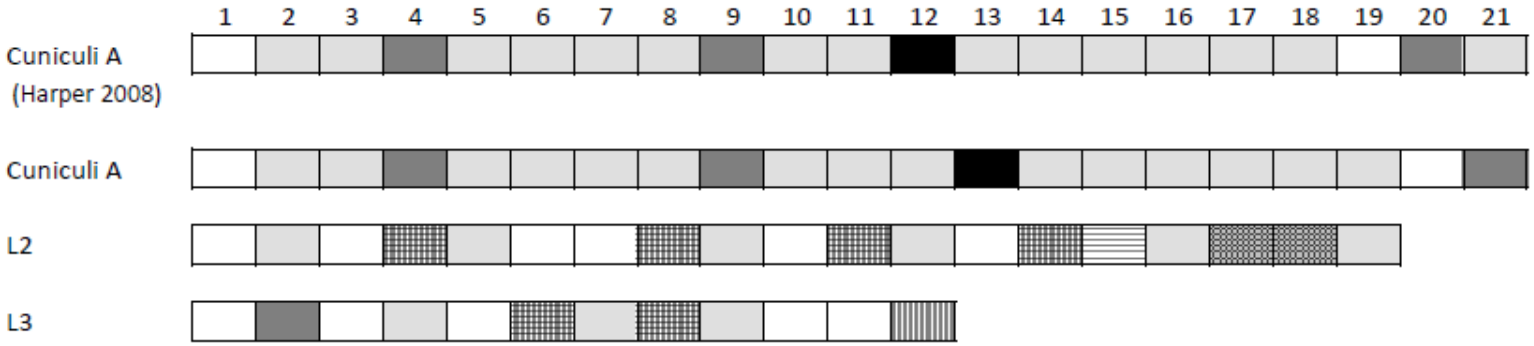
Evolution of genomes caused by inversions. Different colored boxes represent the gene blocks. Grey arrow shows the inverted region of the genome. The identification of the inversions can reveal the evolutionary history of the organisms as depicted here.



A schematic representation of evolution of the TP0136 locus in TPeL and TPeC trepoenemes. The evolution of this regions required several steps including two gene conversion events and one deletion. The part of the TP0136 showing modular structure was not affected during these changes.



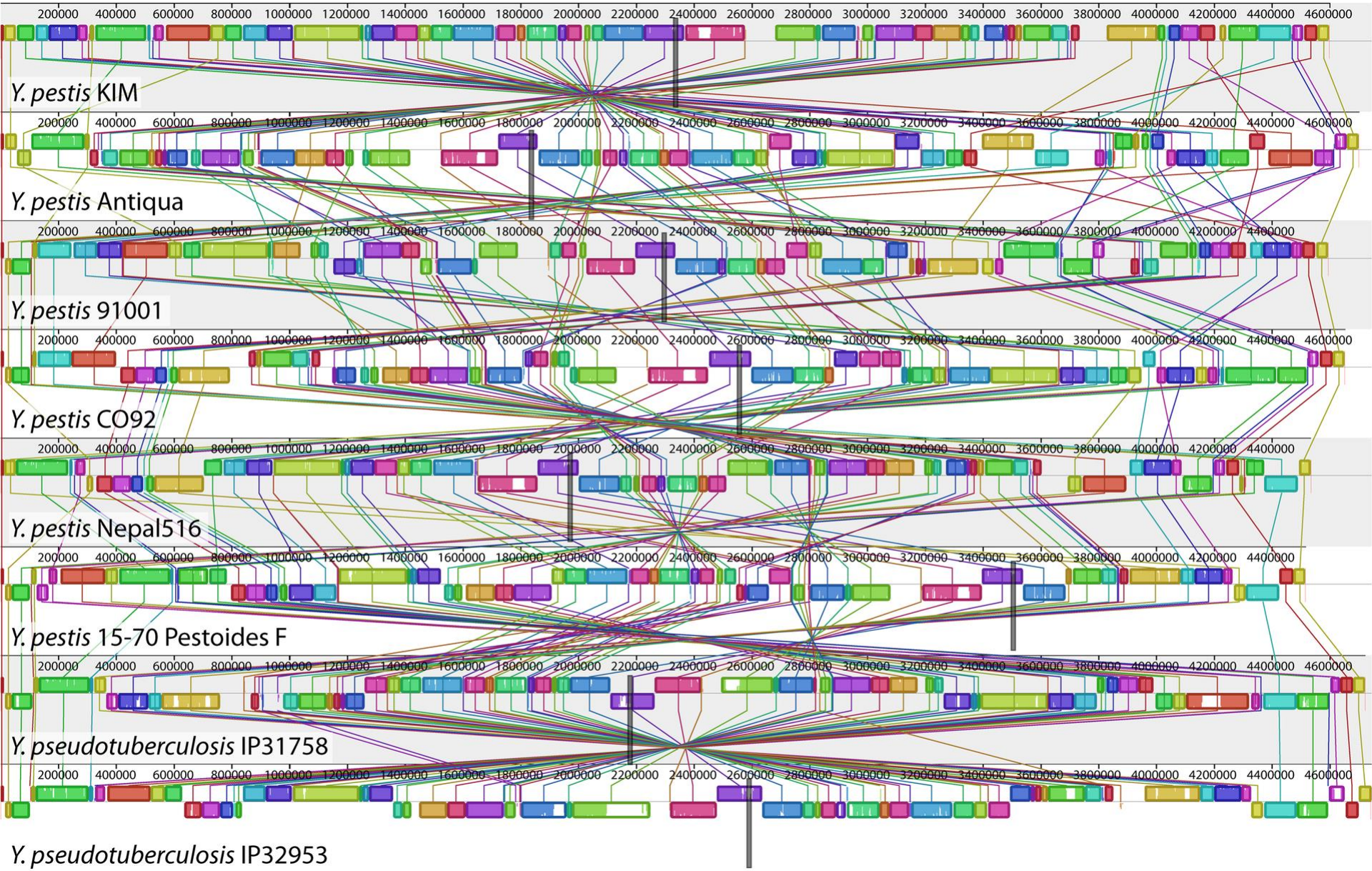
Composition of repeat motif regions observed in the TPeL L2, L3, and TPeC Cuniculi A genomes.



Repeat Units Types and their aminoacid sequences

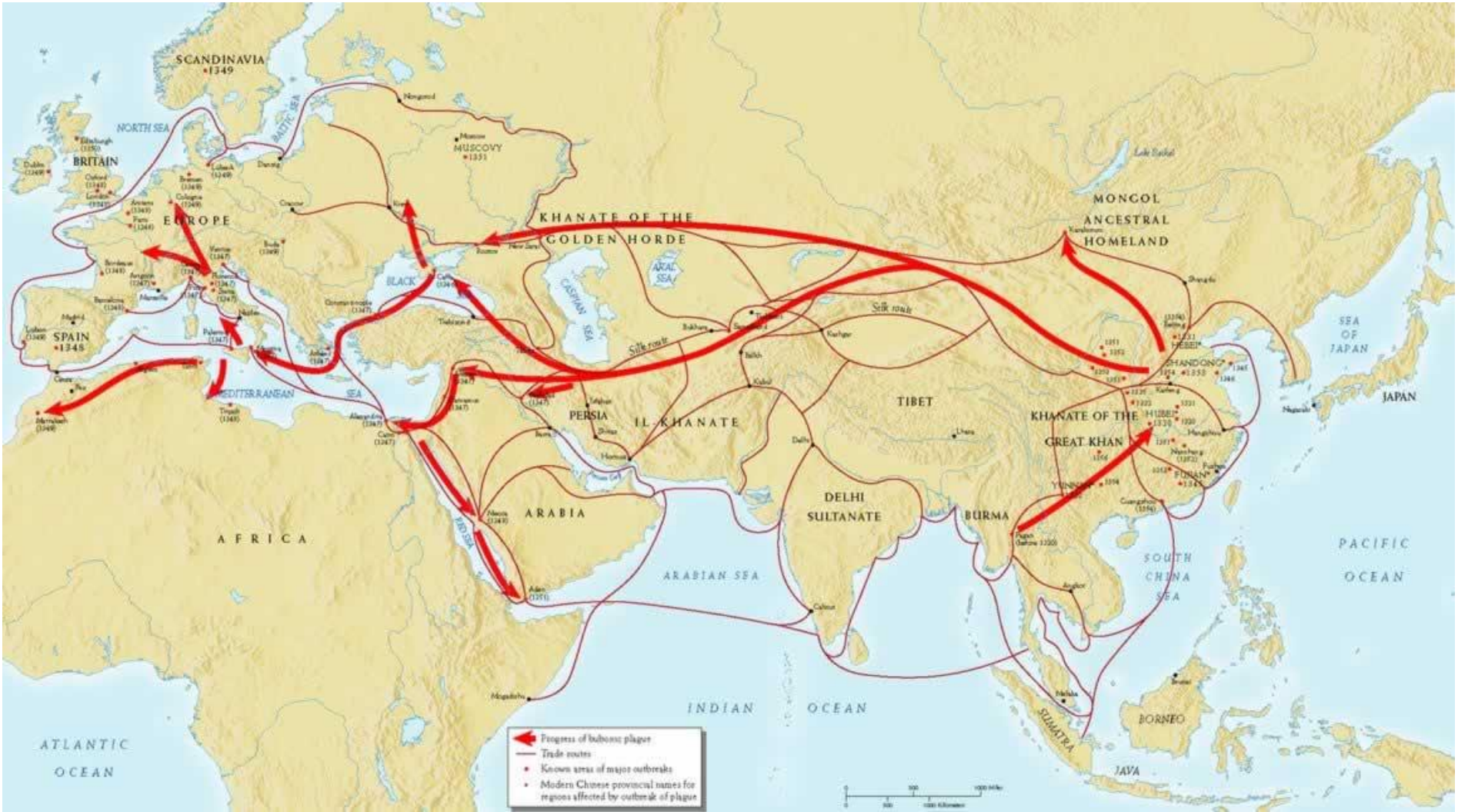
		CA	L2	L3	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	Type IV	■	■	■	R	E	V	E	D	V	P	K	V	V	E	P	A	S	E	R	G	G	R	E
	Type V	■	■	■	R	E	V	E	D	A	P	G	V	V	E	P	A	S	E	R	G	G	R	E
	Type VI	■		■	R	E	V	E	D	V	P	G	V	V	E	P	A	S	E	R	G	G	R	E
	Type VII	■			R	E	V	E	D	A	P	K	V	V	E	P	A	S	E	R	G	G	R	E
	Type VIII		■	■	R	E	V	E	D	V	P	K	V	V	E	P	V	F	E	R	G	G	G	E
	Type IX		■		R	E	V	E	D	A	P	G	V	V	E	P	V	F	E	R	G	G	G	E
	Type X		■		R	E	V	E	D	V	P	K	V	V	E	P	A	S	E	R	G	G	G	E
	Type XI			■	R	E	V	E	D	V	P	K	V	V	E	P	V	F	E	R	G	G	R	E

Genome rearrangements and evolution

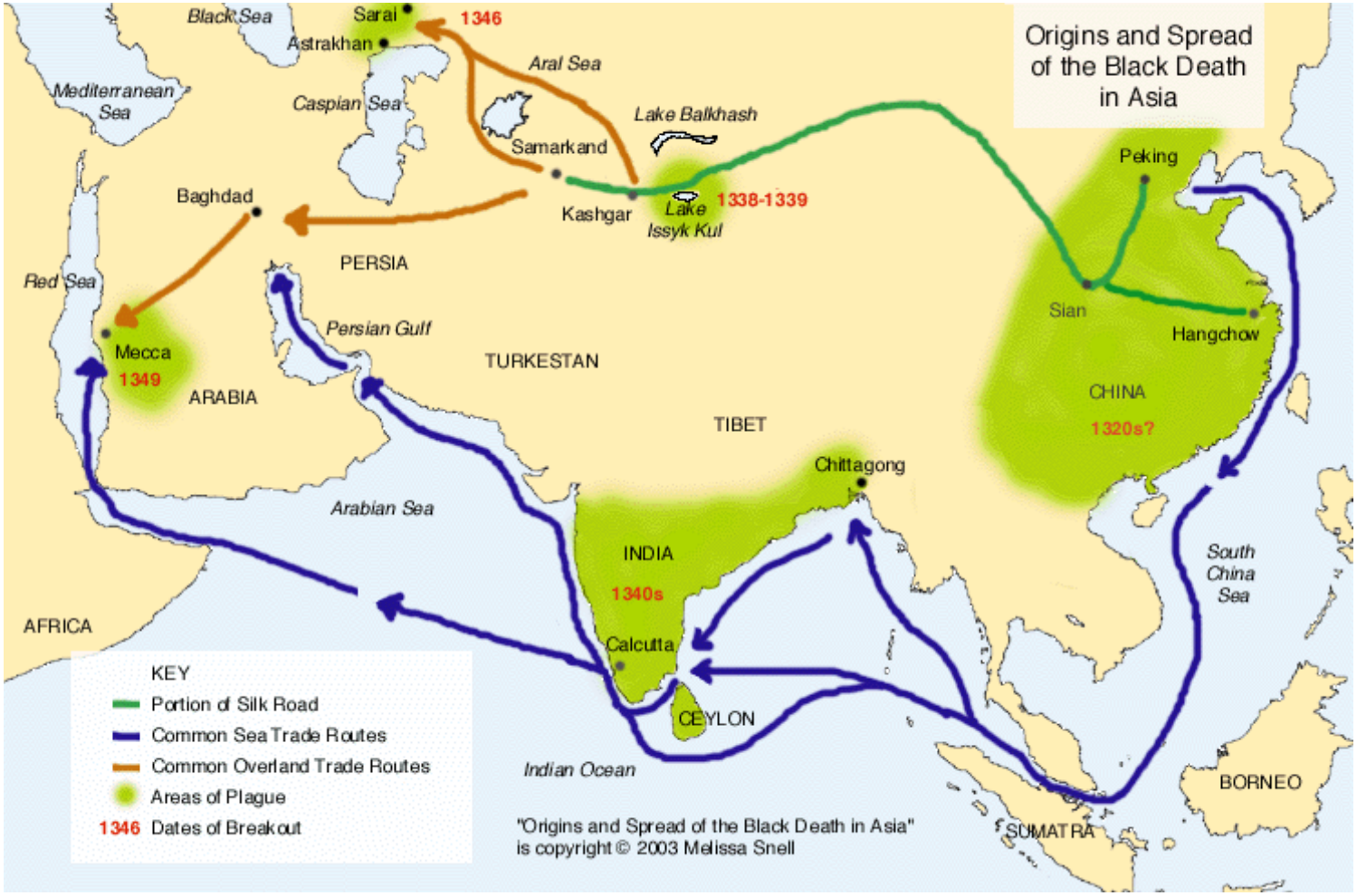


Black death, *Y. pestis*

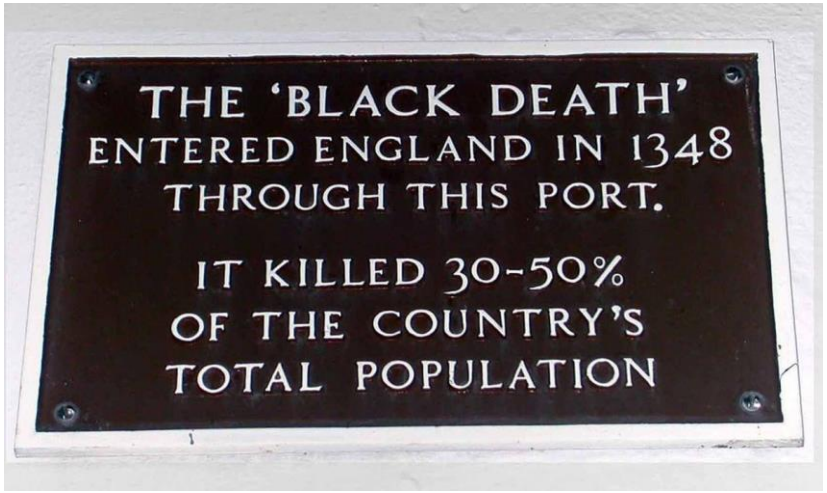
Some of the pre-existing conditions necessary for this occurrence include war, famine, weather and interactions between people. At the end of the 13th century and into the first half of the 14th century, disastrous weather had severe implications worldwide (**Little Ice Age**). The Great Famine struck all of Northern Europe in the 14th century, resulting in hunger and malnutrition.



