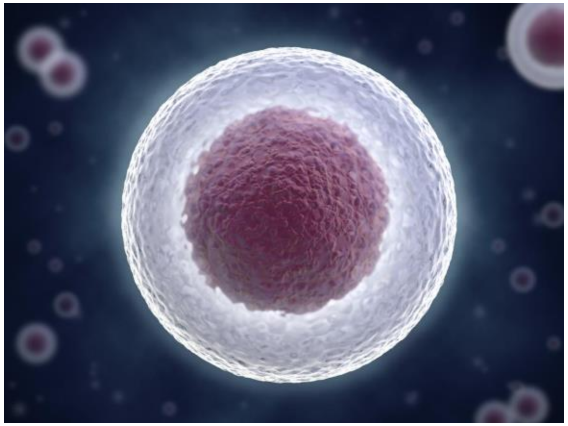
Genome and chromosome structure



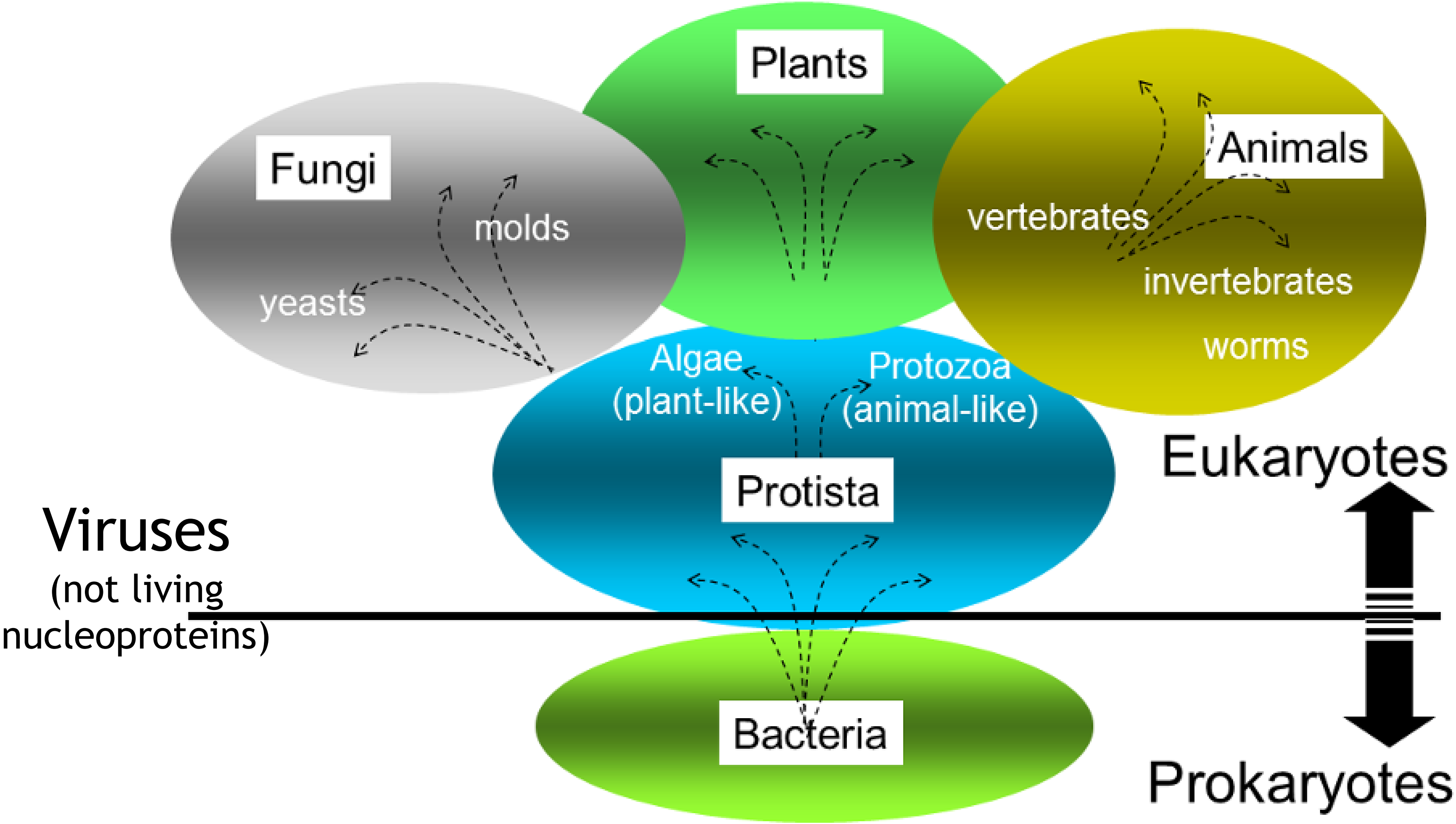
Martin A. Lysák

CEITEC, Masaryk University

Genome (Hans Winkler, 1920)

- 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

Living Things: 5 Kingdoms



Plants

Fungi

molds

yeasts

Animals

vertebrates

invertebrates

worms

Algae
(plant-like)

Protozoa
(animal-like)

Protista

Bacteria

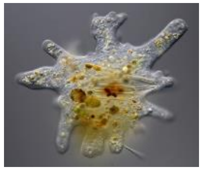
Viruses
(not living
nucleoproteins)

Eukaryotes

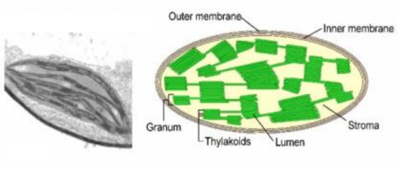
Prokaryotes

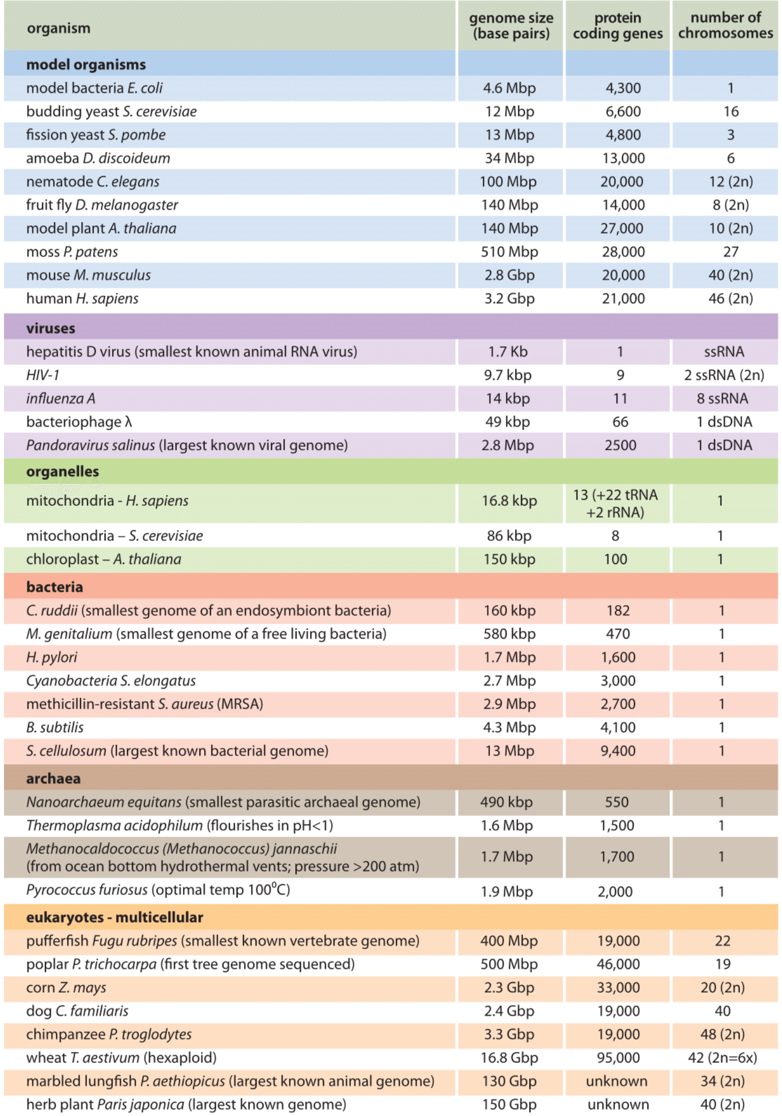
Genome size variation

Polychaos dubium



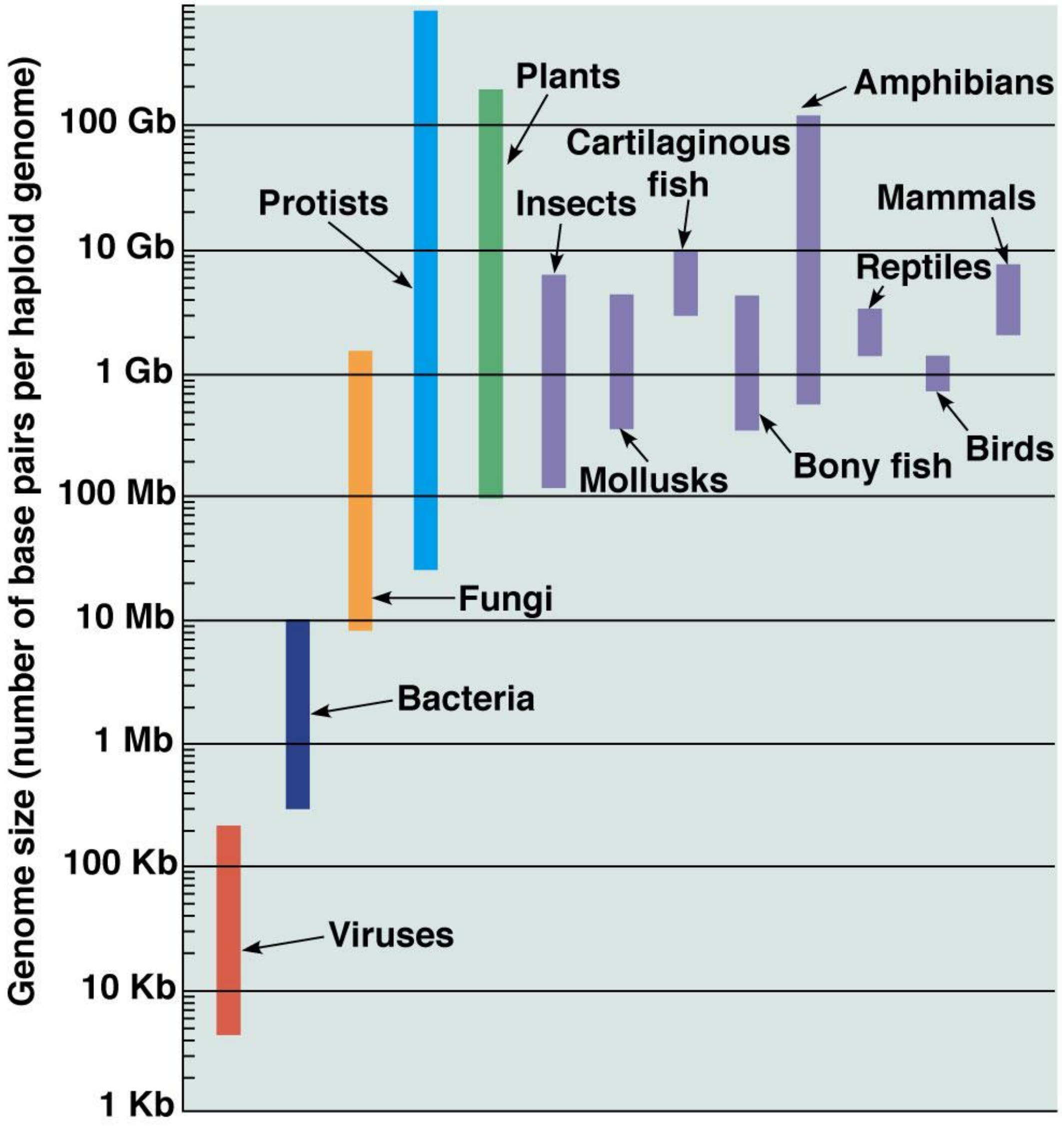
…perhaps the largest known genome – 670 billion base pairs (670 Gb)
(~200-times larger than the human genome, 3.2 Gb; some authors suggest treating the value with caution)



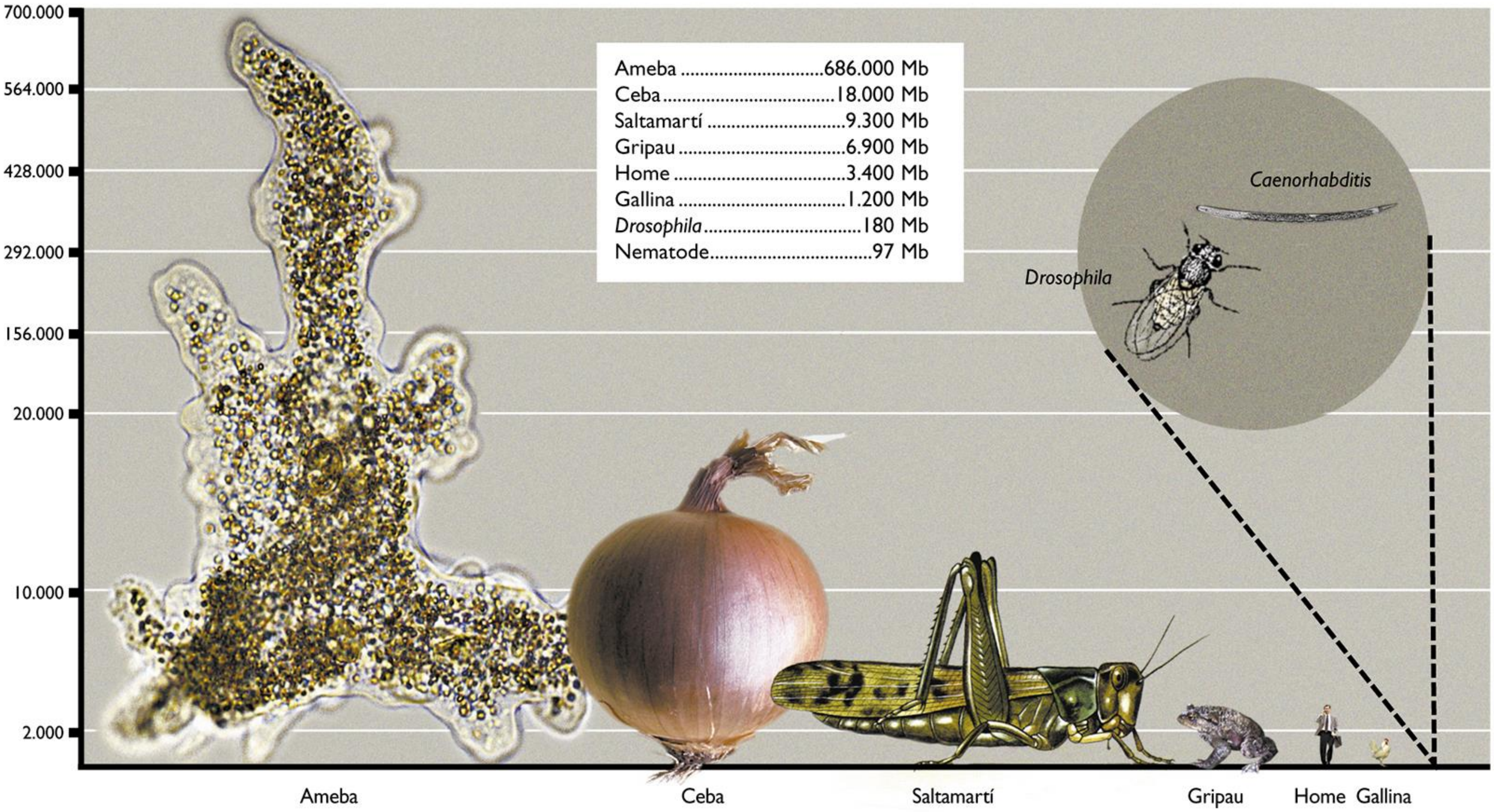


organism	genome size (base pairs)	protein coding genes	number of chromosomes
model organisms			
model bacteria *E. coli*	4.6 Mbp	4,300	1
budding yeast *S. cerevisiae*	12 Mbp	6,600	16
fission yeast *S. pombe*	13 Mbp	4,800	3
amoeba *D. discoideum*	34 Mbp	13,000	6
nematode *C. elegans*	100 Mbp	20,000	12 (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)

C-value paradox (CA Thomas, 1971)



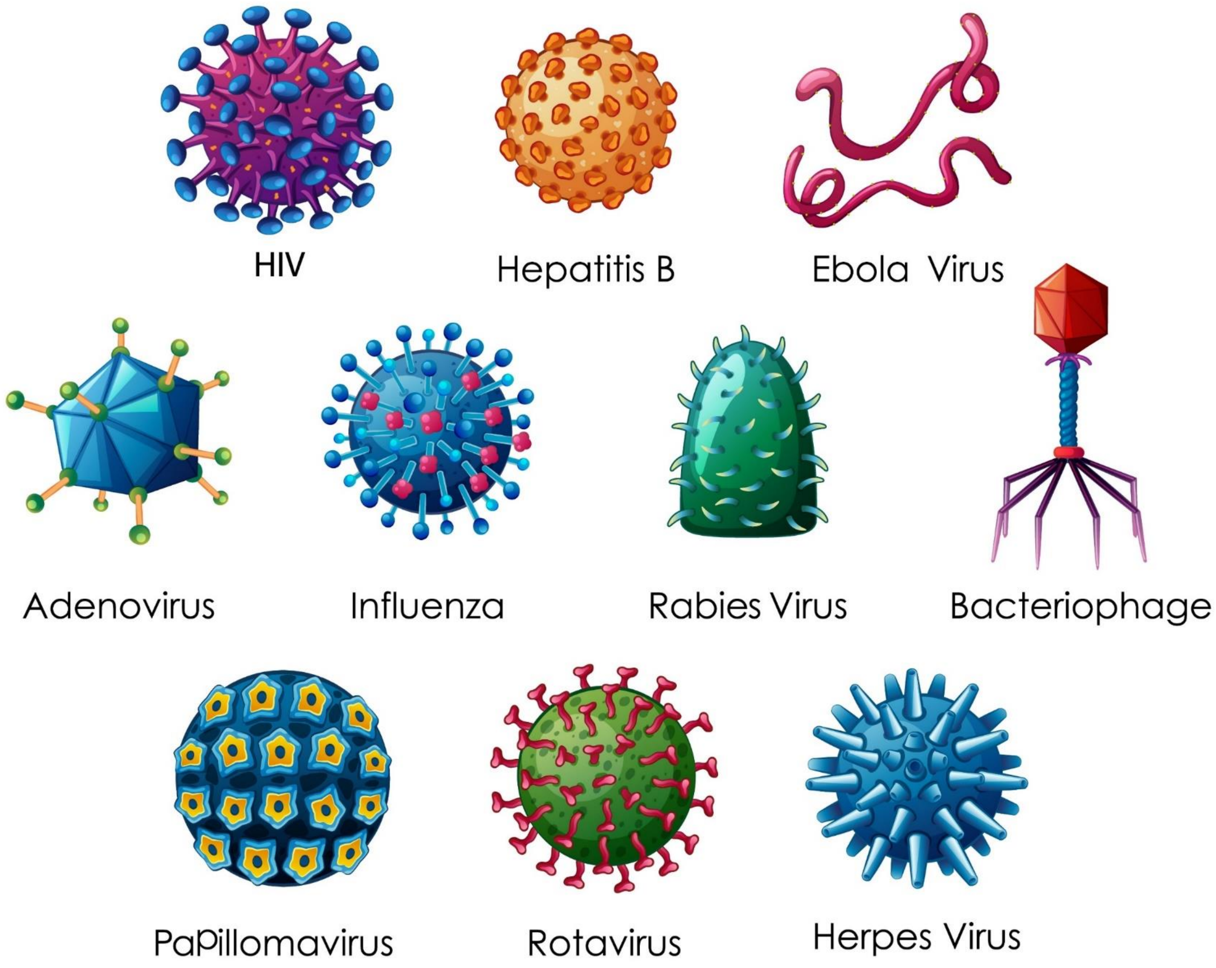
C-value paradox



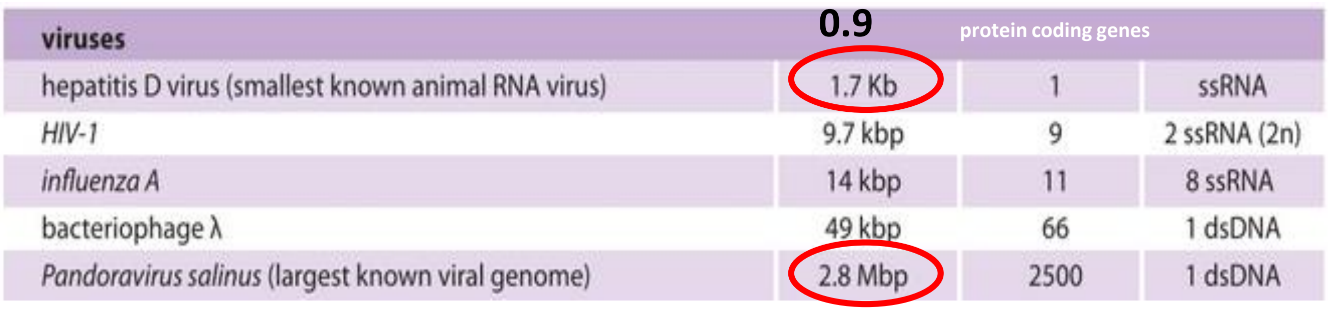
Ameba	686.000 Mb
Ceba	18.000 Mb
Saltamartí	9.300 Mb
Gripau	6.900 Mb
Home	3.400 Mb
Gallina	1.200 Mb
Drosophila	180 Mb
Nematode	97 Mb

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

Viruses

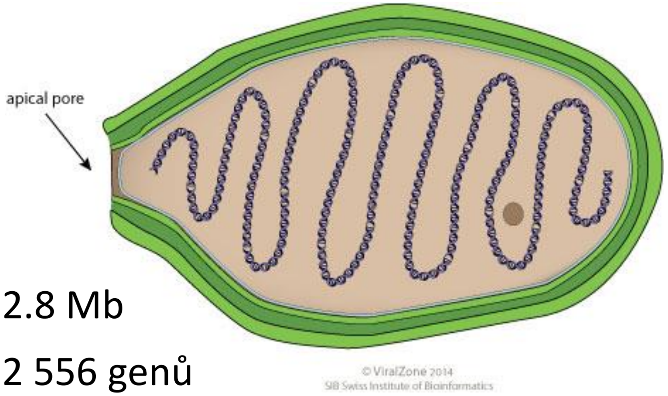

HIV
Hepatitis B
Ebola Virus
Adenovirus
Influenza
Rabies Virus
Bacteriophage
PaPillomavirus
Rotavirus
Herpes Virus

Viruses - physical and genome size

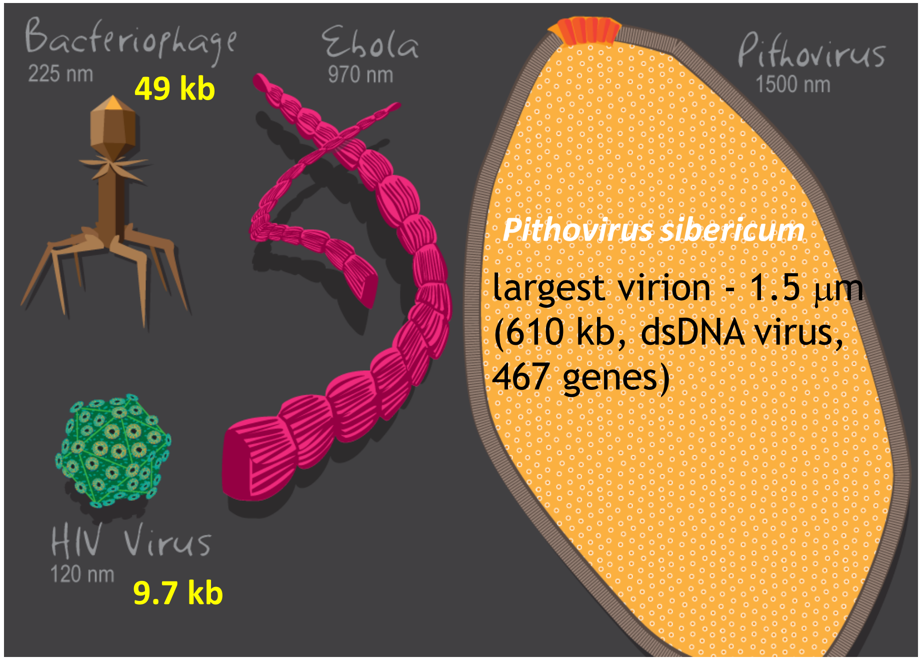


viruses	0.9	protein coding genes	
hepatitis D virus (smallest known animal RNA virus)	1.7 Kb	1	ssRNA
HIV-1	9.7 kbp	9	2 ssRNA (2n)
influenza A	14 kbp	11	8 ssRNA
bacteriophage λ	49 kbp	66	1 dsDNA
Pandoravirus salinus (largest known viral genome)	2.8 Mbp	2500	1 dsDNA

Pandoravirus salinus (dsDNA)



