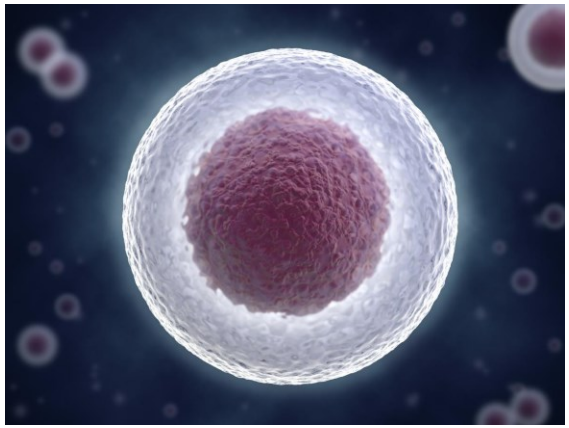
Genome and chromosome structure



Martin A. Lysák

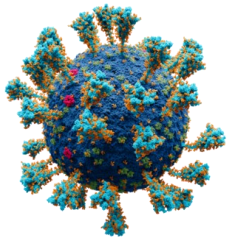
CEITEC and Faculty of Science,
Masaryk University

Genome (Hans Winkler, 1920)

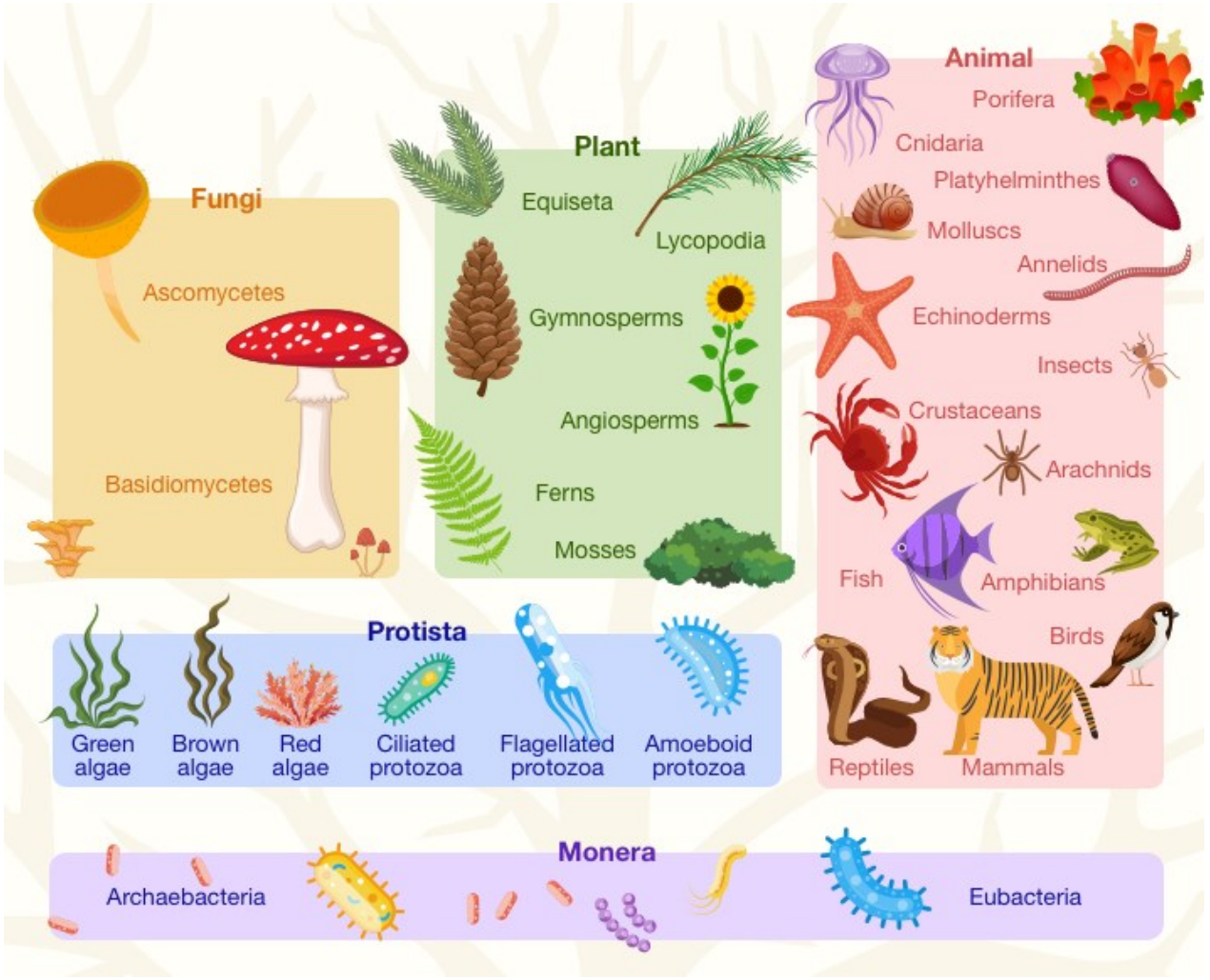
> Ich schlage vor, für den haploiden Chromosomensatz, der im Verein mit dem zugehörigen Protoplasma die materielle Grundlage der systematischen Einheit darstellt, den Ausdruck: das ***Genom*** zu verwenden und Kerne, Zellen und Organismen, in denen ein gleichartiges Genom mehr als einmal in jedem Kern vorhanden ist, homogenomatisch zu nennen, solche dagegen, die verschiedenartige Genome im Kern führen, heterogenomatisch

- Genetic material, i.e. DNA (RNA in RNA viruses)
- By genome we either mean nuclear genome (eukaryotes) or genetic material of prokaryotes, mitochondria and chloroplasts
- Genomes contain coding DNA regions (genes) and non-coding DNA
- DNA (RNA) is associated with proteins, thus genomes are essentially nucleoprotein structures
- Genomes differ by size and complexity
- Genomics studies genomes (their structure and evolution)

Living things: five kingdomes



Viruses
(not living nucleoproteins)



eukaryotes

prokaryotes

Genomes

1. Viral genome

2. Prokaryotic genome

2.1 Genome of Archaea
2.2 Bacterial genome

3. Eukaryotic genome

3.1 Extra-nuclear genome

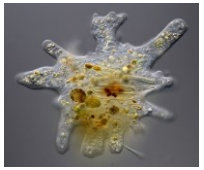
3.1.1 Mitochondrial genome
3.1.2 Plastid genome
3.1.3 Extra-chromosomal DNA

3.2 Nuclear genome

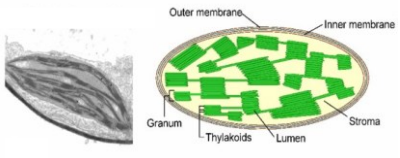
3.2.1 Nuclear architecture
3.2.2 Chromosomes

Genome size variation

Polychaos dubium



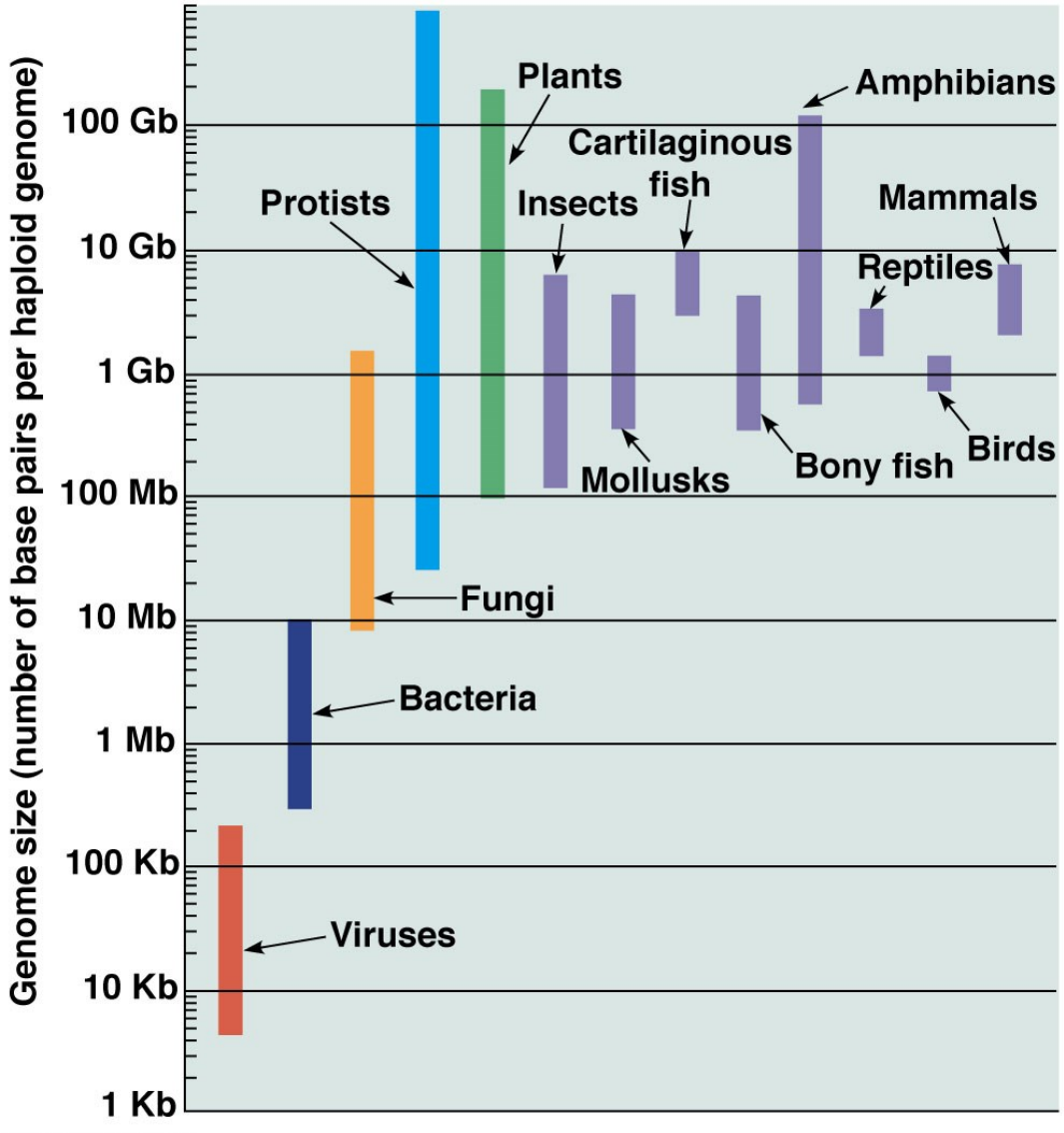
…perhaps the largest known genome – 670 billion base pairs (670 Gb) (~200-times larger than the human genome, 3.2 Gb; some authors suggest treating the value with caution – *Amoeba proteus* has ~34 - 43 Gb…)



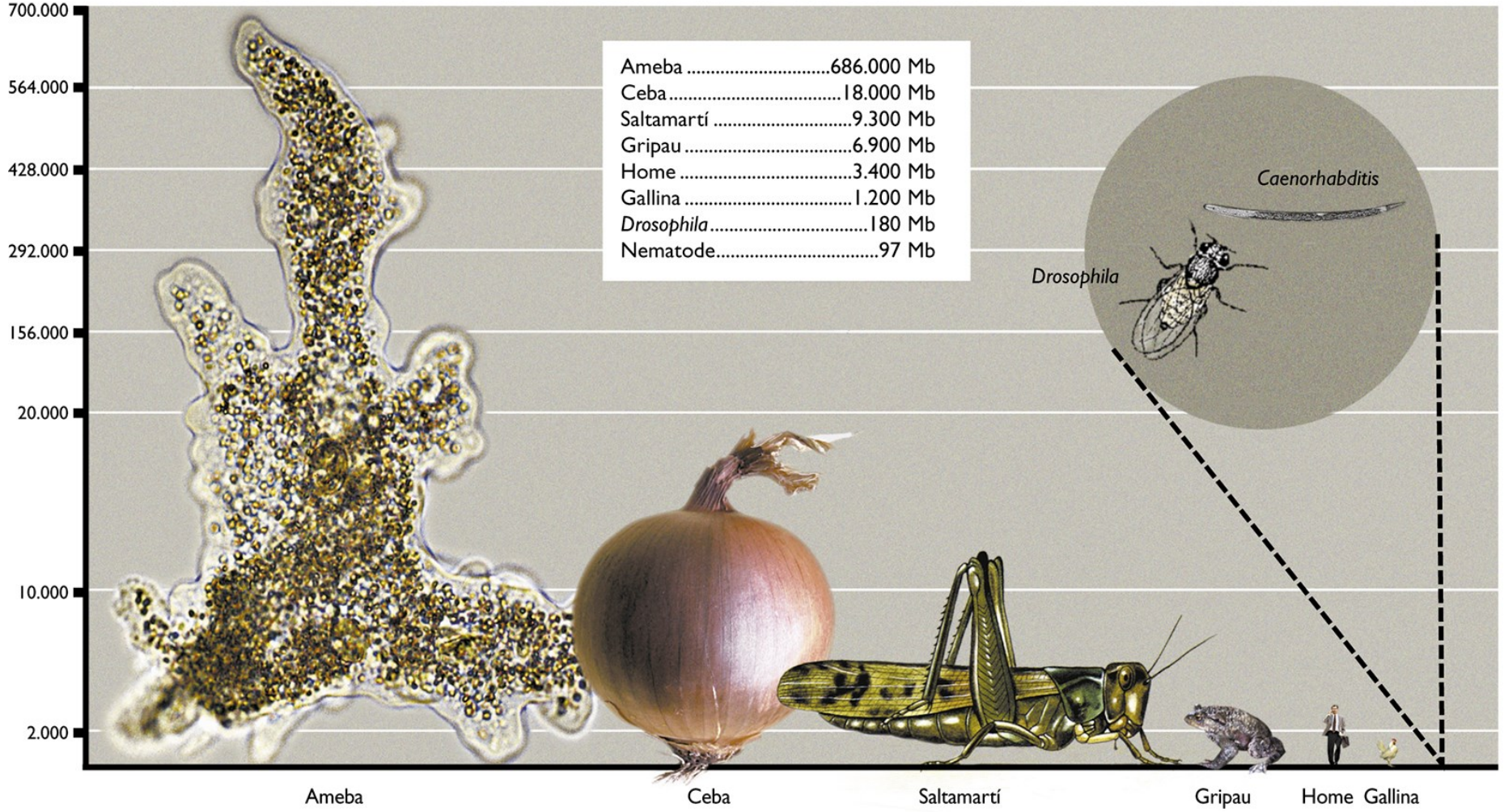
organism	genome size (base pairs)	protein coding genes	number of chromosomes
model organisms			
model bacteria *E. coli*	4.6 Mbp	4,300	1
budding yeast *S. cerevisiae*	12 Mbp	6,600	16
fission yeast *S. pombe*	13 Mbp	4,800	3
amoeba *D. discoideum*	34 Mbp	13,000	6
nematode *C. elegans*	100 Mbp	20,000	12 (2n)
fruit fly *D. melanogaster*	140 Mbp	14,000	8 (2n)
model plant *A. thaliana*	140 Mbp	27,000	10 (2n)
moss *P. patens*	510 Mbp	28,000	27
mouse *M. musculus*	2.8 Gbp	20,000	40 (2n)
human *H. sapiens*	3.2 Gbp	21,000	46 (2n)
viruses			
hepatitis D virus (smallest known animal RNA virus)	1.7 Kb	1	ssRNA
HIV-1	9.7 kbp	9	2 ssRNA (2n)
influenza A	14 kbp	11	8 ssRNA
bacteriophage λ	49 kbp	66	1 dsDNA
Pandoravirus salinus (largest known viral genome)	2.8 Mbp	2500	1 dsDNA
organelles			
mitochondria - *H. sapiens*	16.8 kbp	13 (+22 tRNA +2 rRNA)	1
mitochondria – *S. cerevisiae*	86 kbp	8	1
chloroplast – *A. thaliana*	150 kbp	100	1
bacteria			
C. ruddii (smallest genome of an endosymbiont bacteria)	160 kbp	182	1
M. genitalium (smallest genome of a free living bacteria)	580 kbp	470	1
H. pylori	1.7 Mbp	1,600	1
Cyanobacteria S. elongatus	2.7 Mbp	3,000	1
methicillin-resistant *S. aureus* (MRSA)	2.9 Mbp	2,700	1
B. subtilis	4.3 Mbp	4,100	1
S. cellulosum (largest known bacterial genome)	13 Mbp	9,400	1
archaea			
Nanoarchaeum equitans (smallest parasitic archaeal genome)	490 kbp	550	1
Thermoplasma acidophilum (flourishes in pH<1)	1.6 Mbp	1,500	1
Methanocaldococcus (Methanococcus) jannaschii (from ocean bottom hydrothermal vents; pressure >200 atm)	1.7 Mbp	1,700	1
Pyrococcus furiosus (optimal temp 100°C)	1.9 Mbp	2,000	1
eukaryotes - multicellular			
pufferfish *Fugu rubripes* (smallest known vertebrate genome)	400 Mbp	19,000	22
poplar *P. trichocarpa* (first tree genome sequenced)	500 Mbp	46,000	19
corn *Z. mays*	2.3 Gbp	33,000	20 (2n)
dog *C. familiaris*	2.4 Gbp	19,000	40
chimpanzee *P. troglodytes*	3.3 Gbp	19,000	48 (2n)
wheat *T. aestivum* (hexaploid)	16.8 Gbp	95,000	42 (2n=6x)
marbled lungfish *P. aethiopicus* (largest known animal genome)	130 Gbp	unknown	34 (2n)
herb plant *Paris japonica* (largest known genome)	150 Gbp	unknown	40 (2n)

Protopterus aethiopicus

C-value paradox (CA Thomas, 1971)

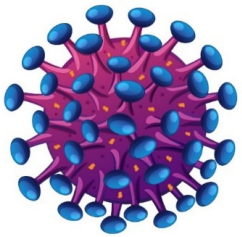


C-value paradox

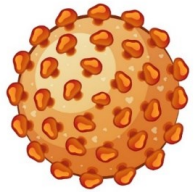


The height of the drawings is proportional to the size of their genome (amoebae, onions, grasshoppers, toads, humans, hens, *Drosophila* and *Caenorhabditis*).