The Black Death first **<u>came</u> to Europe in 1347**, engulfing the continent in sickness and turmoil. Transported through fleas (and, by extension, rats stowed away on seafaring ships) the Black Death killed off as much as **50 percent** of the population in less than a decade, and outbreaks of the plague continued to sicken Europeans until the disease disappeared in the early 19th century.



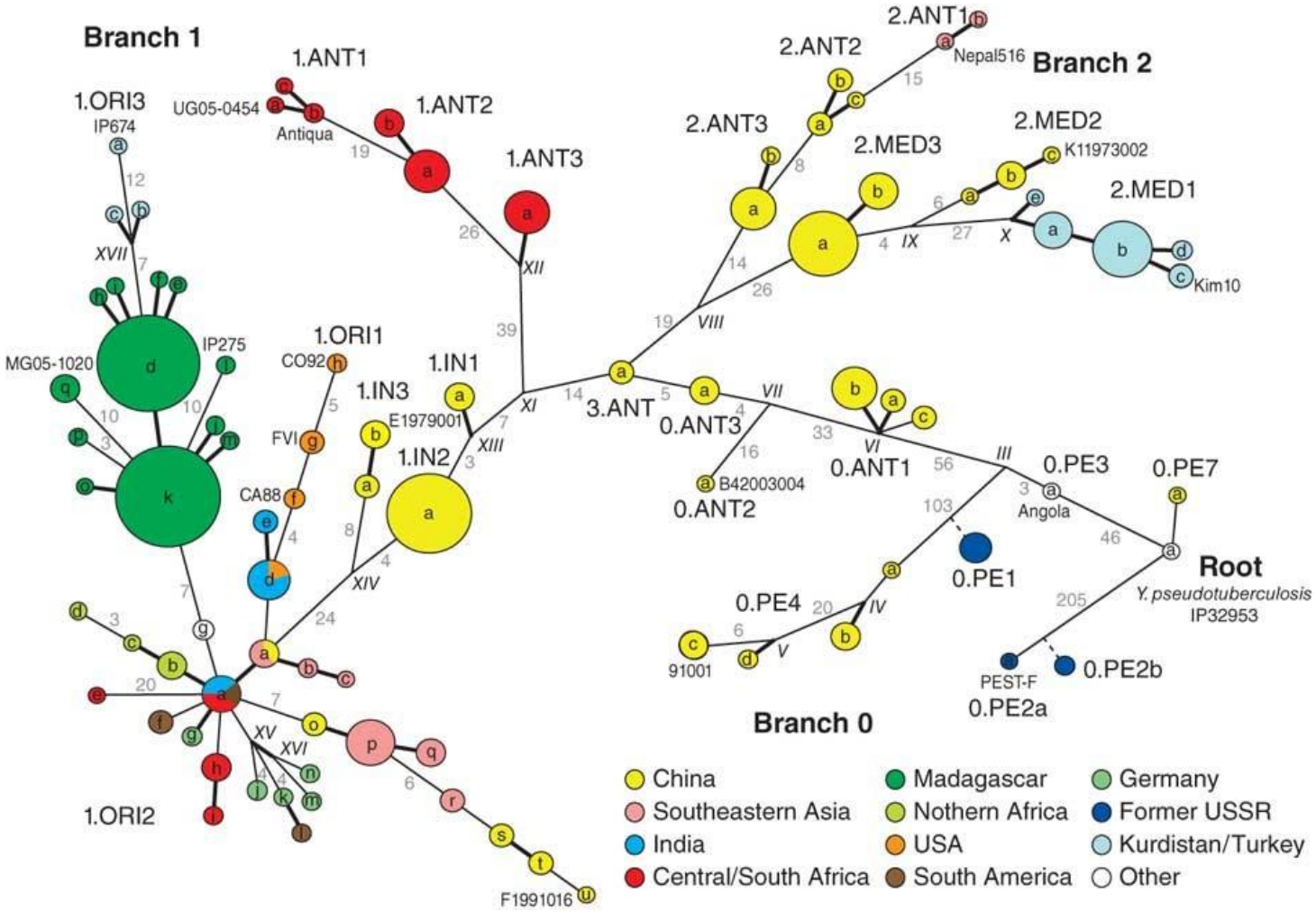
Black death, *Y. pestis*



Scientists think that plague bacteria circulate at low rates within populations of certain rodents without causing excessive rodent die-off. This is called the **enzootic** cycle. Occasionally, other species become infected, causing an outbreak among animals, called an **epizootic**. Humans are usually more at risk during, or shortly after, a plague epizootic.

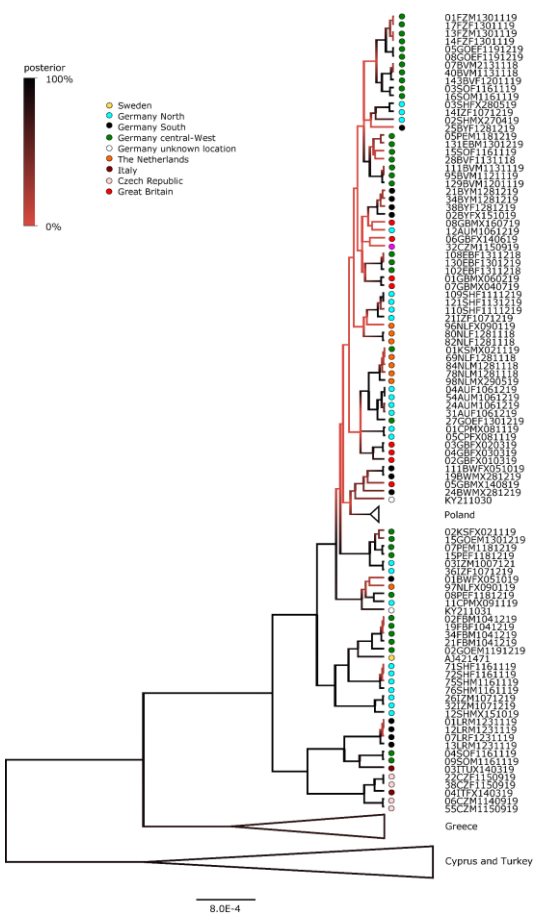
The plague outbreaks in Asia were consistently followed by flare-ups in Europe roughly 15 years later—just enough time for Asia's rodent reservoirs to travel trade routes into Europe. In other words, European rats aren't to blame for the Black Death; it's their furry Asian counterparts.

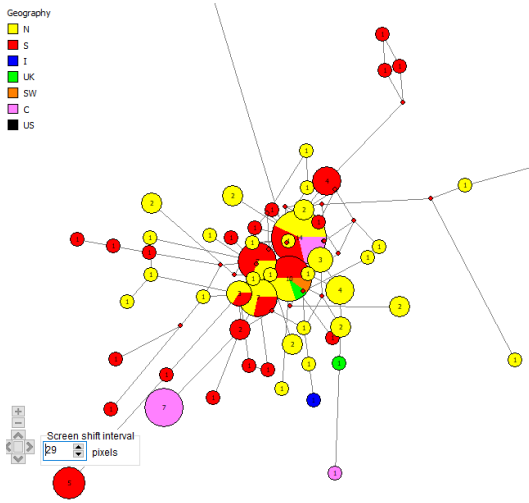
Fully parsimonious minimal spanning tree of 933 SNPs for 282 isolates of *Y. pestis* colored by location.

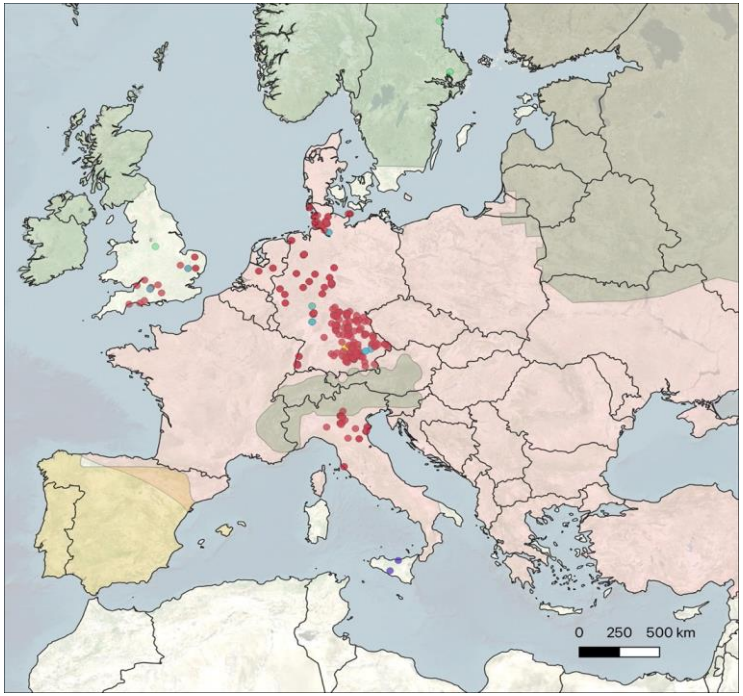


The phylogenetic analysis suggests that ***Y. pestis*** evolved in or near China and spread through multiple radiations to Europe, South America, Africa and Southeast Asia, leading to country-specific lineages that can be traced by lineage-specific SNPs. All 626 current isolates from the United States reflect one radiation, and 82 isolates from Madagascar represent a second radiation. Subsequent local microevolution of *Y. pestis* is marked by sequential, geographically specific SNPs.

Hare syphilis in Europe, more than half of hare population is infected with *T. paraluisleporidarum*







MJ network of all tested samples based on concatenated sequences (TP0548 and TP0488). Only parsimony informative sites (n = 63) were used for construction of network. Color code correspond to geographic region (N - North Germany, S - South Germany). While there is a significant bootstrap support for the group of Swedish samples and Cuniculi A (USA), the bootstrap support for two clusters in the fig. below is very low.

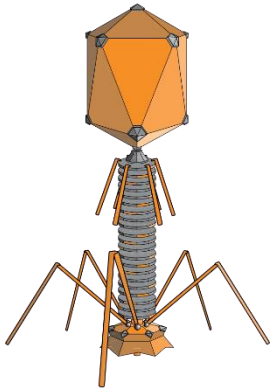
There is no clear clustering of samples based on geographical origin.

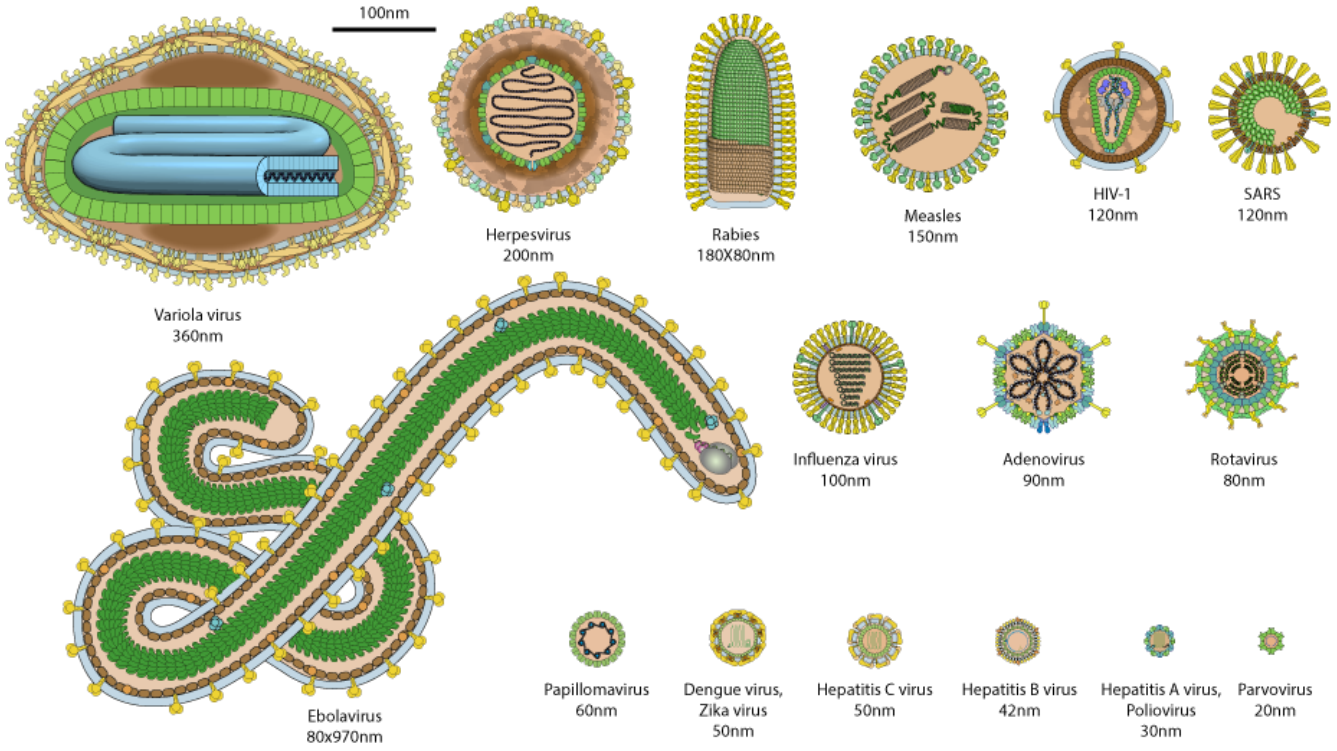
Viral genomics

Viral particles (**virions**) consist of (i) **genetic material** (DNA or RNA), (ii) protein coat (**capsid**), which surrounds and protects the genetic material, and in some cases (iii) lipid **envelope**.