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



Pithovirus sibericum

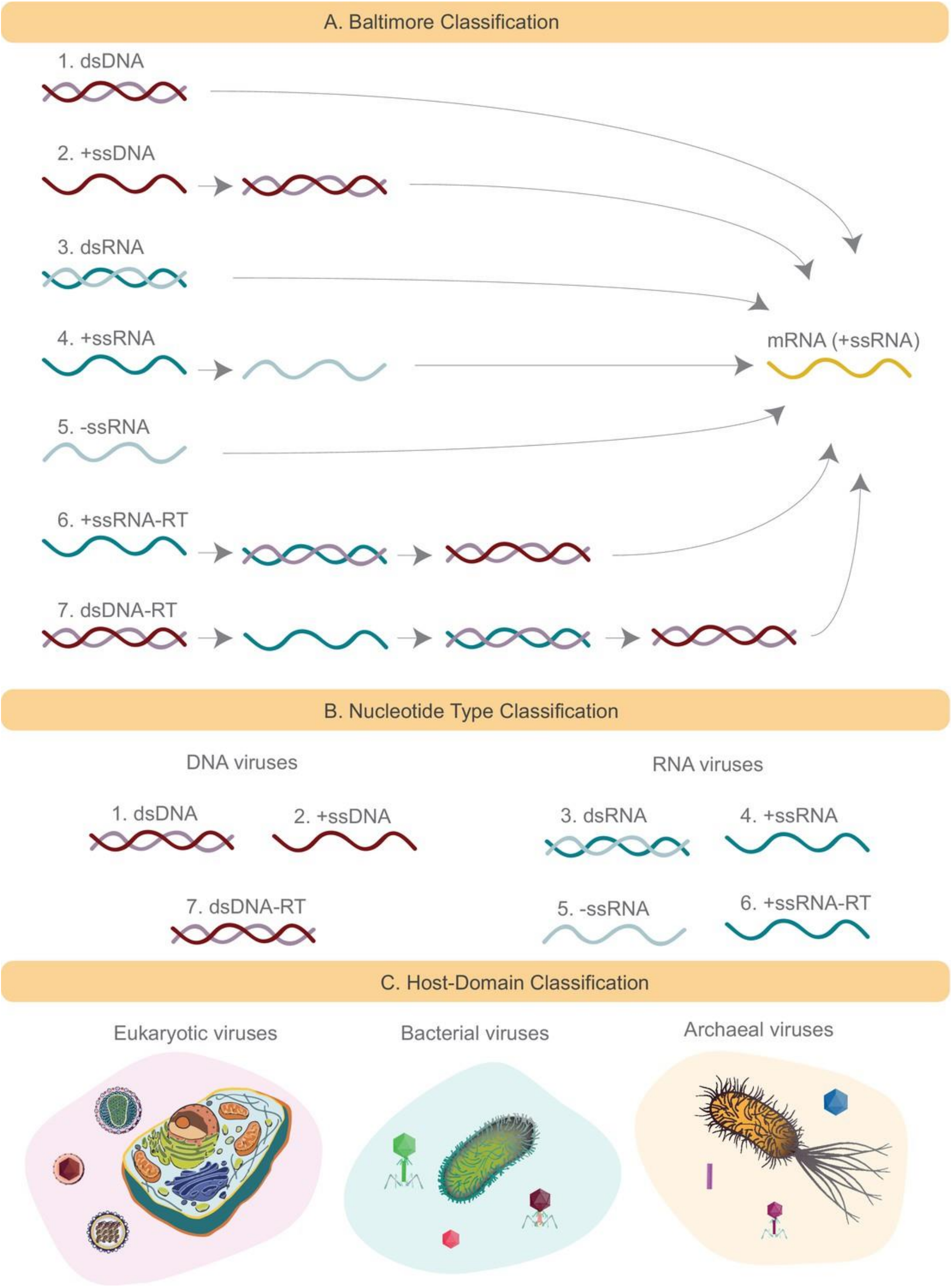
largest virion - 1.5 μm (610 kb, dsDNA virus, 467 genes)

Endogenous viral elements (EVEs)

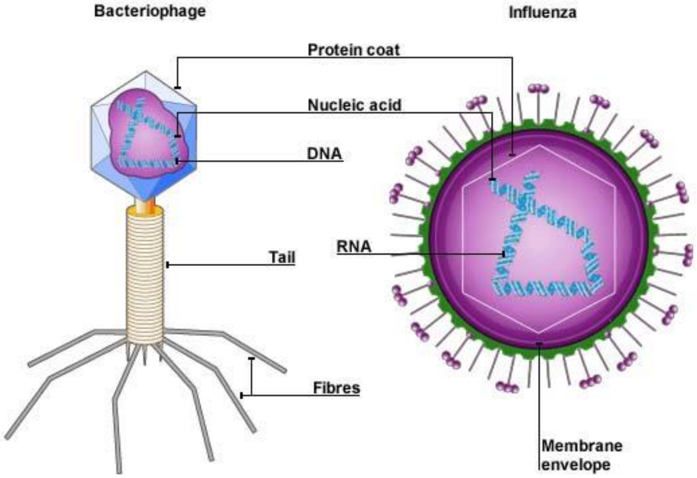
Viruses which integrated their genomes into genomes of their eukaryotic hosts.

- usually small DNA fragments (few genes)
- algae (chlorophytes): large dsDNA viruses can integrate in the host genome
- 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)

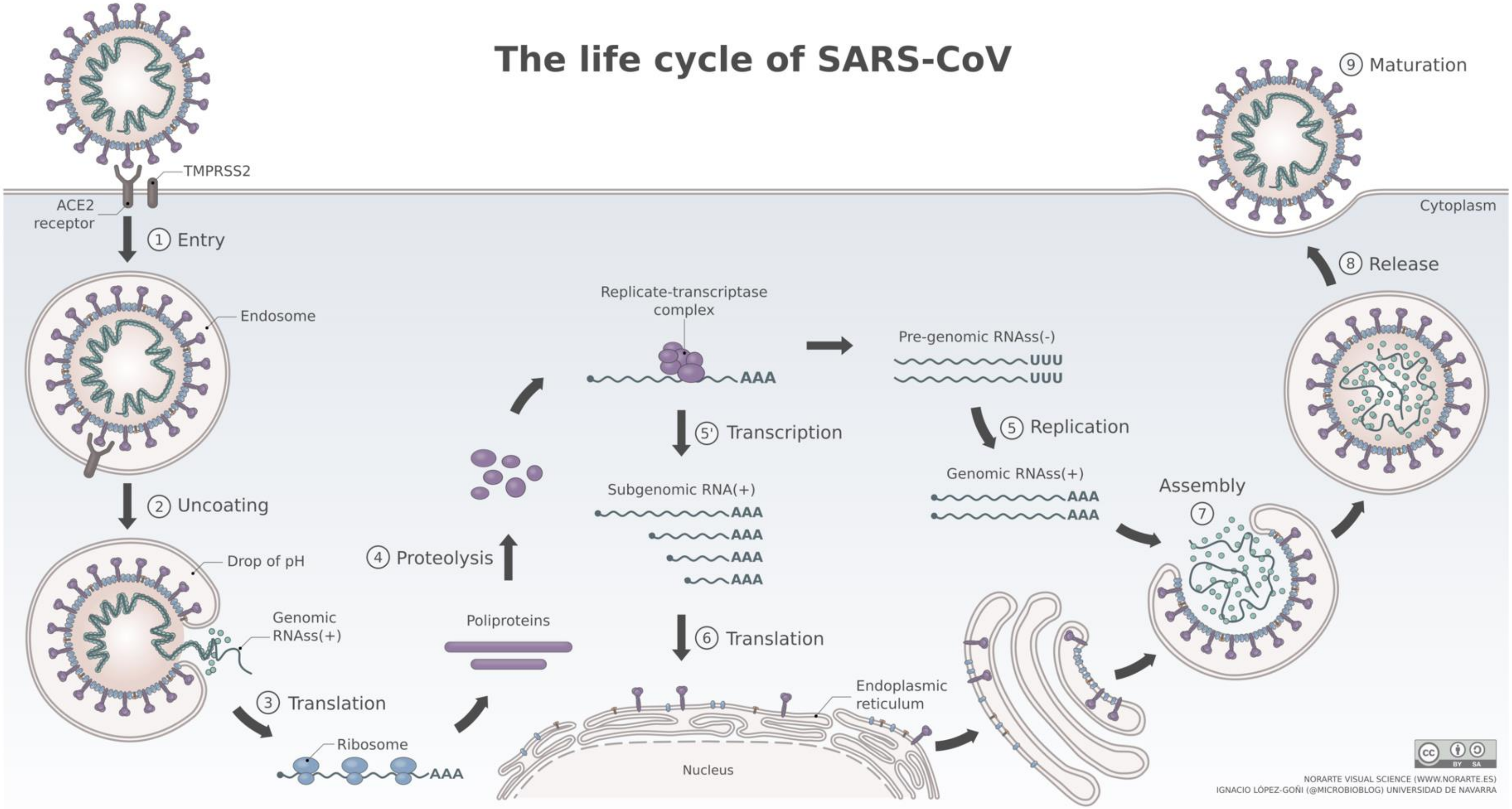
Viruses



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

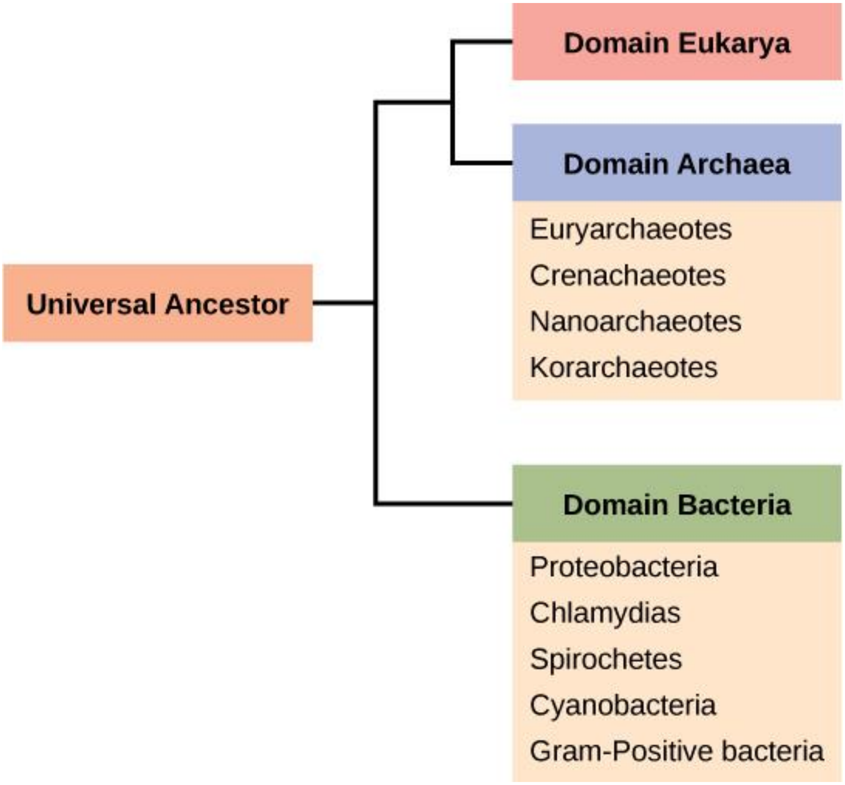


Viral life cycle (coronaviruses)


The life cycle of SARS-CoV

NORARTE VISUAL SCIENCE (WWW.NORARTE.ES)
IGNACIO LÓPEZ-GOÑI (@MICROBIOBLOG) UNIVERSIDAD DE NAVARRA

Prokaryotes and eukaryotes: three domains



Universal Ancestor

Domain Eukarya

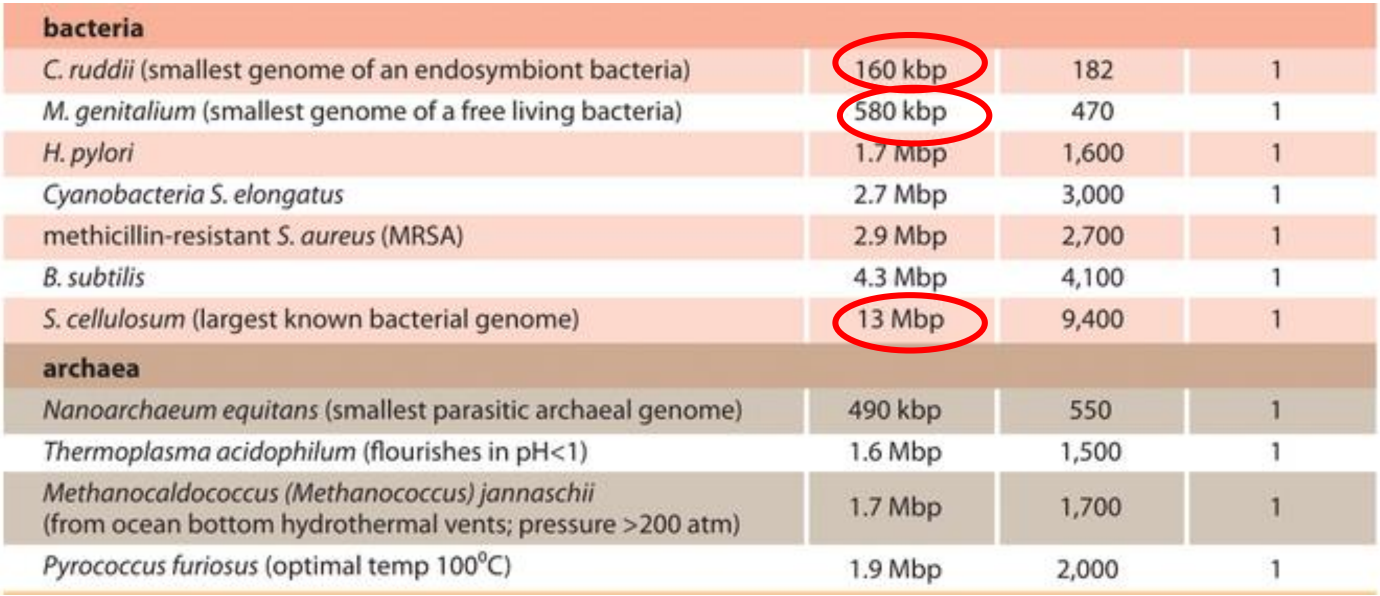
Domain Archaea

Euryarchaeotes
Crenachaeotes
Nanoarchaeotes
Korarchaeotes

Domain Bacteria

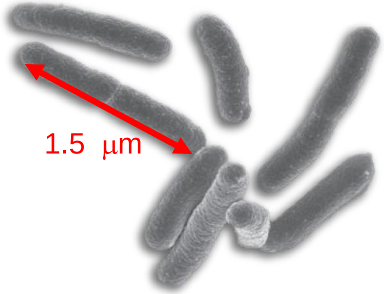
Proteobacteria
Chlamydias
Spirochetes
Cyanobacteria
Gram-Positive bacteria

Genomes of Archaea and Bacteria



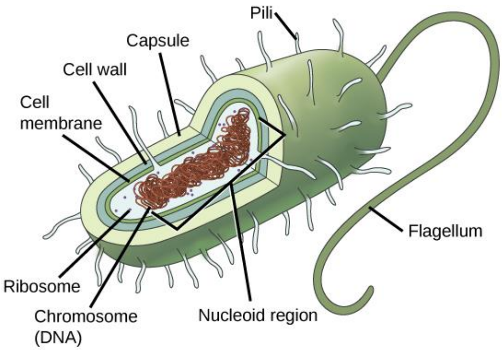
bacteria			
C. ruddii (smallest genome of an endosymbiont bacteria)	160 kbp	182	1
M. genitalium (smallest genome of a free living bacteria)	580 kbp	470	1
H. pylori	1.7 Mbp	1,600	1
Cyanobacteria S. elongatus	2.7 Mbp	3,000	1
methicillin-resistant *S. aureus* (MRSA)	2.9 Mbp	2,700	1
B. subtilis	4.3 Mbp	4,100	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
- 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)

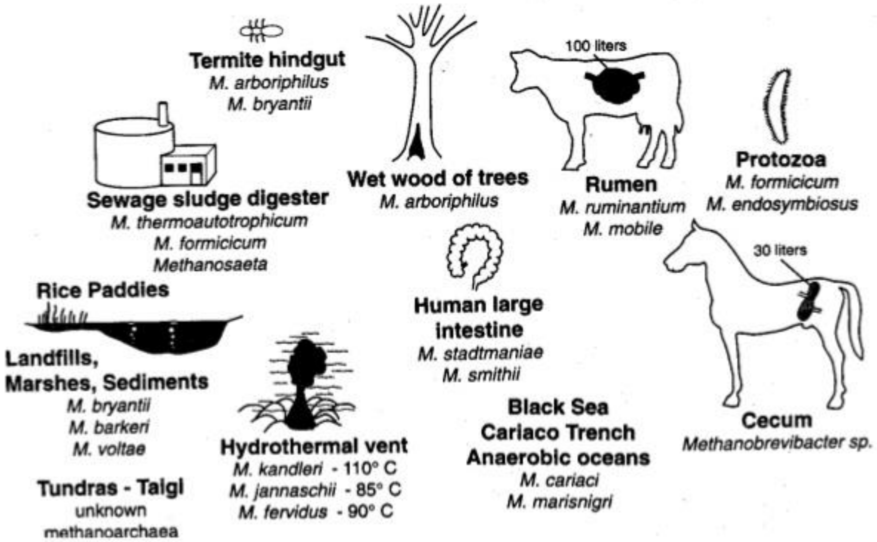
Sorangium cellulosum – **the largest known bacterial genome** (13 Mb, 9 400 genes)

Genomes of Archaea (formerly Archaebacteria)

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

the archaea *Methanosphaera stadtmanae*

Methanogen Habitats



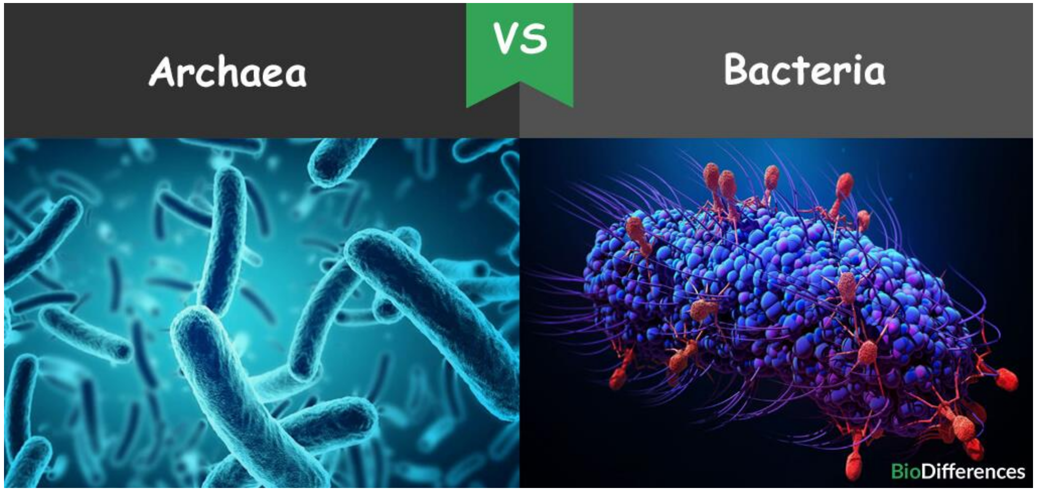
Archaea

- usually a single circular chromosome, plasmids can be found
- 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)
- 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; DNA polymerase similar to eukaryotic DNA polymerases

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.

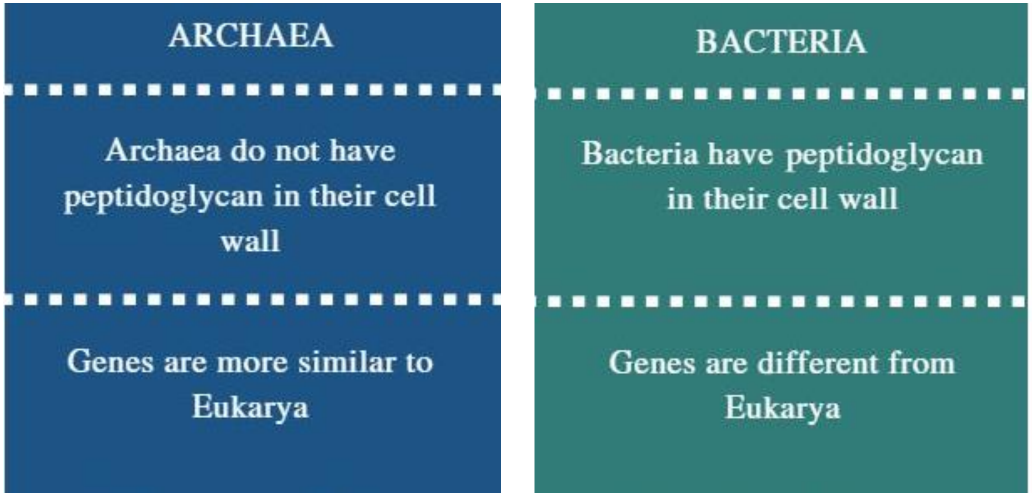


Bacteria might be pathogenic or non-pathogenic.

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

Cell wall is made up of pseudopeptidoglycan and lack D-aminoacids and N-acetylmuramic acid.

Introns are present in the chromosomes of archaea.

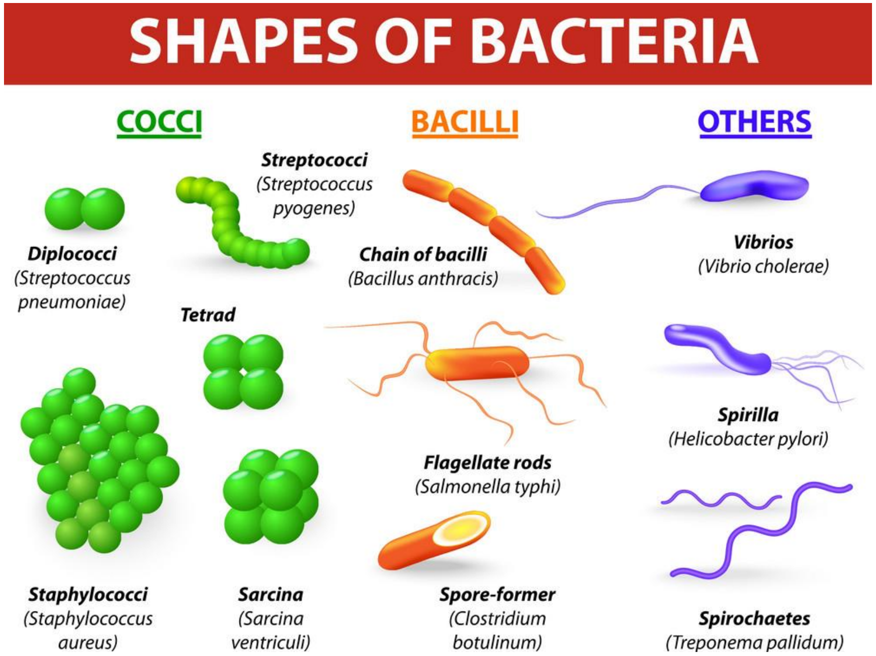


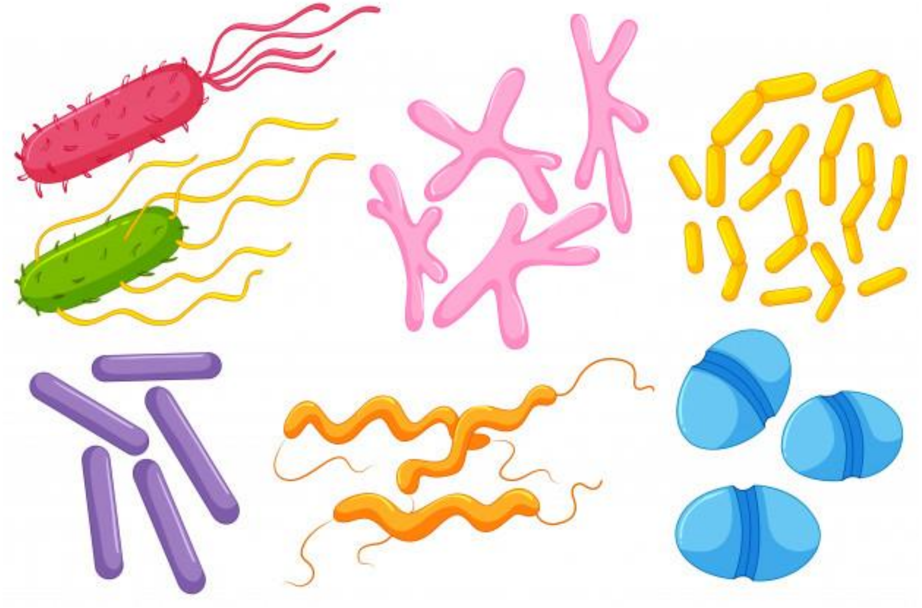
ARCHAEA	BACTERIA
Archaea do not have peptidoglycan in their cell wall	Bacteria have peptidoglycan in their cell wall
Genes are more similar to Eukarya	Genes are different from Eukarya

Cell wall is made up of peptidoglycan consisting of N-acetylmuramic acid and D-amino acids.

Introns are absent in the chromosomes of bacteria.

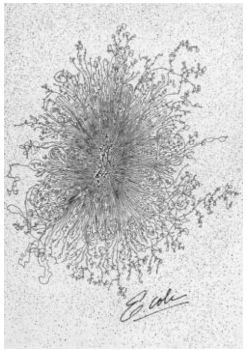
Bacterial genomes

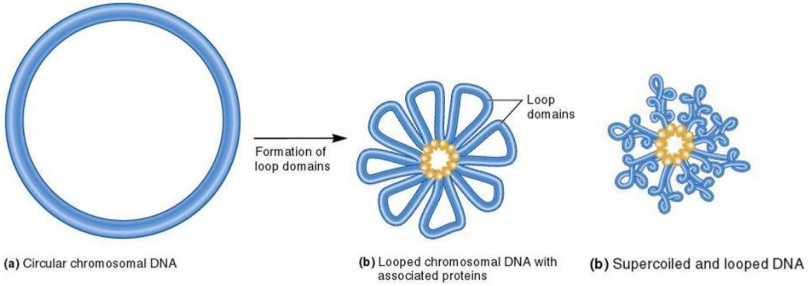




Bacterial chromosome

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





(a) Circular chromosomal DNA

(b) Looped chromosomal DNA with associated proteins

(b) Supercoiled and looped DNA

! 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)

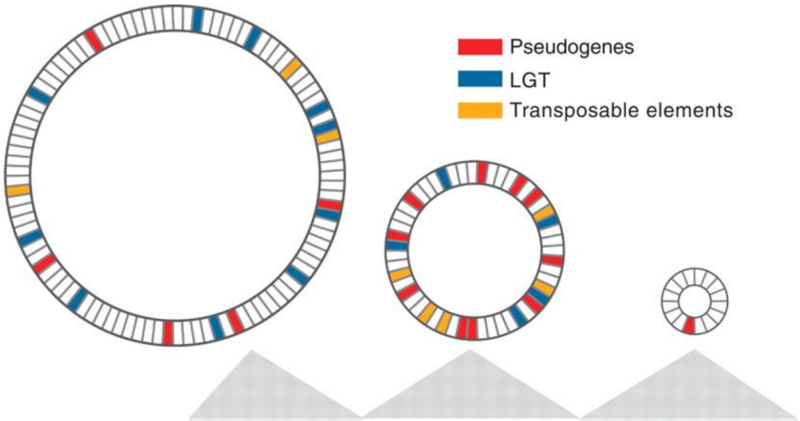
Problematic ends of linear chromosomes:

- palindromic hairpin loops
- invertron telomeres – a protein binds to the 5’-ends

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

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.