Viruses

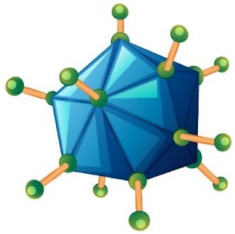


HIV

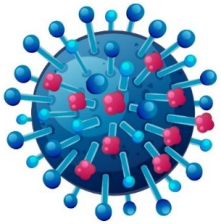


Hepatitis B

Ebola Virus

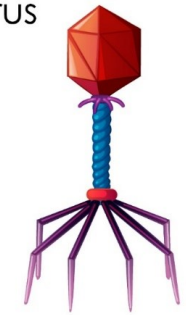


Adenovirus

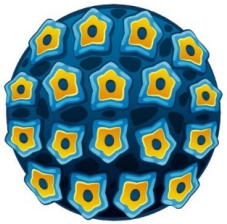


Influenza

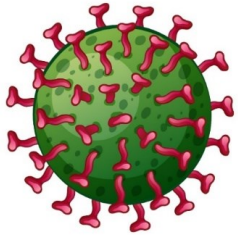
Rabies Virus



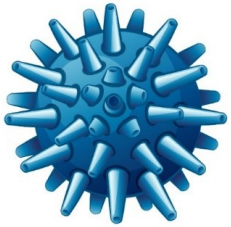
Bacteriophage



PaPillomavirus



Rotavirus

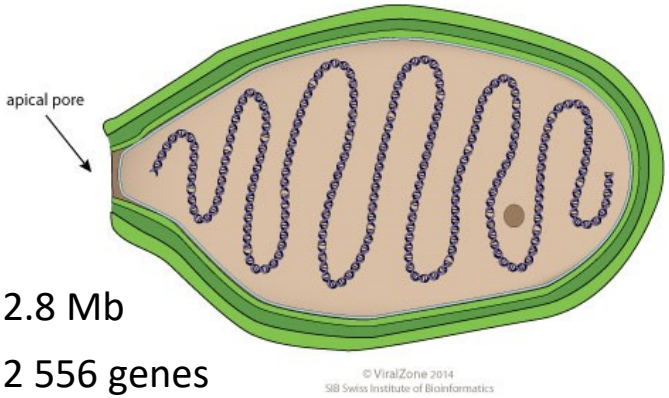


Herpes Virus

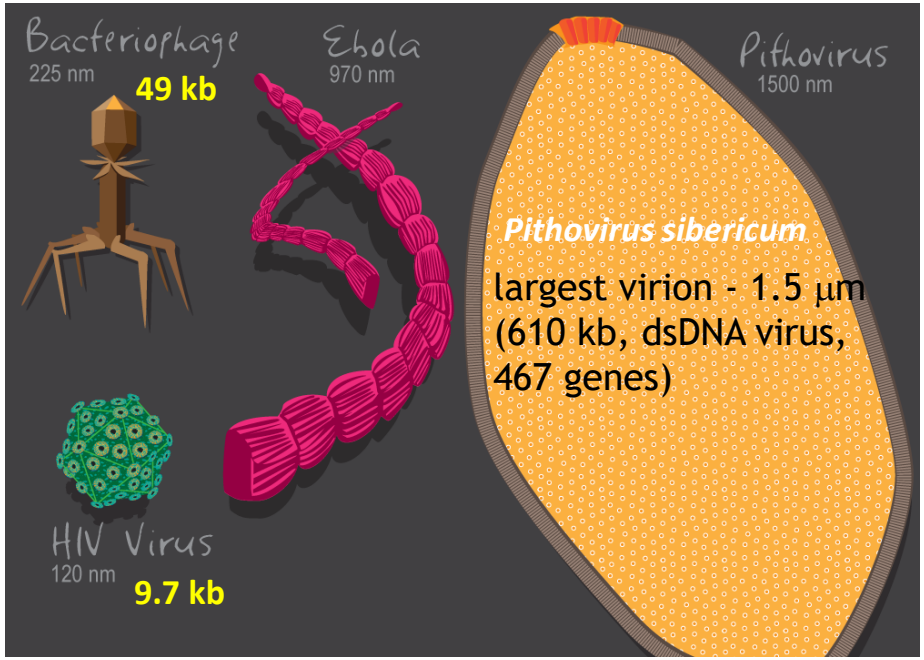
Viruses - physical and genome size

viruses	0.9	protein coding genes	
hepatitis D virus (smallest known animal RNA virus)	1.7 Kb	1	ssRNA
HIV-1	9.7 kbp	9	2 ssRNA (2n)
influenza A	14 kbp	11	8 ssRNA
bacteriophage λ	49 kbp	66	1 dsDNA
Pandoravirus salinus (largest known viral genome)	2.8 Mbp	2500	1 dsDNA

Pandoravirus salinus (dsDNA)



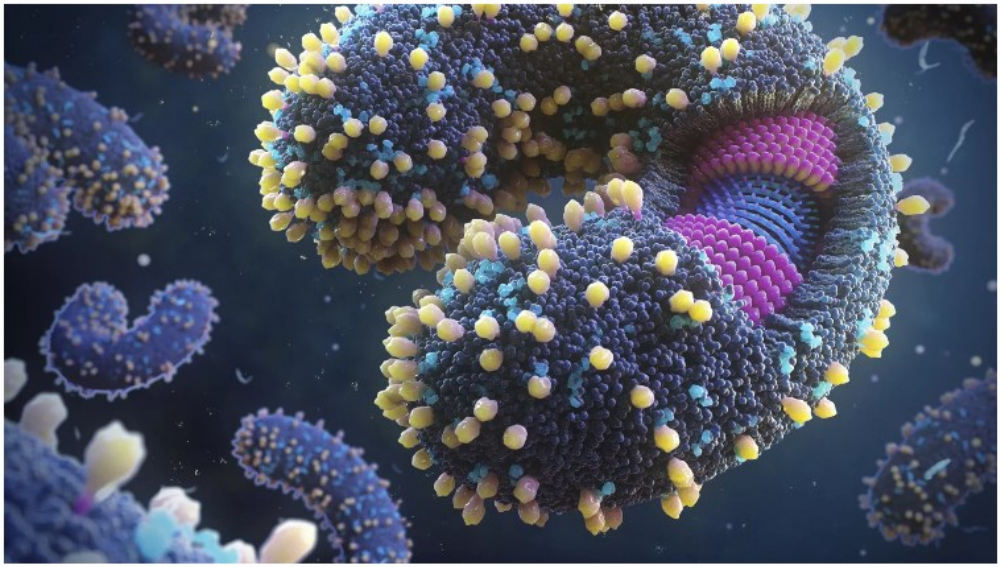
- 2.8 Mb
- 2 556 genes
- „parasites“ of amoebas
- only 6 % of genes match the known genes – unknown part of the tree of life?



Thousands of new viruses discovered in the ocean

By Rachael Rettner published April 10, 2022

More than 5,000 new RNA virus species were identified.



(Image credit: Nanoclustering/Getty Images)

More than 5,000 new virus species have been identified in the world's oceans, according to a new study.

The study researchers analyzed tens of thousands of water samples from around the globe, hunting for RNA viruses, or viruses that use RNA as their genetic material. The novel coronavirus, for instance, is a type of RNA virus. These viruses are understudied compared with DNA viruses, which use DNA as their genetic material, the authors said.

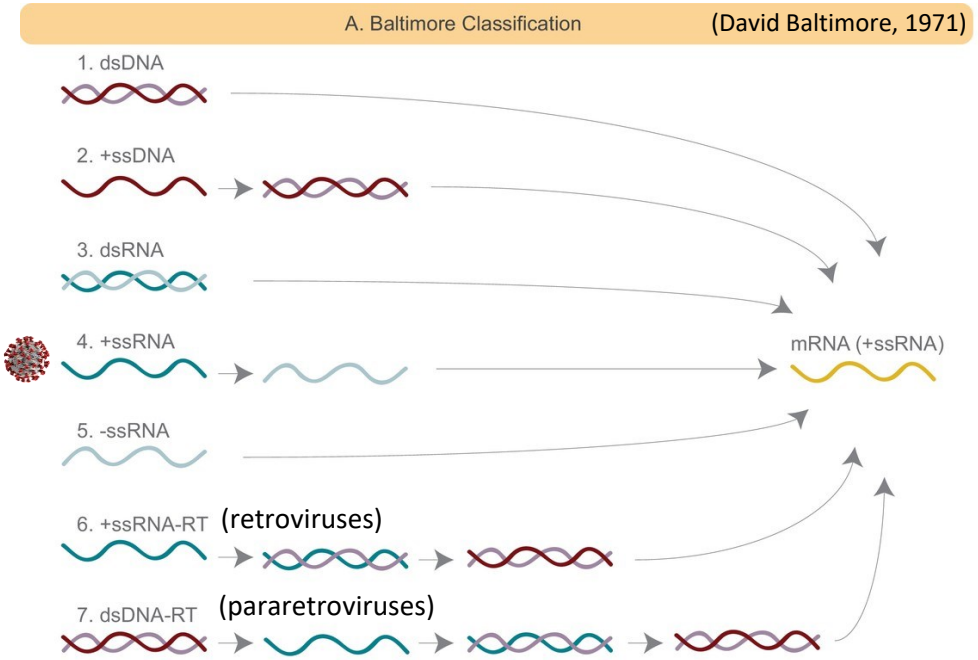
Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome

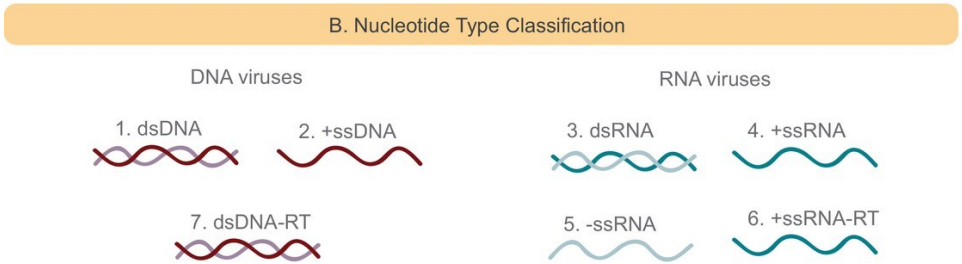
Ahmed A. Zayed[1,2,3]†, James M. Wainaina[1,3]†, Guillermo Dominguez-Huerta[1,2,3]†, Eric Pelletier[4,5], Jiarong Guo[1,2,3], Mohamed Mohssen[1,3,6], Funing Tian[1,3], Akbar Adjie Pratama[1,2], Benjamin Bolduc[1,2,3], Olivier Zablocki[1,2,3], Dylan Cronin[1,2,3], Lindsey Solden[1], Erwan Delage[5,7], Adriana Alberti[4,5]§, Jean-Marc Aury[4,5], Quentin Carradec[4,5], Corinne da Silva[4,5], Karine Labadie[4,5], Julie Poulain[4,5], Hans-Joachim Ruscheweyh[8], Guillem Salazar[8], Elan Shatoff[9], Tara Oceans Coordinators‡, Ralf Bundschuh[6,9,10,11], Kurt Fredrick[1], Laura S. Kubatko[12,13], Samuel Chaffron[5,7], Alexander I. Culley[14], Shinichi Sunagawa[8], Jens H. Kuhn[15], Patrick Wincker[4,5], Matthew B. Sullivan[1,2,3,6,12,16]*

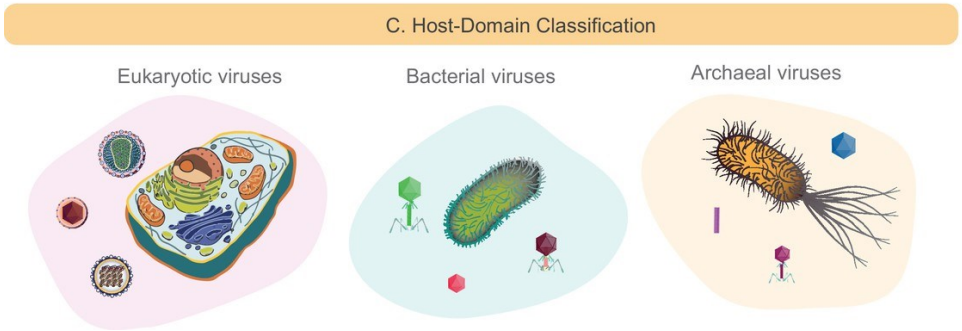
Whereas DNA viruses are known to be abundant, diverse, and commonly key ecosystem players, RNA viruses are insufficiently studied outside disease settings. In this study, we analyzed ≈28 terabases of Global Ocean RNA sequences to expand Earth's RNA virus catalogs and their taxonomy, investigate their evolutionary origins, and assess their marine biogeography from pole to pole. Using new approaches to optimize discovery and classification, we identified RNA viruses that necessitate substantive revisions of taxonomy (doubling phyla and adding >50% new classes) and evolutionary understanding. "Species"-rank abundance determination revealed that viruses of the new phyla "*Taraviricota*," a missing link in early RNA virus evolution, and "*Arctiviricota*" are widespread and dominant in the oceans. These efforts provide foundational knowledge critical to integrating RNA viruses into ecological and epidemiological models.

Science 376: 156-162 (2022)

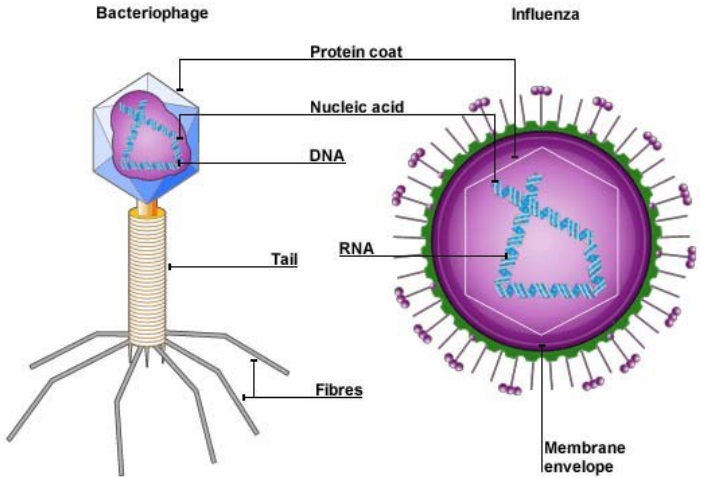
Viruses







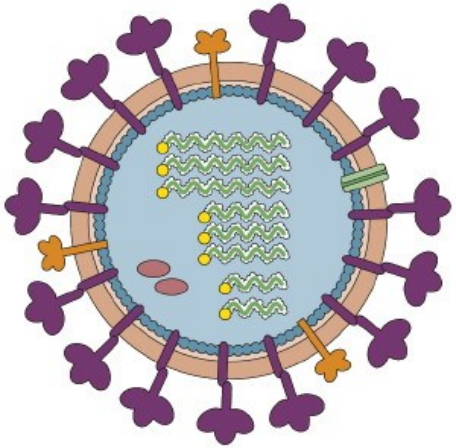
- single- or double-stranded DNA or RNA (DNA and RNA viruses)
- linear or circular
- one molecule or in segments (RNA viruses)
- very few genes (4 to a few hundred)



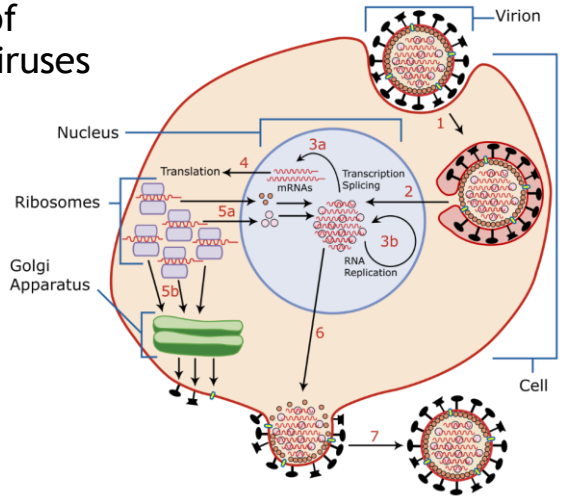
Viruses

- one linear or circular molecule
- segmented RNA virus genomes

(e.g., influenza virus: -ssRNA)