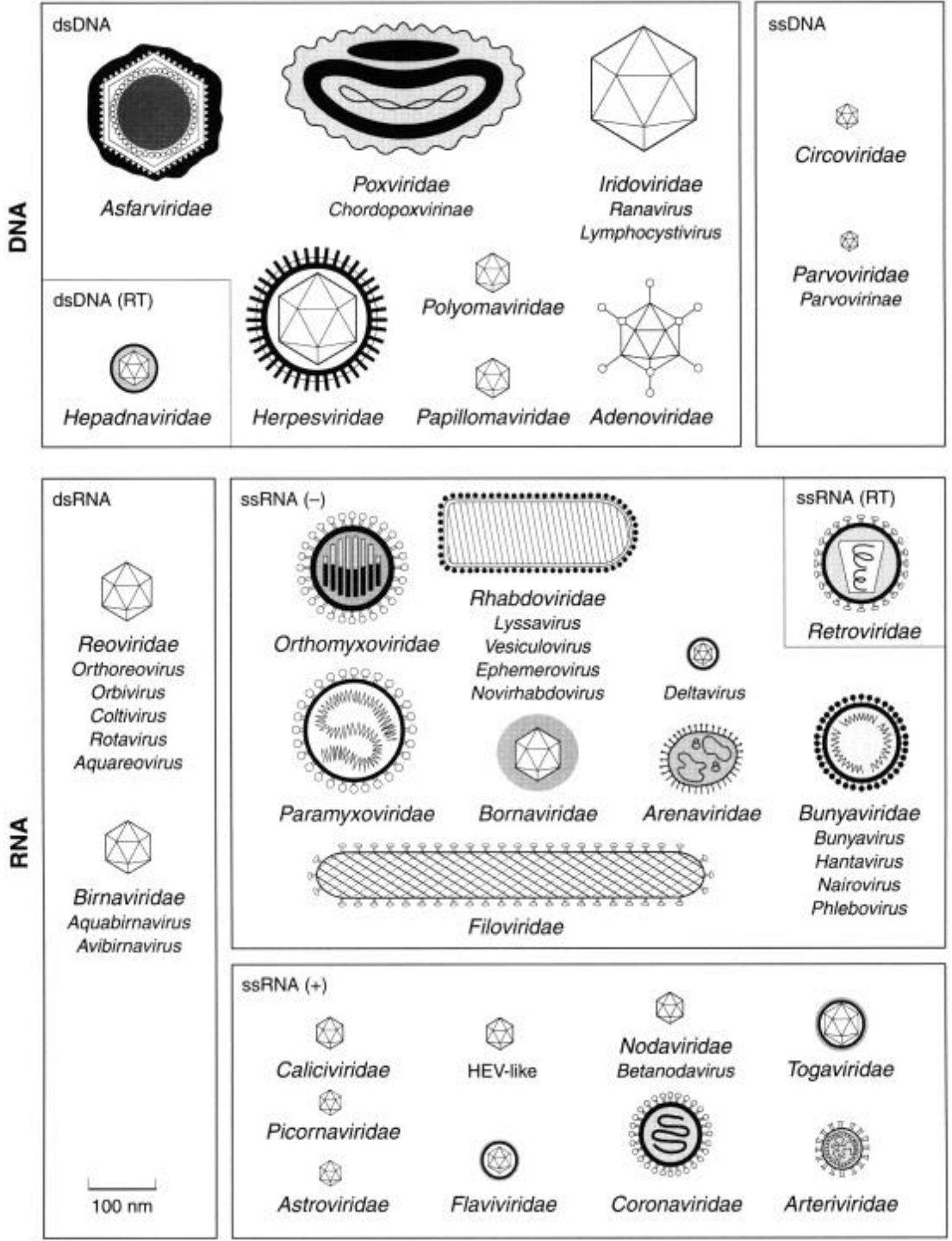
Virion morphology – helical, icosahedral, and complex structures.

Bacteriophage





https://en.wikipedia.org/wiki/Virus#Etymology

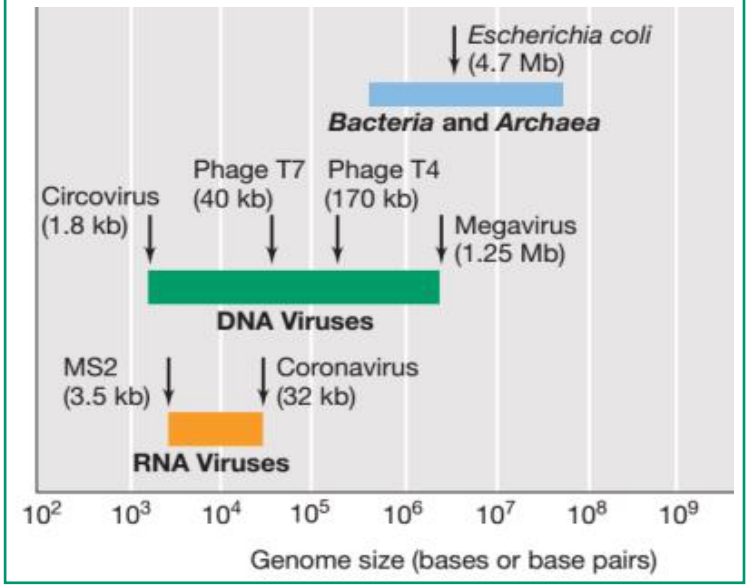


Families and genera of viruses infecting vertebrates.

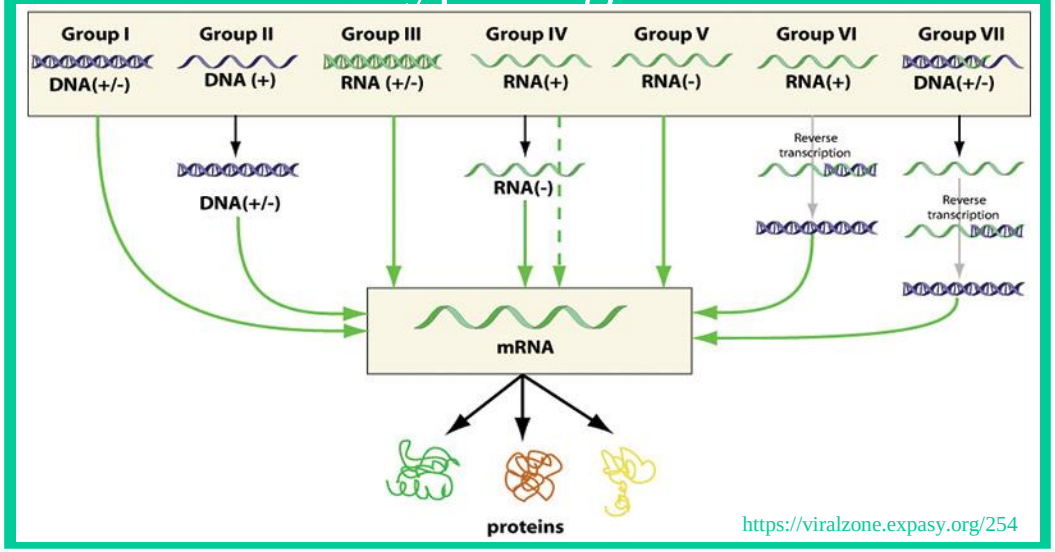
Viral genomes

- DNA or RNA
- Single or double stranded
- Linear or circular

Size comparison of viral and prokaryotic genomes.

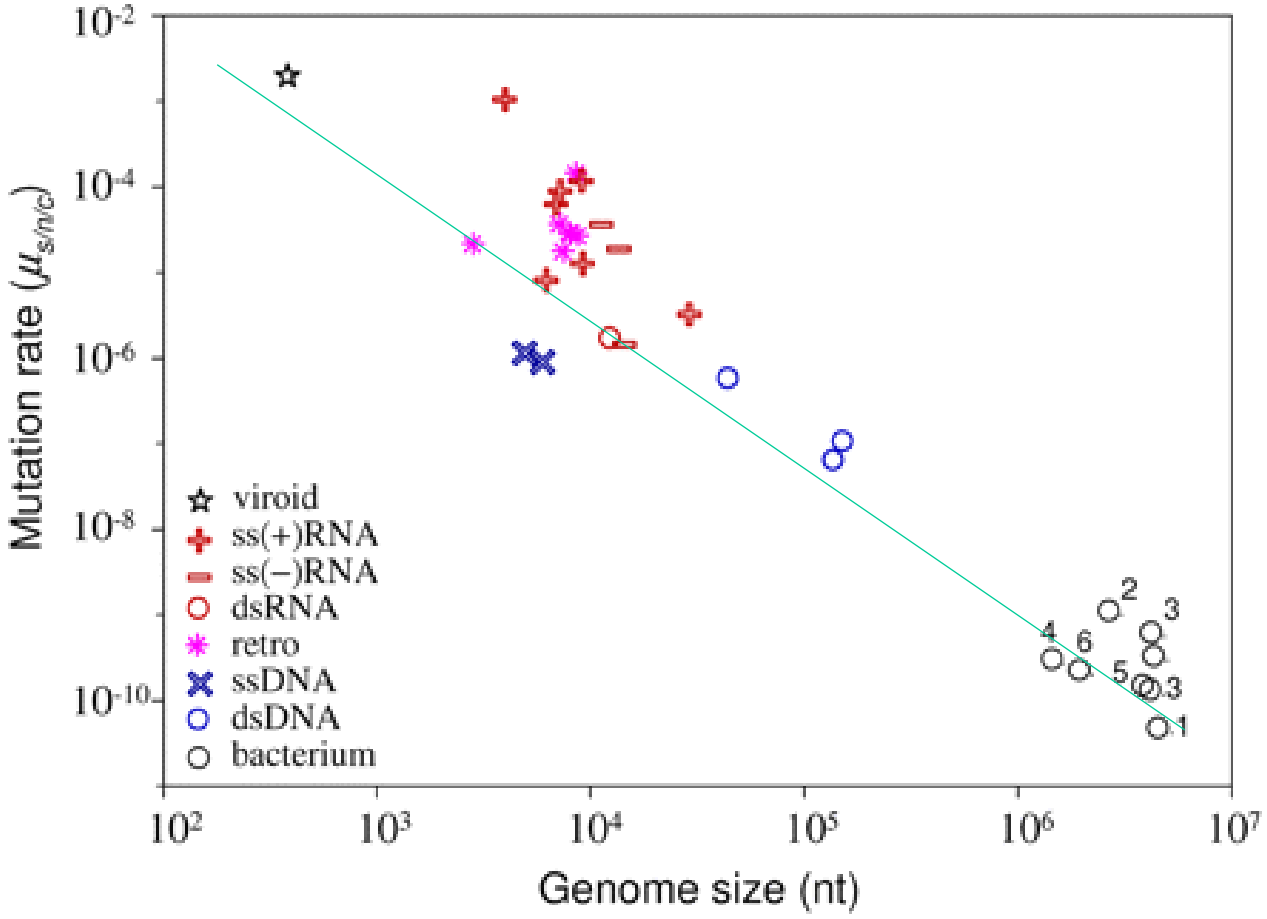


The Baltimore classification clusters viruses into 7 groups depending on their type of genome.



*The **positive-sense genome** acts as mRNA and it is directly translated into viral proteins.*

*The **negative-sense genome** is complementary to mRNA molecules, which are synthesized by the viral RNA-dependent RNA polymerase (RdRp).*

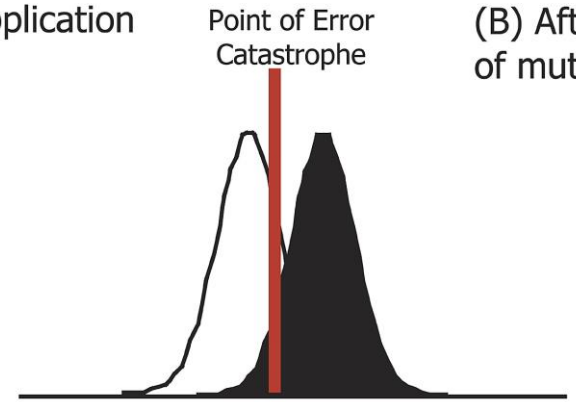


Drake's rule is a notoriously universal property of genomes from microbes to mammals—the number of (functional) mutations per-genome per-generation is approximately constant within a phylum, despite the orders of magnitude differences in genome sizes and diverse populations' properties.

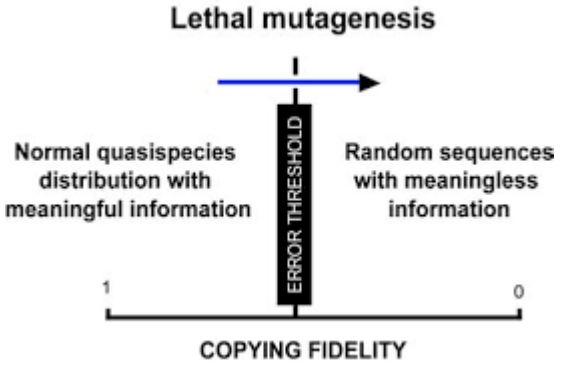
0.0033 per generation

Relationship between mutation rate and genome size, with major virus groups indicated. Values for viroids and bacteria, the two adjacent levels of biological complexity, are also plotted. The mutation rate is expressed as the number of substitutions **per nucleotide per generation**, defined as a cell infection in viruses ($\mu_{s/n/c}$).