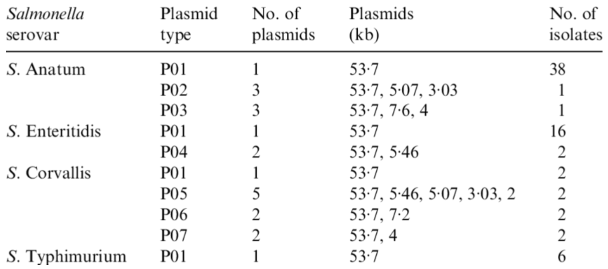
In **long-term host-dependent species**, the ongoing mutational bias toward deletions has removed all superfluous sequences, resulting in a highly reduced genome containing few, if any, pseudogenes or transposable elements.

LGT, lateral gene transfer.

Ochman and Davalos, Science 2006

What is the role of plasmids?

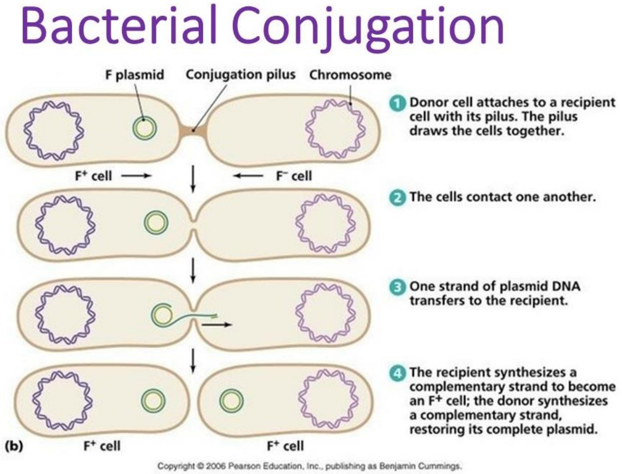
- usually small (1-200 kb), circular DNAs; independent replication



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

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

- Genes providing protection from toxic substances (including antibiotic resistence)
- 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



① Donor cell attaches to a recipient cell with its pilus. The pilus draws the cells together.

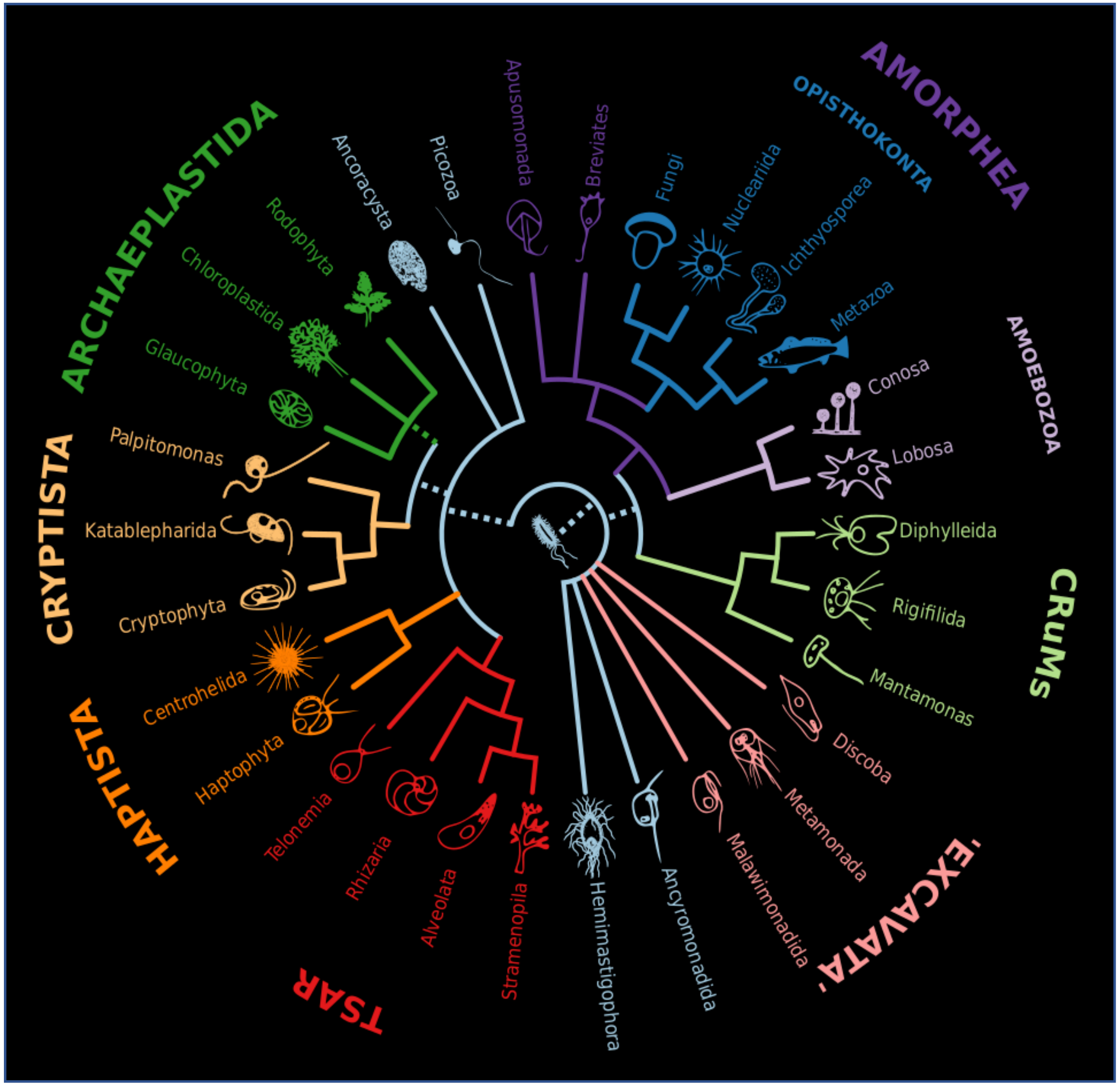
② The cells contact one another.

③ One strand of plasmid DNA transfers to the recipient.

④ The recipient synthesizes a complementary strand to become an F⁺ cell; the donor synthesizes a complementary strand, restoring its complete plasmid.

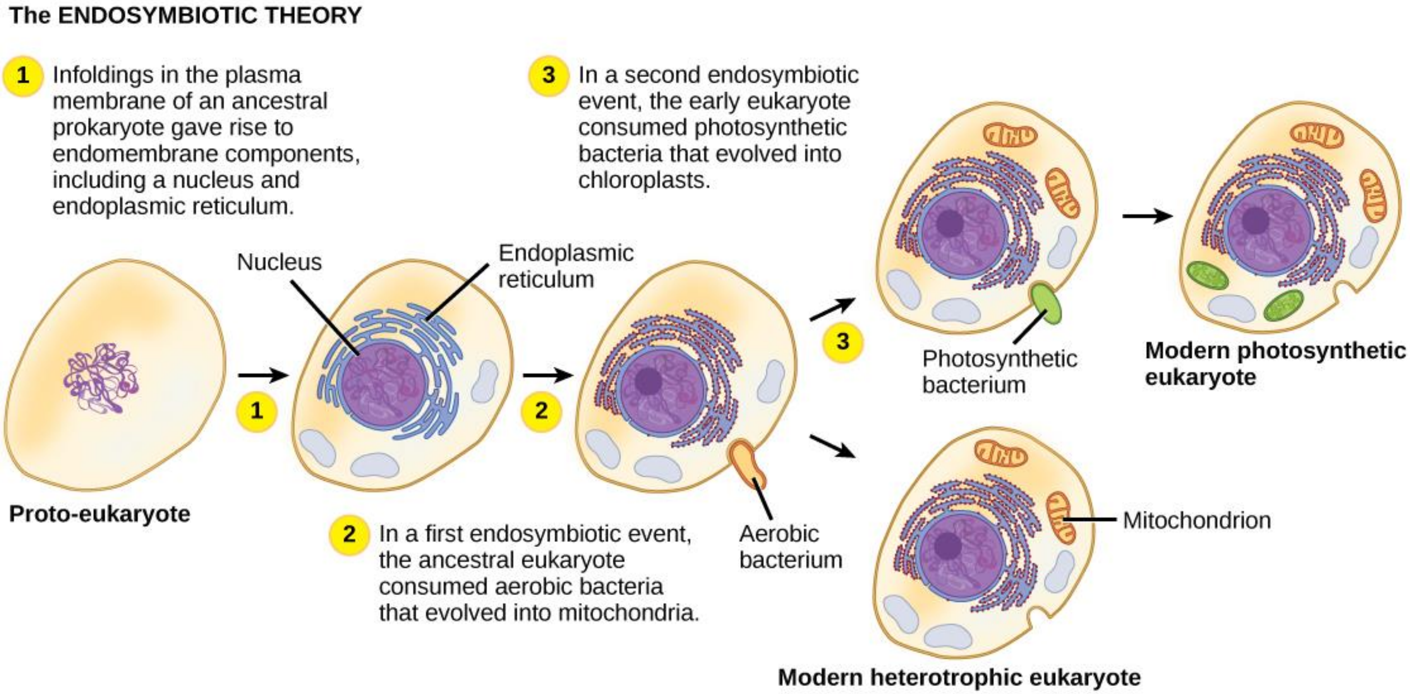
Copyright © 2006 Pearson Education, Inc., publishing as Benjamin Cummings.

The new tree of eukaryotes and eukaryotic genomes



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



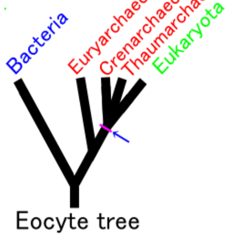
The ENDOSYMBIOTIC THEORY

1 Infoldings in the plasma membrane of an ancestral prokaryote gave rise to endomembrane components, including a nucleus and endoplasmic reticulum.

2 In a first endosymbiotic event, the ancestral eukaryote consumed aerobic bacteria that evolved into mitochondria.

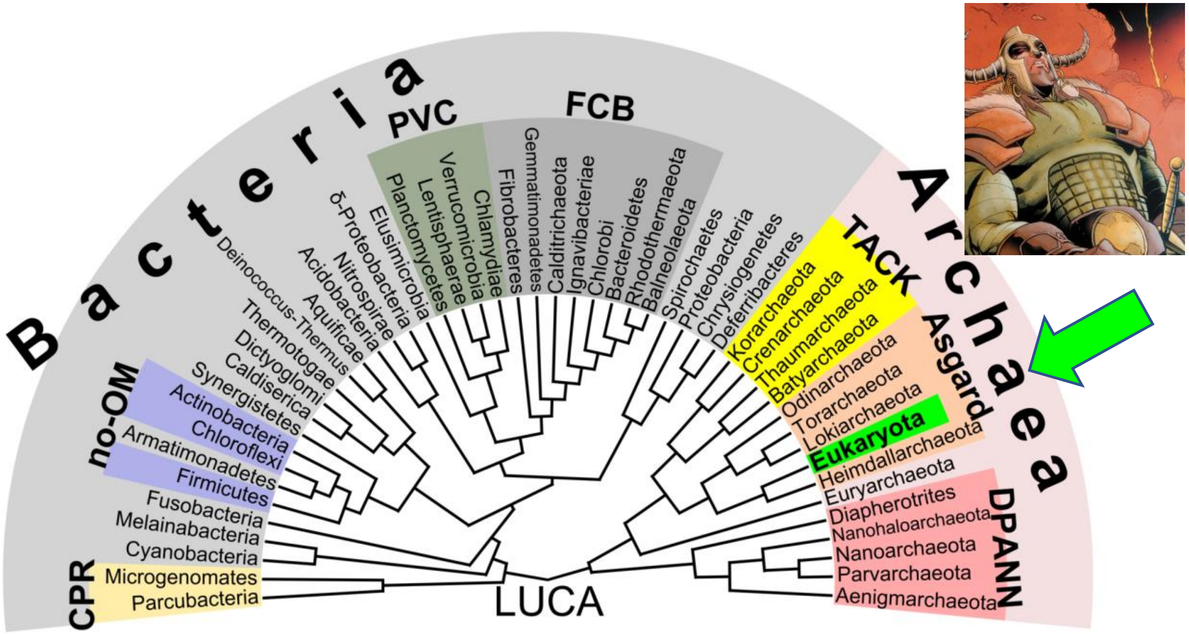
3 In a second endosymbiotic event, the early eukaryote consumed photosynthetic bacteria that evolved into chloroplasts.

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 originat within **Heimdallarchaeota**
- **in cladistic view, eukaryotes are Archaea, similarly as birds are dinosaurs**

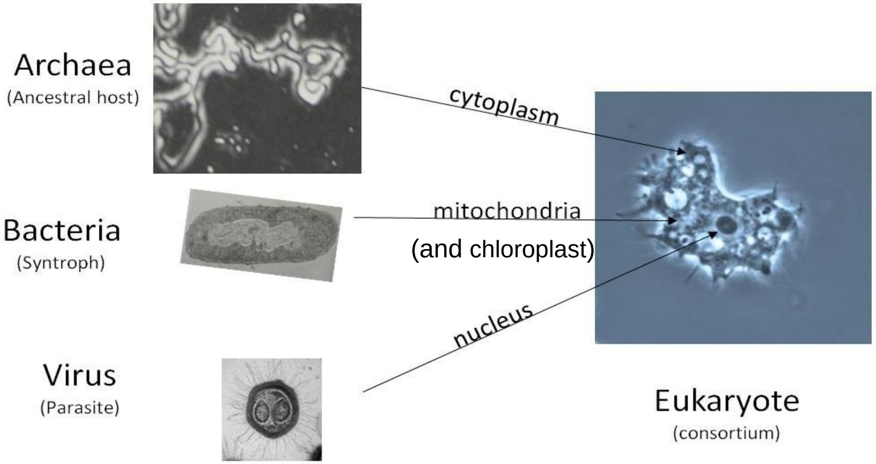


Last **u**niversal **c**ommon **a**ncestor

Origin of Eukaryotes – viral eukaryogenesis

(Philip Bell, 2001)

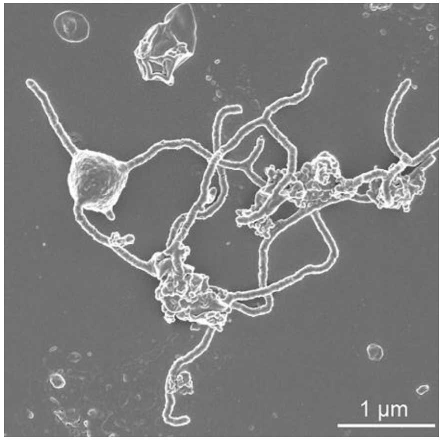
- A hypothesis that the eukaryotic nucleus could originated from a virus
- The virus (= the nucleus) probably acquired some genes from the archaeal host genome and bacterial genome(s)
- This virus(es) could be similar to large, complex DNA viruses (such as Mimivirus) that are capable of protein biosynthesis
- 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



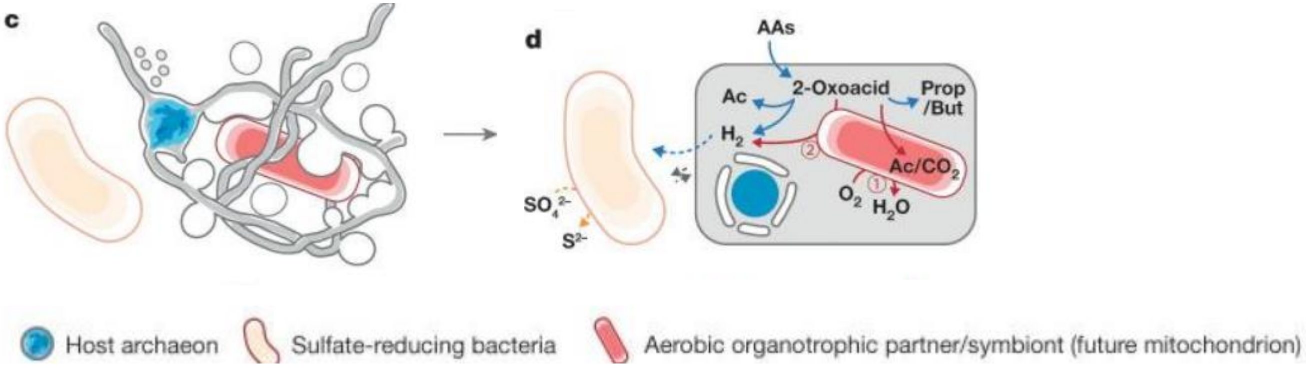
Living potential link between Archaea and eukaryotes

- *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:**

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:

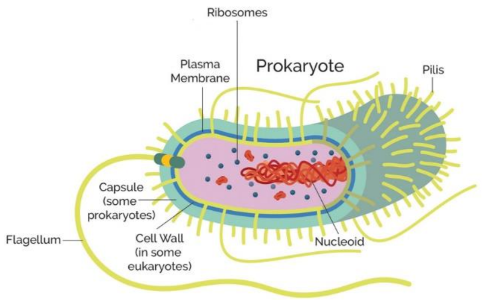


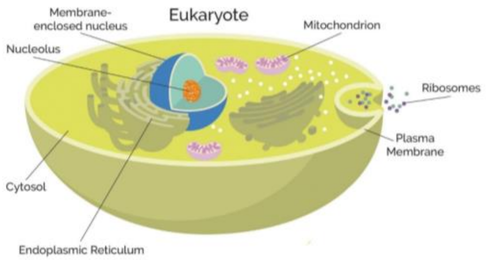
Prometheoarchaeum syntrophicum



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

Prokaryotes vs Eukaryotes

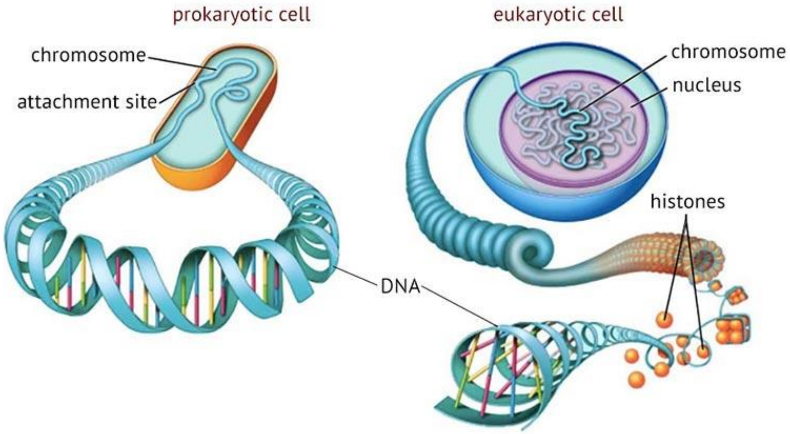




- Simple, small (0.1 – 5 μm)
- Do not have membrane-bound structures (nucleus, mitochondria)
- Nucleoid: DNA
- Cell wall: protection from the outside environment. Most bacteria have a rigid cell wall made from carbohydrates and peptidoglycans.
- Cell membrane (plasma membrane)
- Capsule: Some bacteria 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

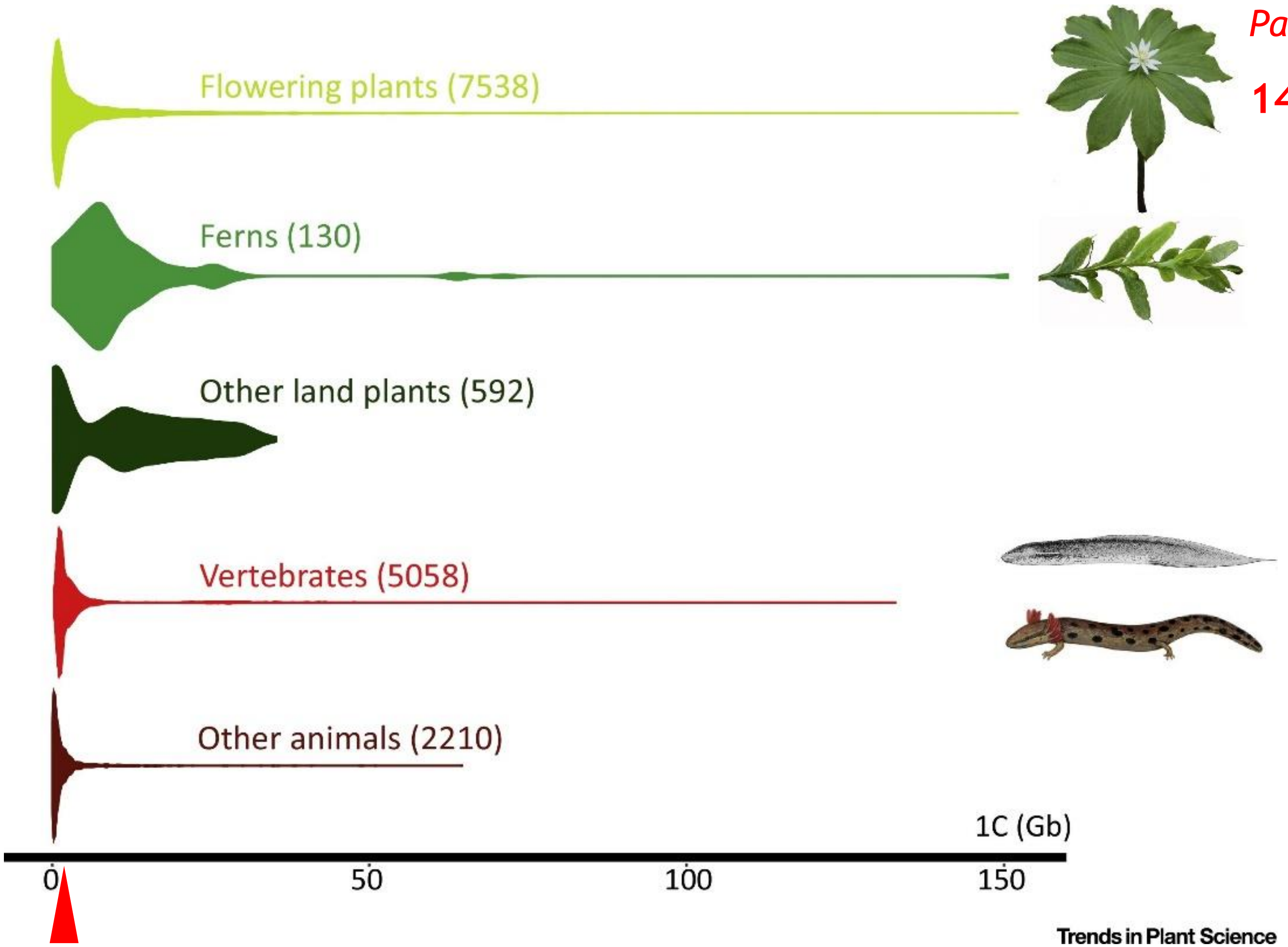
Prokaryotes vs Eukaryotes (more differences)



- 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

Eukaryotes: genome size variation (64 000-fold)



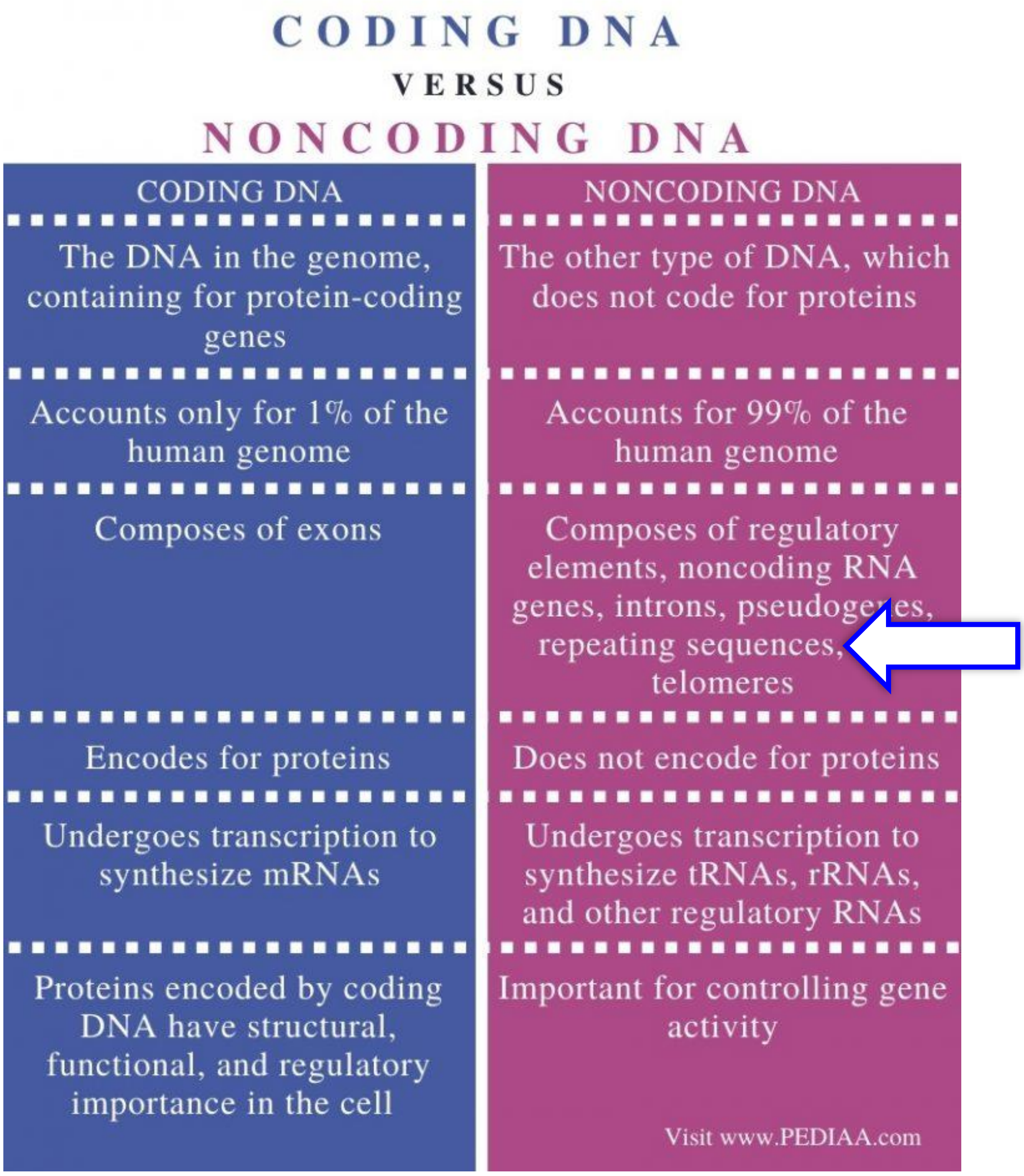
Trends in Plant Science

Paris japonica

149 Gb

Encephalitozoon intestinalis

2.3 Mb (0.0023 Gb)



CODING DNA

VERSUS

NONCODING DNA

CODING DNA	NONCODING DNA
The DNA in the genome, containing for protein-coding genes	The other type of DNA, which does not code for proteins
Accounts only for 1% of the human genome	Accounts for 99% of the human genome
Composes of exons	Composes of regulatory elements, noncoding RNA genes, introns, pseudogenes, repeating sequences, telomeres
Encodes for proteins	Does not encode for proteins
Undergoes transcription to synthesize mRNAs	Undergoes transcription to synthesize tRNAs, rRNAs, and other regulatory RNAs
Proteins encoded by coding DNA have structural, functional, and regulatory importance in the cell	Important for controlling gene activity