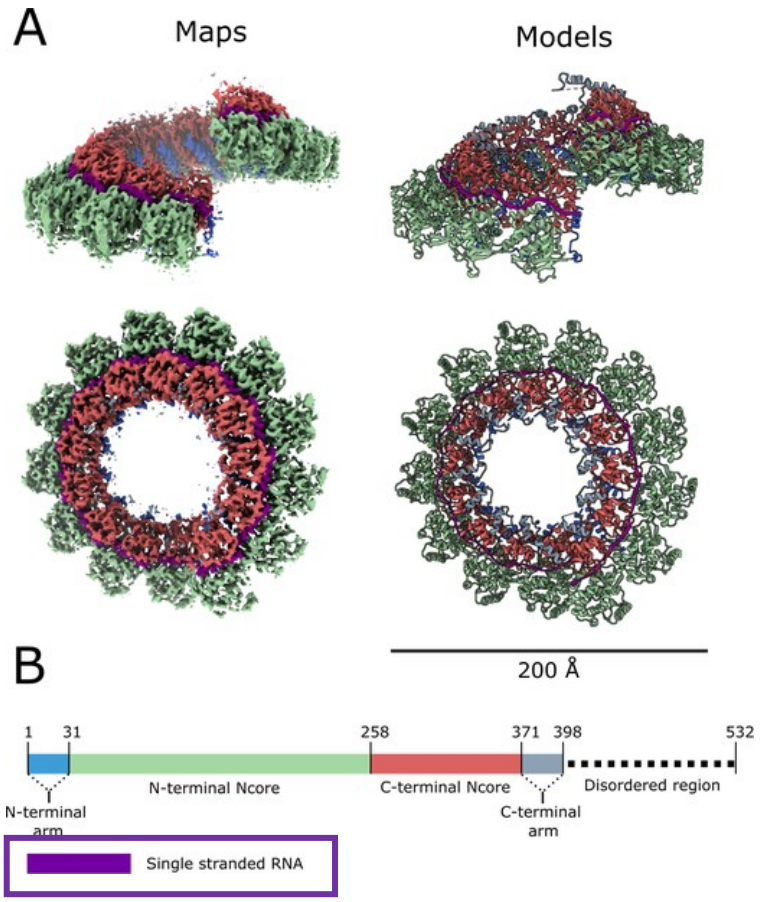


Life cycle of influenza viruses



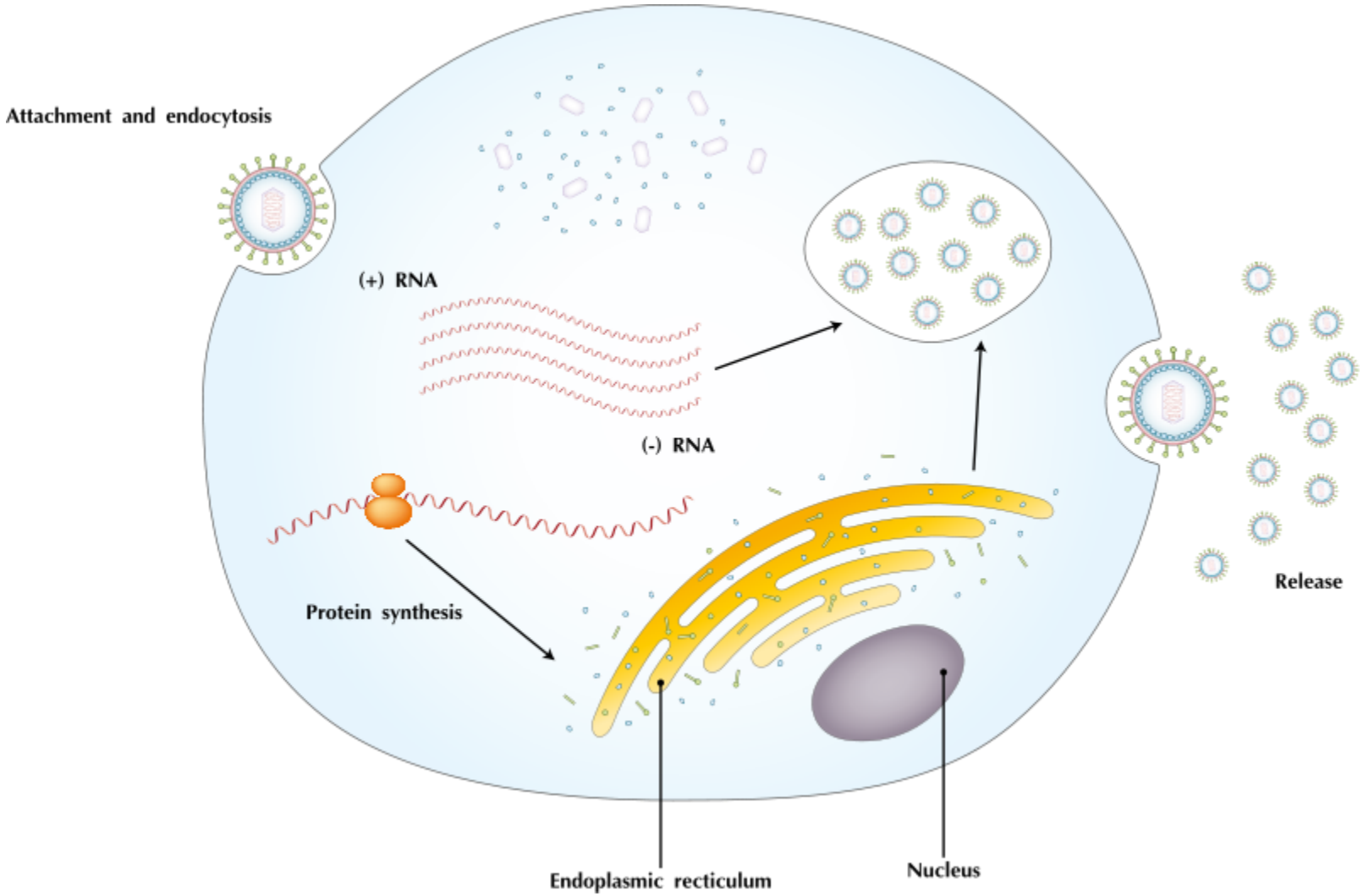
How DNA/RNA is compacted in viral nucleocapsids

Structure of the NiV nucleocapsid protein-RNA complex



Ker DS, Jenkins HT, Greive SJ, Antson AA (2021) CryoEM structure of the Nipah virus nucleocapsid assembly. PLOS Pathogens 17(7): e1009740.

Viral life cycle

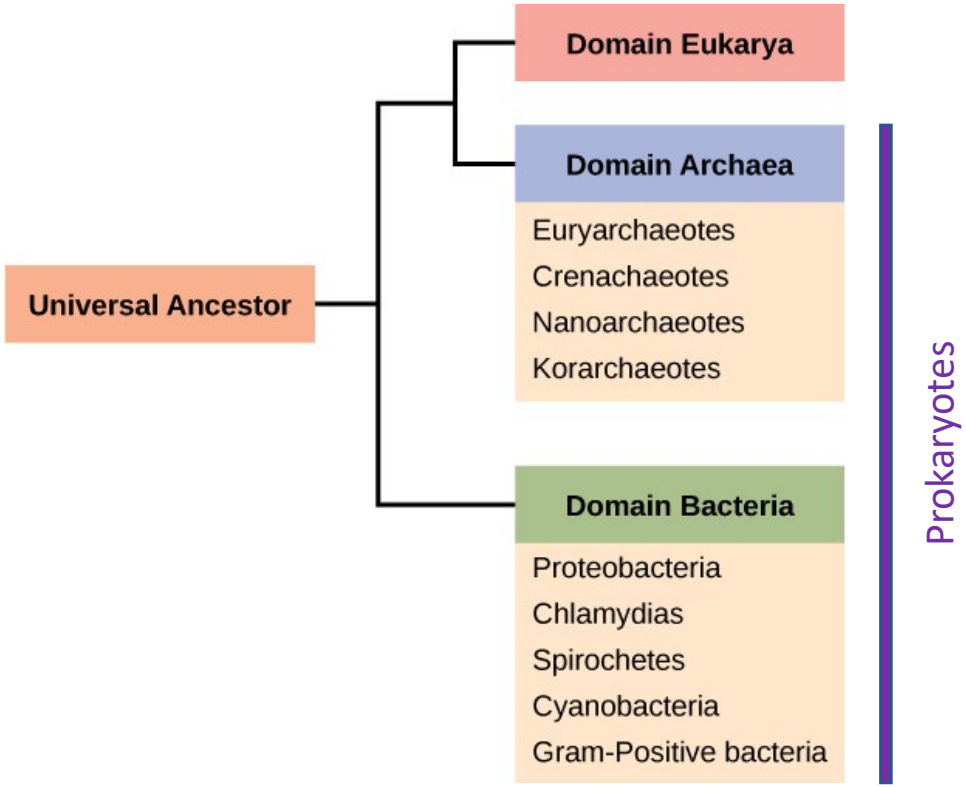


Endogenous viral elements (EVEs)

Viruses which integrated their genomes into genomes of their eukaryotic hosts.

- +ssRNA-RT (retroviruses) or dsDNA-RT (pararetroviruses)
- usually small DNA fragments (few genes) but also (rarely) entire viral genome, can be inherited (fixed)
- some can remain active (rarely)
- paleovirology: EVEs („viral fossils“) but parental viruses have become extinct
- example - algae (chlorophytes): large dsDNA viruses can integrate in the host genome (https://doi.org/10.1038/s41586-020-2924-2)
- between 78 and 1 782 genes from the virus to the algal genome, some algae have the whole genome of a giant virus in their DNA (up to 10% of all genes)
- some genes of the EVEs duplicated, some have introns = long-term „co-evolution“ with the host genome (two-way interaction between the viral and host genome)

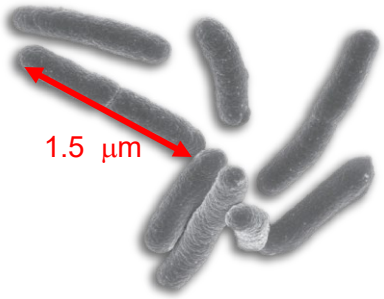
Bacteria, Archaea and Eukarya (Eukaryotes): three domains of life



Genomes of Archaea and Bacteria

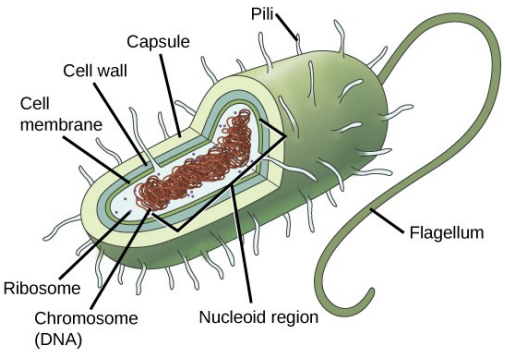
bacteria			
C. ruddii (smallest genome of an endosymbiont bacteria)	160 kbp	182	1
M. genitalium (smallest genome of a free living bacteria)	580 kbp	470	1
H. pylori	1.7 Mbp	1,600	1
Cyanobacteria S. elongatus	2.7 Mbp	3,000	1
methicillin-resistant *S. aureus* (MRSA)	2.9 Mbp	2,700	1
B. subtilis	4.3 Mbp	4,100	1
S. cellulosum (largest known bacterial genome)	13 Mbp	9,400	1
archaea			
Nanoarchaeum equitans (smallest parasitic archaeal genome)	490 kbp	550	1
Thermoplasma acidophilum (flourishes in pH<1)	1.6 Mbp	1,500	1
Methanocaldococcus (Methanococcus) jannaschii (from ocean bottom hydrothermal vents; pressure >200 atm)	1.7 Mbp	1,700	1
Pyrococcus furiosus (optimal temp 100°C)	1.9 Mbp	2,000	1

Escherichia coli



4.6×10^6 bp = 1.5 **mm**
(a 1000-fold compression)

- single-cell organisms
- small compact genomes
- (usually) circular DNA/chromosome (nucleoid) and plasmids
- do not have a nucleus and membrane-bound organelles
- reproduce by fission (after the chromosome is replicated)



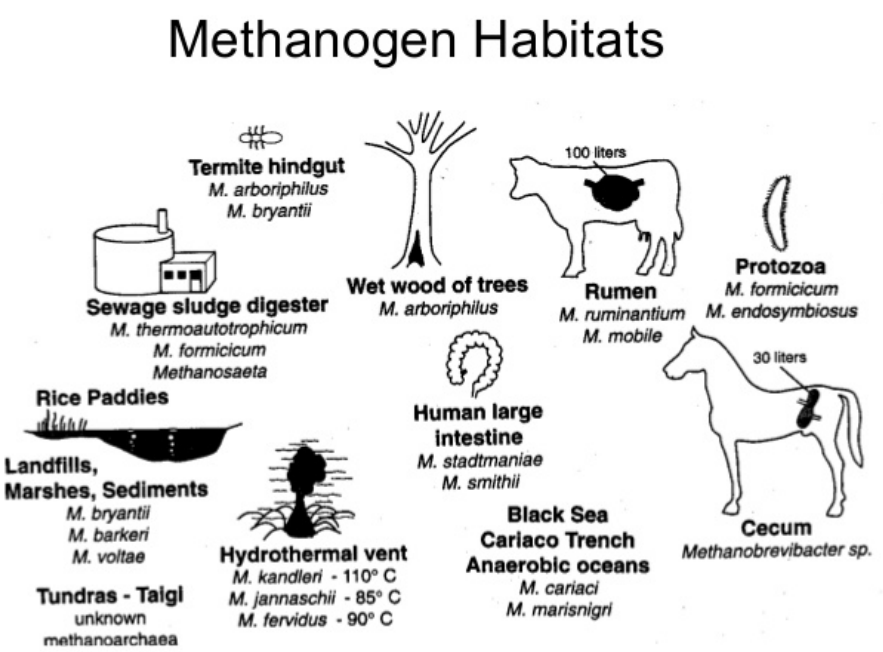
Carsonella ruddii – **smallest genome** of endosymbiotic bacteria (160 kb, 182 genes)
Mycoplasma genitalium – **smallest genome** of free living bacteria (580 kb, 470 genes)

Sorangium cellulosum – **the largest known bacterial genome** (13 Mb, 9 400 genes)

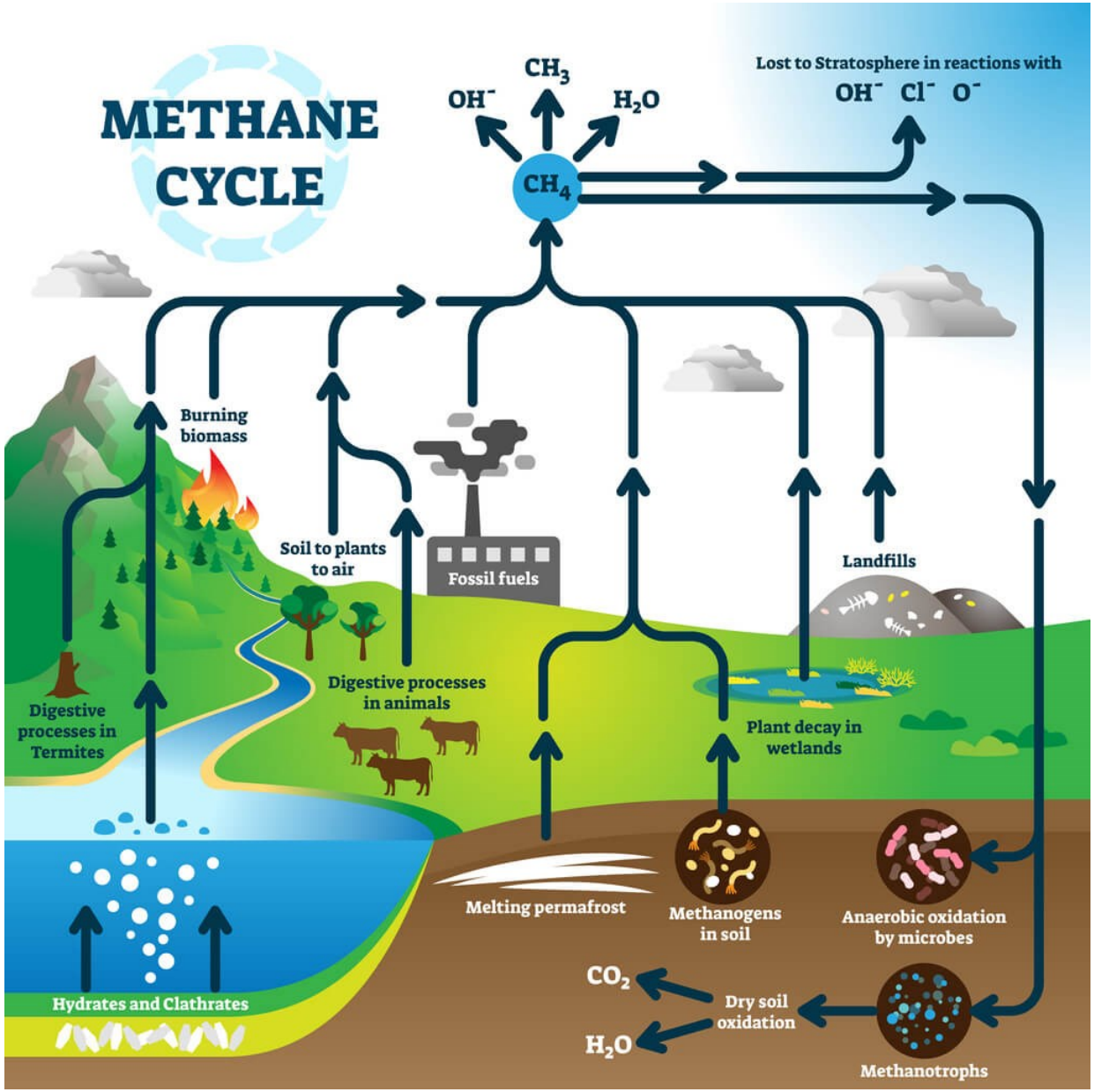
Genomes of Archaea (formerly Archaebacteria)

- Methanogens (methane-producing strains)
- Halophiles
- Thermophiles
- Alkalophiles
- Acidophiles

the archaea *Methanosphaera stadtmanae*



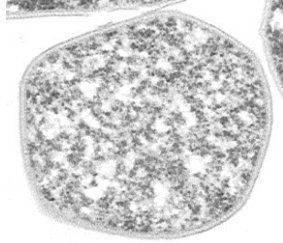
Archaea and methane

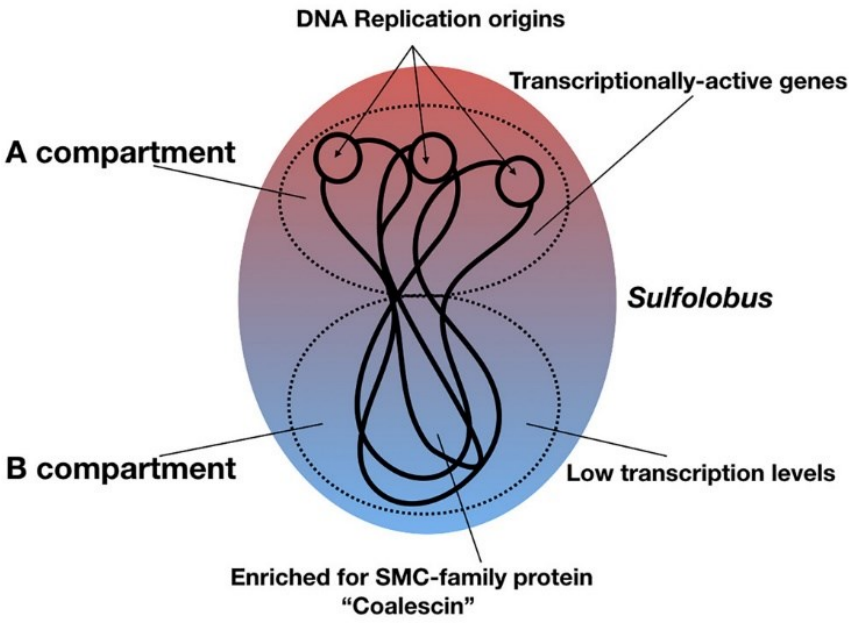


Genomes of Archaea

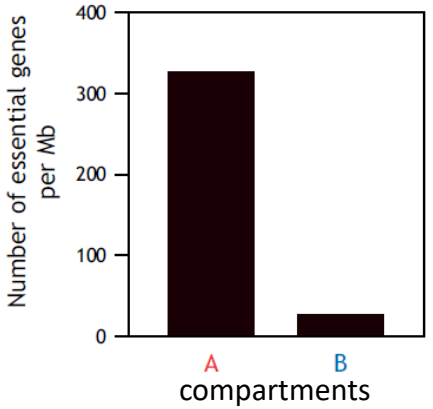
- usually a single circular chromosome, plasmids can be found
- Euryarchaea: multiple copies of their genomes (oligoploidy) – up to 55 copies in *Methanococcus maripaludis*
- smallest genome: 491 kb (*Nanoarchaeum equitans*)
- largest genome: 5.8 Mb (*Methanosarcina acetivorans*), only 537 protein-encoding genes
- some genes common with bacteria and eukaryotes, some unique (mostly unknown function)
- DNA polymerase similar to eukaryotic DNA polymerases, transcription more similar to eukaryotes (one type of RNA polymerase similar to RNA polymerase II in eukaryotes), translation similar to both bacteria and eukaryotes
- reproduction is asexual (fission, fragmentation, budding) after the chromosome is replicated