(A) Before application of mutagen

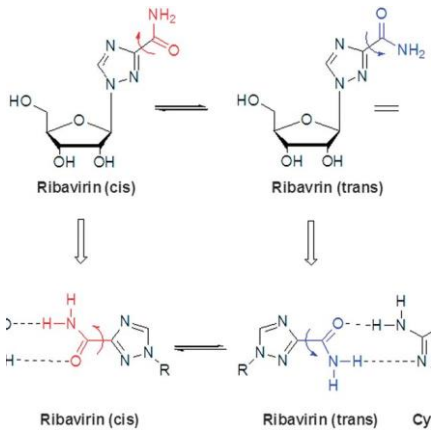


(B) After application of mutagen




```
Seq. 1    TCC TTC CAG ACC TAA
                *
Seq. 2    TCC TTA CAG ACC TAA
```

```
Seq. 1    TCC TTC CAG ACC TAA
            *   *       *   *
Seq. 2    TCA TTA CAG ACT TAG
```



The lethal mutagenesis mechanism of ribavirin. The ribavirin *cis* conformer can pair with uridine by mimicking adenosine, and the trans conformer can pair with cytidine by mimicking guanosine.

Ebola virus disease

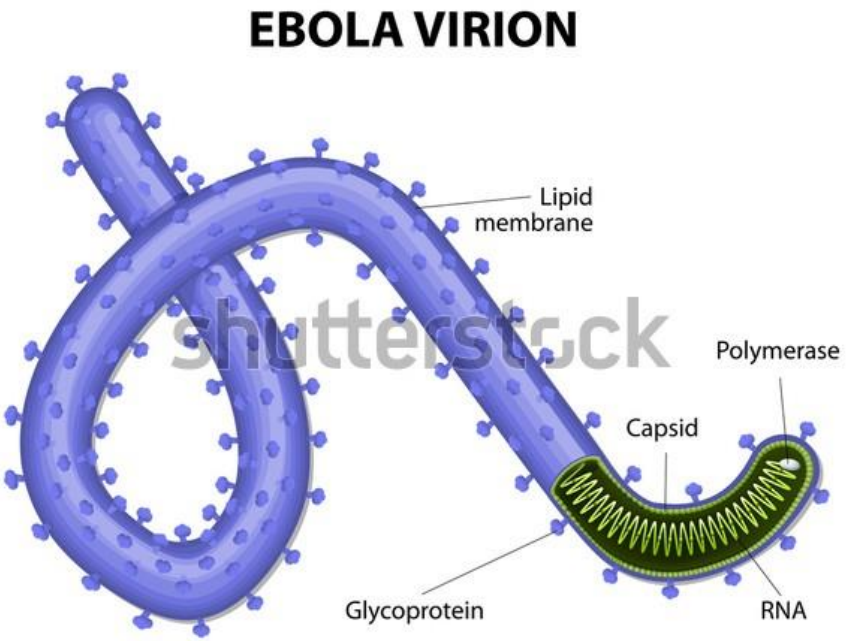
•Ebola virus disease (EVD), formerly known as **Ebola haemorrhagic fever**, is a rare but severe, often fatal illness in humans.

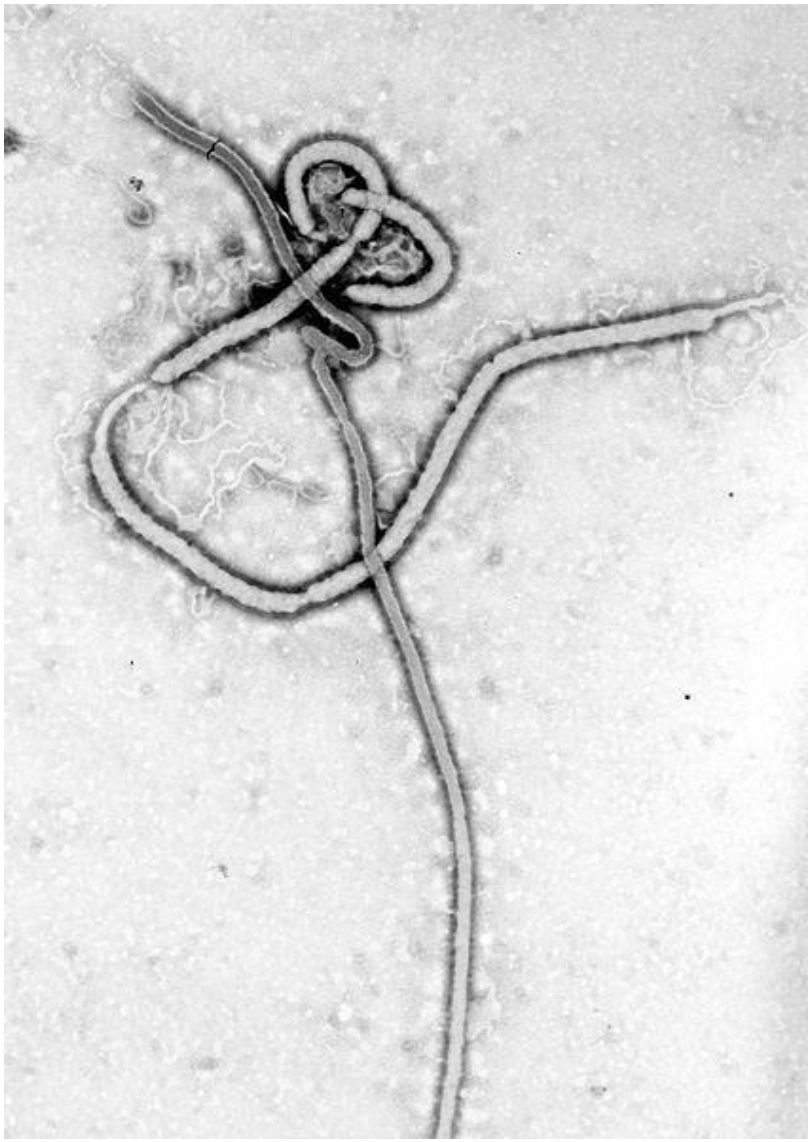
•The virus is transmitted to people **from wild animals** and spreads in the human population through human-to-human transmission.

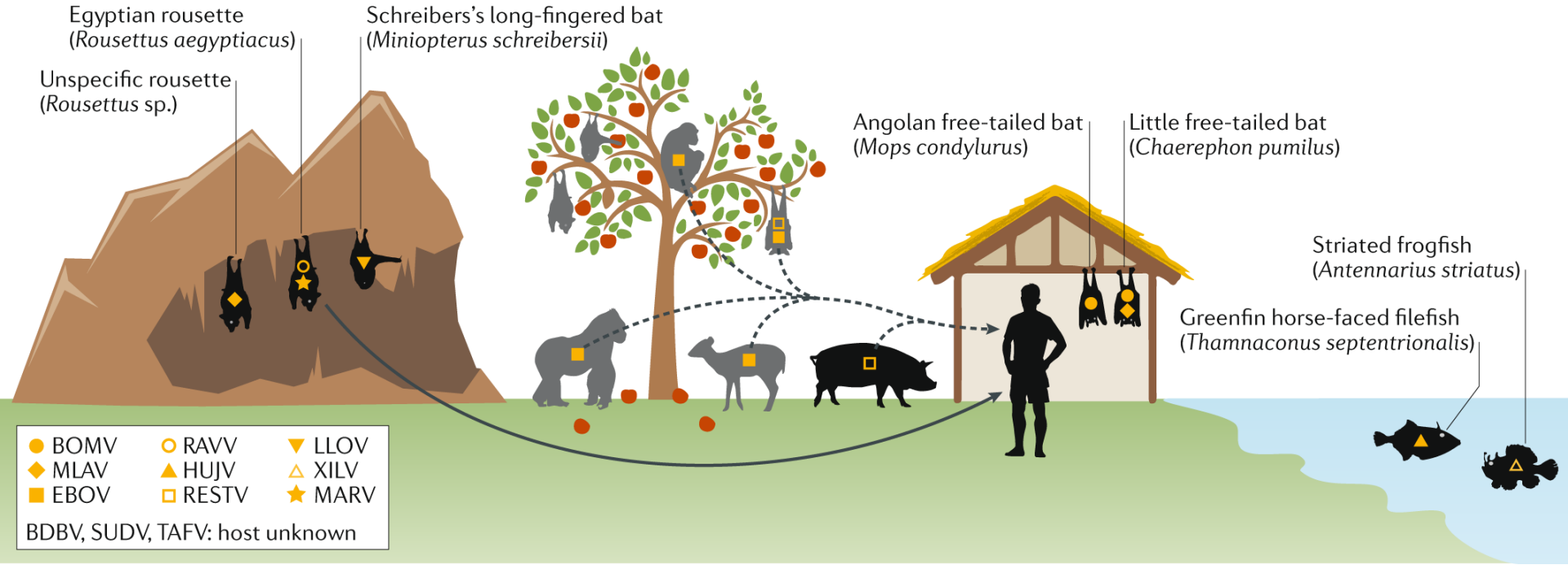
•The average EVD case **fatality rate is around 50%.** Vaccines to protect against Ebola have been developed and have been used to help control the spread of Ebola outbreaks in Guinea and in the Democratic Republic of the Congo (DRC).

•Early supportive care with rehydration, symptomatic treatment improves survival. Two monoclonal antibodies (**Inmazeb and Ebanga**) were approved for the treatment of Zaire ebolavirus (Ebolavirus) infection in adults and children by the US Food and Drug Administration in late 2020.

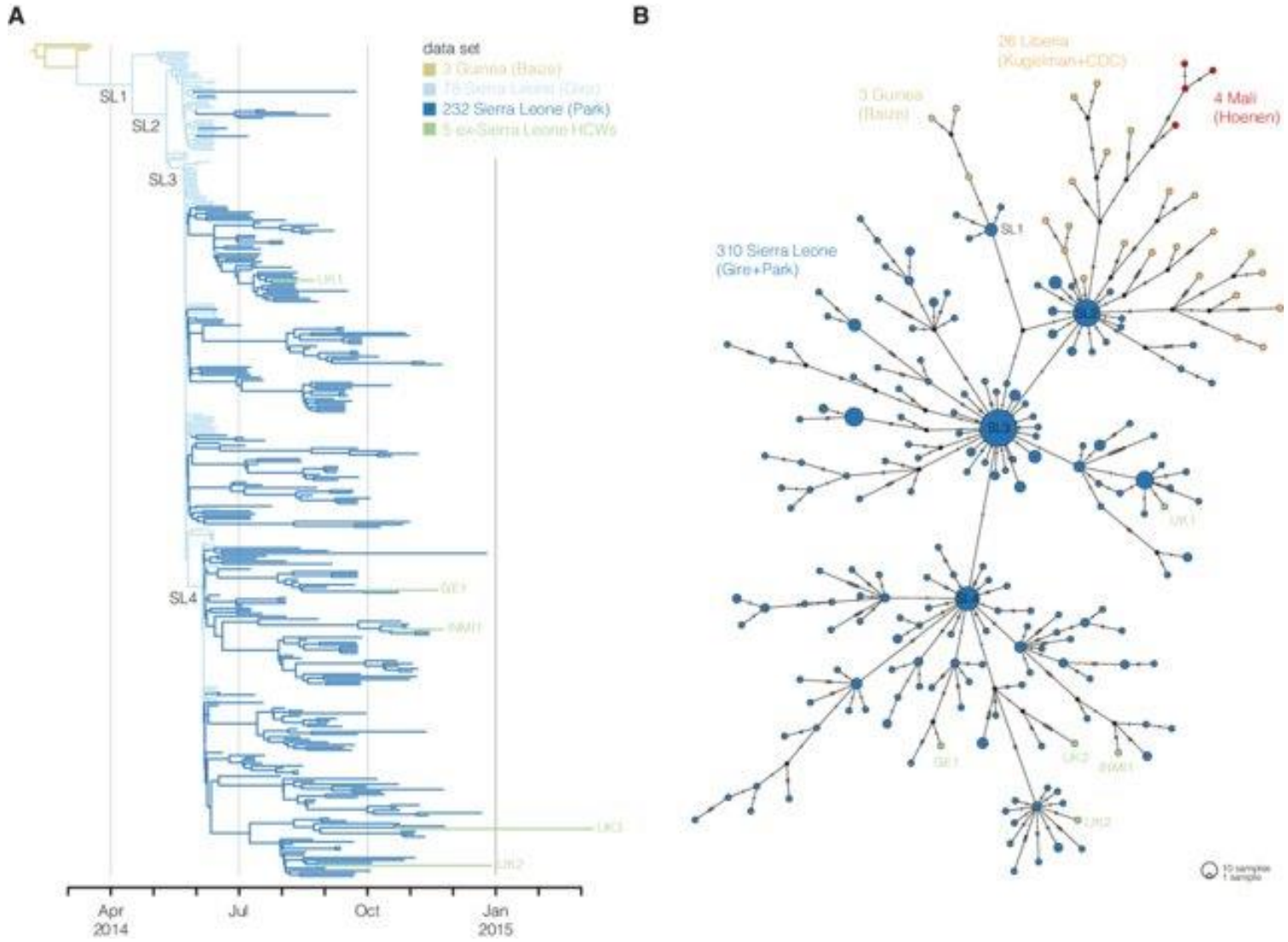
The **2014–2016 outbreak in West Africa** was the largest Ebola outbreak since the virus was first discovered in 1976. The outbreak started in Guinea and then moved across land borders to Sierra Leone and Liberia. The virus family Filoviridae includes three genera: Cuevavirus, Marburgvirus, and Ebolavirus. Within the genus Ebolavirus, six species have been identified: Zaire, Bundibugyo, Sudan, Taï Forest, Reston and Bombali.