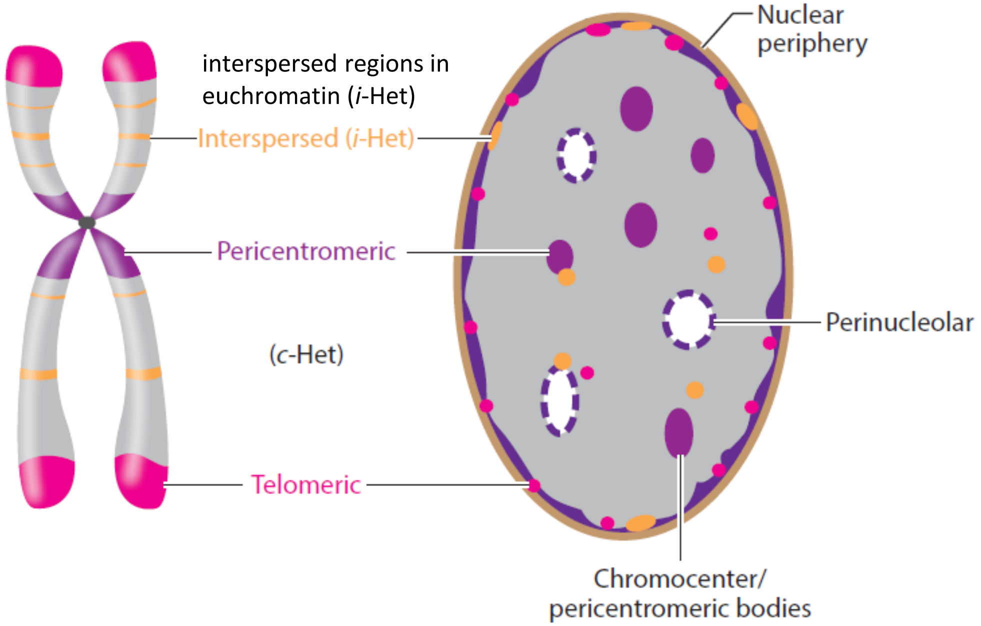
Visit www.PEDIAA.com

Euchromatin and heterochromatin. Repeats frequently building up heterochromatin.

- (simplification) coding DNA: accessible chromatin - **euchromatin** and is active (transcription facilitated)
- non-coding DNA (repeats): **heterochromatin**, generally inactive (thought that regulatory proteins, e.g. transcription factors, cannot access DNA templates)

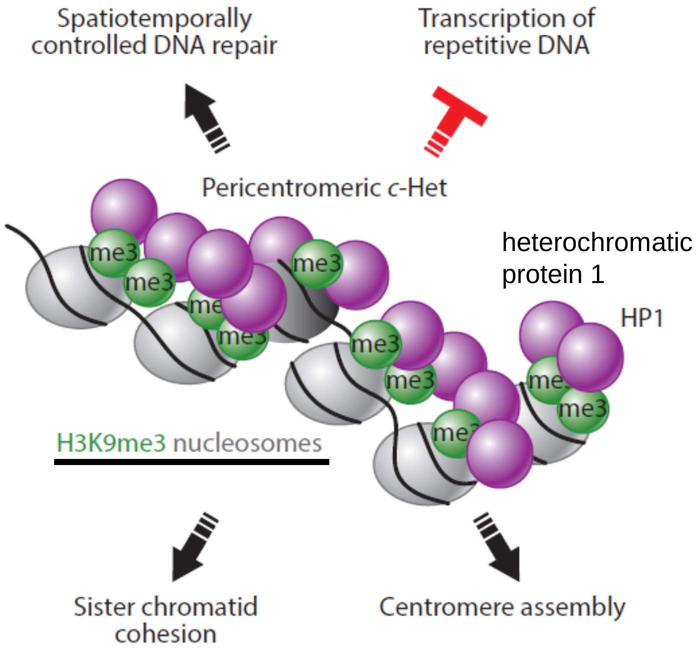
DNA repeats and heterochromatin

on chromsosome/in nucleus

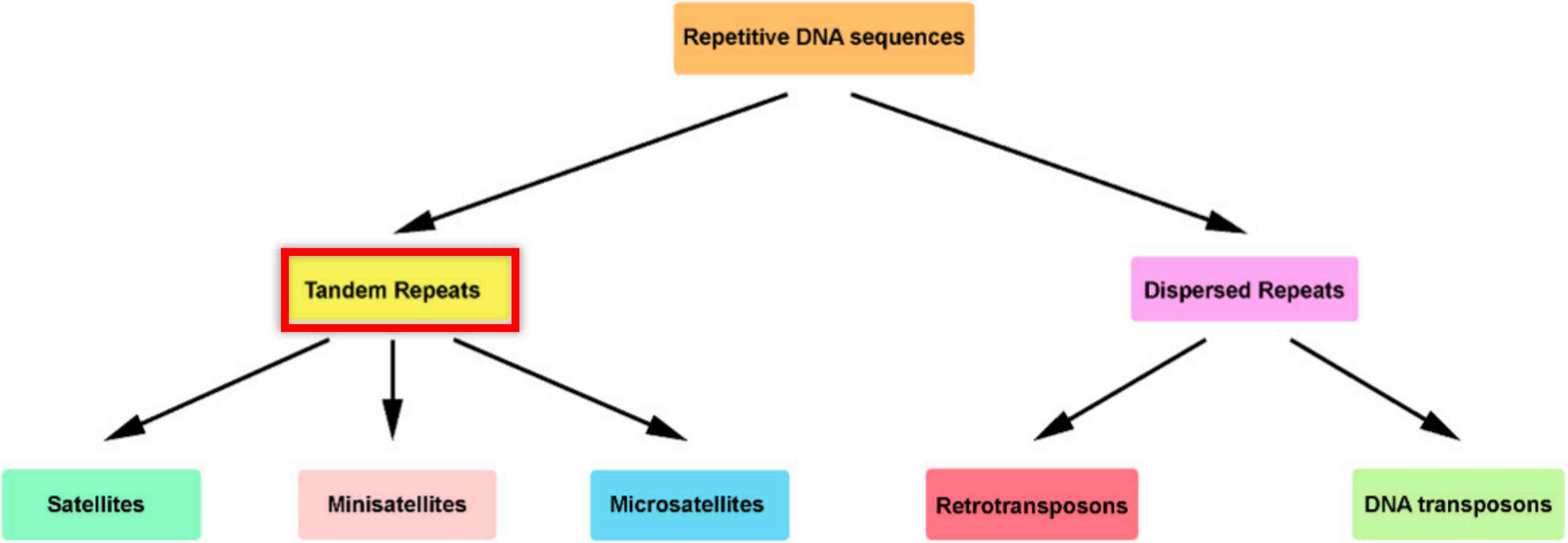


Heterochromatin:

di- and trimethylated histone H3 lysine 9



Repetitive DNA: tandem repeats



- Repetitive DNA sequences
  - Tandem Repeats
    - Satellites
    - Minisatellites
    - Microsatellites
  - Dispersed Repeats
    - Retrotransposons
    - DNA transposons

Microsatellites

- monomer length between 1 and 6 bp (- 10 bp)
- microsatellites are often referred to as short tandem repeats (STRs) or simple sequence repeats (SSRs)

Minisatellites

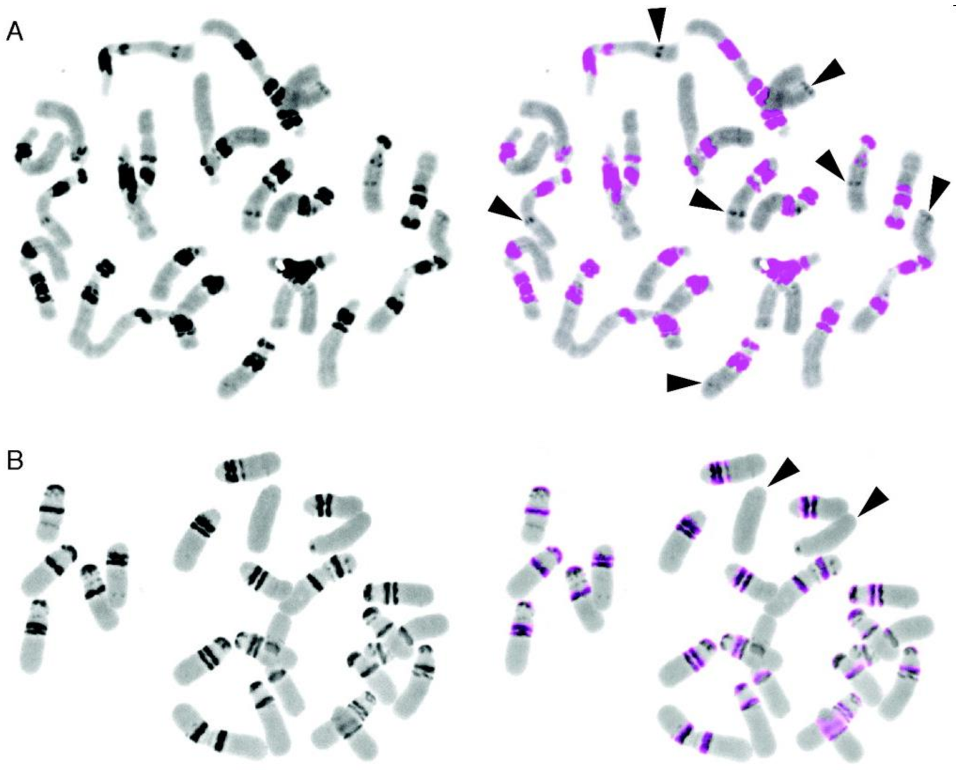
- monomer length between 10 and 60 bp

e.g., telomeric tandem repeats, also in sub-telomeric regions and centromeres

Satellites

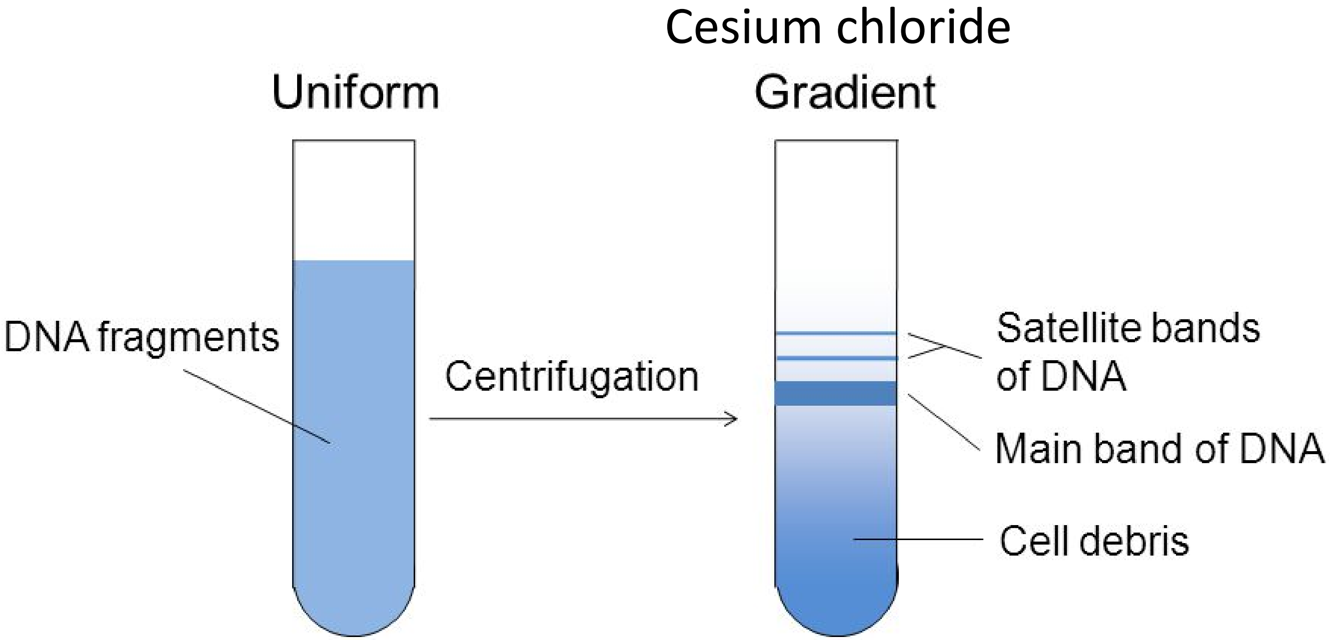
- e.g., 170-bp α (alphoid) DNA at all centromeres of human chromosomes, heterochromatin

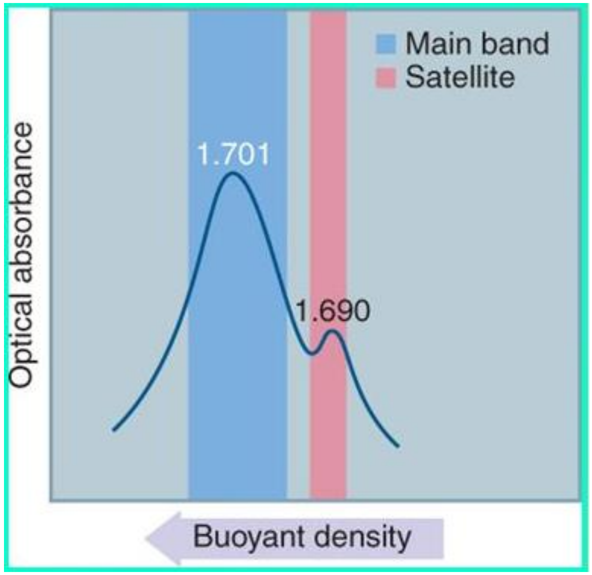
Satellites (tandem repeats)



FriSAT1 tandem repeat on chromosomes of North American *Fritillaria* species (Liliaceae)

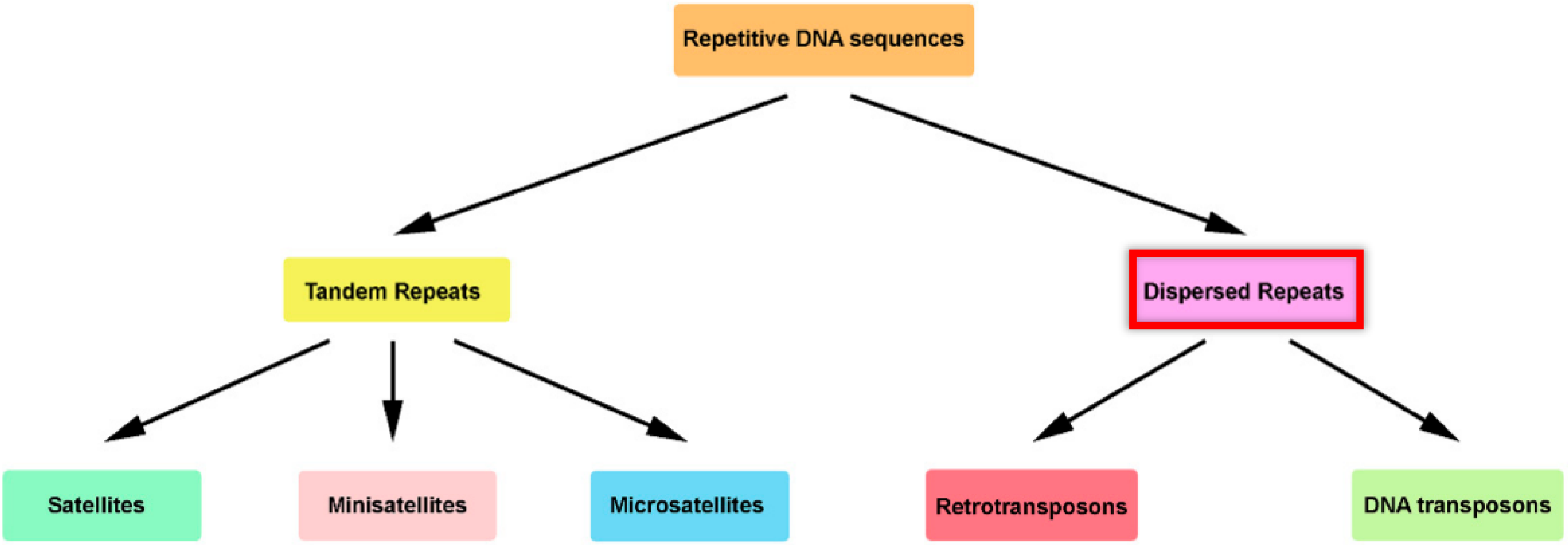
Why „satellite repeats“ are called satellite?





The main band DNA has density of 1.701 g/cm with a G-C content of 42%, and minor band DNA has the buoyant density of 1.690 g/cm with a G-C content of 30%.

Dispersed repetitive DNA



Repetitive DNA sequences

Tandem Repeats

Dispersed Repeats

Satellites

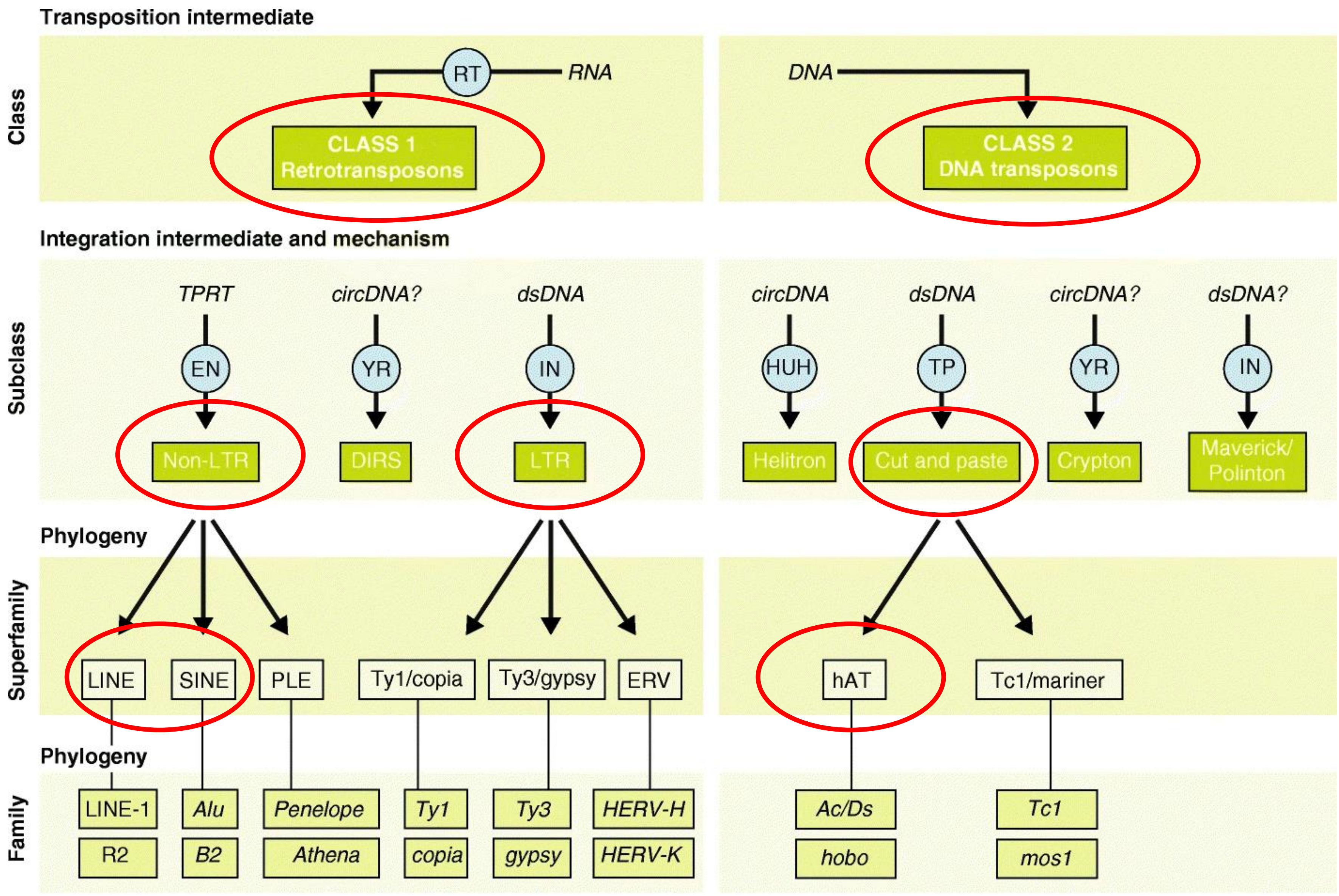
Minisatellites

Microsatellites

Retrotransposons

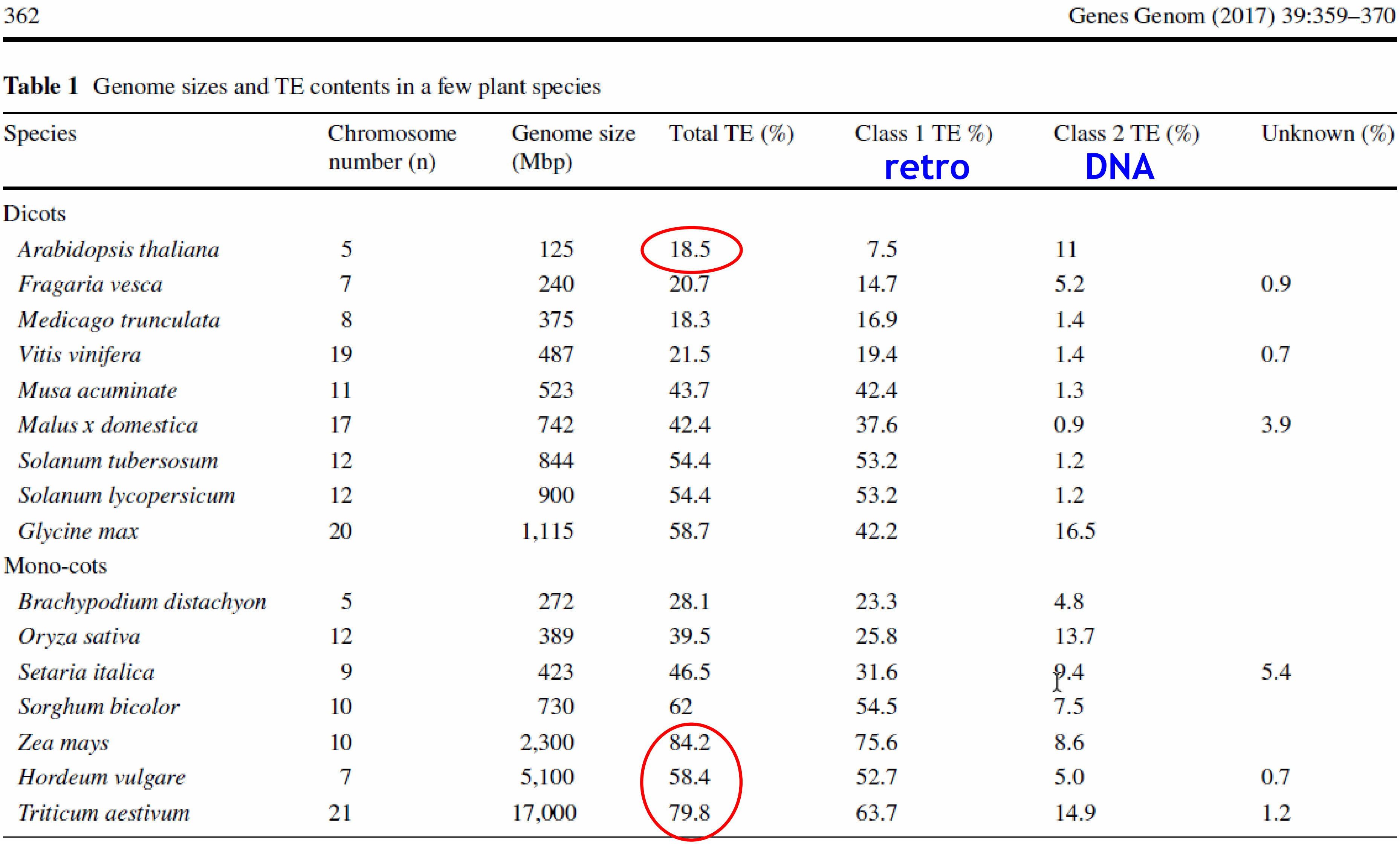
DNA transposons

Classification of eukaryotic transposable elements



Bourque et al. 2018, *Genome Biol* 19

Representation of TEs in plant genomes



362 Genes Genom (2017) 39:359–370

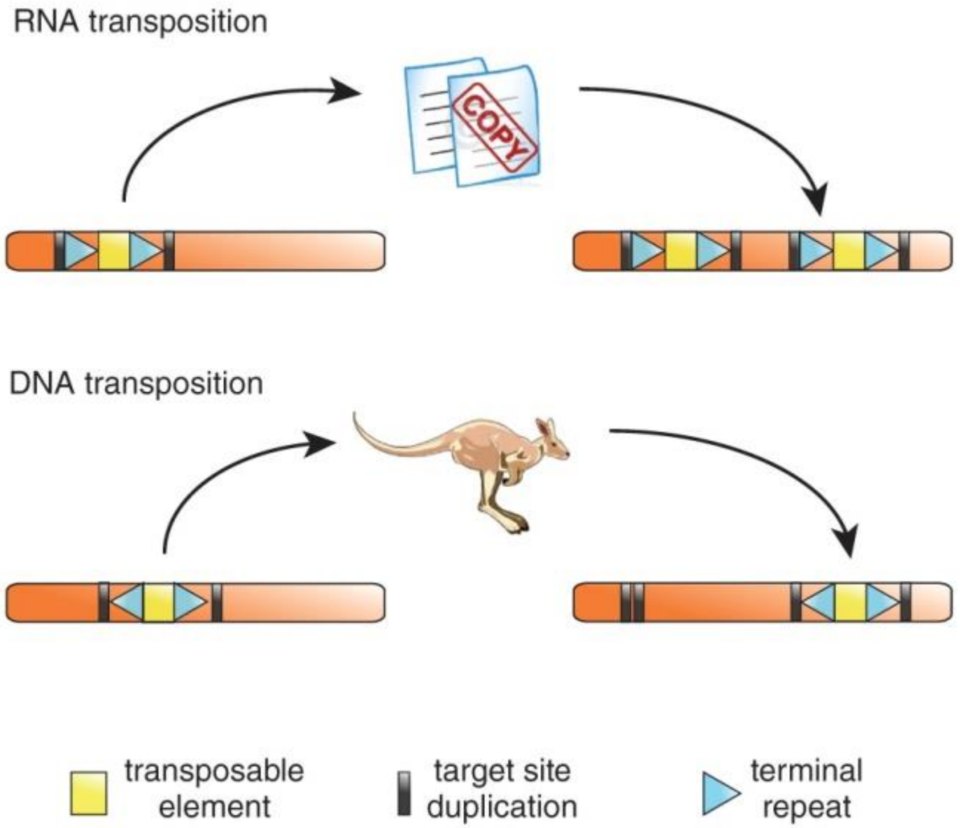
Table 1 Genome sizes and TE contents in a few plant species