Multi-scale architecture of archaeal chromosomes





- *Sulfolobus*: single circular chromosome (2.2 – 3 Mb)
- two-domain compartmentalization that influences gene expression (domain-like organization somewhat similar to eukaryotes)



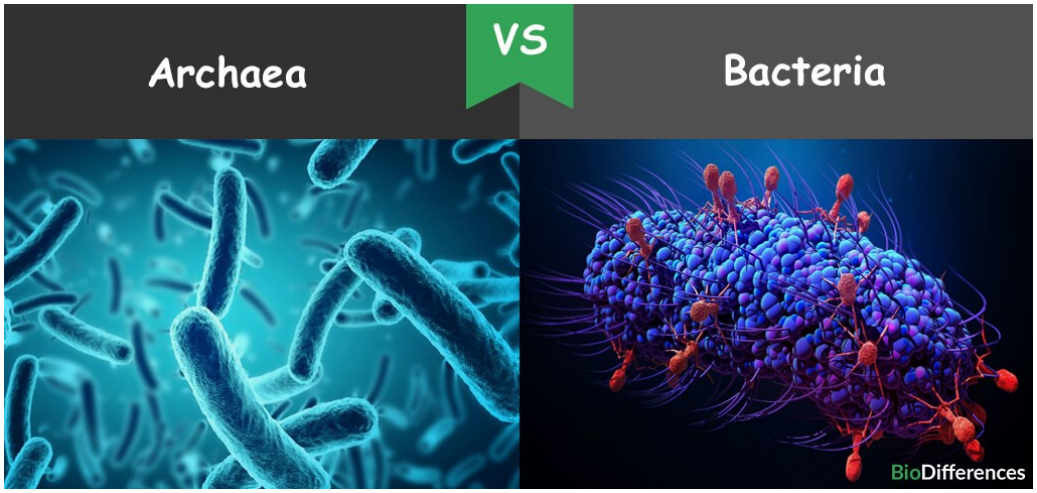
- coalescin (SMC-family protein*) marks the domain with low transcription levels (B compartment)

*SMC- Structure Maitenance of Chromosomes (protein complexes), mainly in eukaryotes (e.g., condensin, cohesin)

Archaea vs Bacteria

Archaea are non-pathogenic.

Do not use glycolysis or Kreb's cycle for glucose oxidation but follow metabolic pathways similar to these.



Bacteria might be pathogenic or non-pathogenic.

Glycolysis and Kreb's cycle are important metabolic pathways in bacteria for glucose oxidation.

Cell wall is made up of pseudopeptidoglycans: N-acetylglucosamine (NAG) and N-acetyltalosaminuronic acid (NAT)

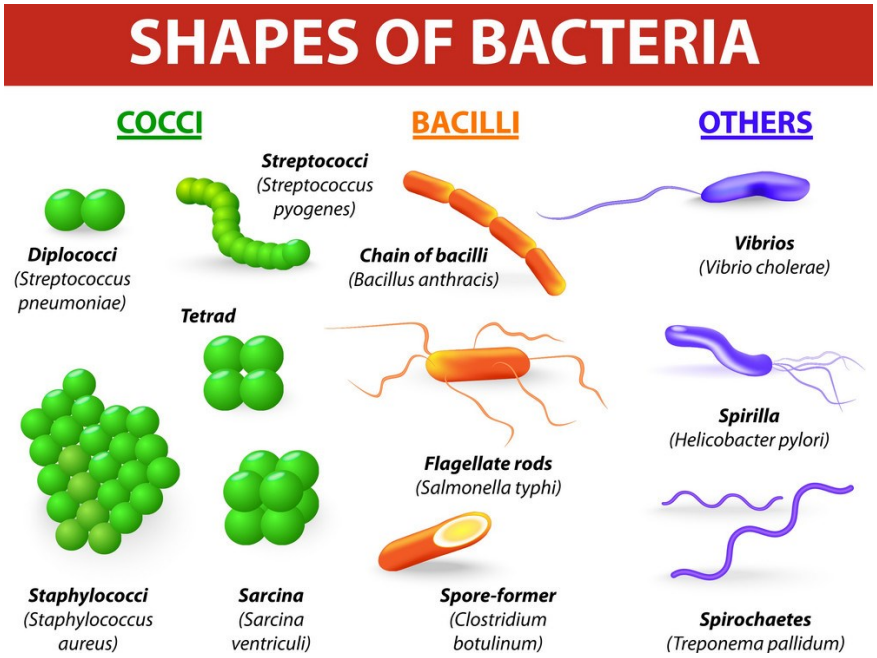
Introns are present in the chromosomes of archaea.

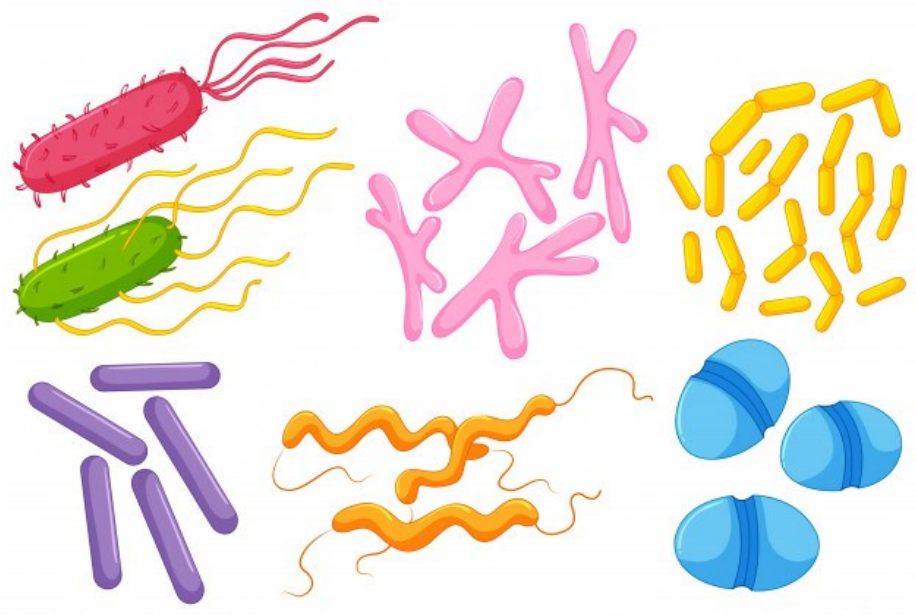
ARCHAEA	BACTERIA
Archaea do not have peptidoglycan in their cell wall	Bacteria have peptidoglycan in their cell wall
Genes are more similar to Eukarya	Genes are different from Eukarya

Cell wall is made up of peptidoglycans consisting of N-acetylglucosamine (NAG) and N-acetylmuramic acid (NAM).

Introns are absent in the chromosomes of bacteria.

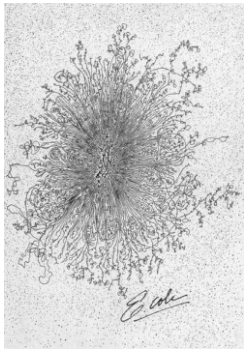
Bacterial genomes

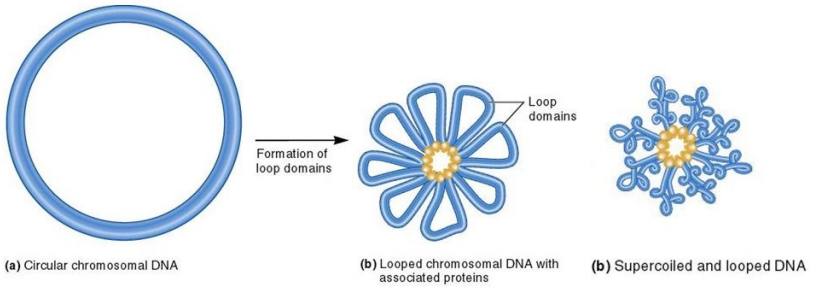


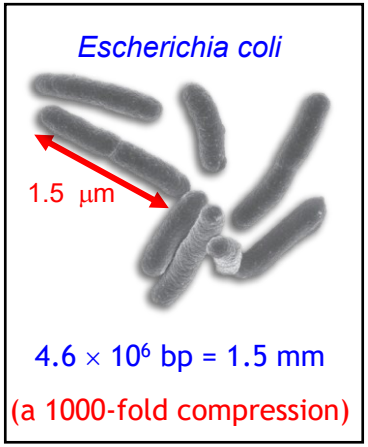


Bacterial chromosome

Escherichia coli - traditional view: single circular chromosome (dsDNA)







! Some bacteria have multiple chromosomes (e.g. 3.1-Mb and 0.9-Mb circular chromosomes in *Rhodobacter sphaeroides*).

! Linear chromosomes in some bacteria (1970, 1989 by PFGE: *Borrelia burgdorferi*, size c. 1 Mb)

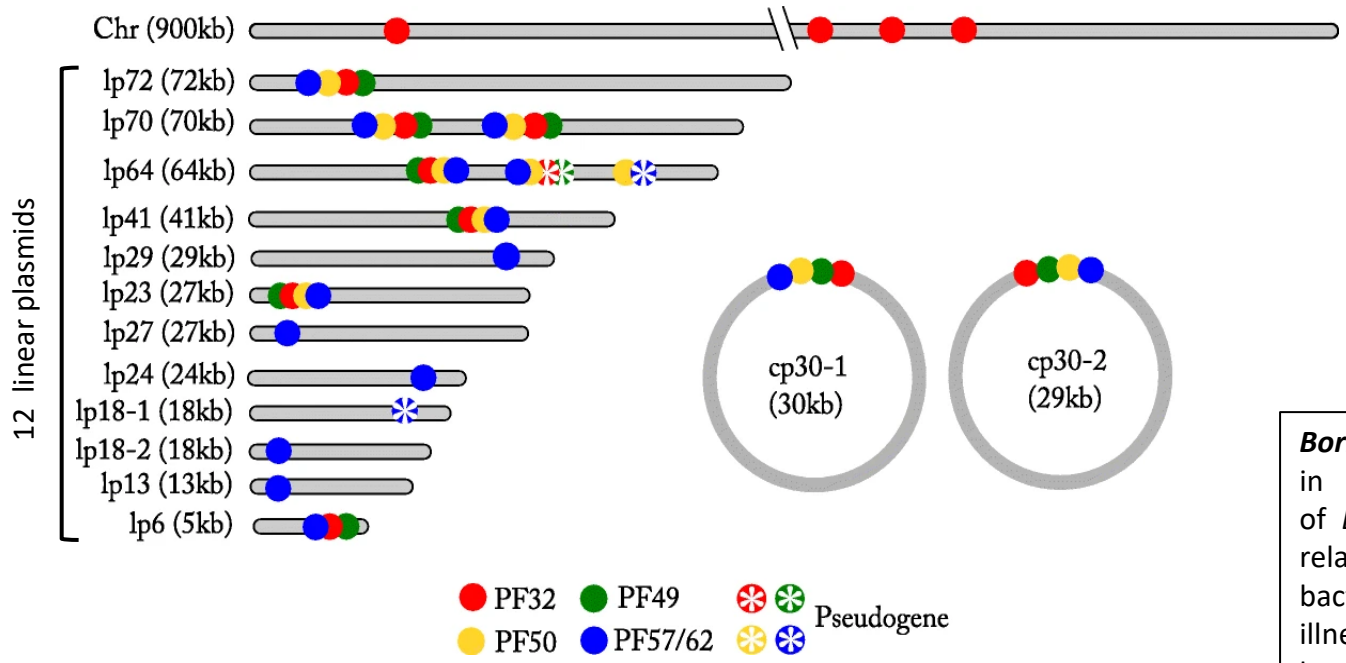
Problematic ends of linear chromosomes:

- palindromic hairpin loops (**hairpin telomeres**)
- **invertron telomeres** – inverted terminal repeats and covalently attached capping (terminal) proteins

Bacteria	**Chromosome organization**
Agrobacterium tumefaciens	One linear and one circular
Bacillus subtilis	Single and circular
Bacillus subtilis	Single and linear
Borrelia burgdorferi	Single and linear
Escherichia coli	Single and circular
Paracoccus denitrificans	Three circular
Pseudomonas aeruginosa	Single and circular
Rhodobacter sphaeroides	Two circular
Streptomyces griseus	Linear
Vibrio cholerae	Two circular
Vibrio fluvialis	Two circular

Genome of *Borrelia miyamotoi* (example)

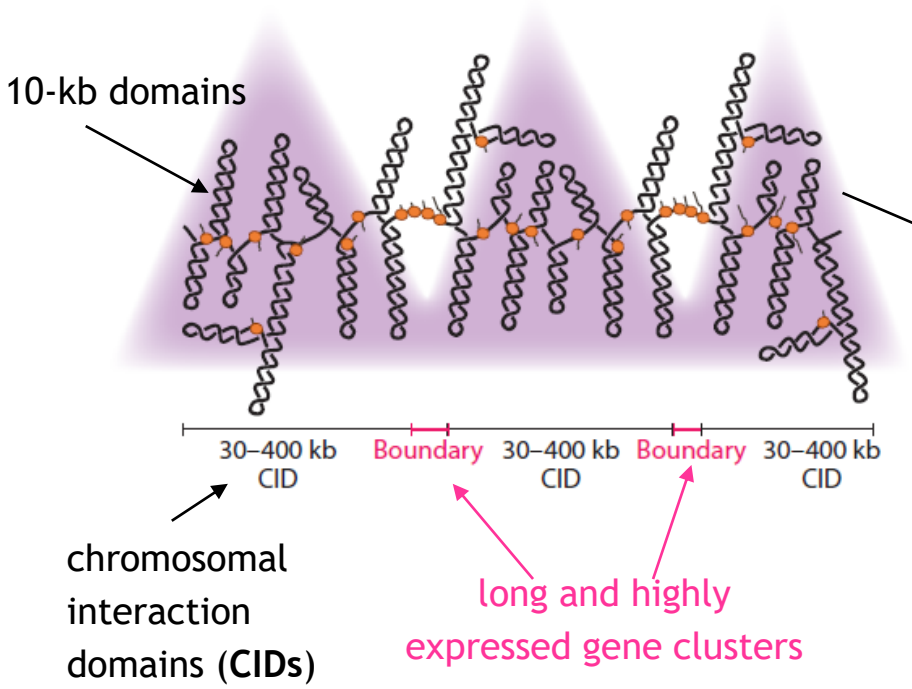
- one linear chromosome (906 kb)
- 12 linear plasmids (6 to 72 kb)
- 2 circular plasmids (each ~30 kb)



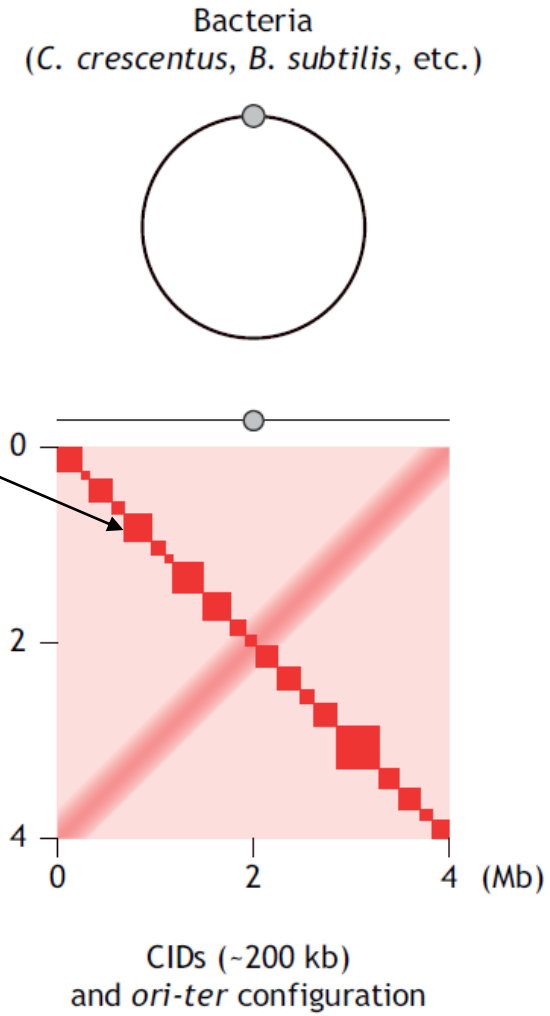
Borrelia miyamotoi is a bacteria is in the relapsing fever group of *Borrelia*. Although it's not closely related to the Lyme disease bacteria, it can cause a Lyme-like-illness. Symptoms include fever, headaches, muscle aches and chills, rash uncommon. Diagnosis is by PCR testing that is now available at several labs. Treatment is doxycycline. *B. miyamotoi* was identified in 1995 in ticks from Japan.