EBOLA VIRION
Lipid membrane
Polymerase
Capsid
Glycoprotein
RNA
shutterstock
www.shutterstock.com · 212007784

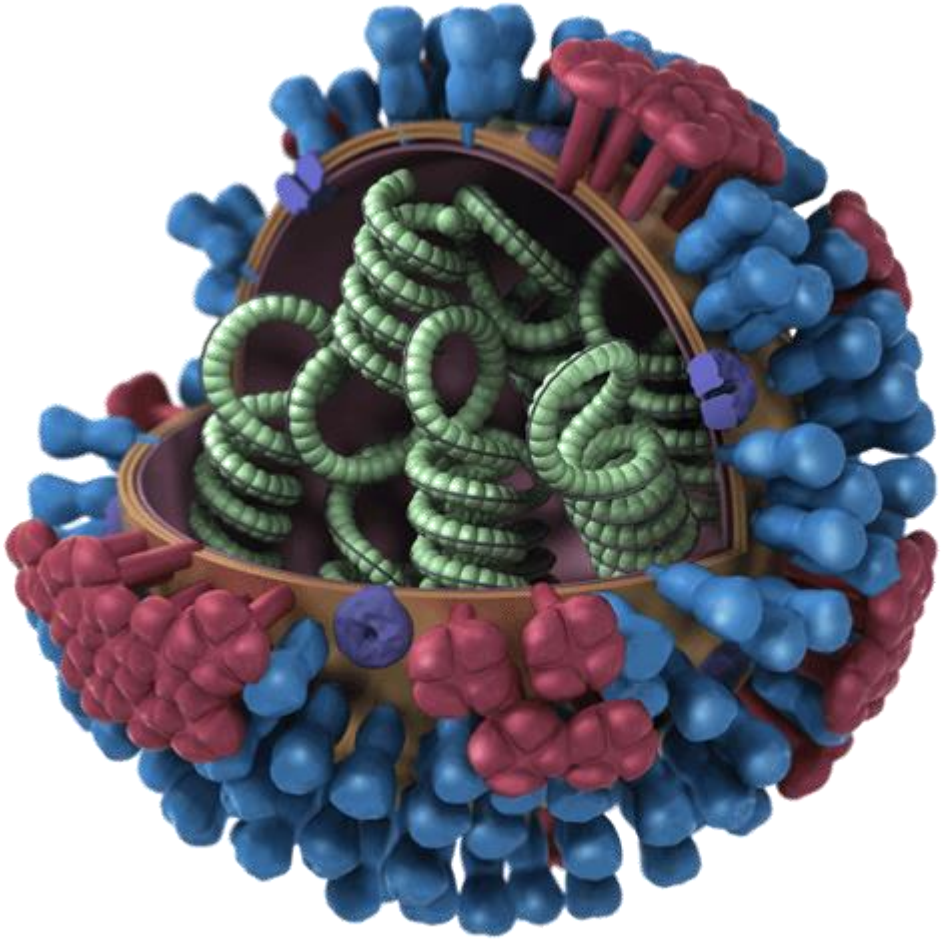




Complete or coding-complete filovirus genome sequences have been obtained from cave-dwelling and house-dwelling bats and highly diverse fish on the African, Asian and European continents. The pathogenic potential of most filoviruses remains unclear, as does the transmission route of pathogenic filoviruses proven to infect humans and pigs or of pathogenic filoviruses suspected to infect **chimpanzees, duikers and gorillas**. Animals that have been proven to be infected by filoviruses are indicated in black; grey animals are suspected but unproven reservoirs of the indicated viruses. Solid arrows indicate highly likely transmission routes; dashed arrows indicate hypothesized transmission routes. BDBV, Bundibugyo virus; BOMV, Bombali virus; EBOV, Ebola virus; HUJV, Huángjiāo virus; LLOV, Lloviu virus; MARV, Marburg virus; MLAV, Měnglà virus; RAVV, Ravn virus; RESTV, Reston virus; SUDV, Sudan virus; TAFV, Taï Forest virus; XILV, Xīlàng virus.

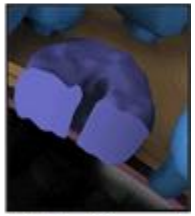


Within and between Country Genomic Relationships of Ebola Virus



Hemagglutinin

Neuraminidase



M2 Ion Channel

RNP

This is a picture of an influenza virus. Influenza A viruses are classified by subtypes based on the properties of their **hemagglutinin (H) and neuraminidase (N)** surface proteins. There are **18 different HA subtypes and 11 different NA** subtypes. Subtypes are named by combining the H and N numbers – e.g., A(H1N1), A(H3N2). Click on the image to enlarge the picture.

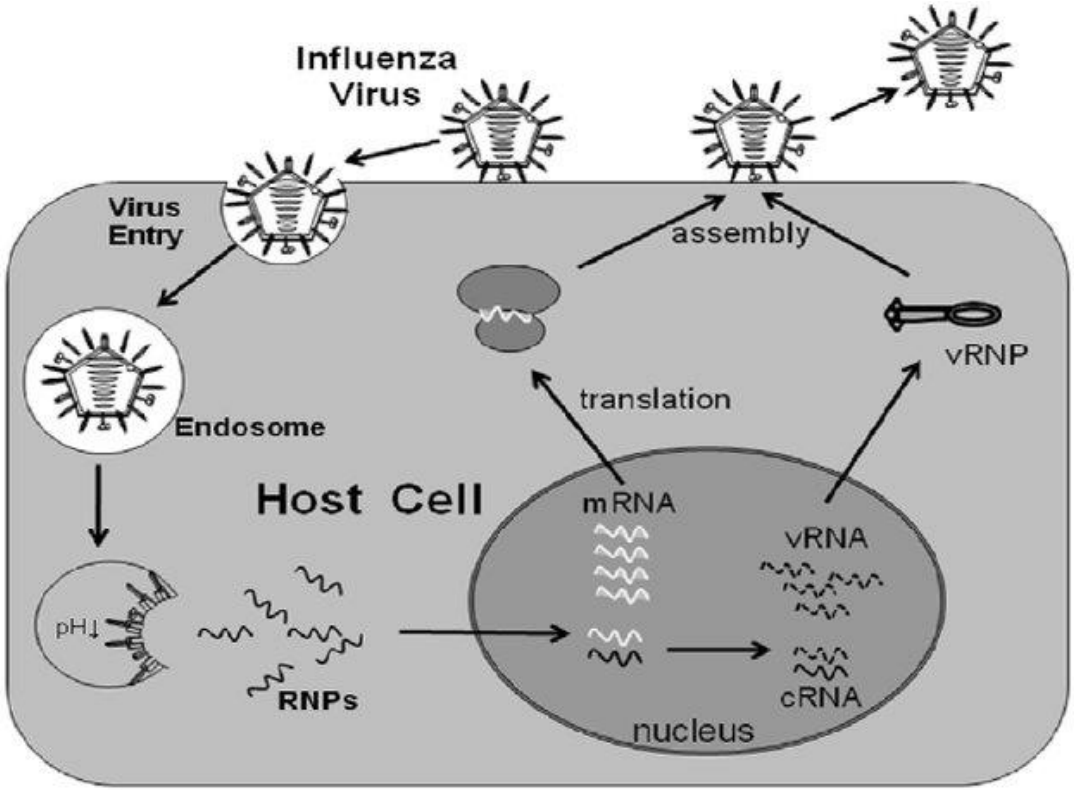
Influenza Virus: 3 Types

- ✓RNA virus
- ✓Antigenically distinct
- ✓No cross-immunity

Influenza	*Type A*	*Type B*	*Type C*
Disease	++	++	-
Epidemics/Pandemics	Epidemics & Pandemics	Milder epidemics	No
Host	Humans & other species !	Humans only	Humans only
Antigenic variation	Frequent	Infrequent	Stable

Influenza virus (*Ortomyxoviridae*)

- Four types – **Influenza A**, B, C, and D
- ssRNA(-) segmented genome (8 segments,~13.5 kb; 11 proteins)

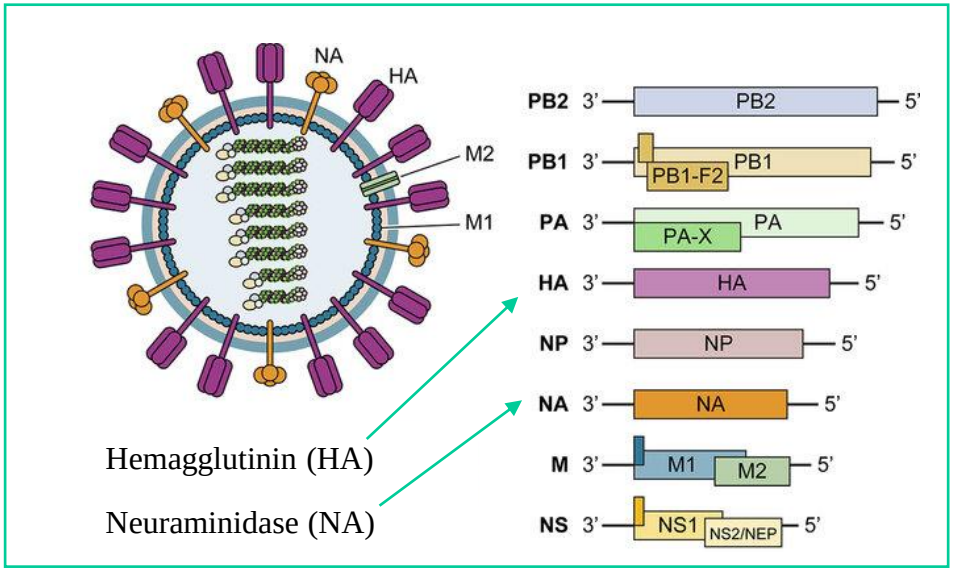


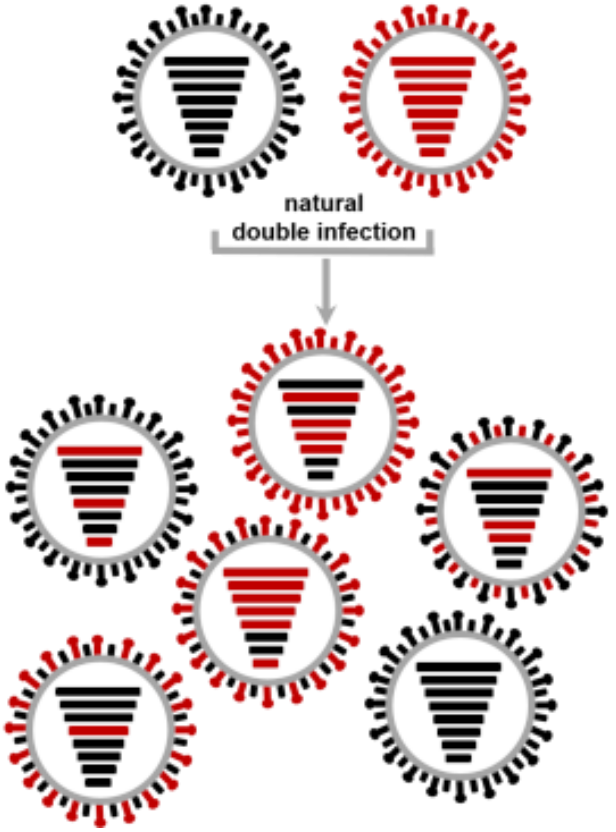
High rate of random mutations (**antigenic drift**) is responsible for a emergence of new influenza variants each season (year).

Yeganeh et al. 2013; PMID: 23454774.

Evolution of influenza through reassortment (antigenic shift)

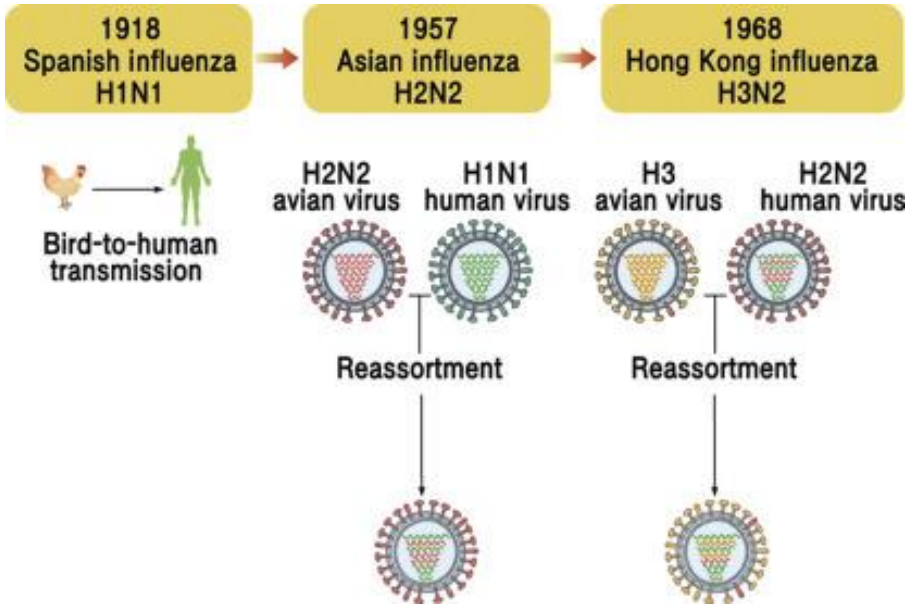
If a host cell is infected with more than one influenza virus, the viral progeny contain random sets of the genome segments.

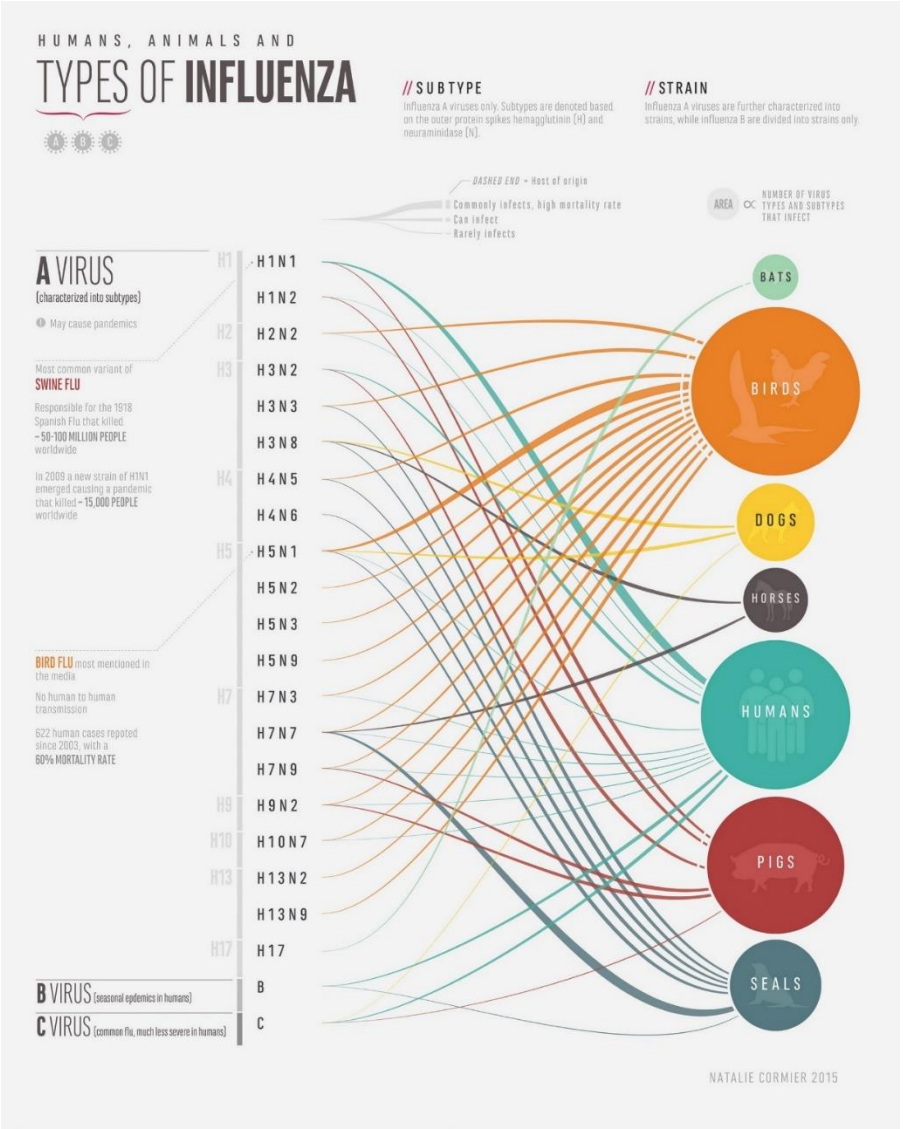




Liang 2023; https://doi.org/10.1080/21505594.2023.2223057
https://www.zkbs-online.de/ZKBS/EN/Meta/focus_topics/Influenzaviruses/Influenza_viruses_node.html

Segment reassortment is responsible for the emergence of **pandemic** influenza variants.