Species	Chromosome number (n)	Genome size (Mbp)	Total TE (%)	Class 1 TE %) retro	Class 2 TE (%) DNA	Unknown (%)
Dicots						
Arabidopsis thaliana	5	125	18.5	7.5	11	
Fragaria vesca	7	240	20.7	14.7	5.2	0.9
Medicago trunculata	8	375	18.3	16.9	1.4	
Vitis vinifera	19	487	21.5	19.4	1.4	0.7
Musa acuminate	11	523	43.7	42.4	1.3	
Malus x domestica	17	742	42.4	37.6	0.9	3.9
Solanum tubersosum	12	844	54.4	53.2	1.2	
Solanum lycopersicum	12	900	54.4	53.2	1.2	
Glycine max	20	1,115	58.7	42.2	16.5	
Mono-cots						
Brachypodium distachyon	5	272	28.1	23.3	4.8	
Oryza sativa	12	389	39.5	25.8	13.7	
Setaria italica	9	423	46.5	31.6	9.4	5.4
Sorghum bicolor	10	730	62	54.5	7.5	
Zea mays	10	2,300	84.2	75.6	8.6	
Hordeum vulgare	7	5,100	58.4	52.7	5.0	0.7
Triticum aestivum	21	17,000	79.8	63.7	14.9	1.2

Transposable elements (TEs)

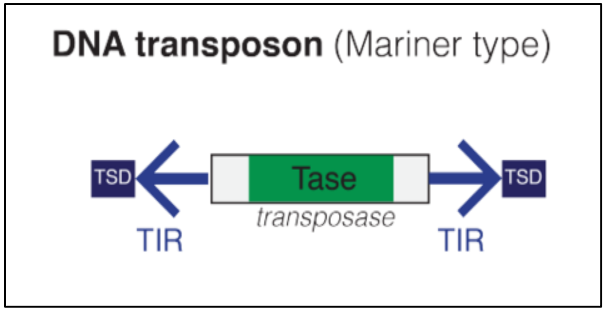
Class I TEs
retrotransposons

Class II TEs
transposons



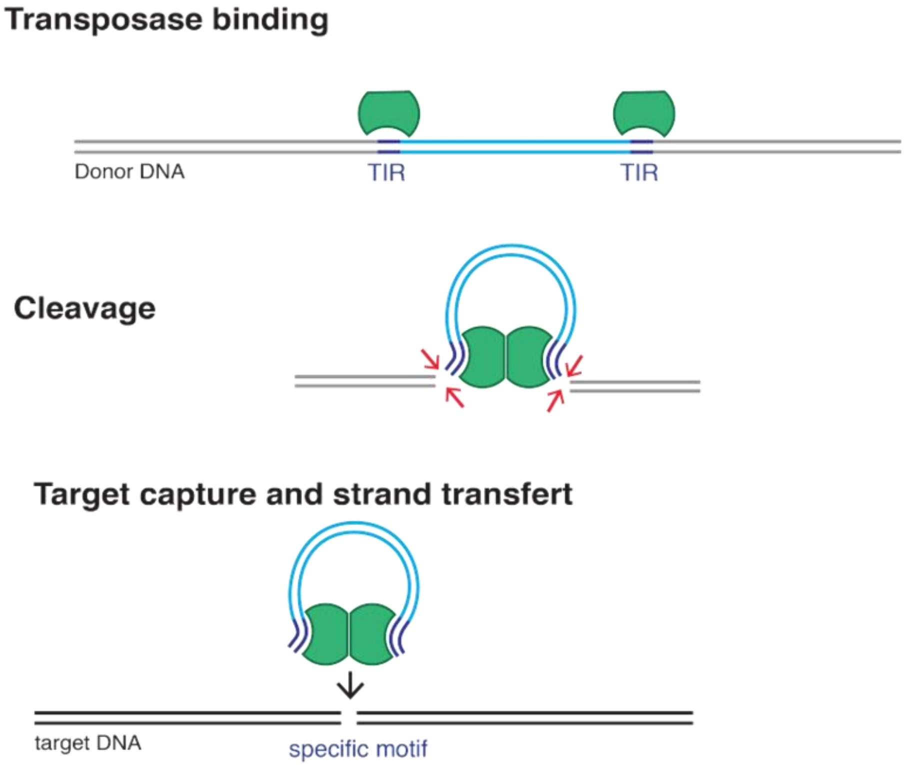
terminal repeats are generally in the same (direct) orientation in retrotransposon but in inverse orientation in transposons

Life cycle of DNA transposons, mechanism of transposition



2 terminal inverted repeats (TIR)

2 short target site duplications (TSD)

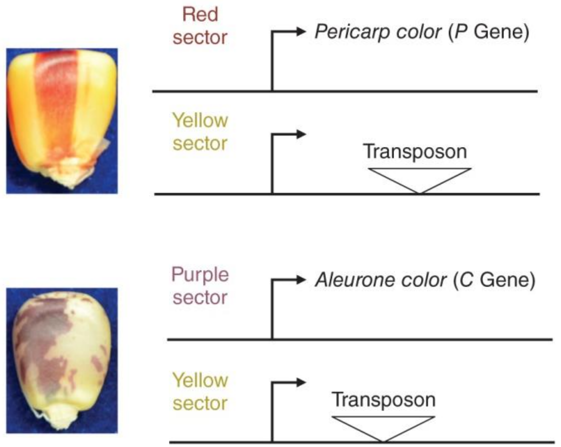


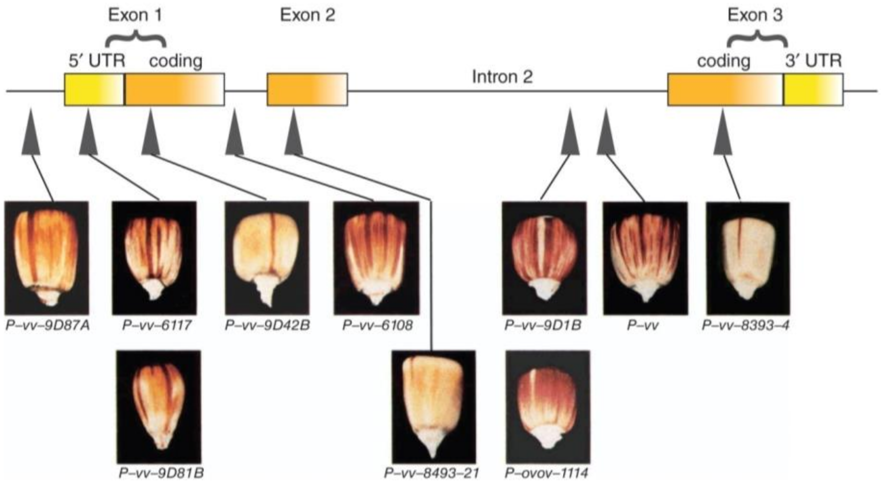
- 2 transposases recognize and bind to TIR sequences, join together (dimer) and promote DNA double-strand cleavage

- the DNA-transposase complex then inserts its DNA cargo at specific DNA motifs elsewhere in the genome (creating short TSDs after integration – target DNA site is duplicated)

DNA transposons (discovery)

Ac/*Ds* (Activator/Dissociator) elements in maize

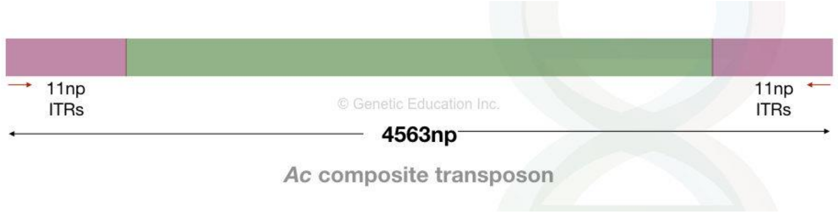




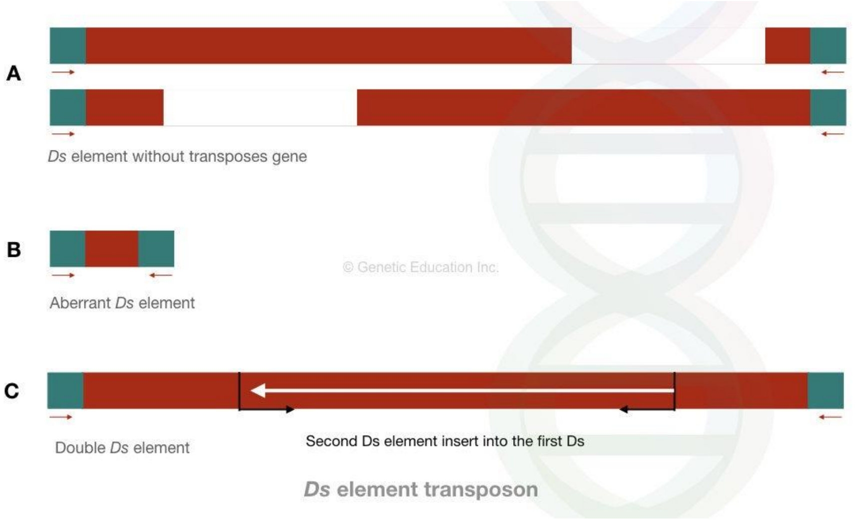
insertion to exons = no protein/pigment = yellowish kernels

Barbara McClintock
(experiments
1947 – 1949)

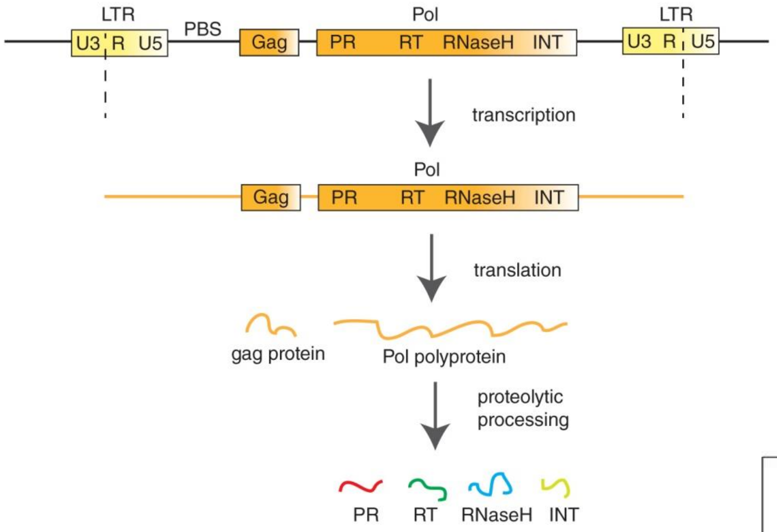
Ac element (active, contains transposase gene)



Ds element (inactive, without transposase gene)



Structure and life cycle of LTR (Long Terminal Repeat) retrotransposons (LTR-RTs)



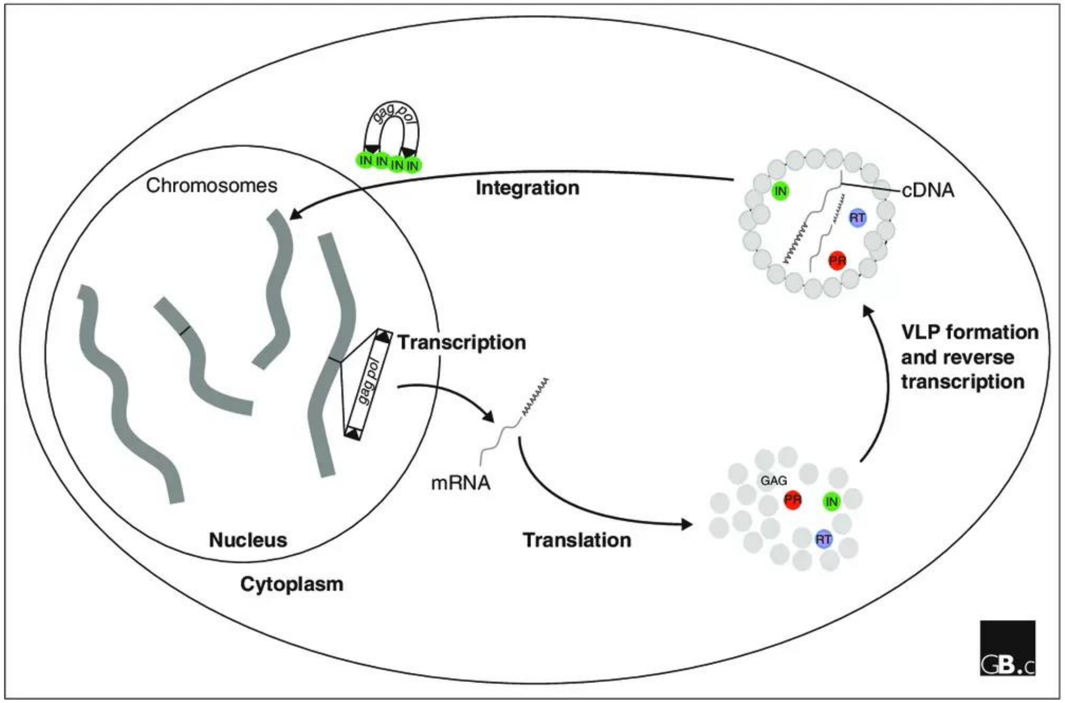
Gag – gene for the Gag protein

INT – integrase

PBS – primer binding site

PR – protease

RT – reverse transcriptase

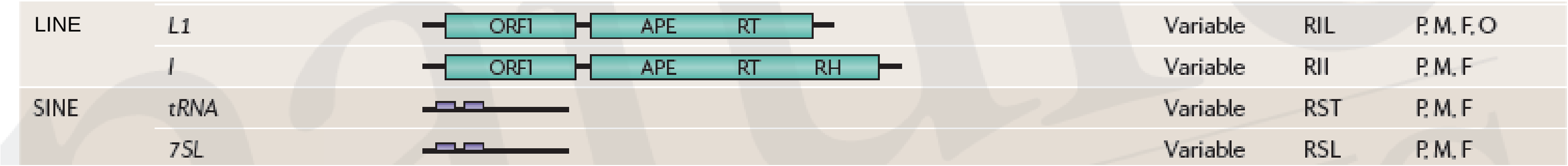


VLP – virus-like particle

Retrotransposons without LTRs (non-LTR retrotransposons)

LINE (long interspersed nuclear elements)

SINE (short interspersed nuclear elements)



LINE	*L1*	ORF1 APE RT	Variable	RIL	P, M, F, O
	I	ORF1 APE RT RH	Variable	RII	P, M, F
SINE	*tRNA*		Variable	RST	P, M, F
	7SL		Variable	RSL	P, M, F

APE – endonuclease, RH – RNase H

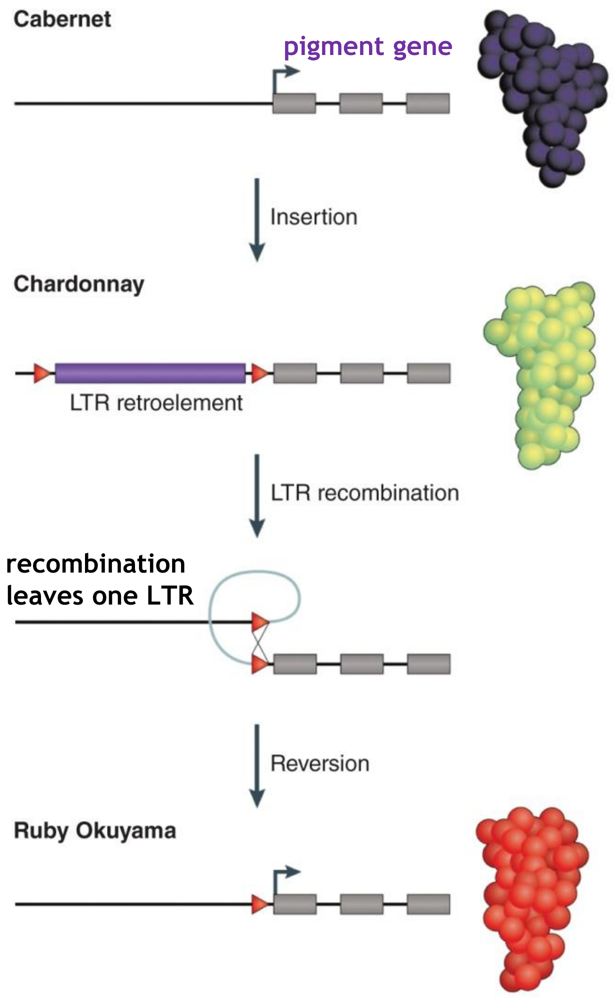
LINE

- autonomous
- LINEs 21% of human genome (one LINE = 7 kb)
- in humans, LINE1 (100,000 truncated and 4,000 full-length LINE-1 elements)

SINE

- non-autonomous – use RT of other elements
- 100 – 700 bp, derived from products tRNA
- Alu elements (300 bp), the most common SINE in humans (>1,000,000 copies, 10% of the genome)

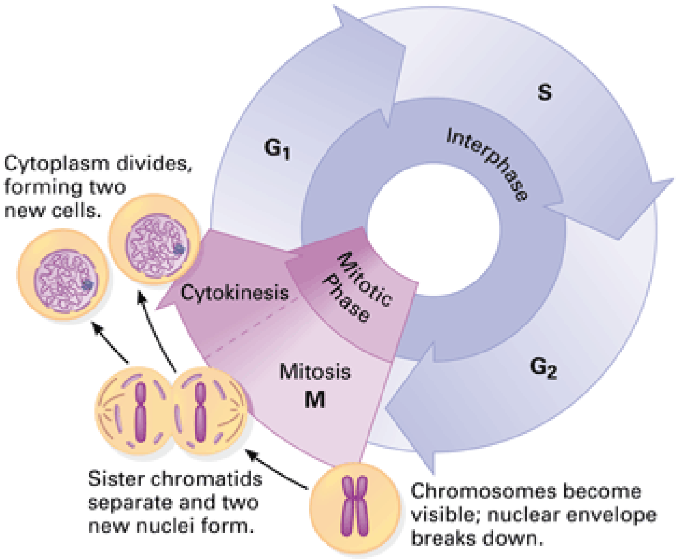
Transposable elements can disrupt or move genes and change their regulation

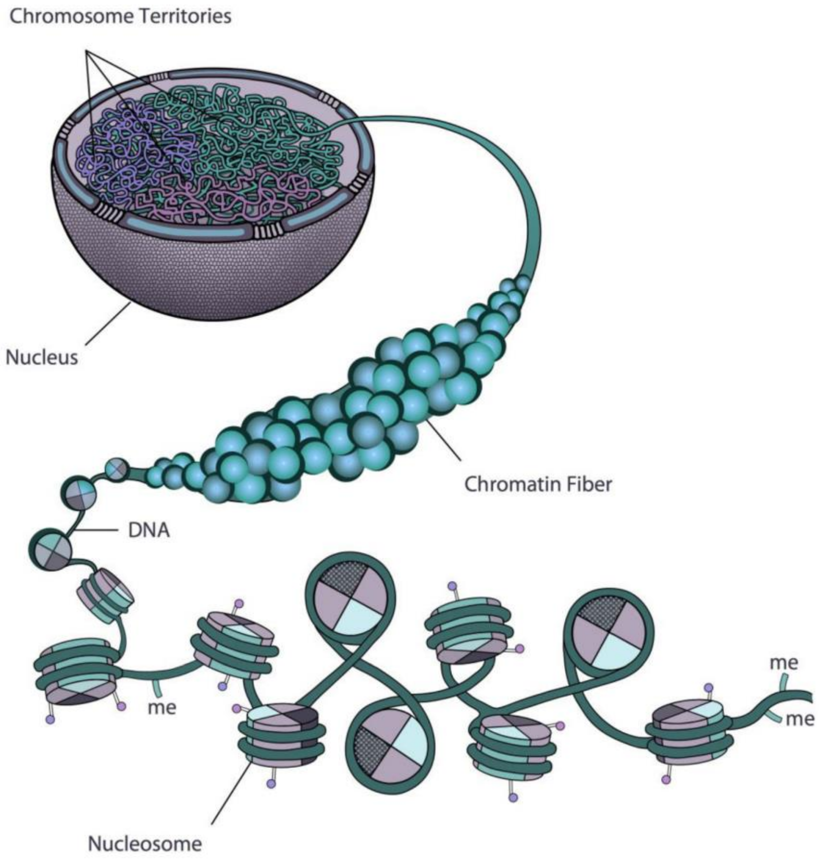


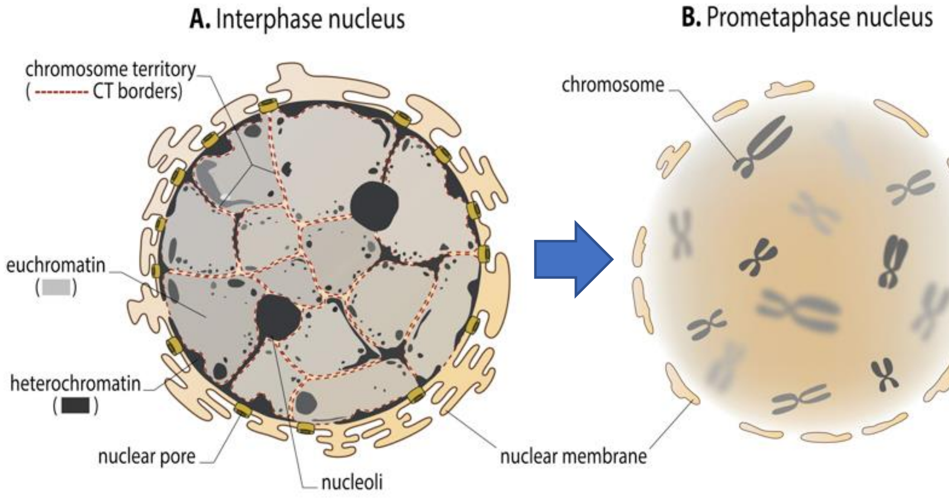
Plant Genes, Genomes and Genetics, First Edition. Erich Grotewold, Joseph Chappell and Elizabeth A. Kellogg.

Companion Website: www.wiley.com/go/grotewold/plantgenes.