3D structuring of bacterial circular chromosome

DNA supercoiling

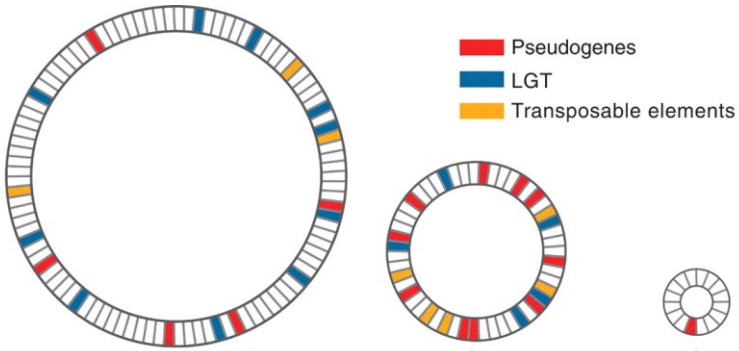


Hi-C analysis (heatmap)



Bacterial genomes – trends in content and size

- 160 kb to 13 Mb
- most of the genome (85-90%) is non-repetitive DNA (coding DNA), while non-coding regions only take a small part
- bacteria have relatively small amounts of junk (non-coding) DNA → a high correlation between the number of genes and the genome size in bacteria
- the lifestyles of bacteria play an integral role in their respective genome sizes. Free-living bacteria have the largest genomes out of the three types of bacteria; however, they have fewer pseudogenes than bacteria that have recently acquired pathogenicity. Parasitic and endosymbiotic bacteria can rely on host environments to provide gene products.



	Free-living	Recent or facultative pathogen	Obligate symbiont or pathogen
Genome size	Large (5-10 MB)	Intermediate (2-5 MB)	Small (0.5-1.5 MB)
Number of pseudogenes	Few	Many	Rare
Incidence of LGT	Frequent	Frequent to rare	Rare to none
Selfish genetic elements	Few	Common	Rare
Genome organization	Stable or unstable	Unstable	Stable
Effective population size	Large	Small	Small

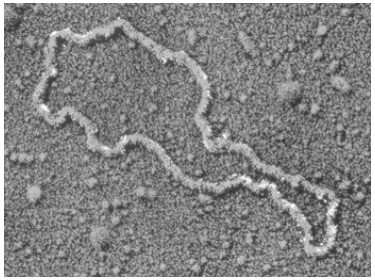
Free-living species— selection effective in removing deleterious sequences → large genomes containing relatively few pseudogenes (red) or mobile genetic elements (yellow).

In **recently derived pathogens**, the availability of host-supplied nutrients combined with decreases in effective population sizes allows for the accumulation of pseudogenes and of transposable elements.

In **long-term host-dependent species**, the ongoing mutational bias toward deletions has removed all superfluous sequences, resulting in a highly reduced genome containing few, if any, pseudogenes or transposable elements.

LGT, lateral gene transfer.

Ochman and Davalos, Science 2006

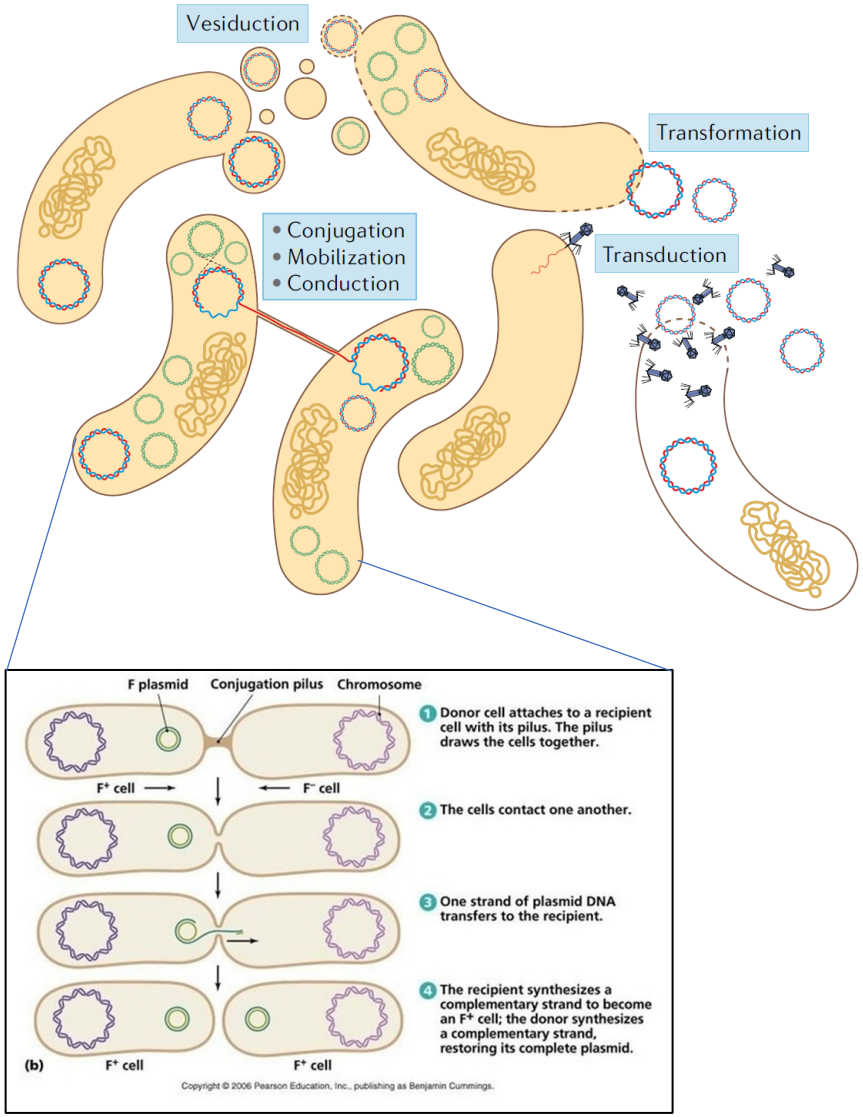


What is the role of plasmid DNA?

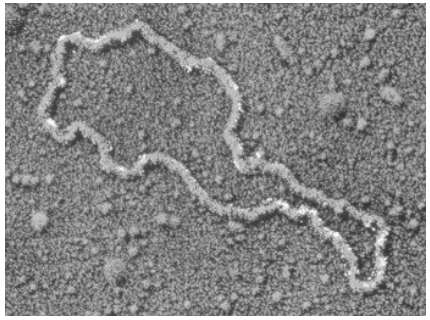
- usually small (1- 200 kb), usually circular DNAs; independent replication

Salmonella serovar	Plasmid type	No. of plasmids	Plasmids (kb)	No. of isolates
S. Anatum	P01	1	53·7	38
	P02	3	53·7, 5·07, 3·03	1
	P03	3	53·7, 7·6, 4	1
S. Enteritidis	P01	1	53·7	16
	P04	2	53·7, 5·46	2
S. Corvallis	P01	1	53·7	2
	P05	5	53·7, 5·46, 5·07, 3·03, 2	2
	P06	2	53·7, 7·2	2
	P07	2	53·7, 4	2
S. Typhimurium	P01	1	53·7	6

Plasmid mobility



What is the role of plasmid DNA?



Plasmids generally contain genes that confer some sort of advantage for survival and reproduction:

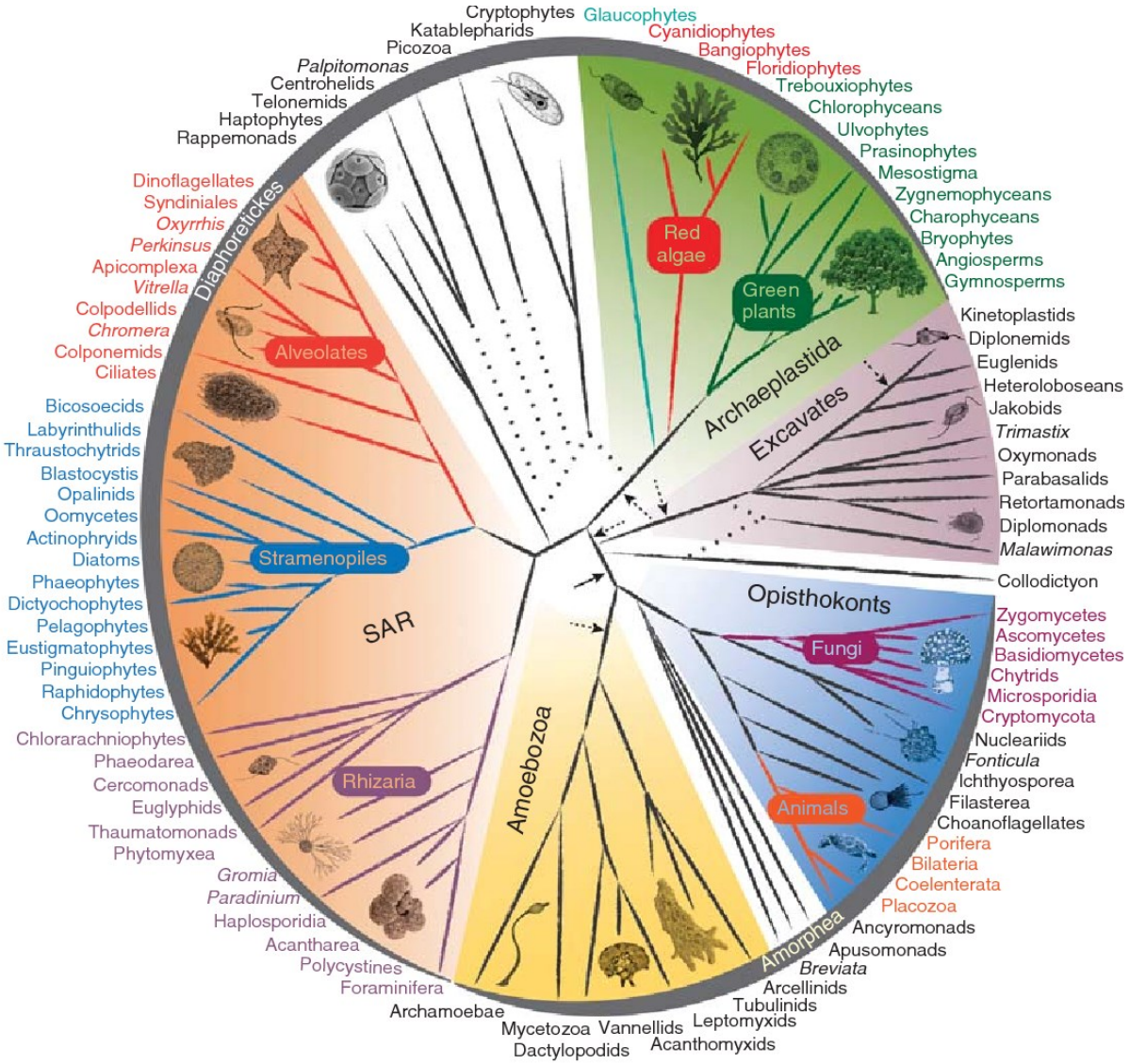
- Genes providing protection from toxic substances (including antibiotic resistence). Plasmids carrying antibiotic resistance genes are key contributors to the uncontrollable spread of bacterial pathogens, particularly in hospitals
- Genes enabling the metabolism of additional sources of energy
- Genes for toxins to kill microbial competitors, enhance pathogenicity
- Genes involved in gene transfer by conjugation
- New evidence that suggests that plasmids might accelerate bacterial evolution, mainly by promoting the evolution of plasmid-encoded genes, but also by enhancing the adaptation of their host chromosome

Plasmids and major transitions in the evolution of bacteria

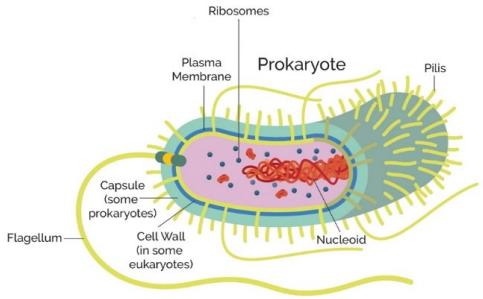
Rhodobacteraceae (marine bacteria) - plasmids responsible for the ability to undergo anoxygenic photosynthesis

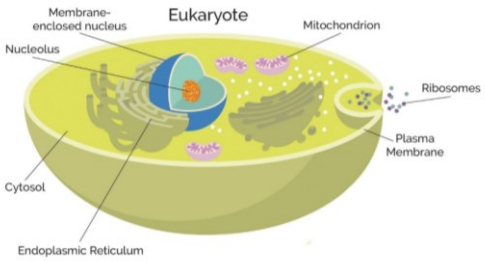
the aphid endosymbiont *Buchnera aphidicola* – plasmids responsible for synthesizing essential amino acids and vitamins required for the bacterium-aphid symbiotic relationship (advantage for aphid)

Eukaryotes


Cryptophytes
Katablepharids
Picozoa
Palpitomonas
Centrohelids
Telonemids
Haptophytes
Rappemonads
Glaucophytes
Cyanidiophytes
Bangiophytes
Floridiophytes
Trebouxiophytes
Chlorophyceans
Ulvophytes
Prasinophytes
Mesostigma
Zygnemophyceans
Charophyceans
Bryophytes
Angiosperms
Gymnosperms
Red algae
Green plants
Archaeplastida
Kinetoplastids
Diplonemids
Euglenids
Heterolobosean
Jakobids
Trimastix
Oxymonads
Parabasalids
Retortamonads
Diplomonads
Malawimonas
Excavates
Collodictyon
Opisthokonts
Fungi
Zygomycetes
Ascomycetes
Basidiomycetes
Chytrids
Microsporidia
Cryptomycota
Nucleariids
Fonticula
Ichthyosporea
Filasterea
Choanoflagellates
Animals
Porifera
Bilateria
Coelenterata
Placozoa
Amorphea
Ancyromonads
Apusomonads
Breviata
Arcellinids
Tubulinids
Leptomyxids
Acanthomyxids
Vannellids
Mycetozoa
Dactylopodids
Archamoebae
Amoebozoa
Dinoflagellates
Syndiniales
Oxyrrhis
Perkinsus
Apicomplexa
Vitrella
Colpodellids
Chromera
Colponemids
Ciliates
Diaphoretickes
Alveolates
Bicosoecids
Labyrinthulids
Thraustochytrids
Blastocystis
Opalinids
Oomycetes
Actinophryids
Diatoms
Phaeophytes
Dictyochophytes
Pelagophytes
Eustigmatophytes
Pinguiophytes
Raphidophytes
Chrysophytes
Stramenopiles
Chlorarachniophytes
Phaeodarea
Cercomonads
Euglyphids
Thaumatomonads
Phytomyxea
Gromia
Paradinium
Haplosporidia
Acantharea
Polycystines
Foraminifera
Rhizaria
SAR

Prokaryotes vs Eukaryotes

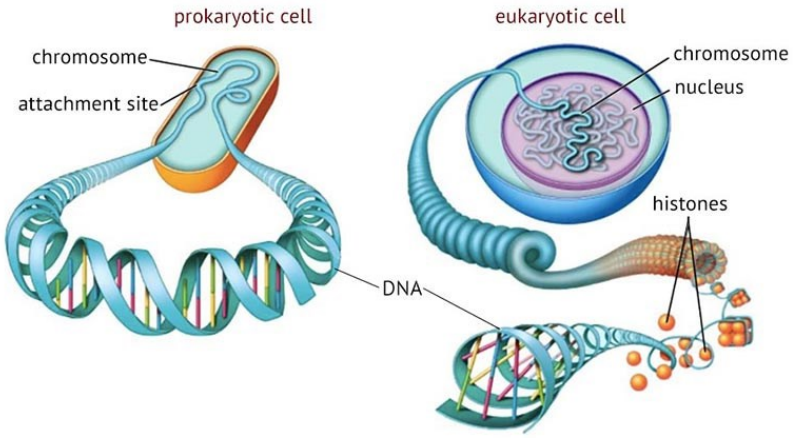




- Simple, small (0.1 – 5 μm)
- Do not have membrane-bound structures (nucleus, mitochondria)
- Nucleoid: DNA
- Cell wall: protection from the outside environment. Most bacteria have a rigid cell wall made from carbohydrates and peptidoglycans.
- Cell membrane (plasma membrane)
- Capsule: Some bacteria have a layer of carbohydrates that surrounds the cell wall called the capsule. The capsule helps the bacterium attach to surfaces.
- Fimbriae: thin, hair-like structures that help with cellular attachment.
- Pili: rod-shaped structures involved in multiple roles, including attachment and DNA transfer.
- Flagella: thin, tail-like structures that assist in movement
- Transcription and translation are coupled (translation begins during mRNA synthesis)

- Complex, cell bigger (10 – 100 μm)
- Multicellular, some single-cell eukaryotes
- Nucleus and other organelles enclosed by a plasma membrane
- Nucleolus: production of ribosomal RNA molecules
- Plasma membrane: a phospholipid bilayer that surrounds the entire cell and encompasses the organelles within.
- Cytoskeleton or cell wall: provides structure, allows for cell movement, and plays a role in cell division.
- Mitochondria: responsible for energy production.
- Endoplasmic reticulum: an organelle dedicated to protein maturation and transportation.
- Vesicles and vacuoles: membrane-bound sacs involved in transportation and storage.
- Transcription in the nucleus (mRNA), translation in cytoplasm

Prokaryotes vs Eukaryotes (more differences)