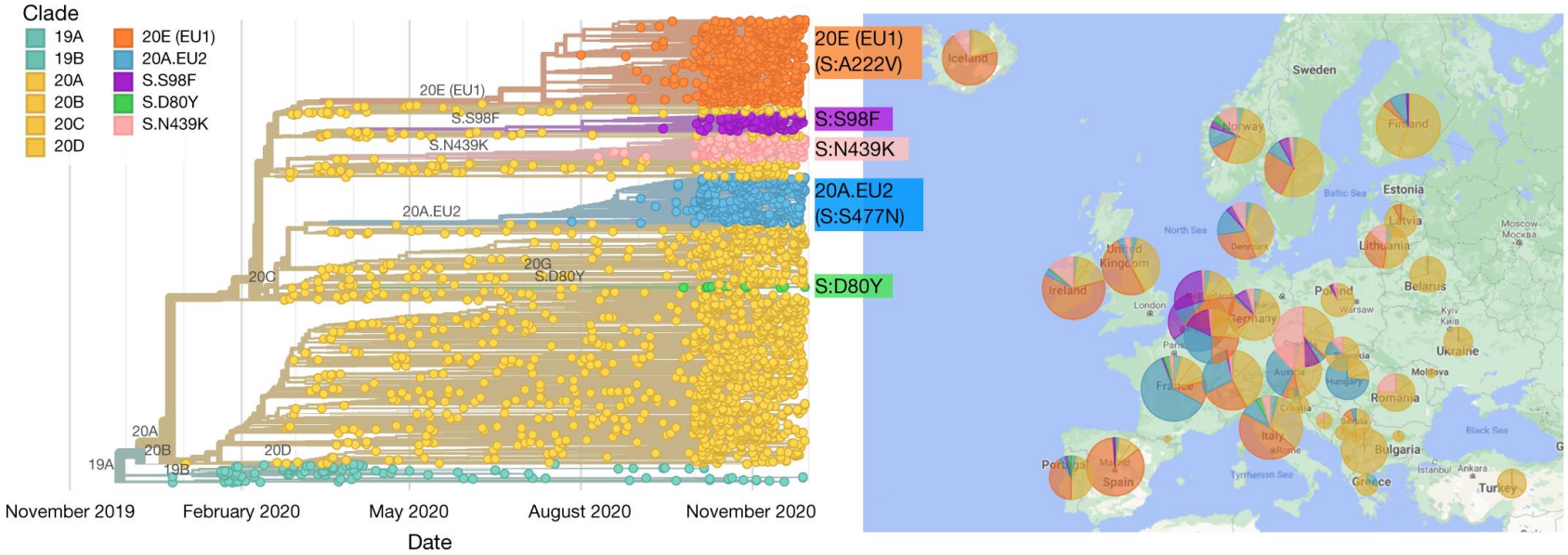




https://bmc1.utm.utoronto.ca/~natalie/flufacts/

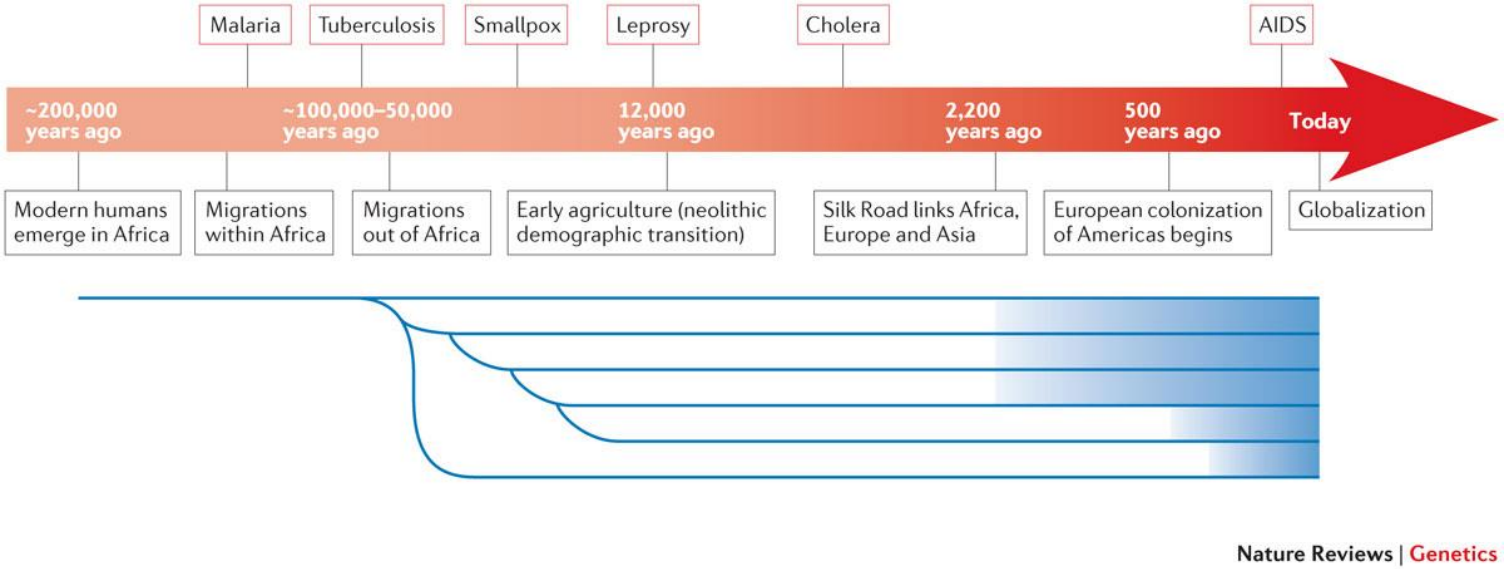
Evolution of SARS-CoV-2 variants

- ssRNA(+) nonsegmented genome (~30 kb; 29 proteins)
- High rate of mutations (antigenic drift)

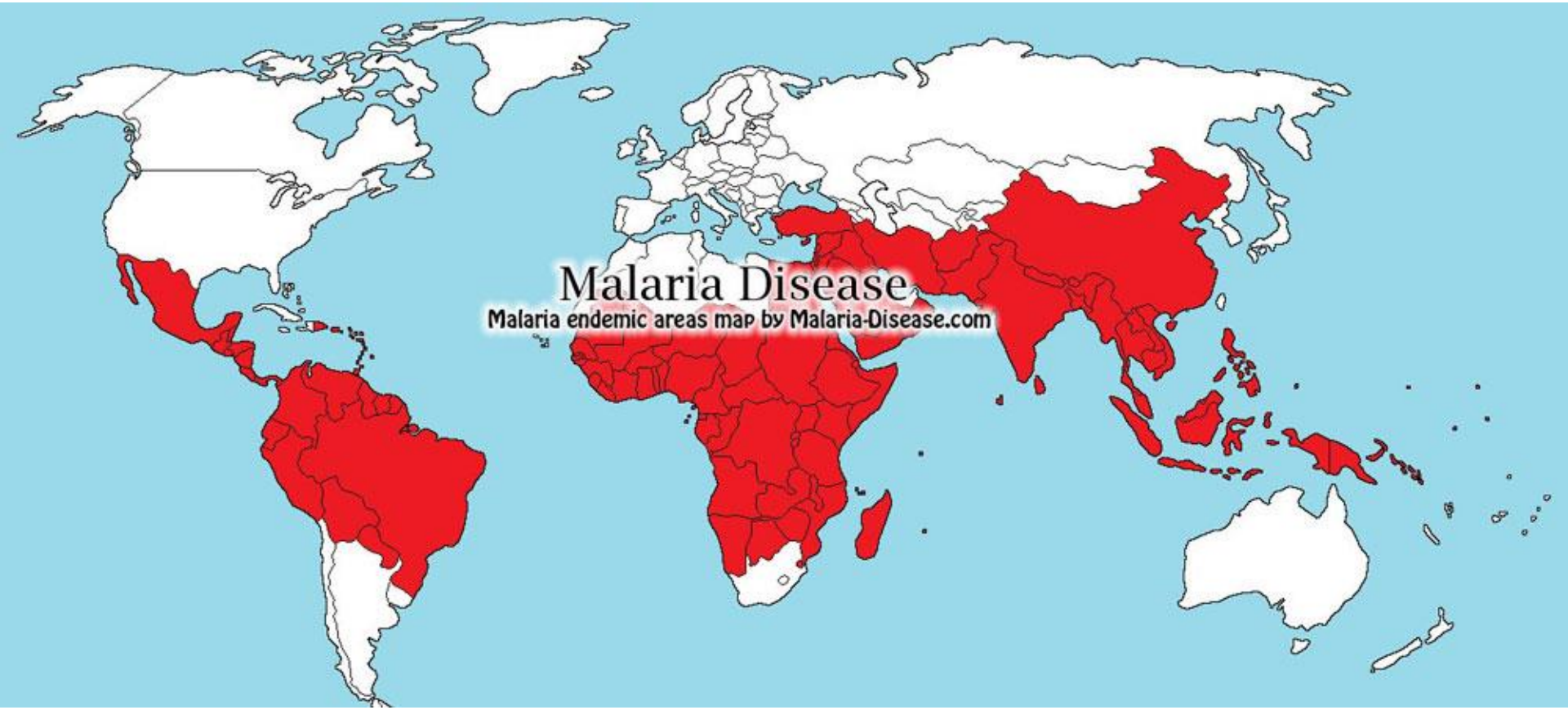


Left, the tree shows a representative sample of isolates from Europe coloured by clades/variants. Clade 20A and its daughter clades 20B and 20C (yellow) carry mutations S:D614G. Variant 20E (EU1) (orange), with mutation S:A222V on a S:D614G background, emerged in early summer 2020 and became common in many European countries in autumn 2020. A separate variant (20A.EU2; blue) with mutation S:S477N became prevalent in France. **Right,** the proportion of sequences belonging to each variant per country.

Changes in human genome selected by pathogens



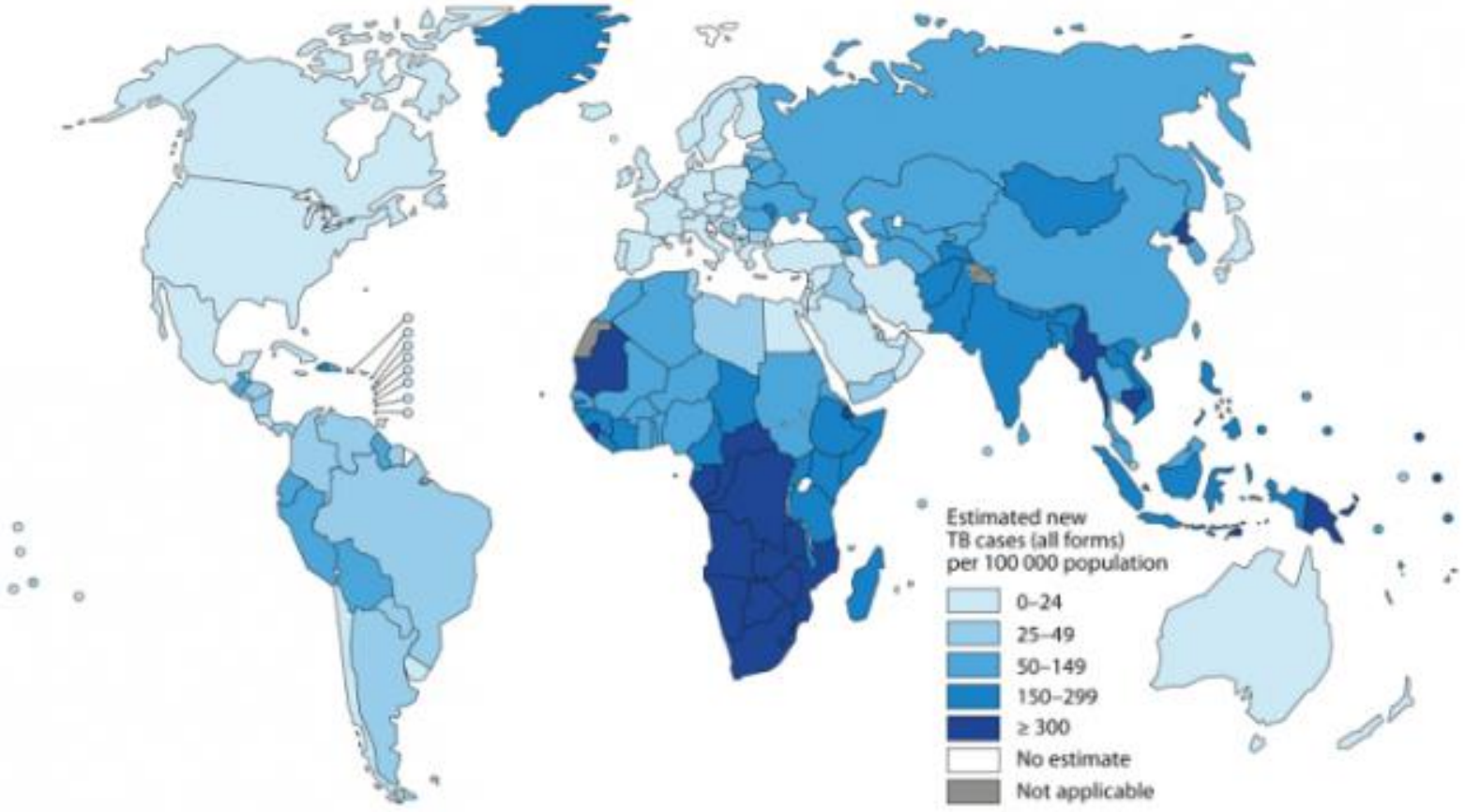
Key events in recent human evolution (boxes outlined in black) are juxtaposed with **the estimated ages of infectious disease emergence** (boxes outlined in red). The fragmentation of the human lineage into genetically and geographically distinct populations (blue lines) accelerates with migration out of Africa. Later, these populations started mixing more (blue shaded regions between the populations) along trade routes (such as the Silk Road), through colonization and through high rates of global travel nowadays. Smallpox-variola.