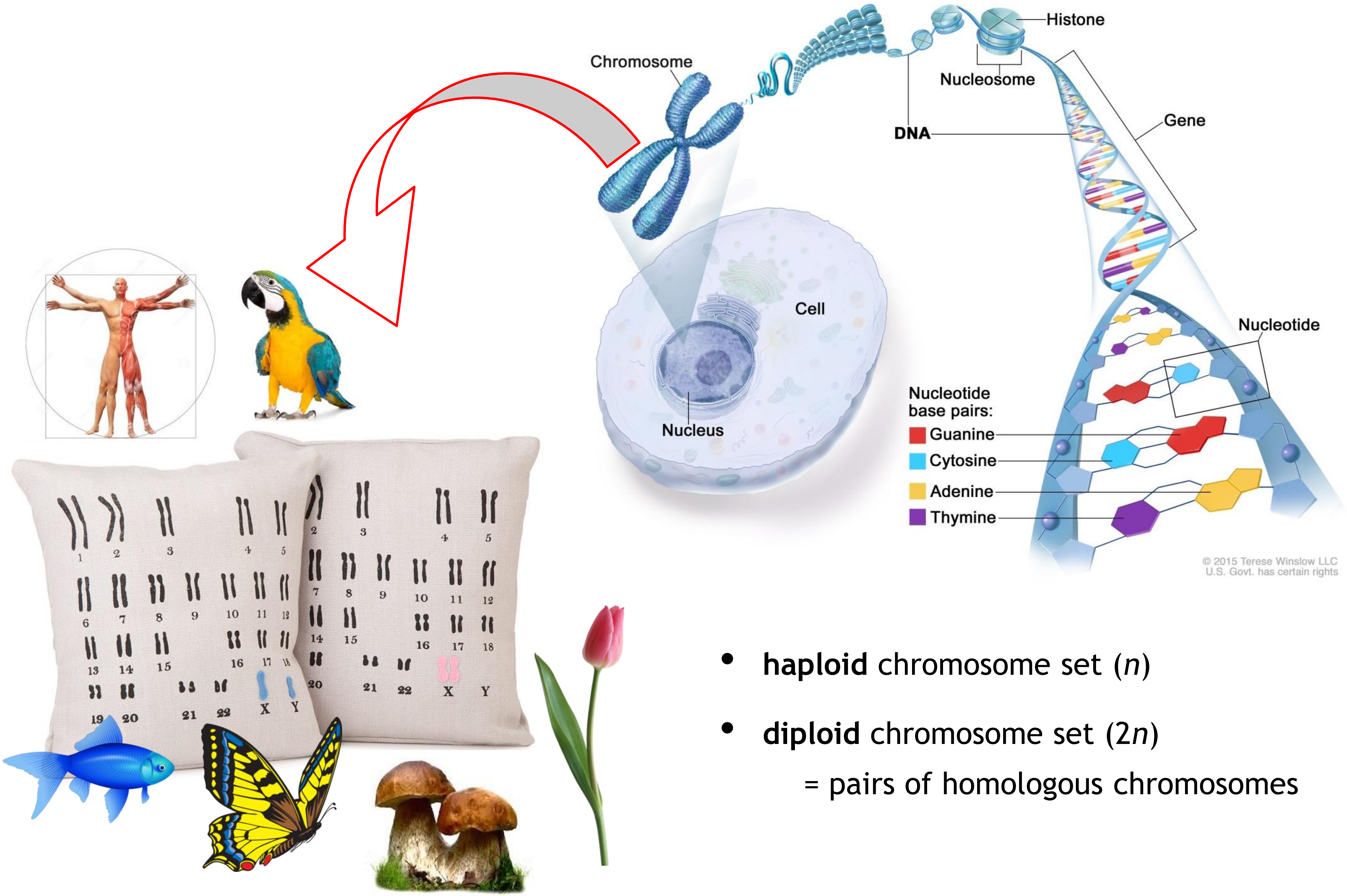
Eukaryotic chromosome







„Species-specific“ chromosome sets = karyotypes



- **haploid** chromosome set (*n*)
- **diploid** chromosome set (2*n*)

  = pairs of homologous chromosomes

Eukaryotes: minimal chromosome number

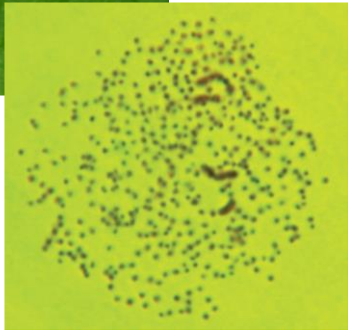
Myrmecia pilosula „Jack jumper ant“, Australia

males (haploid) *n* = 1, females (diploid) 2*n* = 2

five angiosperm species

e.g., *Haplopappus gracilis*, Asteraceae, *n* = 2

Eukaryotes: highest chromosome number



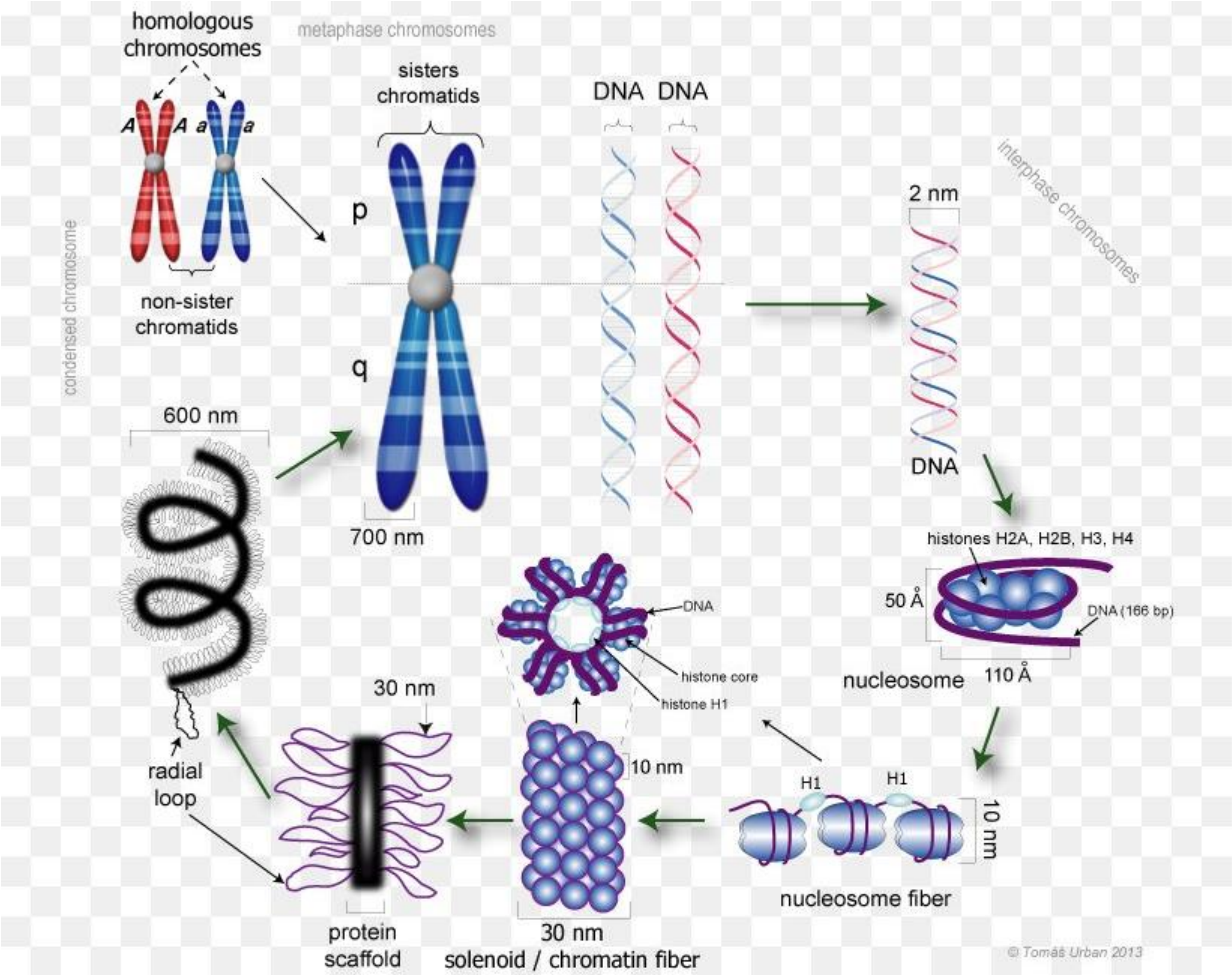
Polyommatus atlanticus

n = c. 220

fern *Ophioglossum reticulatum*

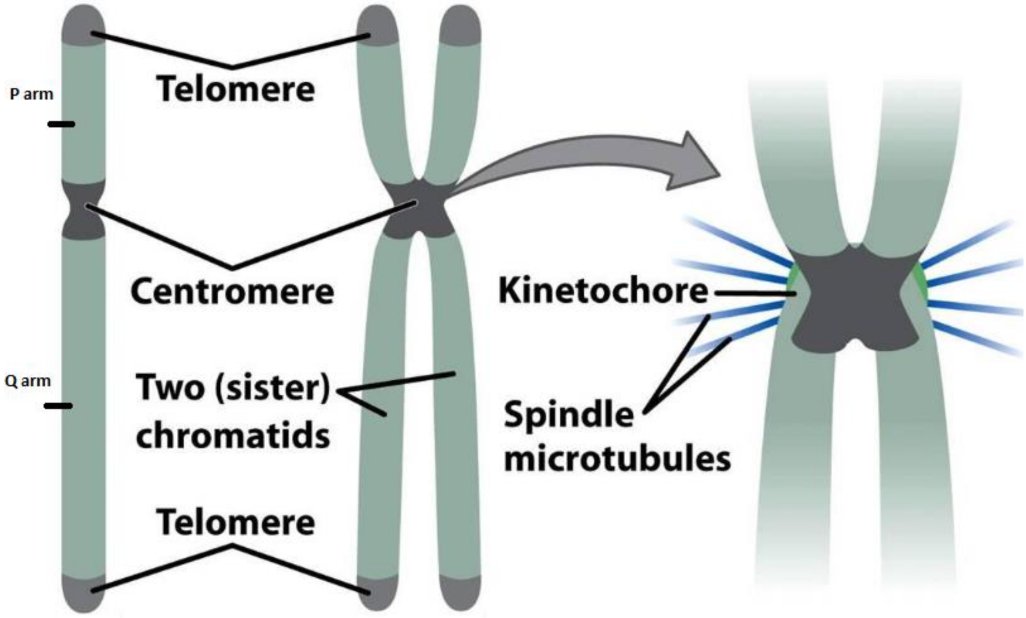
n = c. 530

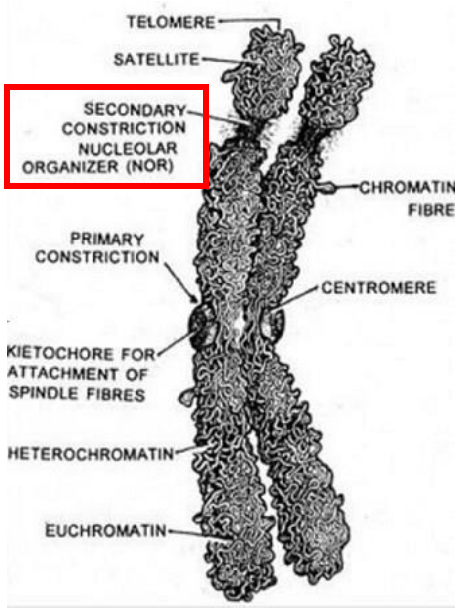
Eukaryotic chromosome

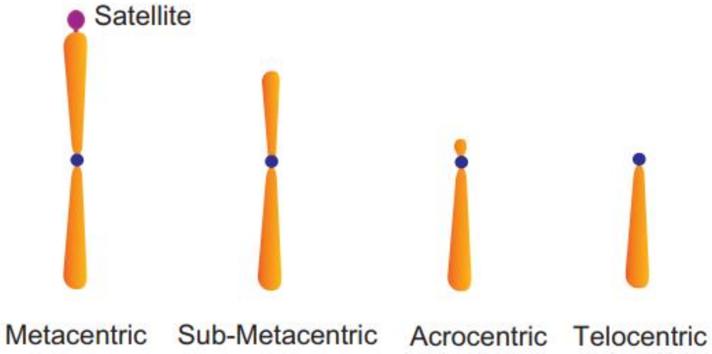


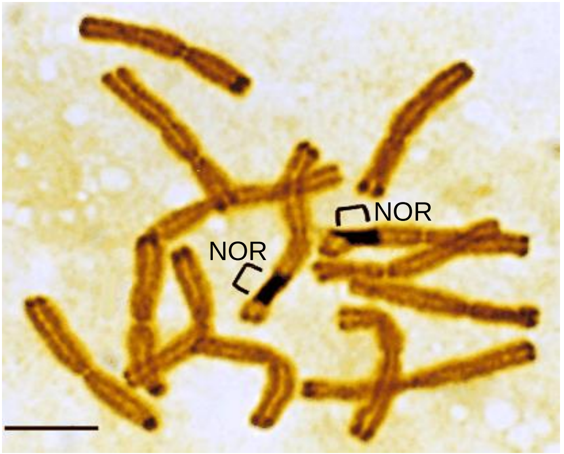
© Tomáš Urban 2013

Eukaryotic chromosome



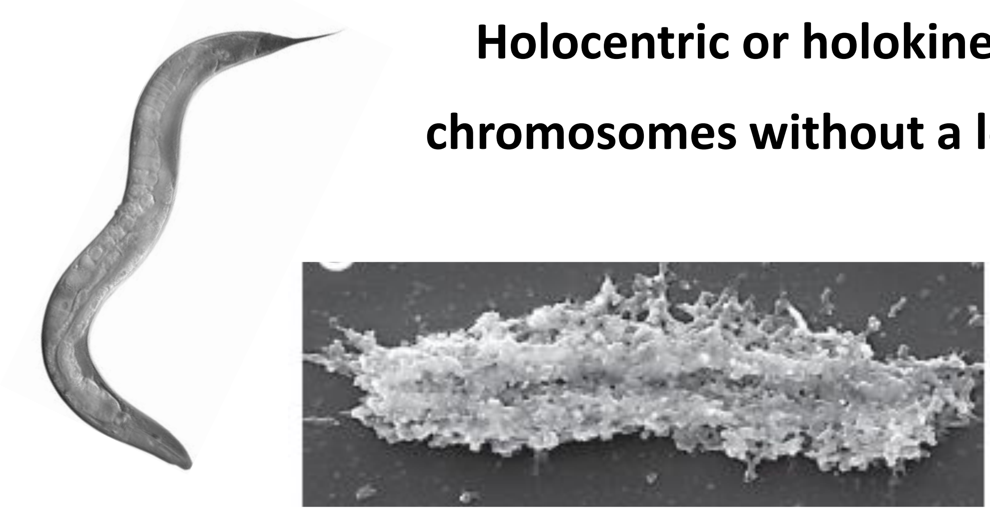


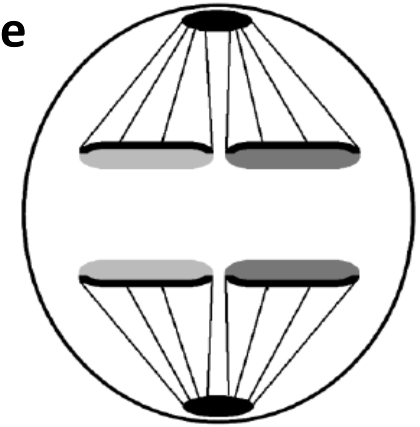




Holocentric or holokinetic chromosomes:

chromosomes without a localized centromere





difuse kinetochor →

chromosome segregation
in anaphase

- chromosome fission (agmatoploidy) and fusion (symploidy) → extensive chromosome number variation
- holocentrics: huge variation in chromosome numbers [the largest number of chromosomes in animals ($2n = 446$) is found in the blue butterfly *Polyommatus atlantica* with holokinetic chromosomes]
- in c. 5,500 angiosperm species
- chromosome numbers from $n = 2$ up to $n = 110$

Angiosperm species with holokinetic chromosomes

Juncaceae

Cyperaceae

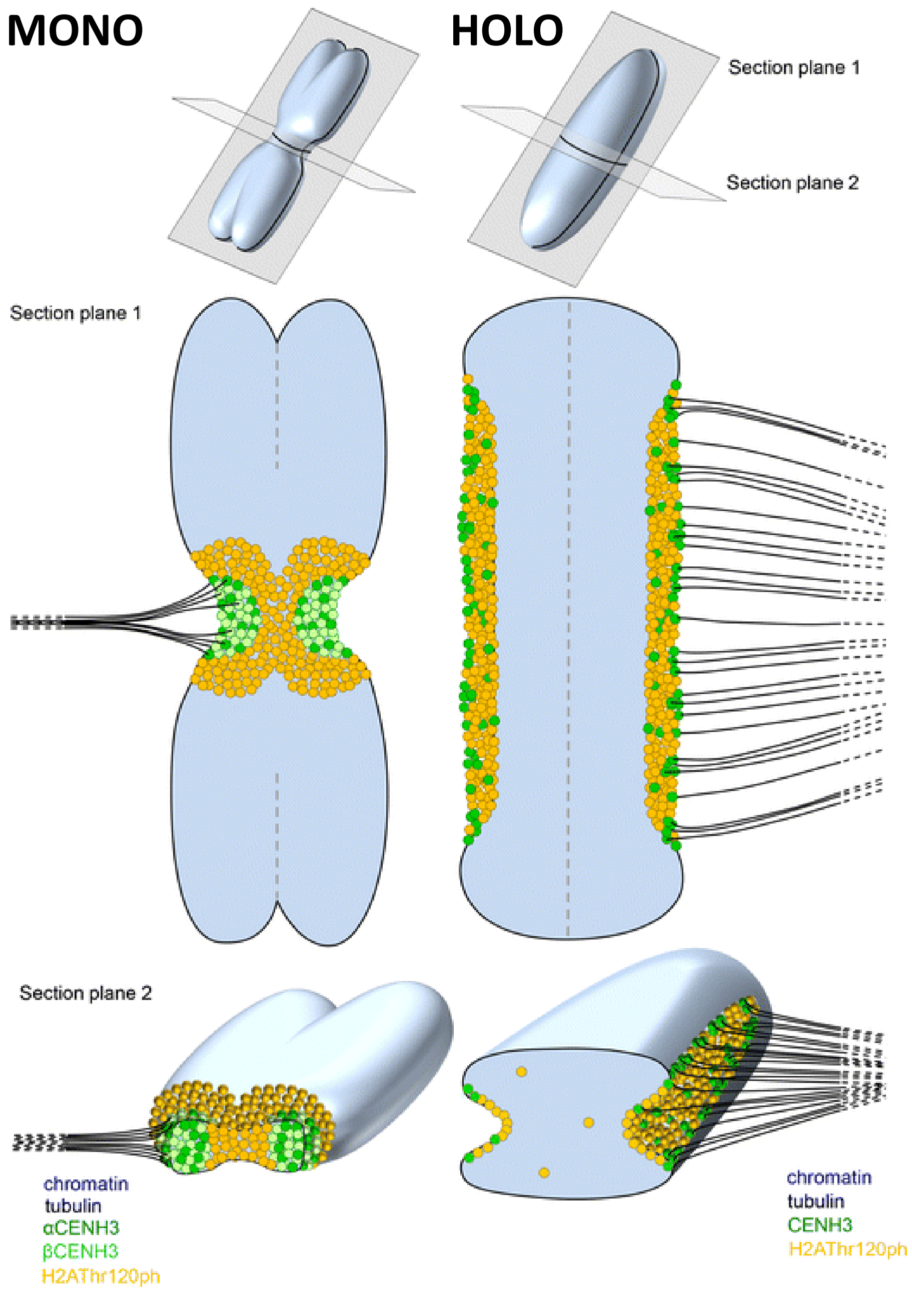
Myristica fragrans (Myristicaceae)

Drosera (Droseraceae)

…

Chionographis (Melanthiaceae)

Model of the centromere organization of mono- and holocentric plant chromosomes

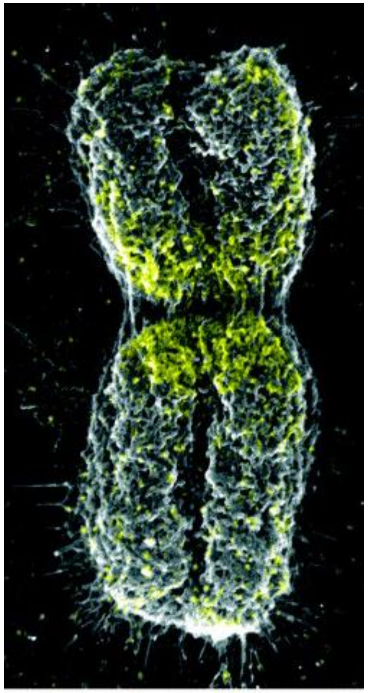


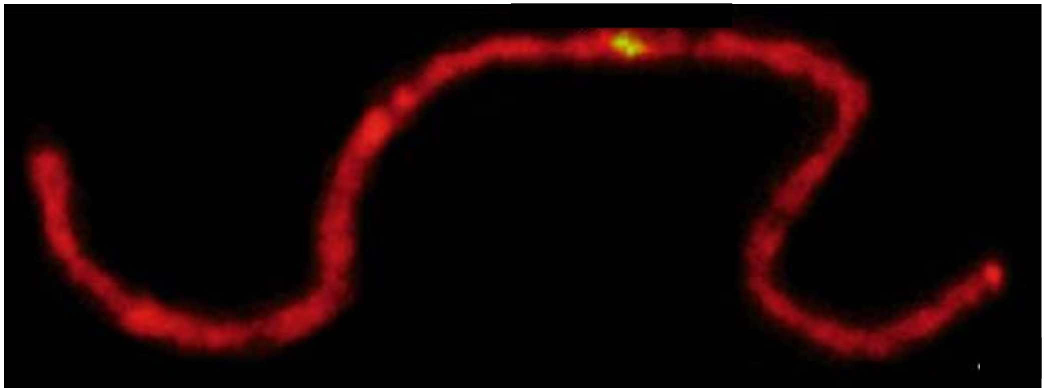
Microtubules (tubulin) attach at CENH3, but not at H2AThr120ph.

The microtubule bundle formation is less pronounced at holocentromeres.

active centromeres have H2AThr120ph - phosphorylation of threonine 120 of histone H2A

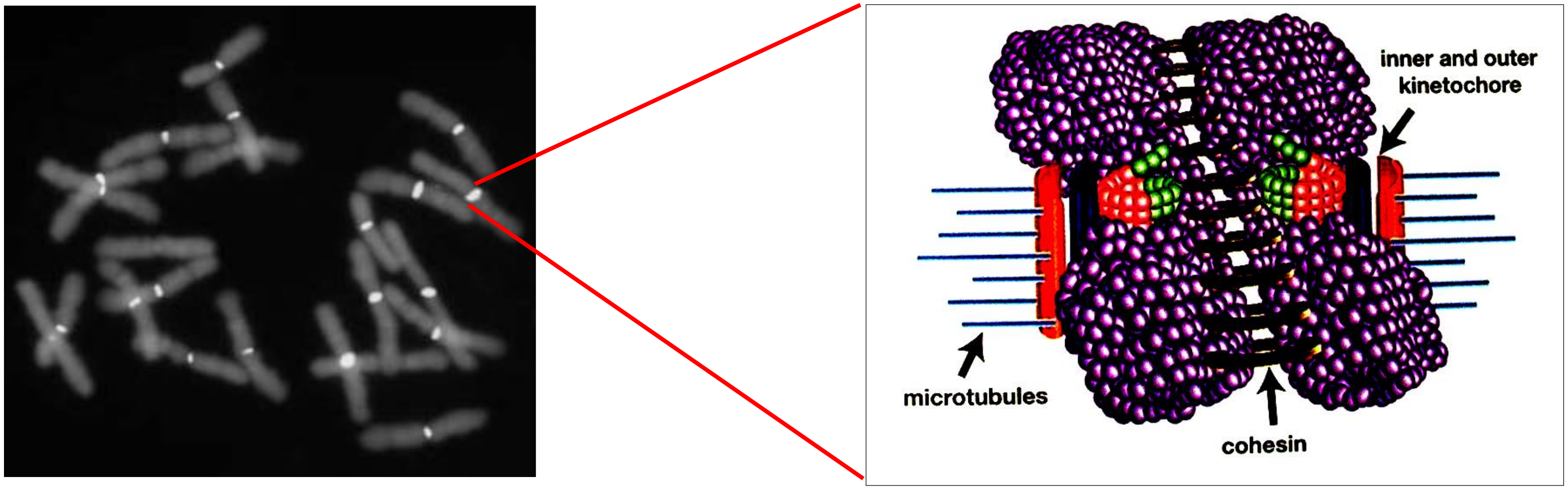
Centromere structure and function



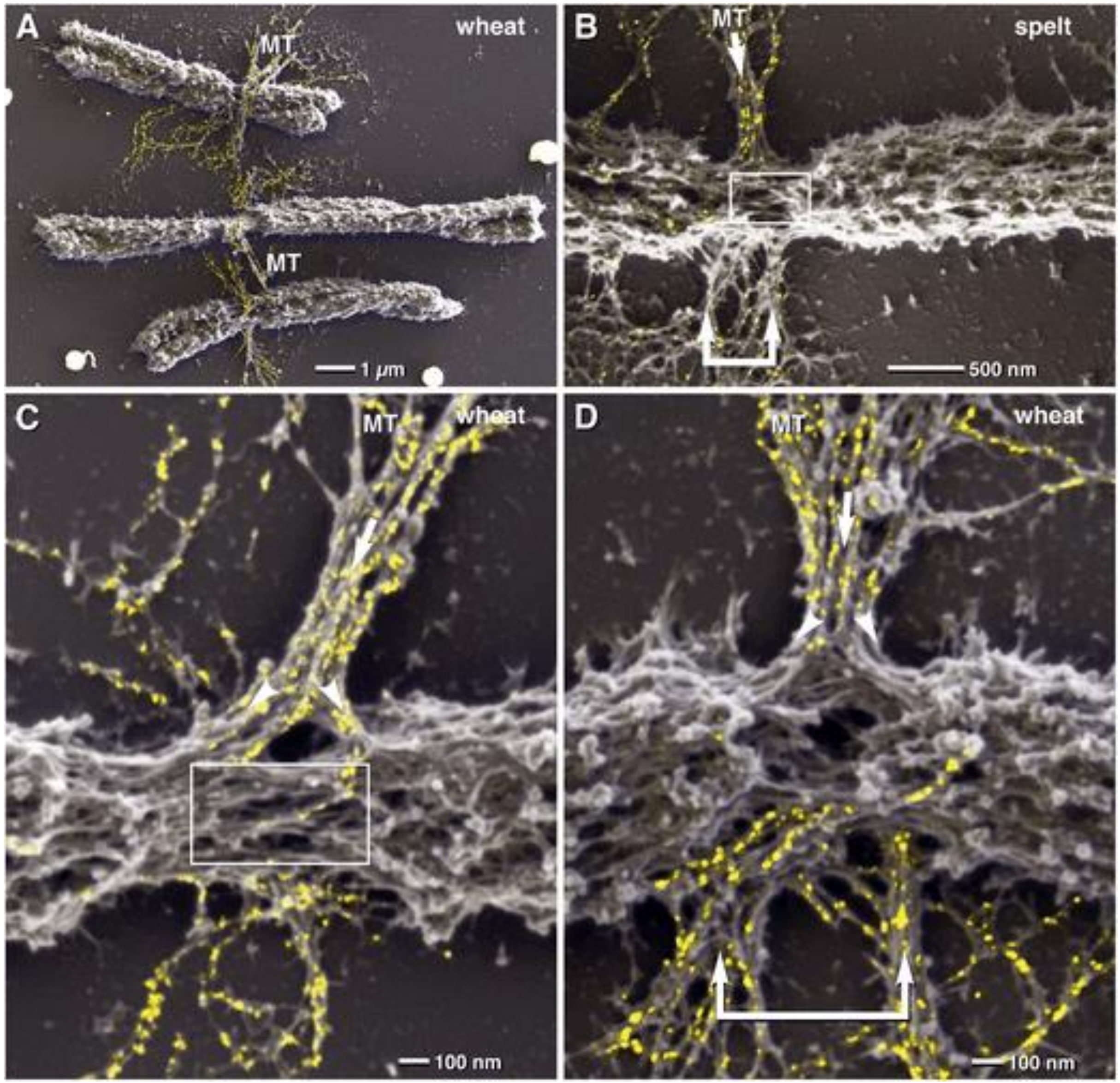


Centromere function

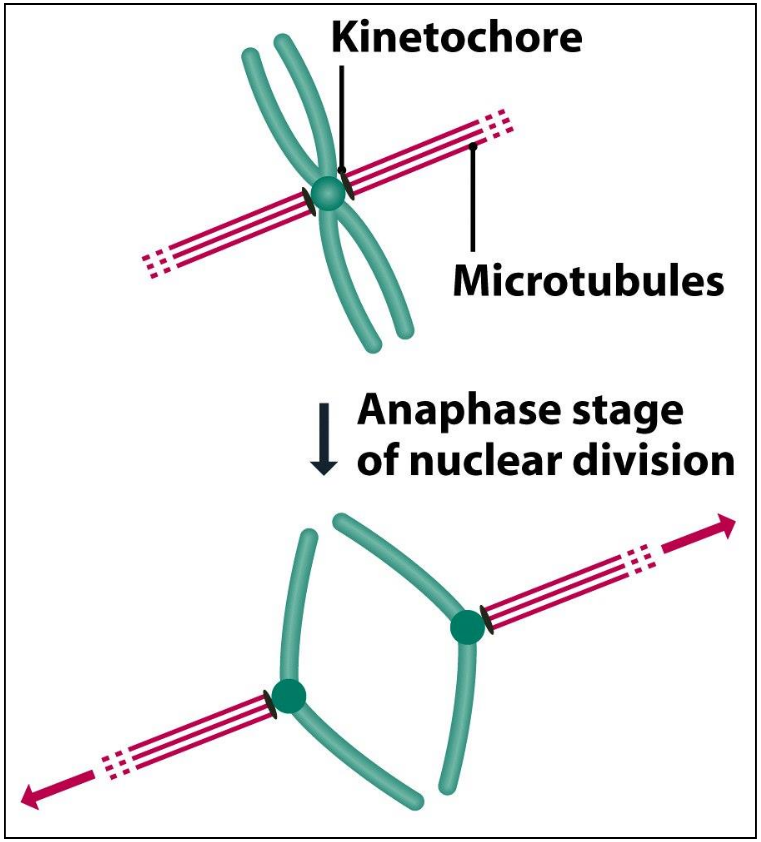
- chromosomes can be monocentric or holocentric (*Luzula*, *Eleocharis*, some insects)
- dicentric chromosomes usuallly unstable (anaphase bridges >> breakage)
- acentric chromosome fragments are unstable at mitosis/meiosis and lost
- sister chromatid cohesion throughout cell cycle until sister chromatid segregation at mitosis/meiosis II
- sites of kinetochore formation ensuring correct chromosome position on mitotic/meiotic spindle (spindle microtubules attached to kinetochores)