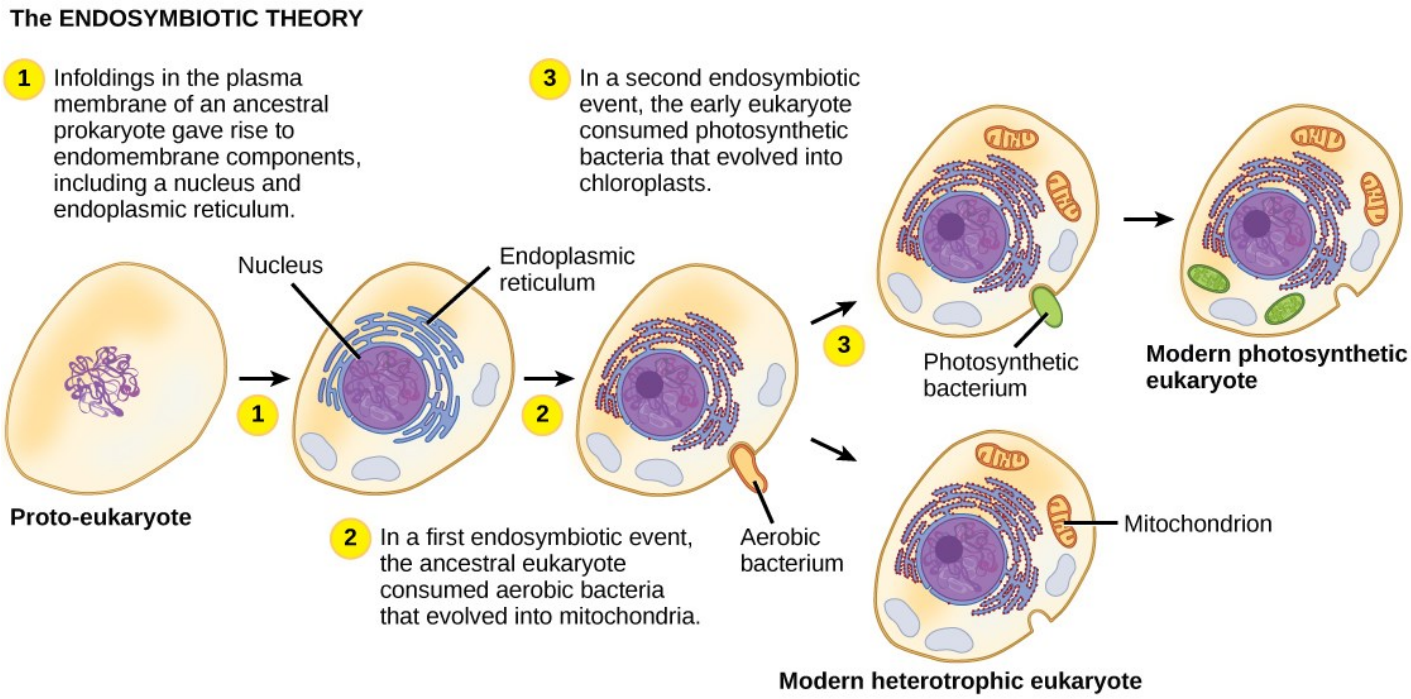


- the circular/linear DNA is packaged → nucleoid (50 or more loops/domains bound to a central protein scaffold, attached to the cell membrane) = DNA is negatively supercoiled, that is, it is twisted upon itself
- several DNA-binding proteins (the most common HU, HLP-1 and H-NS; these are histone-like proteins)

- chromosomes contain both DNA and proteins (mostly histones, but also non-histone proteins)
- each chromosome is a single linear double-stranded DNA molecule
- the extensive packaging of DNA in chromosomes results from three levels of folding (nucleosomes, „30-nm fibres“ and radial loops)
- the length of the packaged DNA molecule varies. In humans, the shortest DNA molecule in a chromosome is about 1.6 cm and the longest is about 8.4 cm

Origin of eukaryotic genomes (eukaryogenesis)

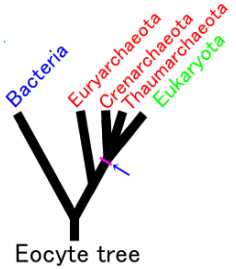
- prokaryotic cells occurred c. 1 billion years after the Earth was formed – i.e. about 3.5 billion years ago
- eukaryotic cells emerged about 2.5 billion years ago
- Lynn Margulis (in the 1960s): endosymbiotic theory of the origin of an eukaryotic cell
- eukaryotic **nuclear genes** appear to originated from the Archaea, **mitochondria** appear to be of the bacterial origin



Paramecium bursaria with *Zoochlorella* algae

Origin of Eukaryotes within the Archaea

- **Eocyte hypothesis** (James A. Lake and others, 1984): eukaryotes emerged within **Crenarchaeota** (formerly eocytes), a phylum of the Archaea; based on the shapes of ribosomes in the Crenarchaeota and eukaryotes being more similar than ribosomes of eukaryotes and bacteria (or other Archaea)



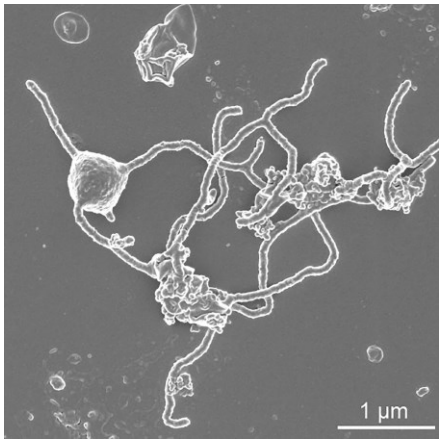
- later studies suggested that eukaryotes might have originated within **Thaumarchaeota** (today Crenarcheaota and Thaumarchaeota belong to the superphylum TACK)

- **Asgard** - another superphylum of the Archea was not known in the 1980s
- it appears that eukaryotes originated within **Heimdallarchaeota**
- **in cladistic view, eukaryotes are Archaea, similarly as birds are dinosaurs**

Last universal common ancestor

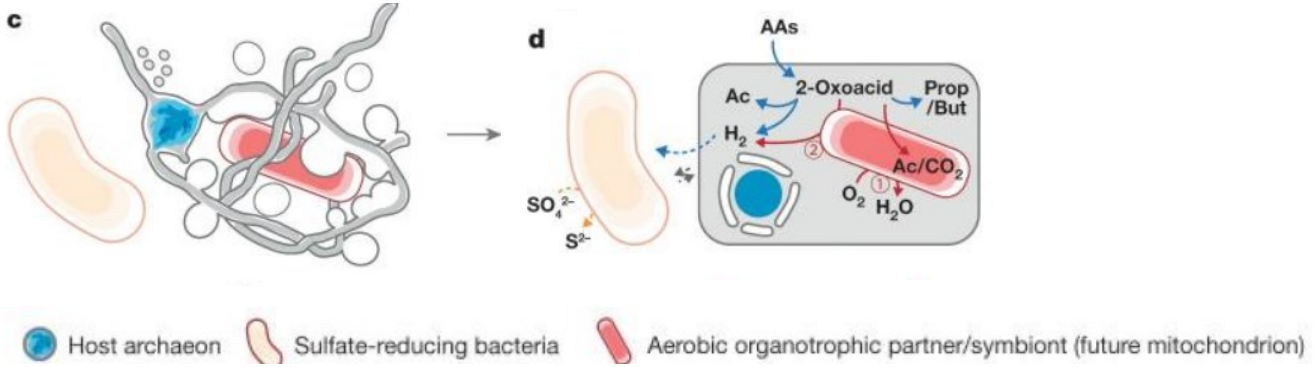
Potential living link between Archaea and eukaryotes

- *Prometheoarchaeum syntrophicum* - an archeon of the **Asgard** superphylum
- from the ocean floor (2 533 m water depth, Japan)
- support for the hypothesis of **eukaryogenesis via endosymbiosis:**



Prometheoarchaeum syntrophicum

the host archaeon engulfed the metabolic partner/bacteria (future mitochondrion) using extracellular structures and simultaneously formed a primitive chromosome-surrounding structure similar to the nuclear membrane:

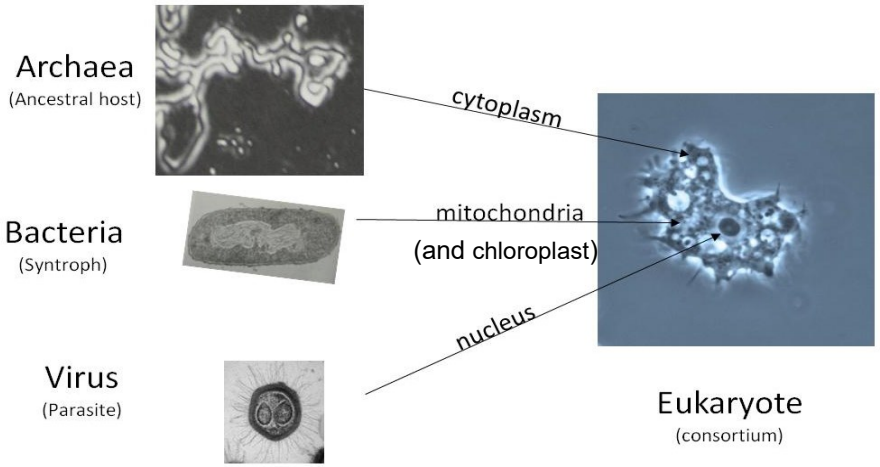


Imachi H, Nobu MK, Nakahara N, et al. Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature*. 2020;577(7791): 519-525.

Origin of Eukaryotes – viral eukaryogenesis

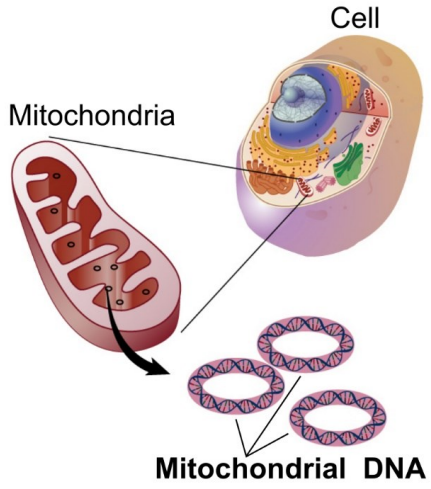
(Philip Bell, 2001, 2020)

- A hypothesis that the eukaryotic nucleus could originated from a virus infecting an archaeal ancestor (donor of cytoplasm)
- This virus(es) could be similar to large, complex DNA viruses (such as Mimivirus) that are capable of protein biosynthesis
- The **virus-derived nucleus probably acquired some genes from the archaeal host genome** and bacterial genome(s)
- A similar proces, when a bacteriophage hijacks bacterial cell‘s machinery and forms a nucleus-like structure, was observed by Chaikeeratisak et al. (2017, Science): https://www.youtube.com/watch?v=0xM5BhQ2kc8&feature=emb_title

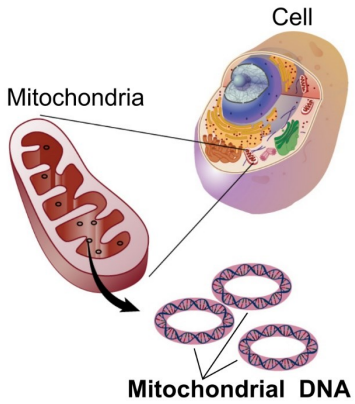


Extra-nuclear genomes and extra-chromosomal DNA in eukaryotes

(outside the chromosomes and typically also outside the nucleus)

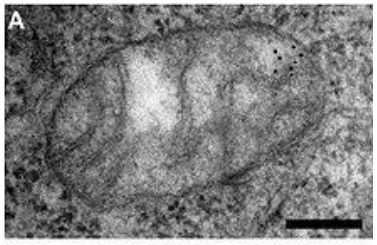


Mitochondrial genome (mtDNA)



- human mtDNA includes 16,569 base pairs and encodes 13 proteins, 2 rRNAs, 22 tRNAs
- animals: usually circular DNA molecule, but also linear genome
- plants and fungi (circular, rarely linear), 3 types of mt genome:
  - a circular genome that has introns (19 to 1 000 kb)
  - a circular genome (20 – 1 000 kb) that also has a plasmid-like structure (1 kb)

  (plasmids sometimes integrated in the circular chromosome)
  - a mt genome consisting of linear DNA molecules



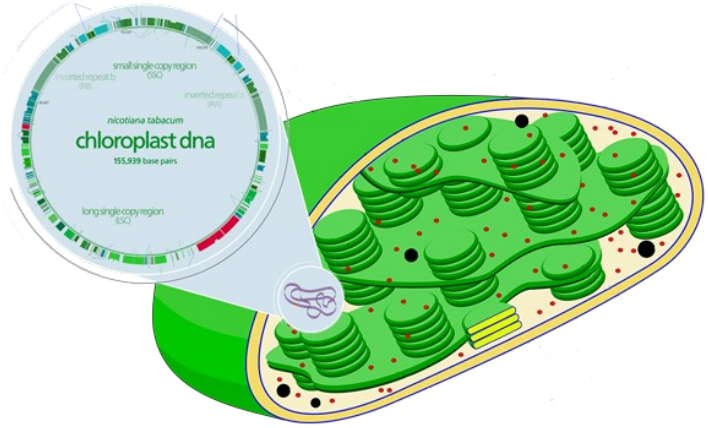
- *Silene conica*: enormous mtDNA genome – 11.3 Mb
- mitochondrion of the cucumber (*Cucumis sativus*): 3 circular chromosomes (1 556, 84 and 45 kb)
- female inheritance (rerely male inheritance)

Kingdom	Introns	Size	Shape	Description
Animal	No	11–28 kb	Circular	Single molecule
Fungi, Plant, Protista	Yes	19–1000 kb	Circular	Single molecule
Fungi, Plant, Protista	No	20–1000 kb	Circular	Large molecule and small plasmid like structures
Protista	No	1–200 kb	Circular	Heterogeneous group of molecules
Fungi, Plant, Protista	No	1–200 kb	Linear	Homogeneous group of molecules
Protista	No	1–200 kb	Linear	Heterogeneous group of molecules

The remains of King Richard III were identified by comparing his mtDNA with that of two matrilineal descendants of his sister.

Chloroplast genome (plastome)

- each chloroplast contains ~100 copies of DNA in young leaves, declining to 15 - 20 copies in older leaves. These usually cluster into nucleoids containing several identical chloroplast DNA rings; many nucleoids in each chloroplast
- usually circular DNA molecule, but frequently also in a linear shape; 120 – 170 kp long
- quadripartite structure: small (SSC) and large single copy (LSC) section, 2 inverted repeats (IRs)
- IRs contain 3 rRNA genes, 2 tRNA genes; loss of one IR multiple times
- land plants (129 genes in average, min. 64, max. 313), parasitic plants (no photosynthesis): reduced no. of genes (63 genes) vs. gene no. increase (*Pelargonium*): 180 genes (243 kb)
- land plants: coding 4 ribosomal RNAs, 30–31 tRNAs, 21 ribosomal proteins, and 4 RNA polymerase subunits; genes important for photosynthesis

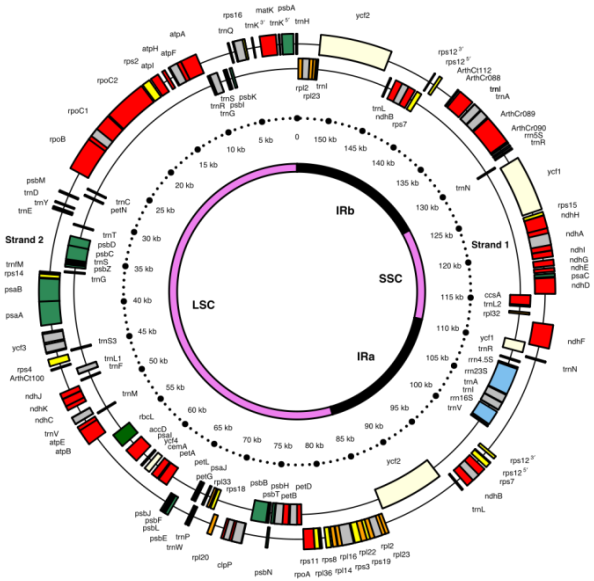


Arabidopsis thaliana
154-kb plastome genome

SSC short single copy section
LSC long single copy section
IR inverted repeats

Chloroplast genome (plastome)

- prokaryotic origin (cyanobacterium), endosymbiosis; chloroplast ribosomes are similar to bacterial ribosomes
- less genes than prokaryotic ancestors: transfer of thousands of genes to the nucleus (e.g., c. 18% of Arabidopsis nuclear DNA (4500 protein-coding genes) originated in chloroplast)
- ~95% of chloroplast proteins are encoded by nuclear genome
- positive correlation between nuclear genome size and length of transferred cp DNA fragments (the largest in rice, 131 kb, almost entire cp genome), integrated mainly to pericentromeric regions in rice (many removed during evolution)
- chloroplast genome evolves about 10-times slower than the nuclear genome
- mostly uniparental maternal inheritance, less common uniparental paternal and biparental inheritance; gymnosperms inherit plastids from male parent (pollen); interspecies hybrids: plastid inheritance can be mixed; 20% of angiosperms (e.g. Alfalfa, *Medicago sativa*) have biparental inheritance

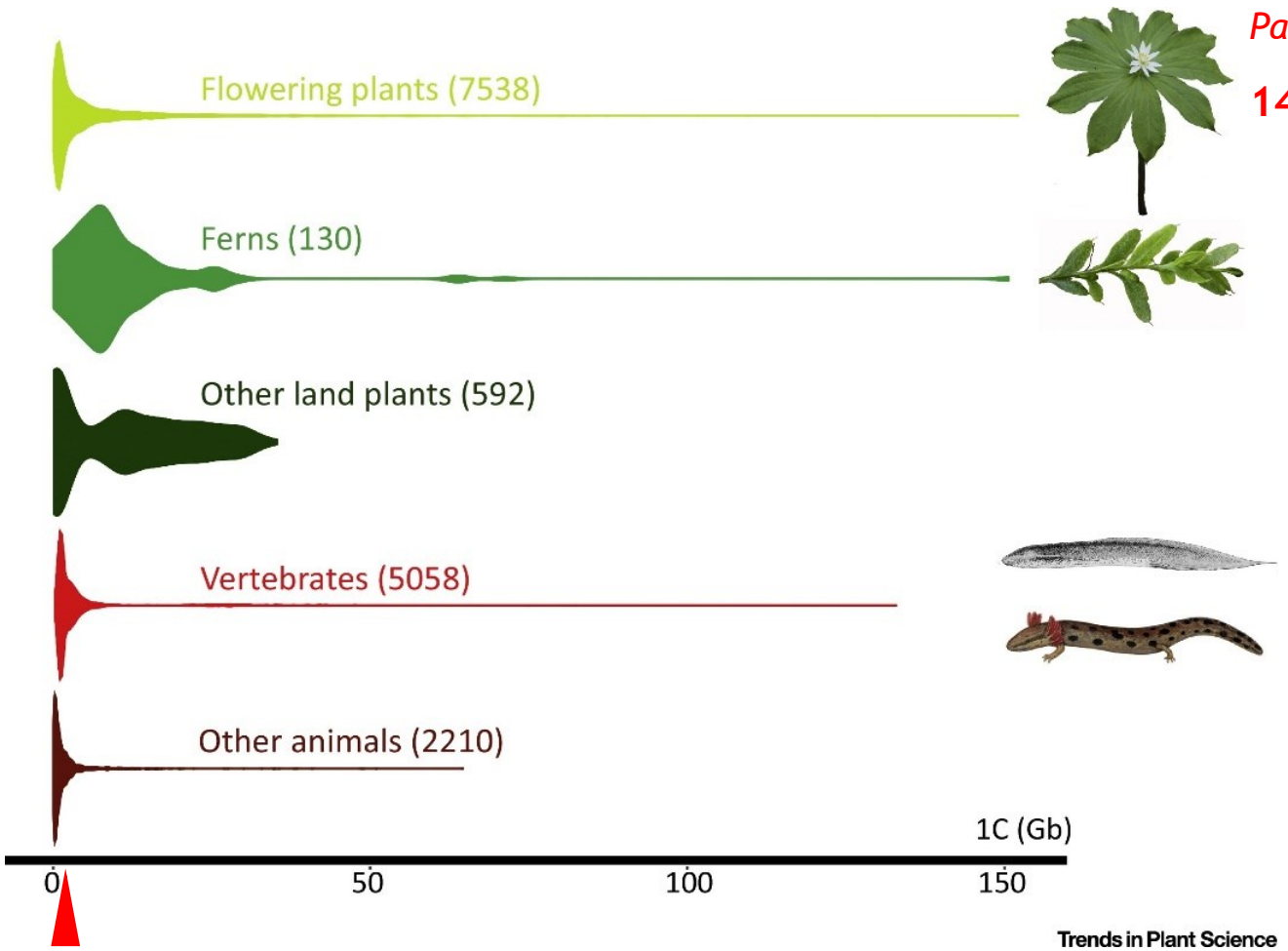


Extrachromosomal circular DNA (eccDNA)