Malaria Disease
Malaria endemic areas map by Malaria-Disease.com

Common Erythrocyte Variants That Affect Resistance to **Malaria**

Gene	Protein	Function	Reported Genetic Associations with Malaria
FY	Duffy antigen	Chemokine receptor	FY*O allele completely protects against *P. vivax* infection.
G6PD	Glucose-6-phosphatase dehydogenase	Enzyme that protects against oxidative stress	G6PD deficiency protects against severe malaria.
GYPA	Glycophorin A	Sialoglycoprotein	GYPA-deficient erythrocytes are resistant to invasion by *P. falciparum.*
GYPB	Glycophorin B	Sialoglycoprotein	GYPB-deficient erythrocytes are resistant to invasion by *P. falciparum.*
GYPC	Glycophorin C	Sialoglycoprotein	GYPC-deficient erythrocytes are resistant to invasion by *P. falciparum.*
HBA	α-Globin	Component of hemoglobin	α^+ Thalassemia protects against severe malaria but appears to enhance mild malaria episodes in some environments.
HBB	β-Globin	Component of hemoglobin	HbS and HbC alleles protect against severe malaria. HbE allele reduces parasite invasion.
HP	Haptoglobin	Hemoglobin-binding protein present in plasma (not erythrocyte)	Haptoglobin 1-1 genotype is associated with susceptibility to severe malaria in Sudan and Ghana.
SCL4A1	CD233, erythrocyte band 3 protein	Chloride/bicarbonate exchanger	Deletion causes ovalocytosis but protects against cerebral malaria.

Estimated tuberculosis (TB) incidence rates, 2011



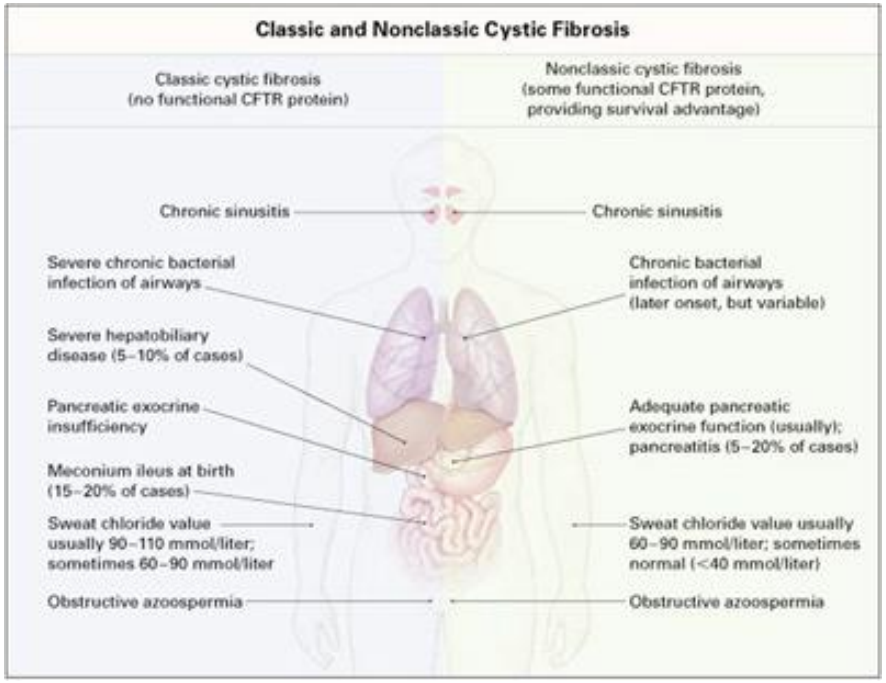
The boundaries and names shown and the designations used on this map do not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate border lines for which there may not yet be full agreement.

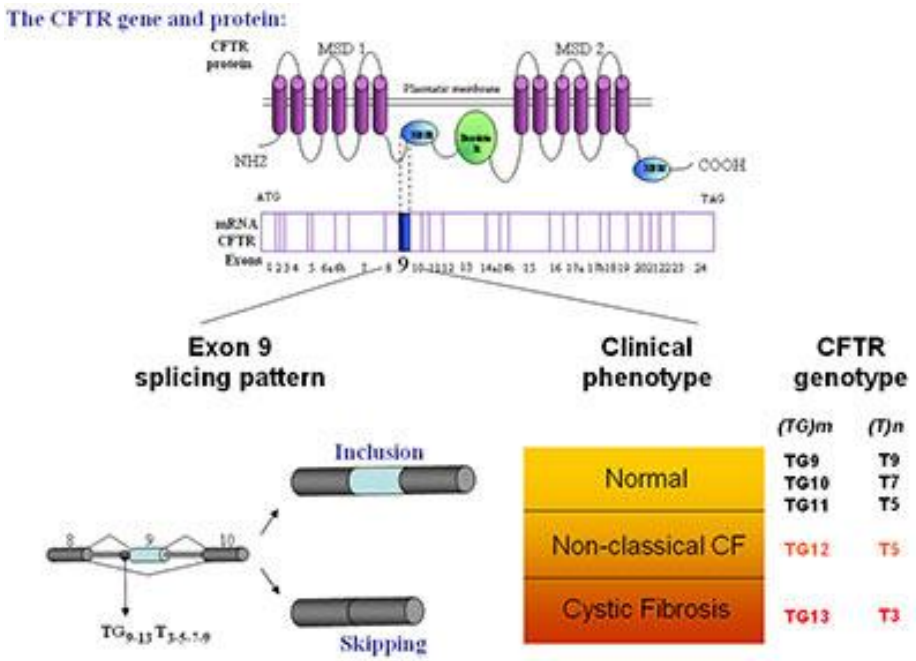
Source: *Global Tuberculosis Report 2012*. WHO, 2012.

World Health Organization

Cystic fibrosis

1 in 3,000 children are born with CF, and 2% of people carry one mutant gene



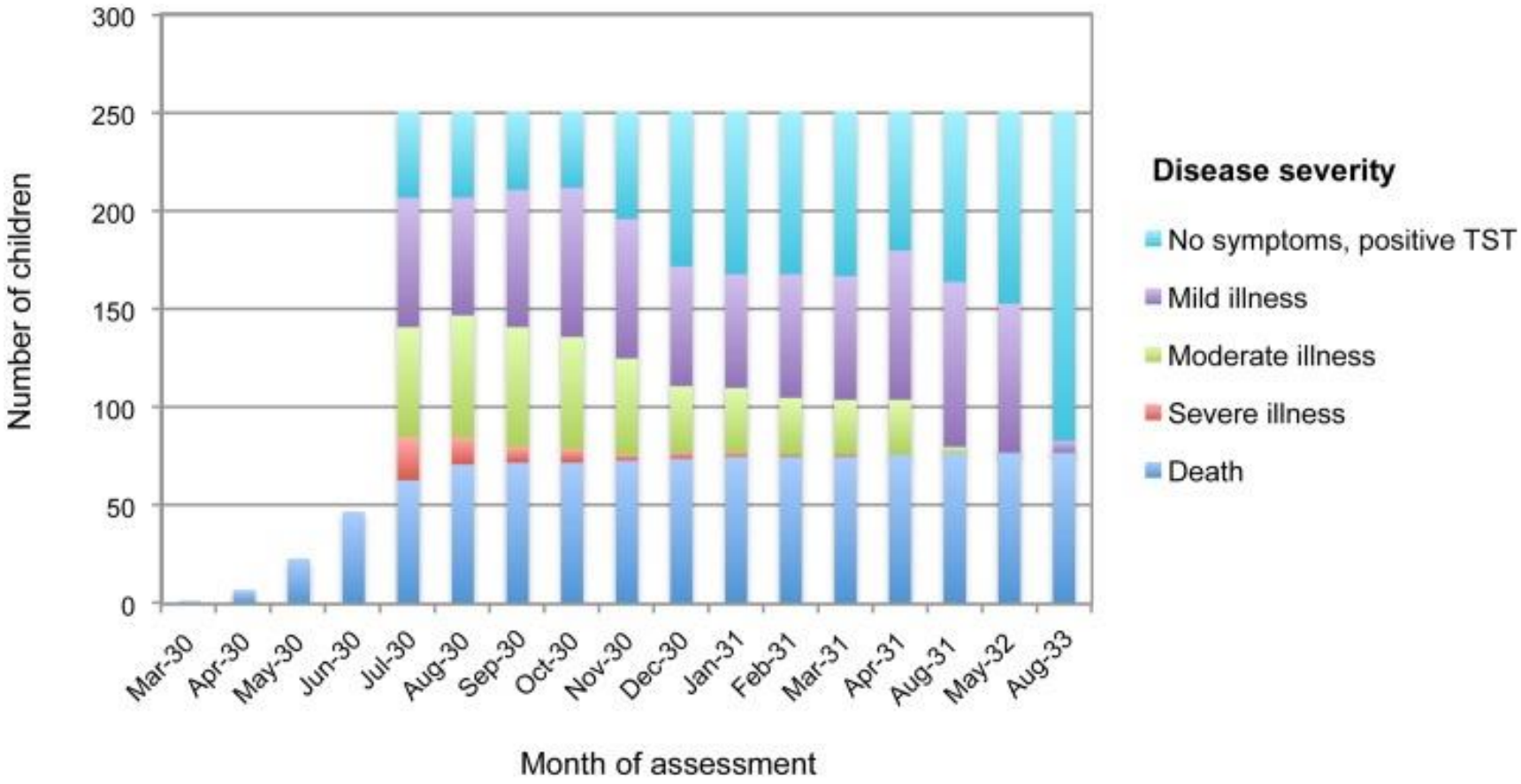


Cystic fibrosis gene protects against tuberculosis

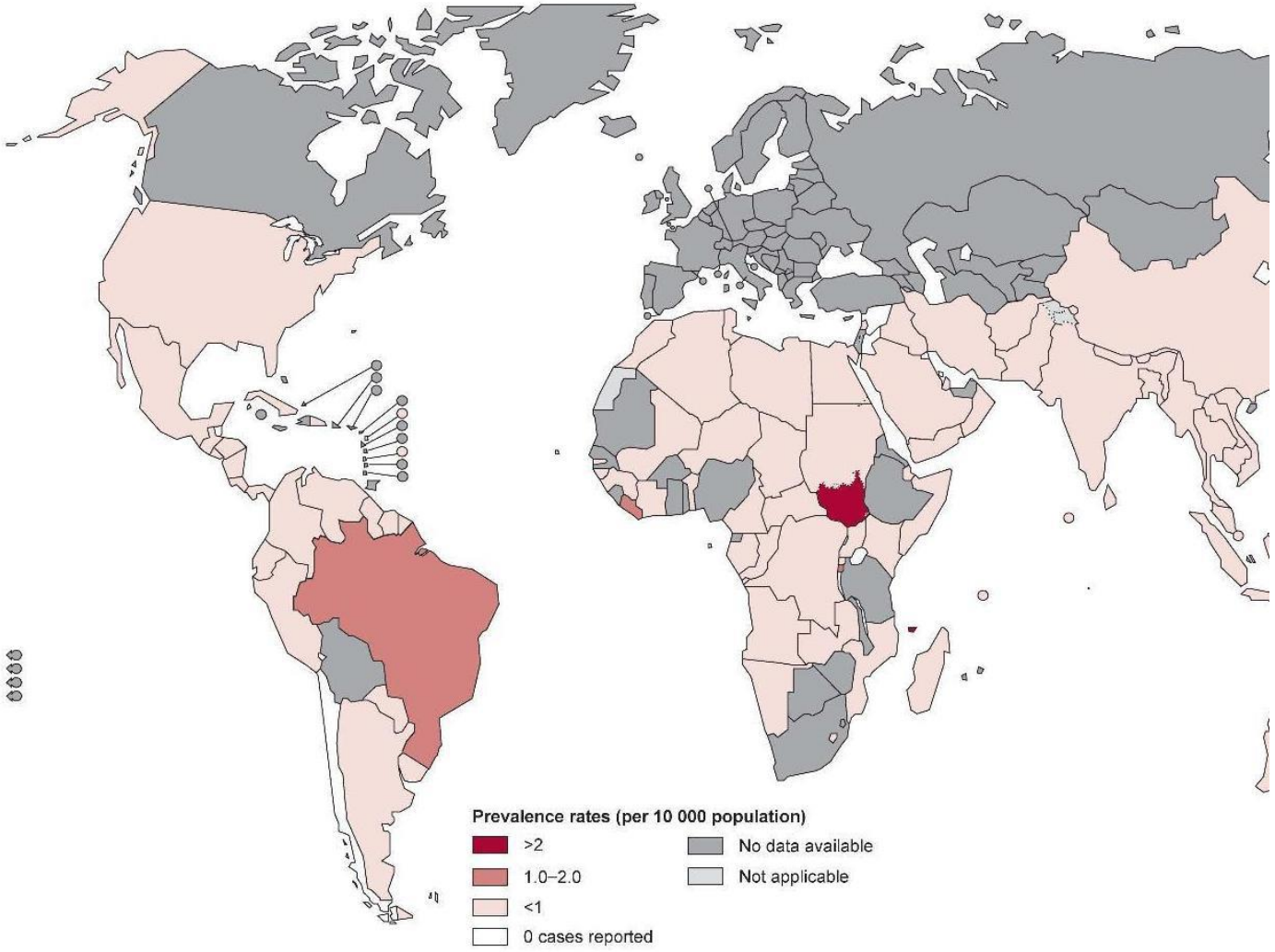
Between 1600 and 1900, TB caused 20% of all deaths in Europe

The Lübeck disaster (1929–1933)

is a unique event in the history of tuberculosis when 251 newborns were accidentally infected with a virulent strain of ***Mycobacterium tuberculosis***. The disaster happened while BCG was introduced as an anti-TB vaccine. In an exemplary multidisciplinary investigation, the disaster was shown to be due to the accidental contamination of BCG vaccine preparations with virulent *M. tuberculosis* and not BCG itself.