Centromeres and microtubules (monocentric chromosomes)

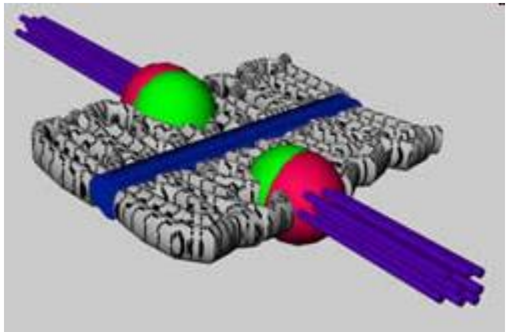


Wanner et al. (2015) Chromosoma



Kinetochore

inner kinetochore - associated with the centromere DNA; specialized form of chromatin persistent throughout the cell cycle

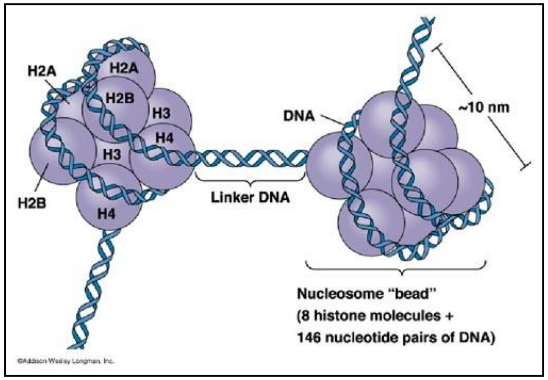


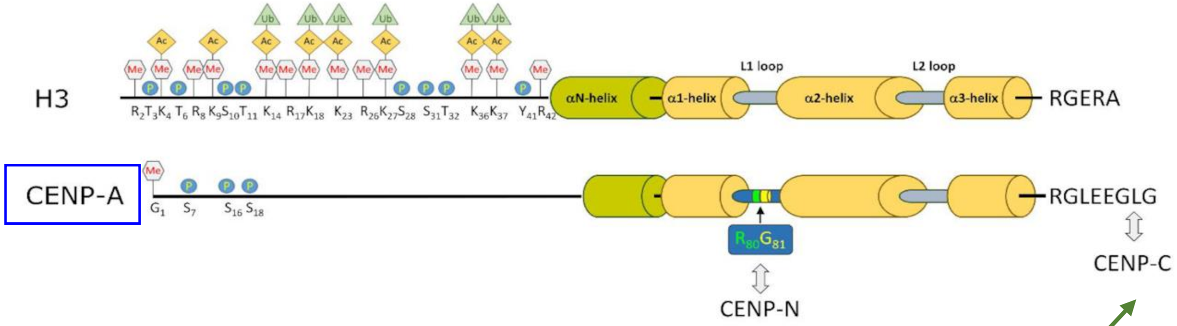
outer kinetochore - interacting with microtubules; functional only during cell division.

Even the simplest kinetochores consist of more than 45 different proteins!

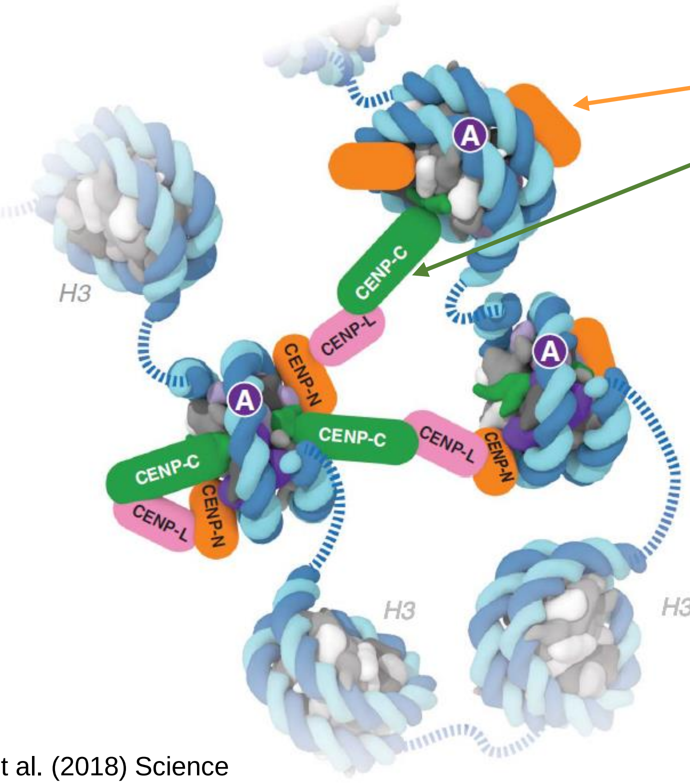
Many proteins conserved between eukaryotic species, including a specialized histone H3 variant (called CENP-A or CenH3) which helps the kinetochore associate with DNA.

Centromeric histone H3.
CENP-A (called CenH3 in plants) determines centromere location/activity



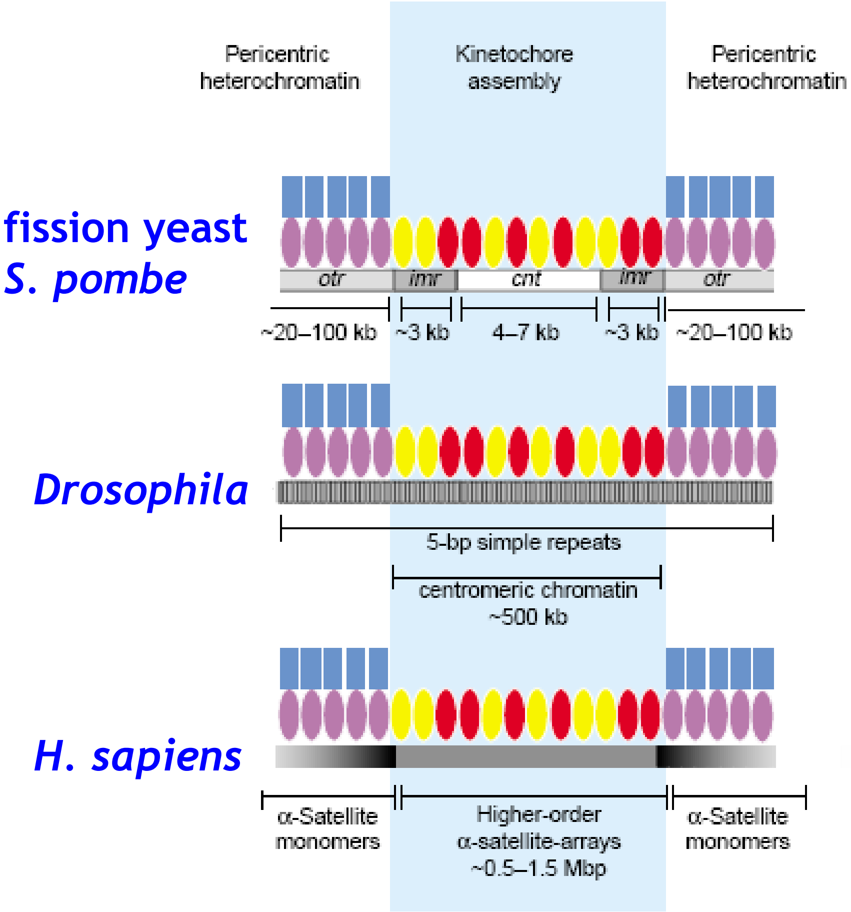


kinetochore proteins



Chittori et al. (2018) Science

The overall chromatin structure of the centromere is conserved among different eukaryotic species



🟦 HP1 protein

● Me-K9 H3 (methylation of H3 on lysine 9)

● diMeK4-H3 (di-methylation of H3 on lysine 4)

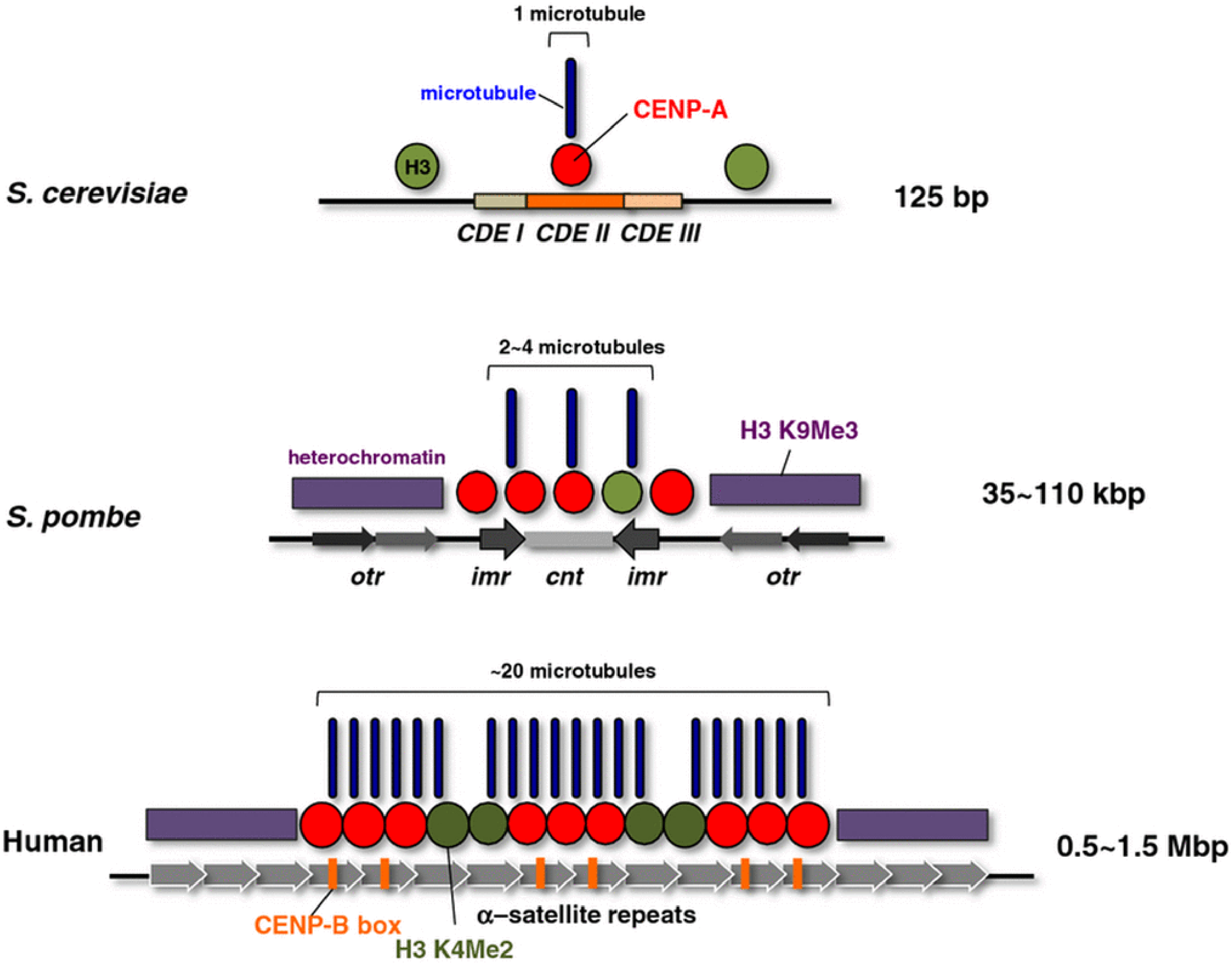
● CENP-A

otr : outer repeat

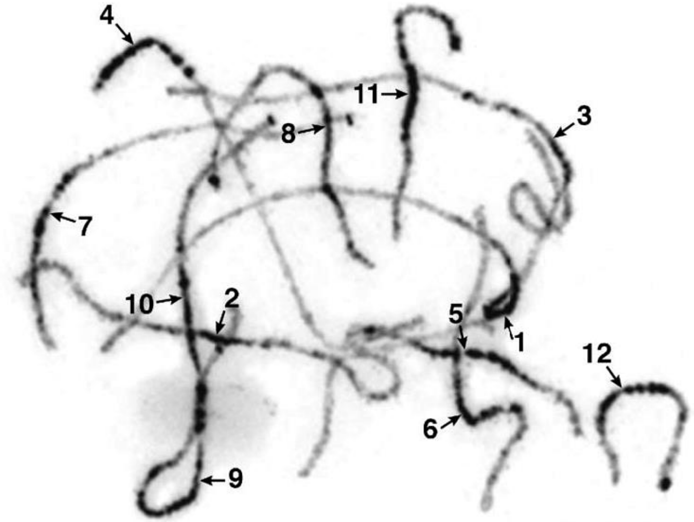
imr : innermost repeat

cnt : central sequence

Centromere size and „strength“

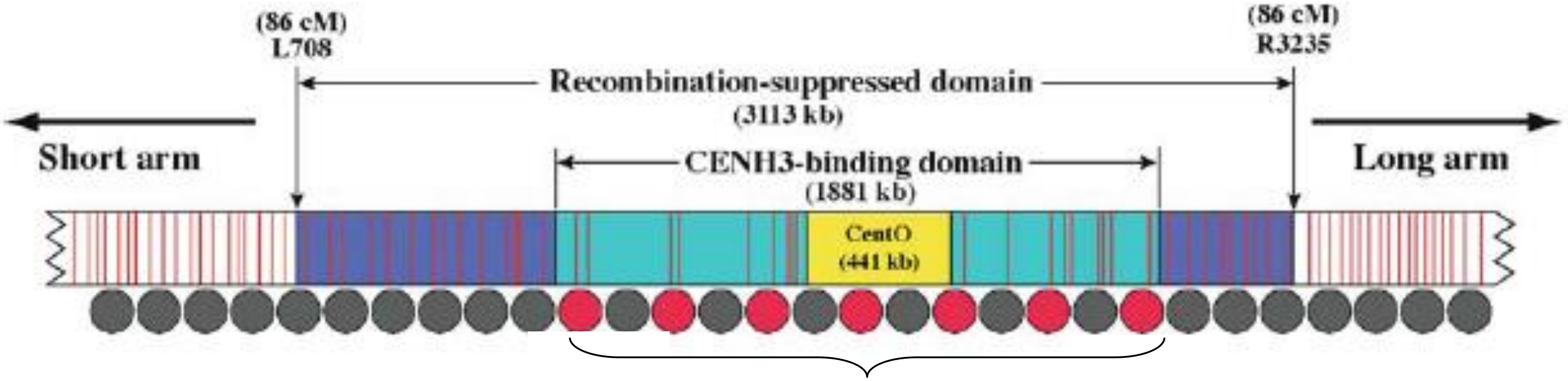


Structure of plant centromeres



Centromere regions can span up to a few Mb, composed mainly of centromere-specific satellite DNA.

centromere of rice chromosome 3

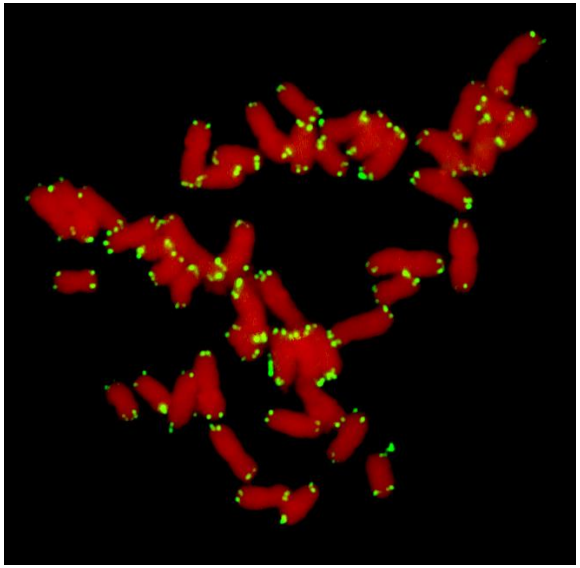


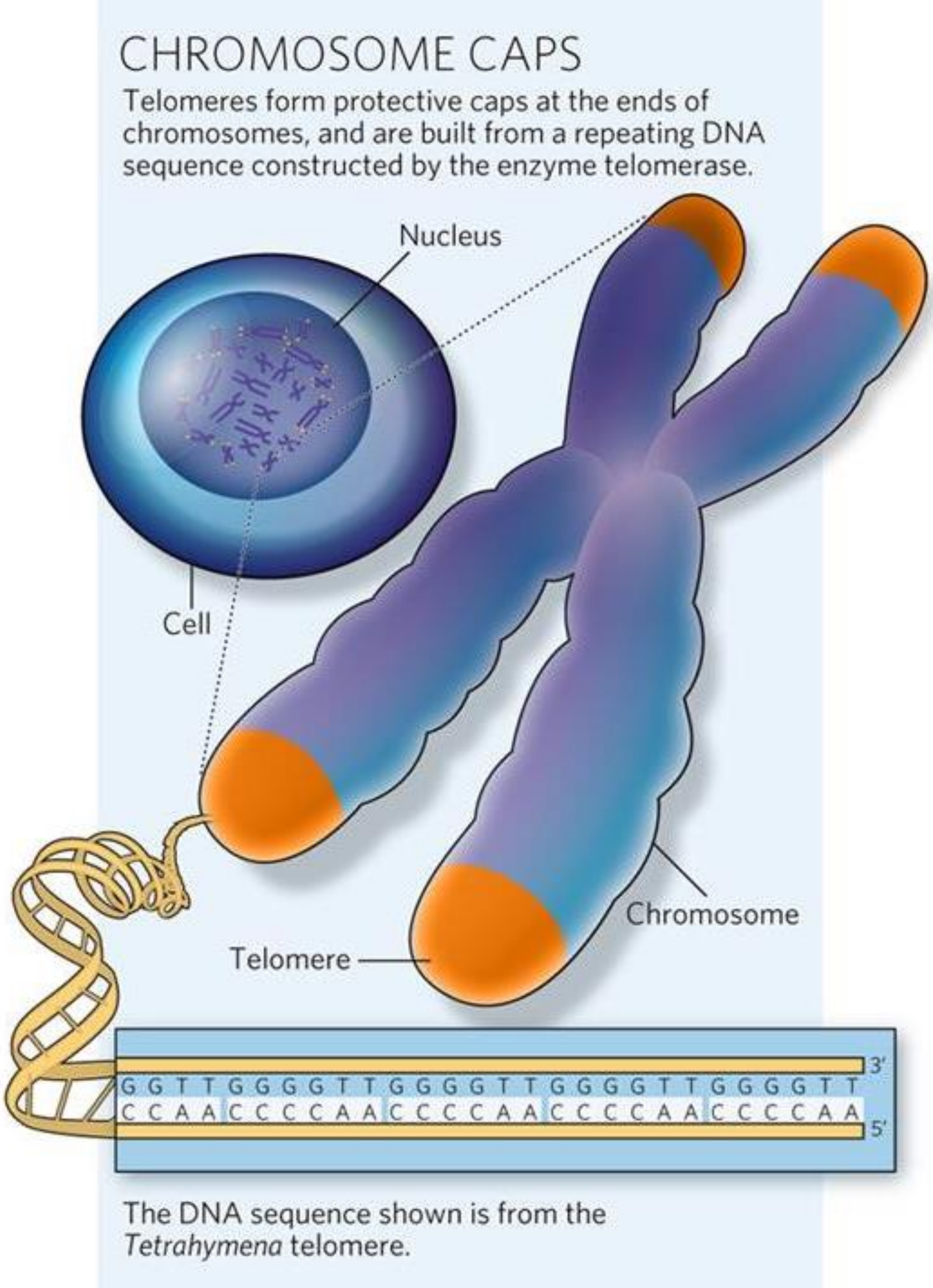
Rice centromeres contain a satellite repeat CentO and centromere-specific retrotransposon (CRR).

CENH3 (CENP-A)-associated and H3-associated nucleosomes

The CENH3-binding domain contains active genes (red bars), but with a lower density than the flanking domains.

Telomeres





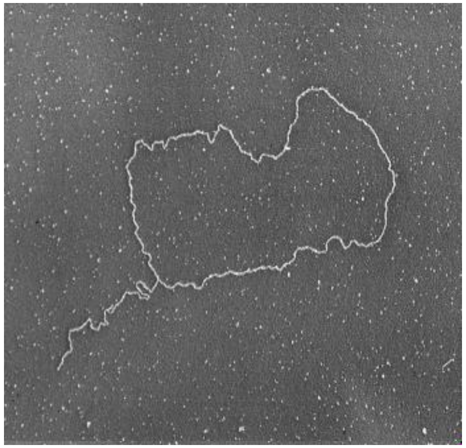
CHROMOSOME CAPS

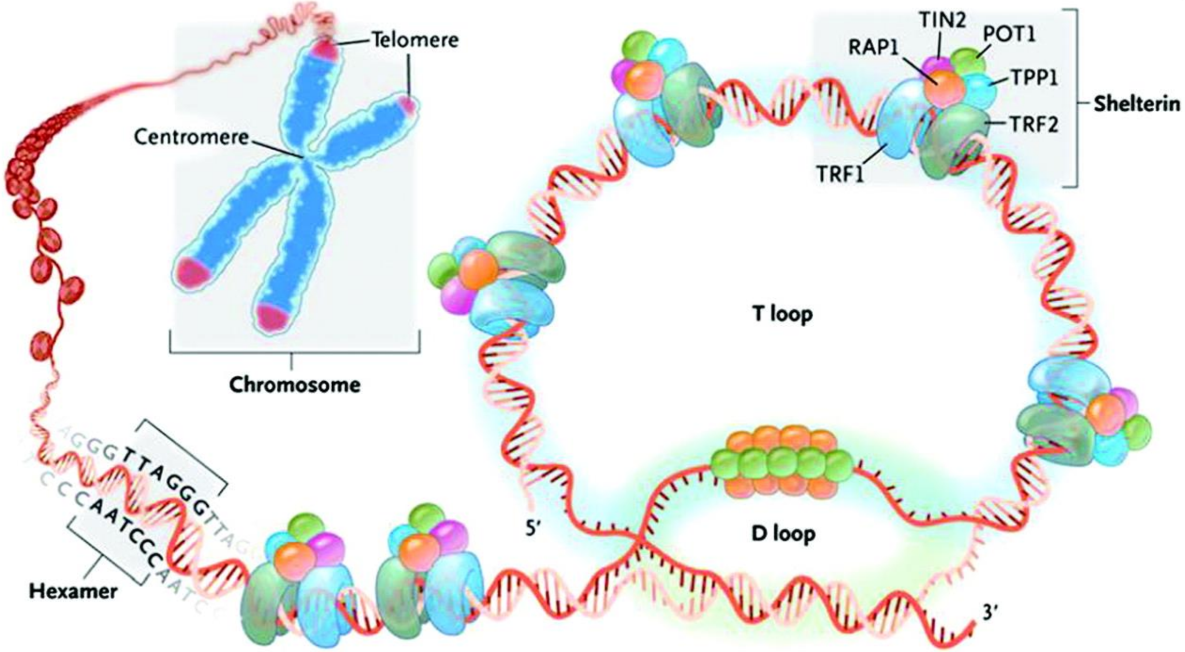
Telomeres form protective caps at the ends of chromosomes, and are built from a repeating DNA sequence constructed by the enzyme telomerase.