- yeast, plants, animals
- size from a few hundred base pairs to hundreds of kilobases
- orgin from chromosomes
- can be „re-inserted“ into chromosomes

- glyphosate – synthetic herbicide patented by Monsanto in 1974
- known as Roundup
- Roundup Ready crops (GMOs)
- glyphosate: inhibition of a critical gene involved in amino acid synthesis, 5-ENOLPYRUVYLSHIKIMATE-3-PHOSPHATE SYNTHASE (EPSPS)
- emergence of glyphosate-resistant weed species (48), such as *Amaranthus palmeri*
- principle of the resistence: increase of EPSPS copy number due to the origin of a self-replicating eccDNA replicon (contains other genes, transposable elements)
- 399 435 bp in length, 59 genes
- the replicon contains elements controling its self-replication
- probably inherited through chromosome tethering (association with dividing chromosomes)
- the replicon can be used for crop improvement (engineering of synthetic replicons)

Eukaryotes: nuclear genome size variation (64 000-fold)



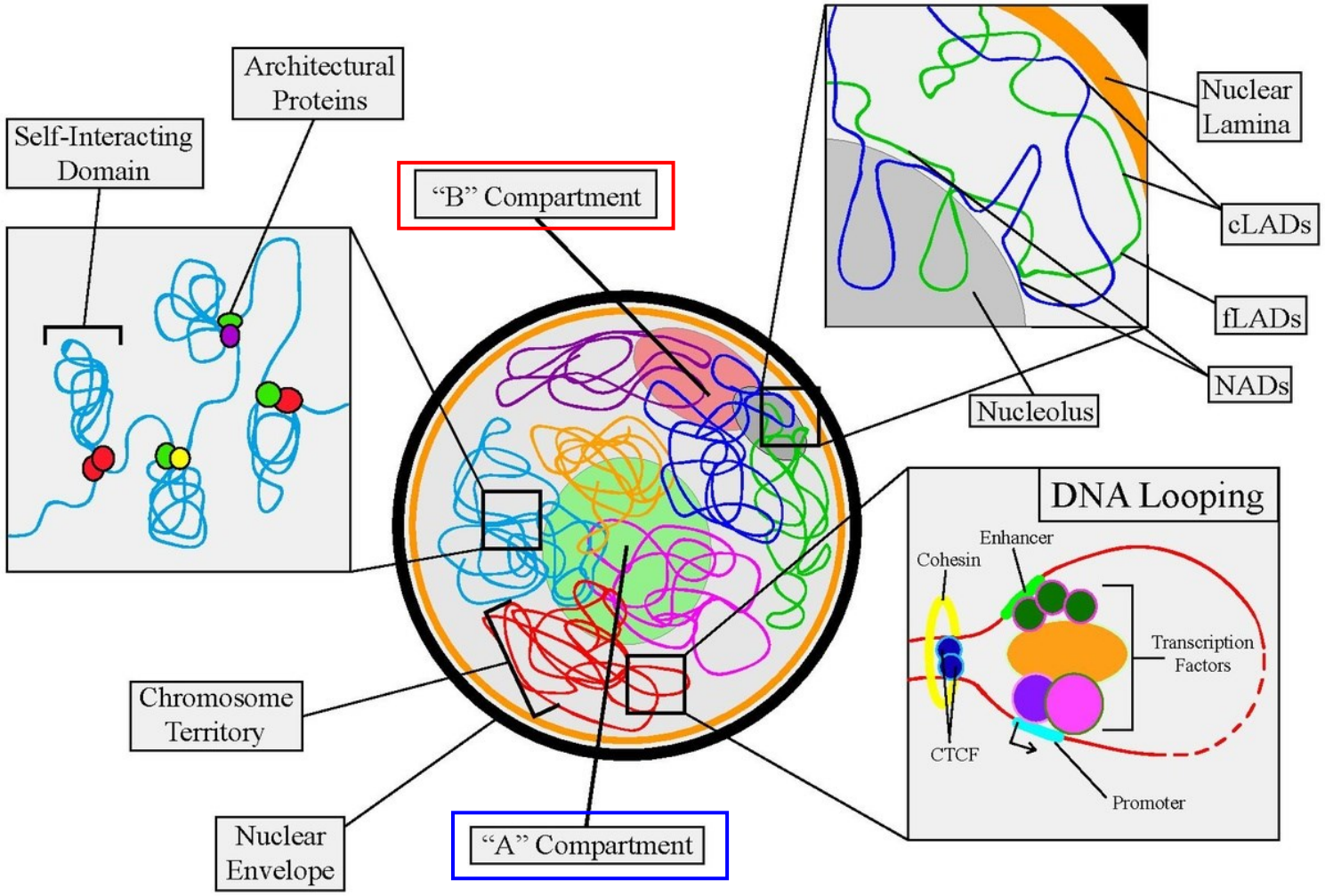
Paris japonica

149 Gb

Encephalitozoon intestinalis

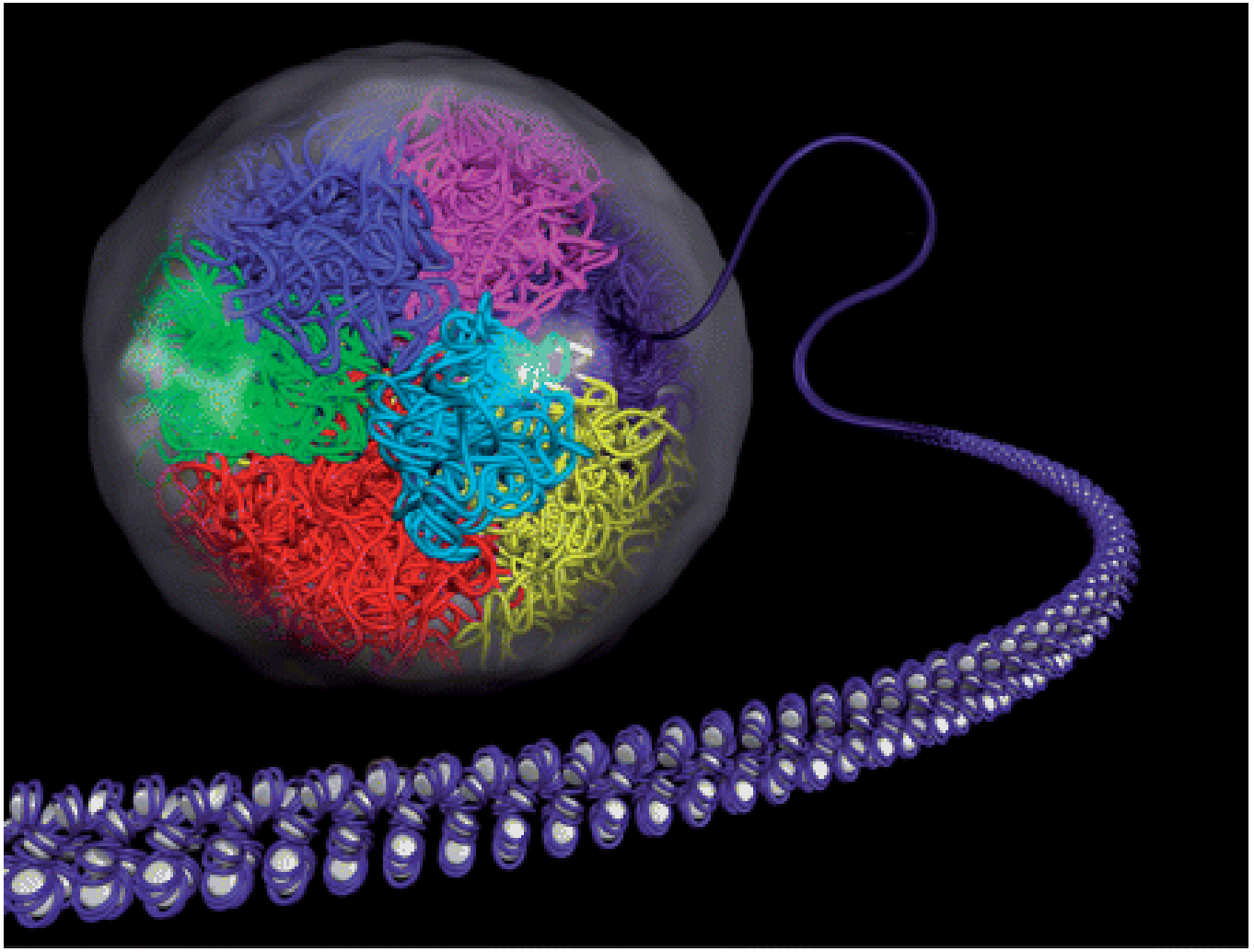
2.3 Mb (0.0023 Gb)

Eukaryotic nuclear genome - nuclear architecture



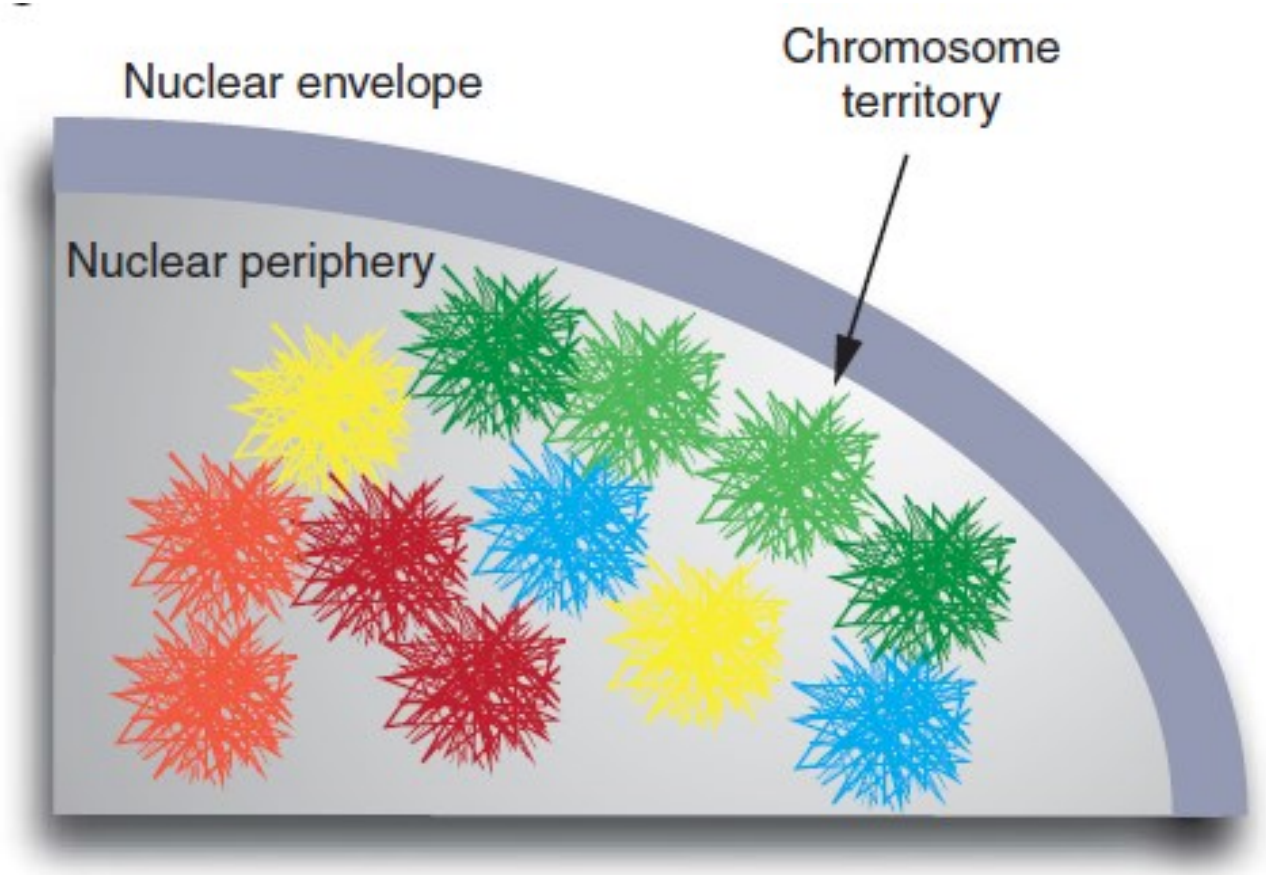
A/B compartment-associated regions are on the multi-Mb scale and correlate with either open and expression-active chromatin **(A compartments)** or closed and expression-inactive chromatin **(B compartments)**

Interphase chromosomes – chromosome territories

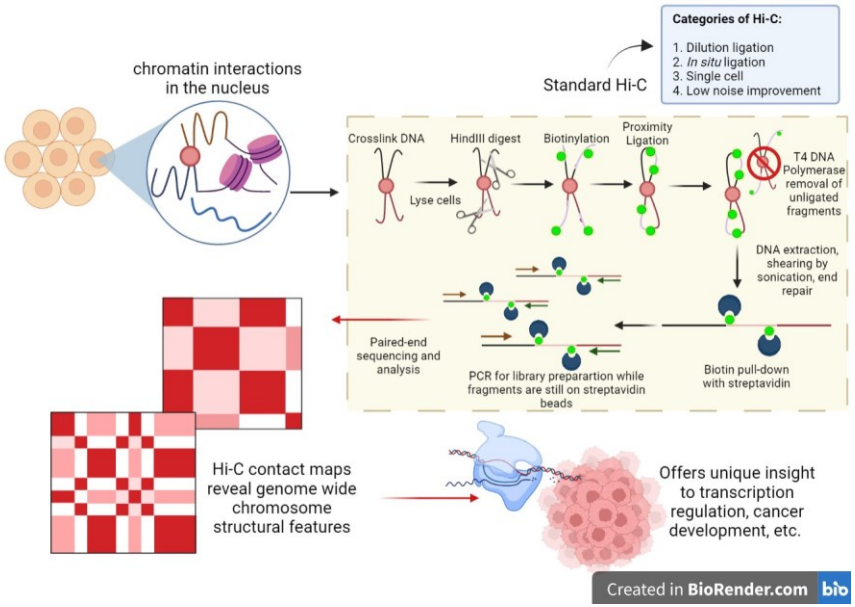


Chromosome territories

The distribution of chromosomes and genes is nonrandom with some chromosomes preferentially occupying internal positions and others occupying peripheral positions.

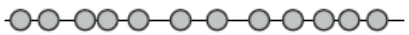


Hi-C maps: high-throughput Chromosome Conformation Capture

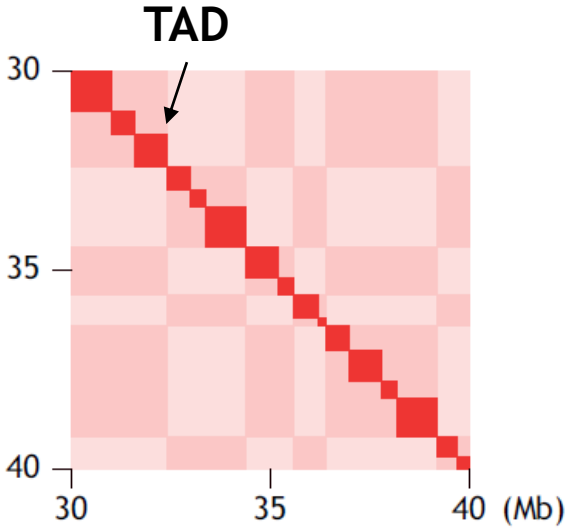


Eukaryotes
(human, mouse, etc.)