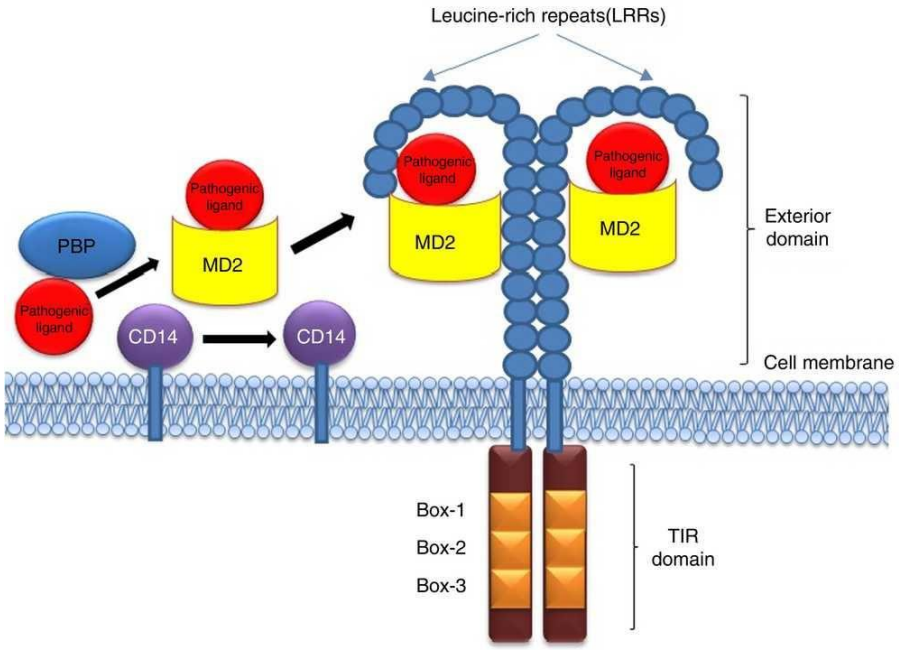
There were nearly 200,000 cases of leprosy around the world at the beginning of 2012.



Leprosy and the Adaptation of Human Toll-Like Receptor 1. Population differentiation at TLR1 I602S.

The **protective dysfunctional 602S** allele is rare in Africa but expands to become the dominant allele among individuals of **European descent**. This supports the hypothesis that this locus may be under selection from mycobacteria or other pathogens that are recognized by TLR1 and its co-receptors.

Wong SH, Gochhait S, Malhotra D, Pettersson FH, Teo YY, et al. (2010) Leprosy and the Adaptation of Human Toll-Like Receptor 1. PLoS Pathog 6(7): e1000979. doi:10.1371/journal.ppat.1000979
http://127.0.0.1:8081/plospathogens/article?id=info:doi/10.1371/journal.ppat.1000979

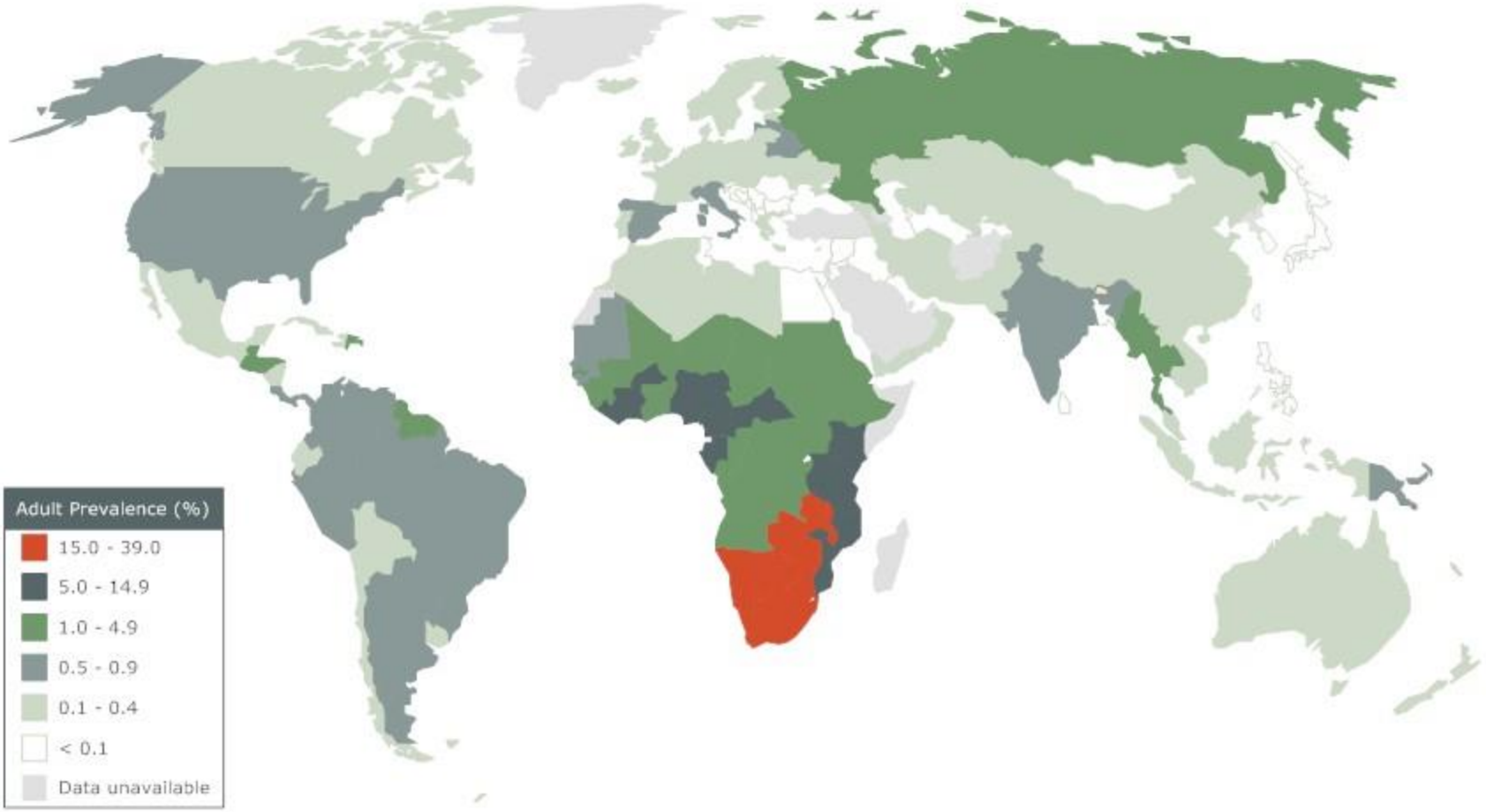


TLR2 is one of the toll-like receptors and plays a role in the immune system. TLR2 is a membrane protein, a receptor, which is expressed on the surface of certain cells and recognizes foreign substances and passes on appropriate signals to the cells of the immune system.

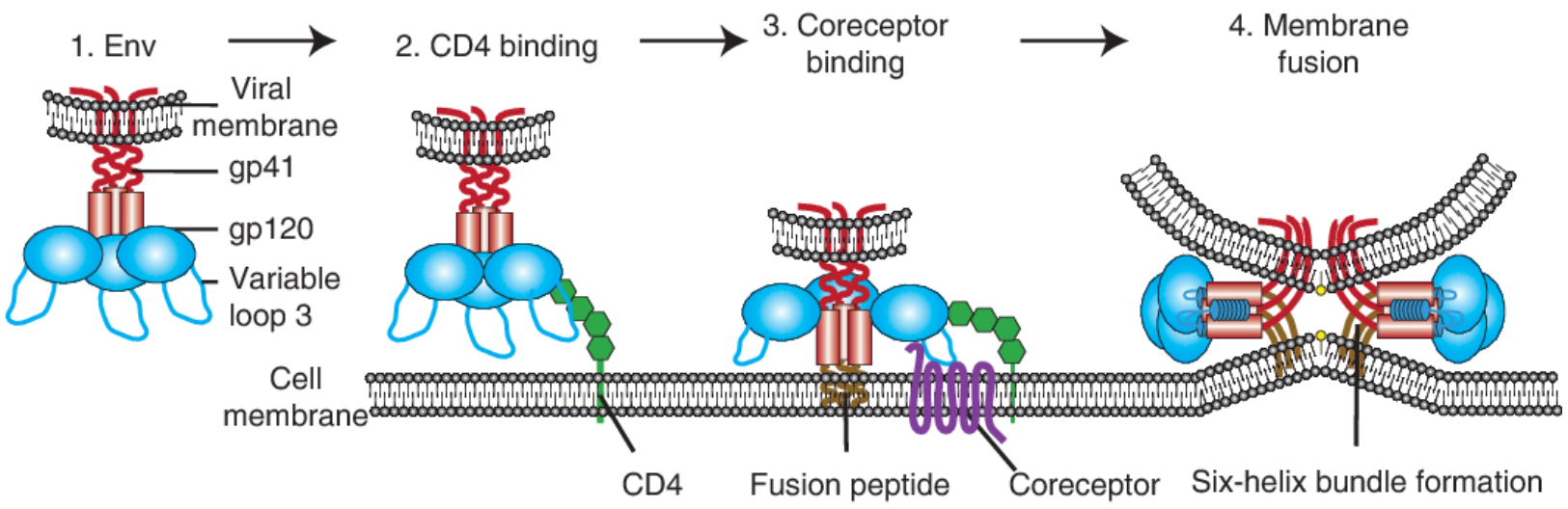
The presence of the TLR2 Arg753Gln polymorphism was significantly **associated with pneumonia** in AML patients.

The presence of the TLR2 Arg753Gln polymorphism was significantly **associated with resistence to syphilis.**

Adults (15-49 years of age) **Living with HIV**



HIV epidemic



The **CCR5 locus (coreceptor)** shows that historical epidemics have been important
in shaping the genomes of humans and other primate species. It has been projected that if the HIV epidemic continues for another 100 years, it will leave a signature on the human genome at the **CCR5 locus** and related HIV resistance loci.

Although higher HLA-C expression protects against HIV progression, it also increases risk of the inflammatory disorder **Crohn's disease**, which highlights the potential for health repercussions of pathogen-driven selection.

The CCR5 chemokine receptor is exploited by HIV-1 to gain entry into CD4+ T cells. A deletion mutation (Δ32) confers resistance against HIV by obliterating the expression of the receptor on the cell surface.. **The allele exists at appreciable frequencies only in Europe**, and within Europe, the frequency is higher in the north.

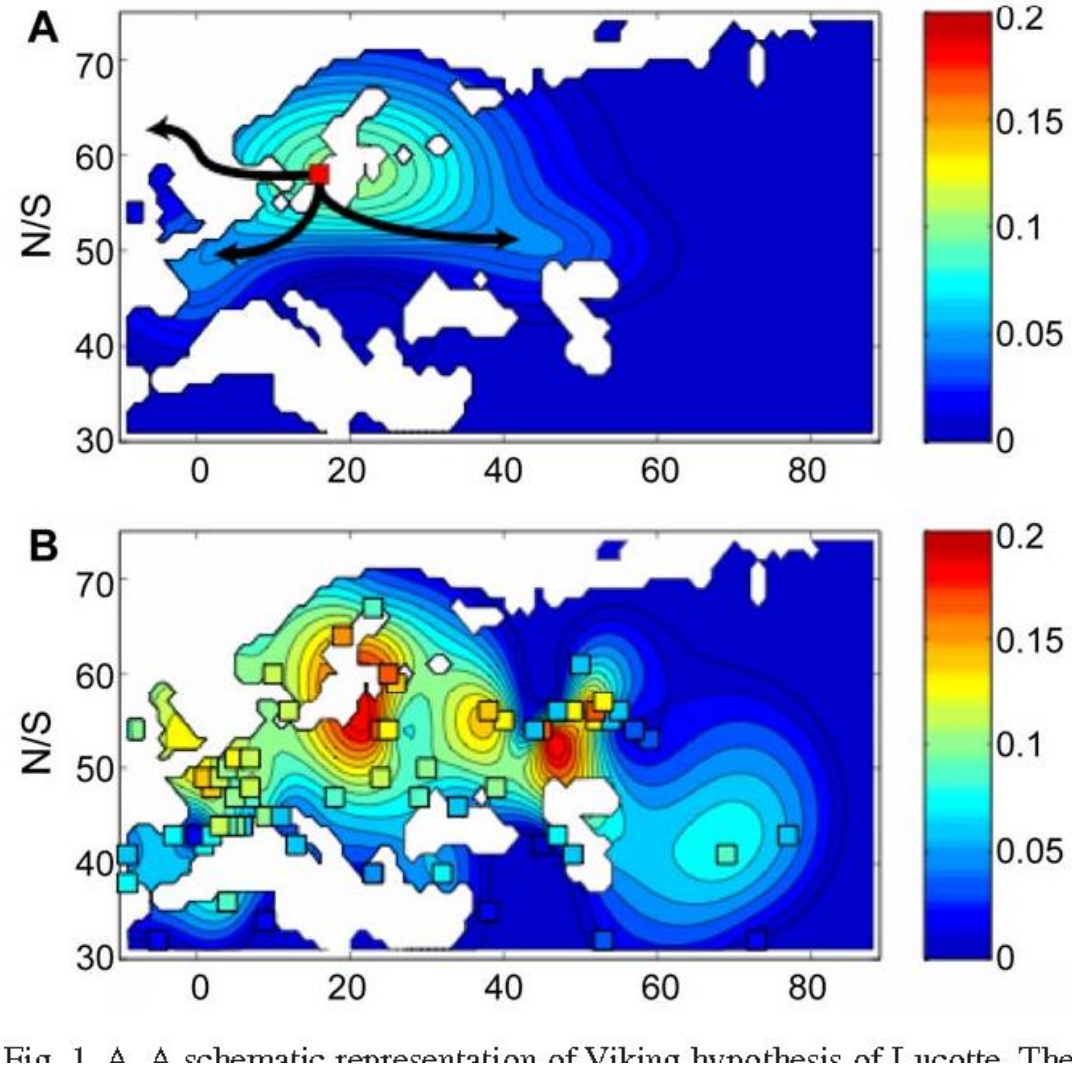


Fig. 1. A. A schematic representation of **Viking hypothesis**. The red square represents a Scandinavian origin of the allele. The black arrows represent dissemination of CCR5-D32 by Vikings southwards towards France and the Mediterranean, eastwards towards Russia, and northwest towards Iceland. Contour lines and color represent the frequency in Europe at an intermediate stage of the allele's migration out of Scandinavia. B. The modern-day observed allele frequencies. Squares mark locations of sampled allele frequencies, and color within the squares denotes the observed frequencies. Contour lines represent interpolated allele frequencie

The selective raise of the CCR5-Δ32 allele was proposed to be attributed to smallpox *(Variola major)* caused by the poxvirus.

Their estimates suggest that the CCR5-Δ32 deletion arose around 2000 years ago with a range from 375 to 4800 years