The DNA sequence shown is from the *Tetrahymena* telomere.

Eukaryotic telomeres

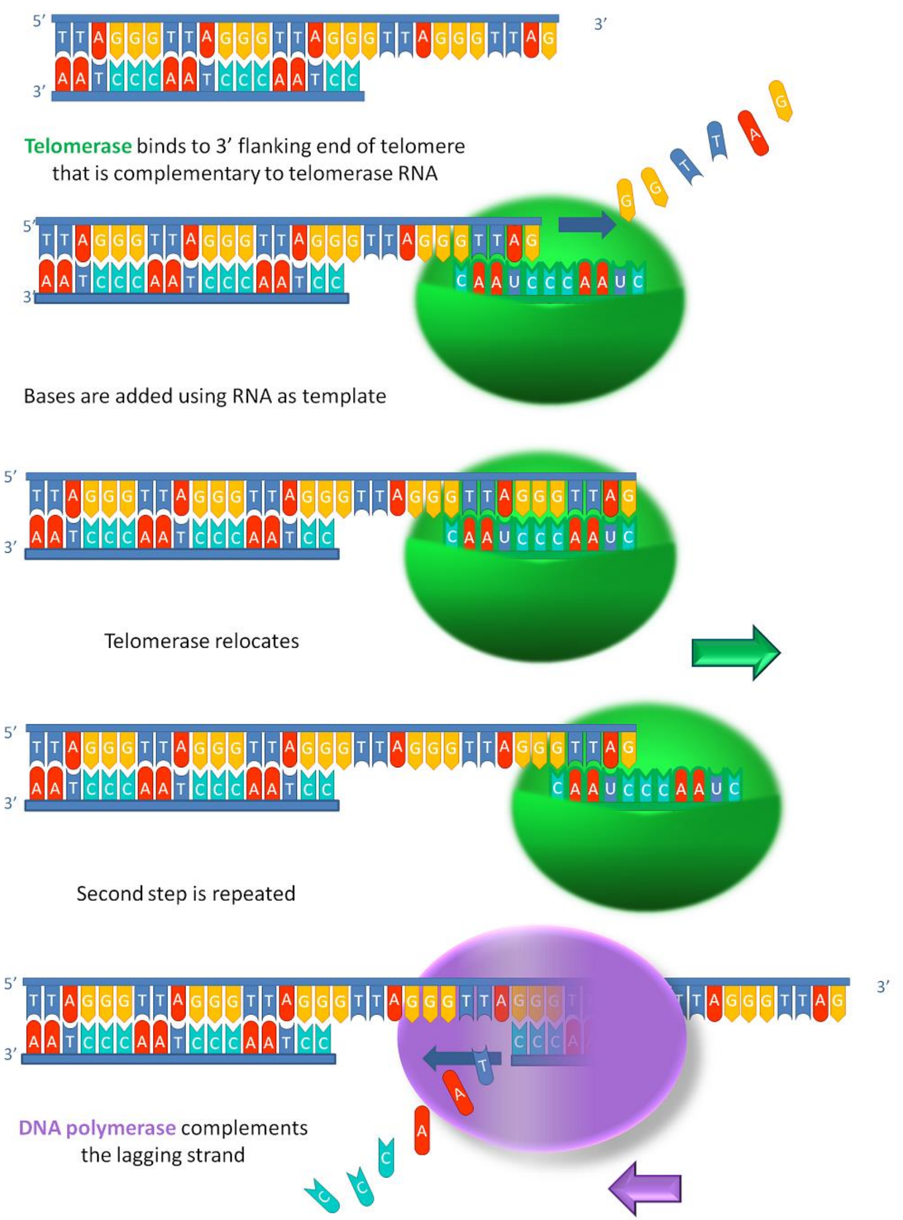
- solving chromosome shortening (loss of DNA sequences)
- protects against DNA repair (repair of double-strand breaks)
- evolutionary conserved telomeric repeats (e.g., TTAGGG)
- telomere-binding proteins (shelterin complex)
- synthesis by the telomerase enzyme





Telomeres are made by telomerase

- ribonucleoprotein, enzyme
- composed of own RNA and reverse transcriptase (TERT)
- adds telomeric repeats (e.g. TTAGGG in all vertebrates) to the 3‘ end of DNA strands at the ends of eukaryotic chromosomes
- preventing constant loss of DNA sequences from chromosome ends



Telomerase binds to 3′ flanking end of telomere that is complementary to telomerase RNA

Bases are added using RNA as template

Telomerase relocates

Second step is repeated

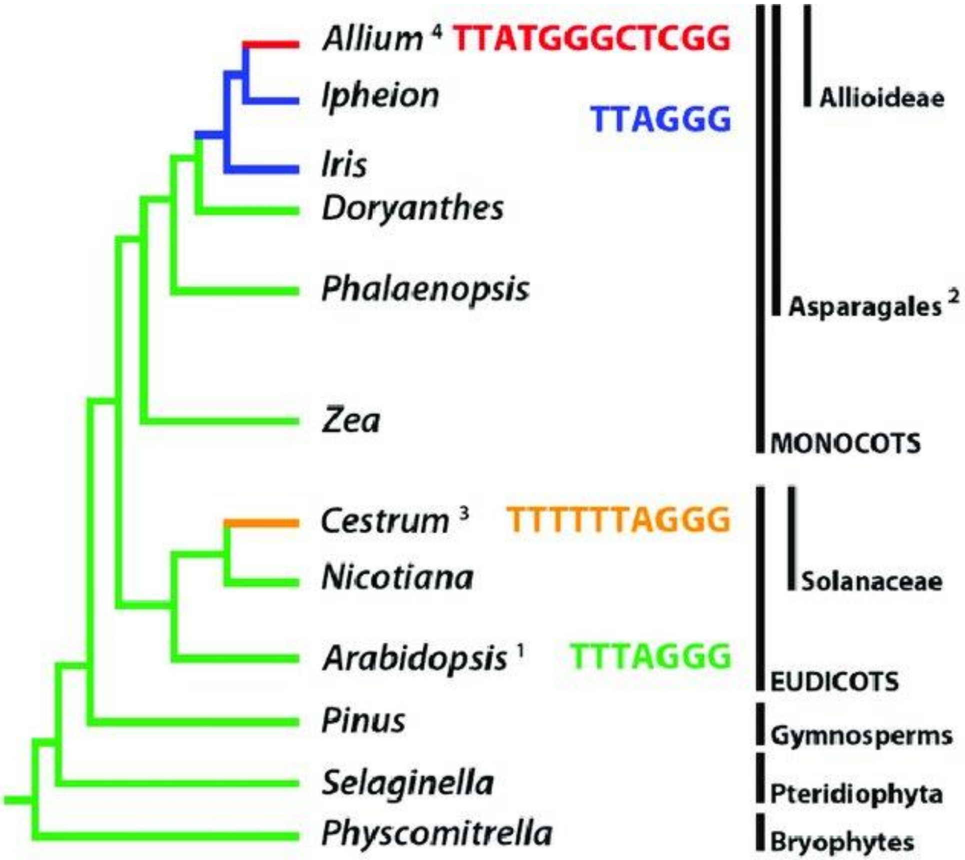
DNA polymerase complements the lagging strand

Telomeres of plants

Sequences of telomeric minisatellites

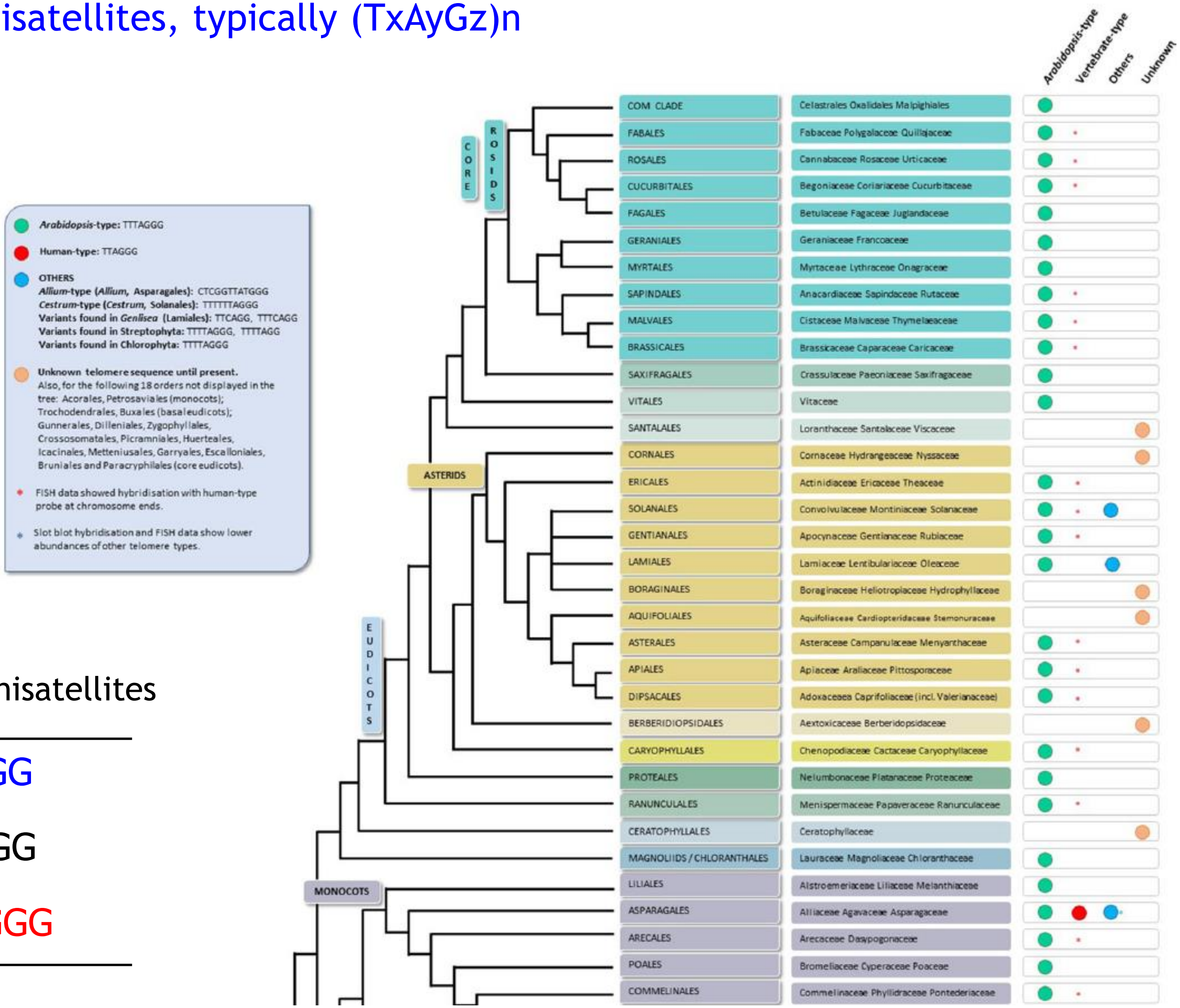
human	TTAGGG
Tetrahymena	TTGGGG
Arabidopsis	TTTAGGG

Human-type telomeric repeat and unusual telomeric motifs in land plants



Telomere sequences in land plants: Arabidopsis-type with some exceptions

- tandemly arranged minisatellites, typically (TxAyGz)n



Sequences of telomeric minisatellites

human	TTAGGG
Tetrahymena	TTGGGG
Arabidopsis	TTTAGGG

Peska V, Garcia S (2020), Front. Plant Sci. 11

Nucleolar Organizing Region
(ribosomal RNA genes on eukaryotic chromosomes)

rDNA = ribosomal DNA = genes coding ribosomal RNAs

- terminally on chromosomes or as the secondary constriction

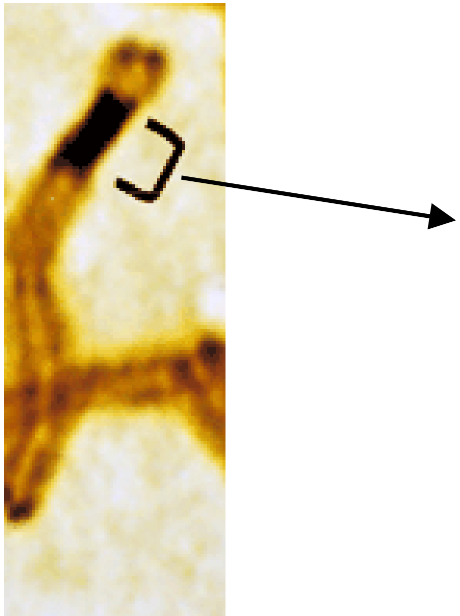
- routinely detected by FISH

- diagnostic value, position and the number usually species-specific

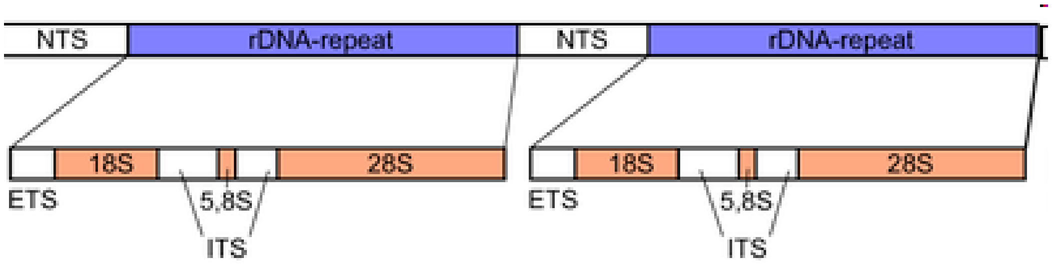
- NORs (45S rDNA) usually in different position on chromosome(s) than 5S rDNA

Physical mapping of 45S rDNA (red) and 5S rDNA (green) to metaphase chromosomes of *Larix leptolepis*. Chromosomes counterstained with DAPI (blue) (Zhang *et al*. 2010)

45S and 5S ribosomal DNA (rDNA)



structure of the 45S rDNA tandem repeat



18S, **5.8S**, and **28S** - genes coding **18S**, **5.8S**, and **28S** RNA molecules

NTS - nontranscribed spacer

ETS - external transcribed spacer

ITS - internal transcribed spacers 1 and 2

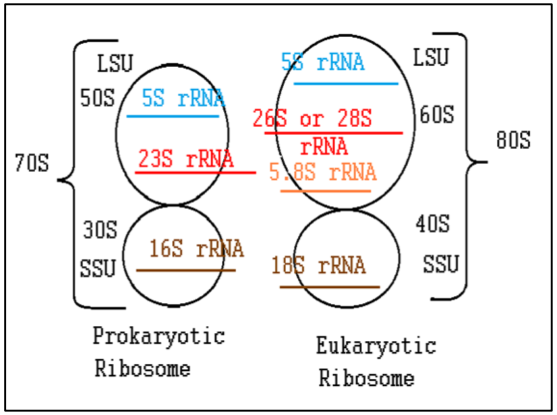
transcription of rDNA→ 45S pre-rRNA→ processing→ 18S RNA, 5.8S and 28S RNA molecules

Ribosomes – proteins and RNA molecules.

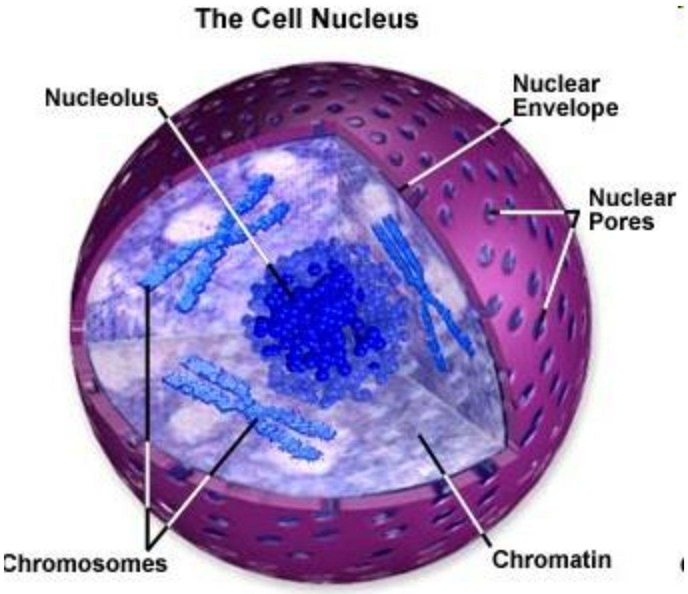
In eukaryotes: small ribosomal subunit (40S): 18S rRNA

large subunit (60S): 5.8S, 28S rRNA and 5S rDNA

In eukaryotes, the 5S rRNA gene is separated from the 45S rRNA genes. But together in *Artemisia*, gymnosperms, and some other plants.



Nucleolus



- ribosomal DNA (rRNA genes) is transcribed and ribosomes are assembled within the nucleolus

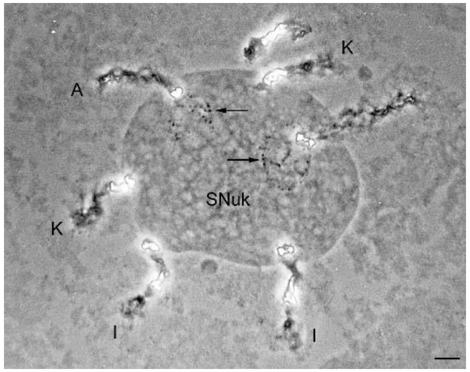
- ribosomes are exported to the cytoplasm. They remain free or associate with the endoplasmic reticulum (rough endoplasmic retictulum)

- one or several nucleoli in a nucleus

- after a cell division, a nucleolus is formed around nucleolar organizing region (NOR) on some chromosomes (chromosomes are brought together by nucleolar organizing regions)

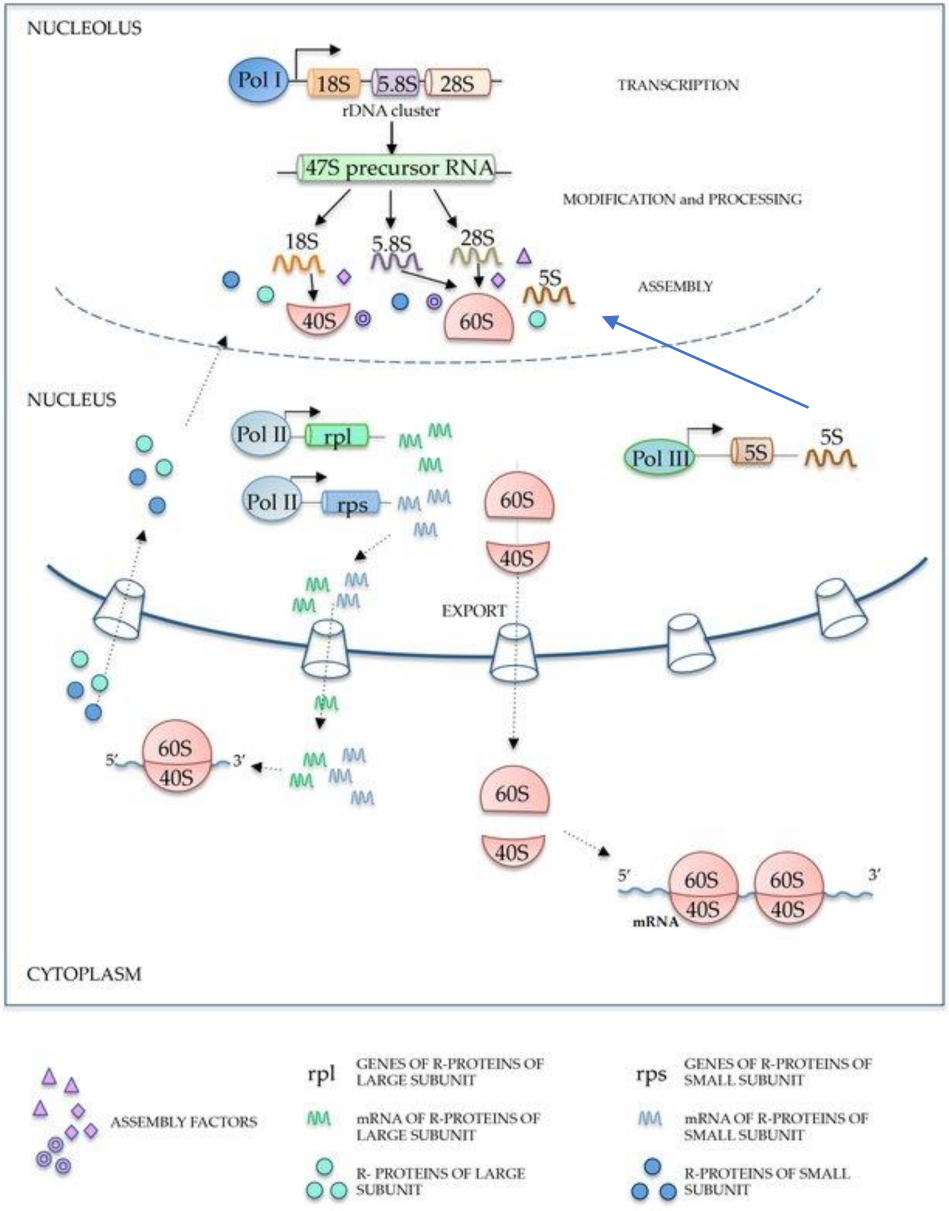
- cell division: nucleolus disappears

Spatial context of rDNA transcription and ribosome assembly



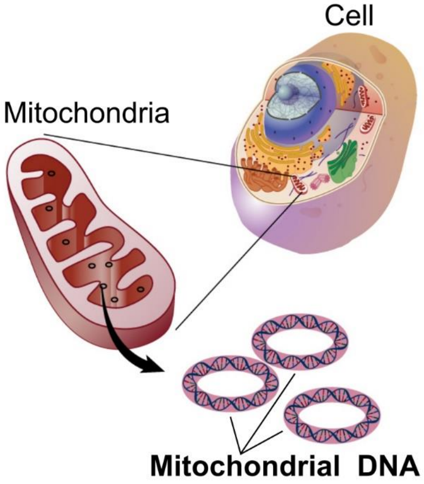
Phaseolus. NORs of 6 chromosomes (pair A, I and K) and the nucleolus.

The small dot-like structures (arrows) are telomeric heterochromatin.

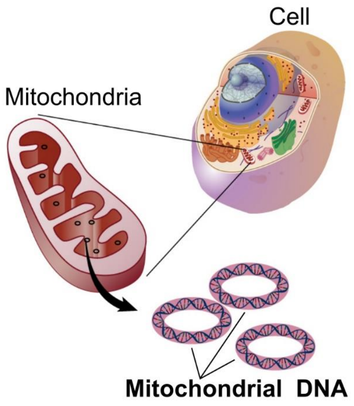


Extra-nuclear genomes and extra-chromosomal DNA in eukaryotes

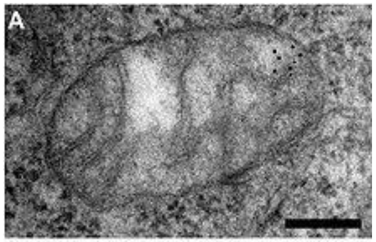
(outside the chromosomes and typically also outside the nucleus)



Mitochondrial genome (mtDNA)



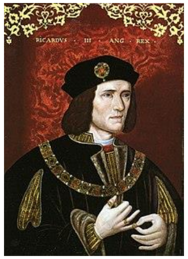
- human mtDNA includes 16,569 base pairs and encodes 13 proteins, 2 rRNAs, 22 tRNAs
- animals: usually circular DNA molecule, but also linear genome
- plants and fungi (circular, rarely linear), 3 types of mt genome:
  - a circular genome that has introns (19 to 1 000 kb)
  - a circular genome (20 – 1000 kb) that also has a plasmid-like structure (1 kb)
  - a linear genome made up of homogeneous DNA molecules



- *Silene conica*: enormous mtDNA genome - 11,300,000 bp
- 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

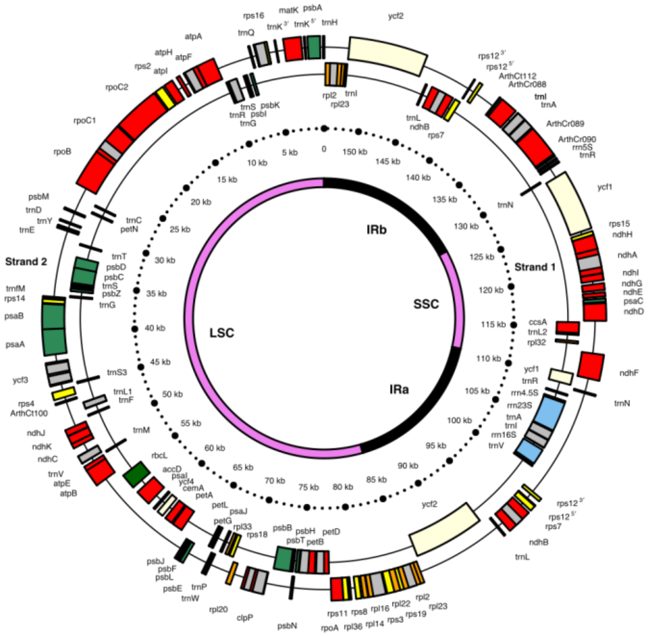
The remains of King Richard III were identified by comparing his mtDNA with that of two matrilineal descendants of his sister.



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; 120 000 – 170 000 bp 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

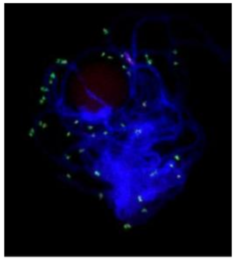
Chloroplast genome (plastome)

- 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 genes 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, 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 →



- the replicon can be used for crop improvement (engineering of synthetic replicons)

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