Chromosome topology and origin distribution



Hi-C heatmap

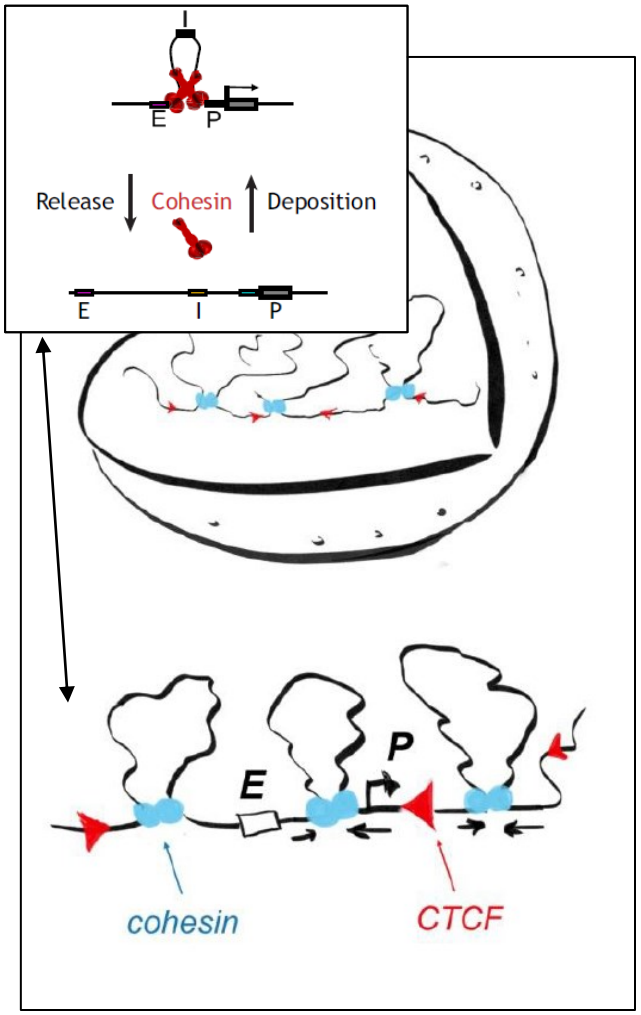


Structural features

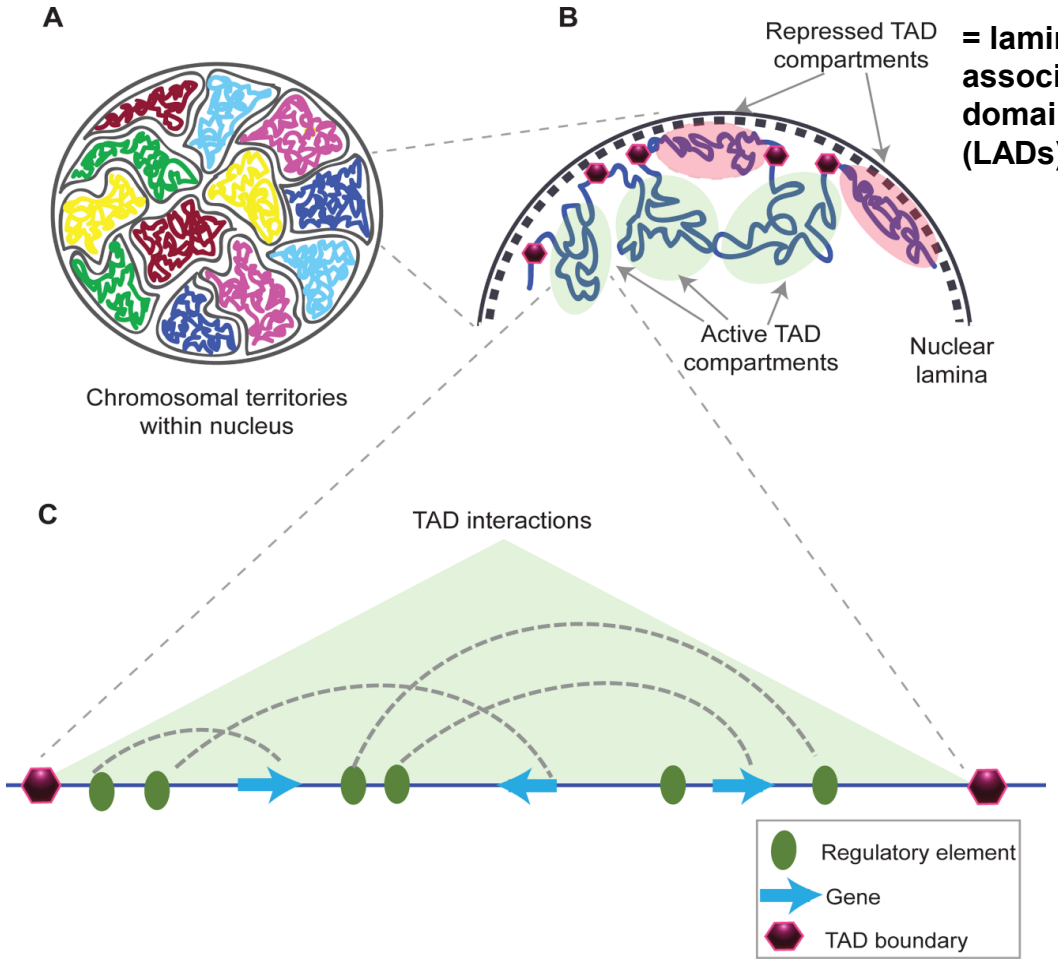
Compartmental domains (multi Mb) and TADs (~1 Mb)

Interphase territories consist of loops.

Usually a loop = topologically associated domain (TAD)



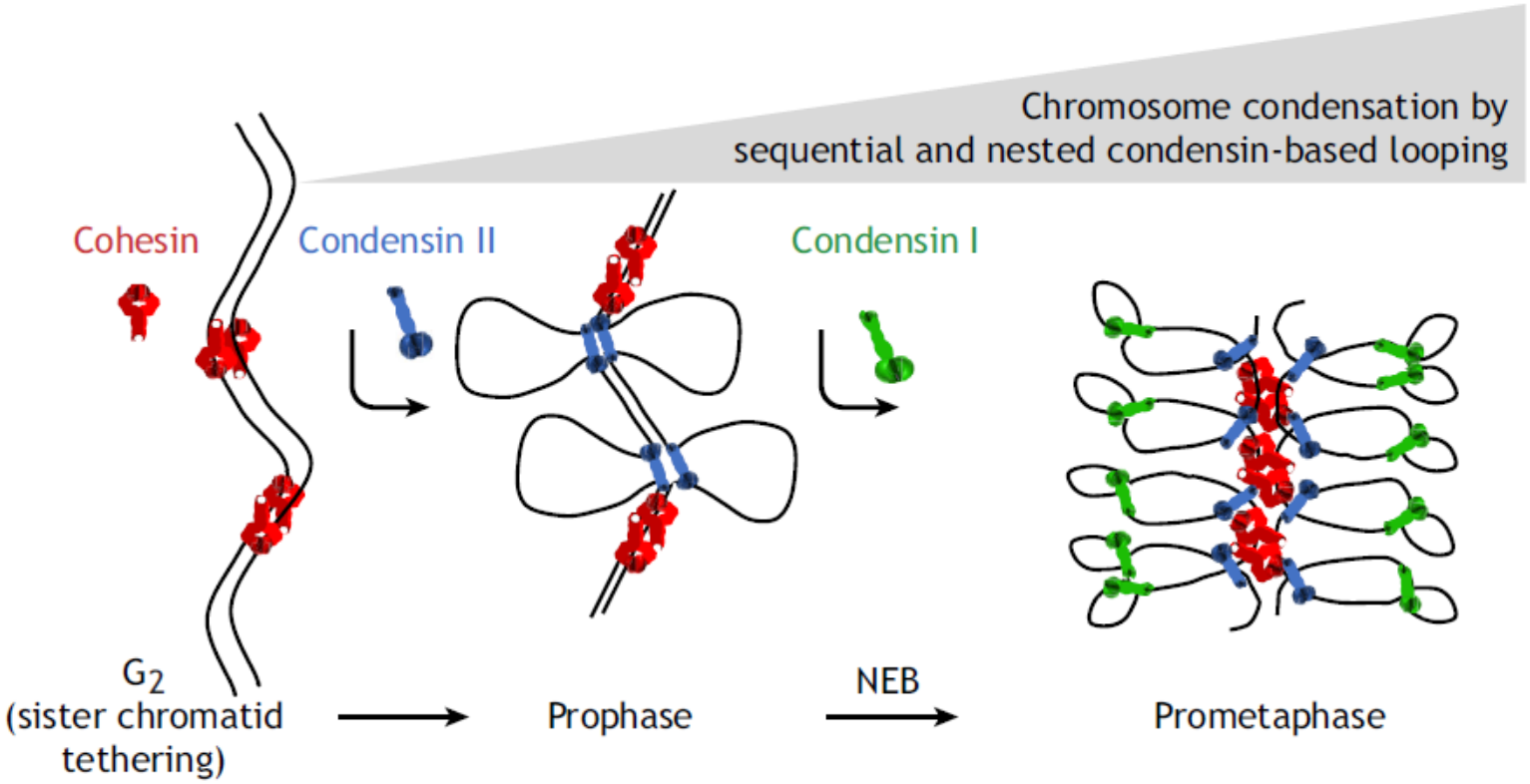
Promoter (***P***) of a gene can be pulled together with an enhancer (***E***) if they are not separated by a CTCF boundary



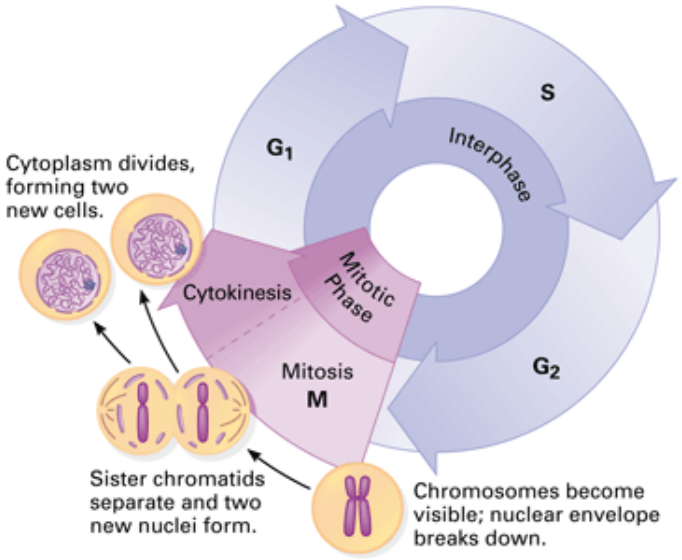
= mainly CTCF („insulator“) protein

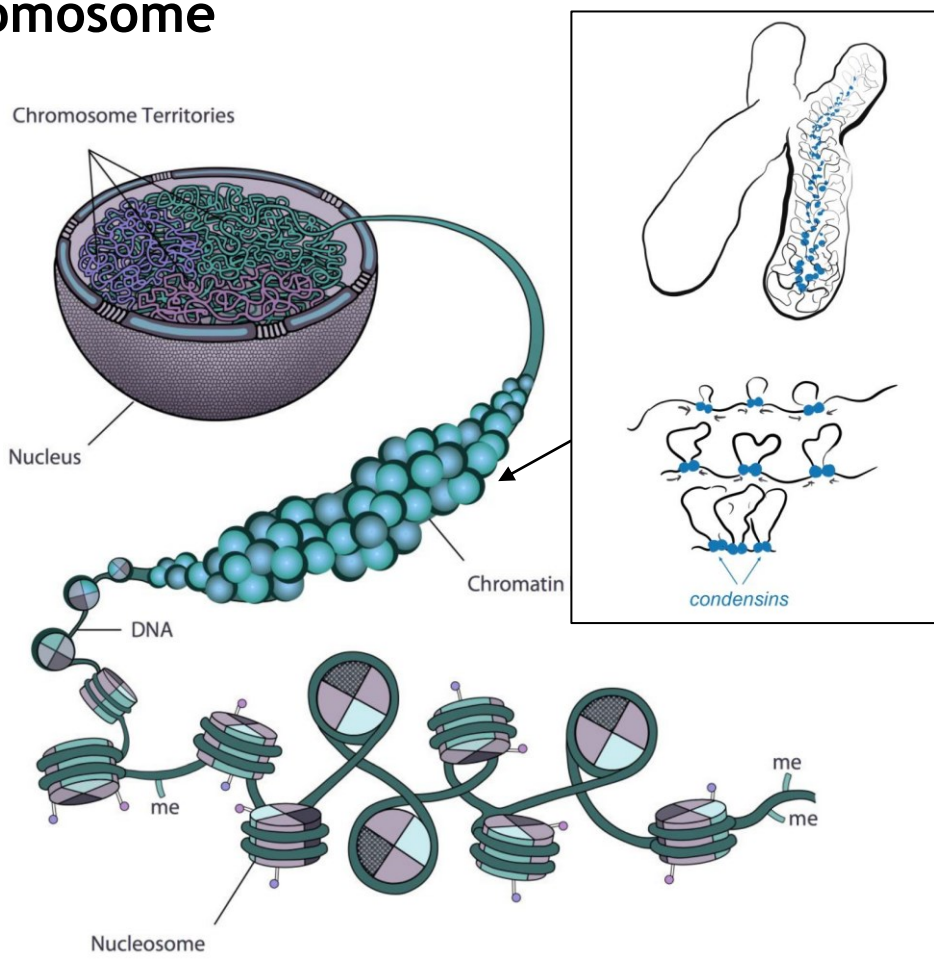
Cohesins and condensins

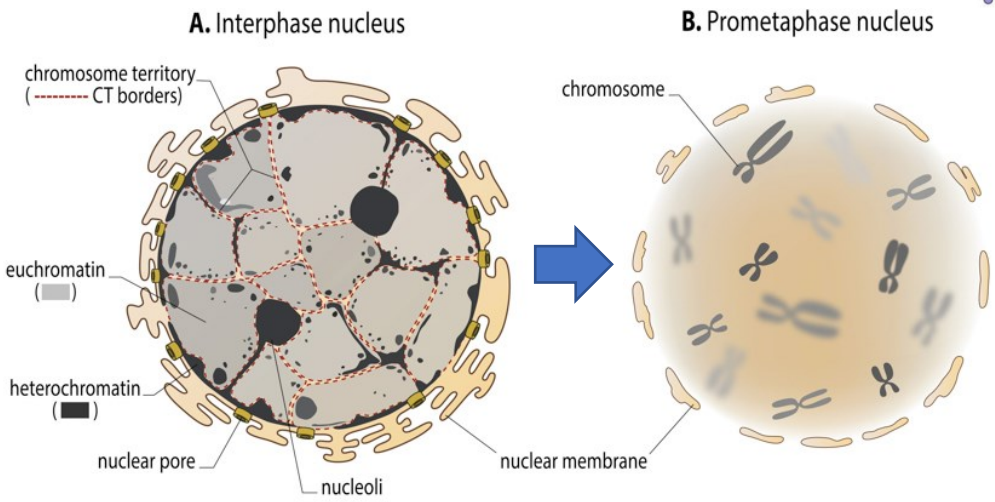
(SMC proteins – Structural Maintenance of Chromosomes proteins)



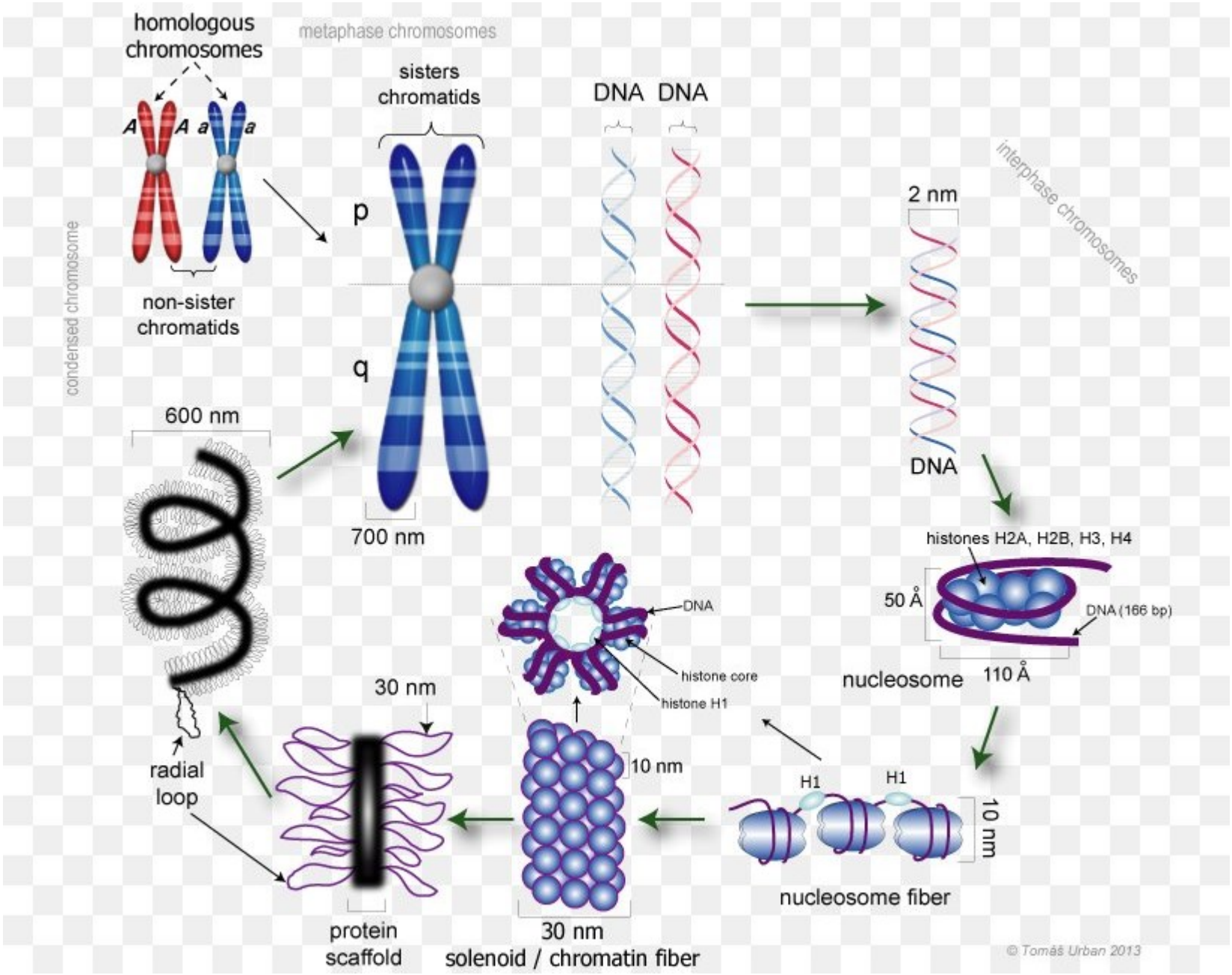
Eukaryotic chromosome







Eukaryotic chromosome



Genome structure: brief summary

- genomes are variable in size, gene number and proportion of non-coding DNA
- genome size is generally not correlated with organismal complexity (C-value paradox)
- viral genomes cannot replicate without a host, composed of either RNA or DNA
- prokaryotes are typically haploid, usually having a single circular chromosome (nucleoid); eukaryotes are diploid, DNA is organized into multiple linear chromosomes found in the nucleus
- protein-based supercoiling and packaging of DNA to fit inside a cell; eukaryotes and archaea use histone proteins, bacteria use different proteins with similar function
- prokaryotic and eukaryotic genomes both contain non-coding DNA (introns, repetitive DNA tandemly repeated or dispersed = transposable elements)
- prokaryotes: extrachromosomal DNA is maintained as plasmids
- eukaryotes: extrachromosomal DNA within organelles of prokaryotic origin (mitochondria and chloroplasts) - origin by endosymbiosis; plus eccDNA
- eukaryotic chromosomes: essential structures – centromere and telomeres