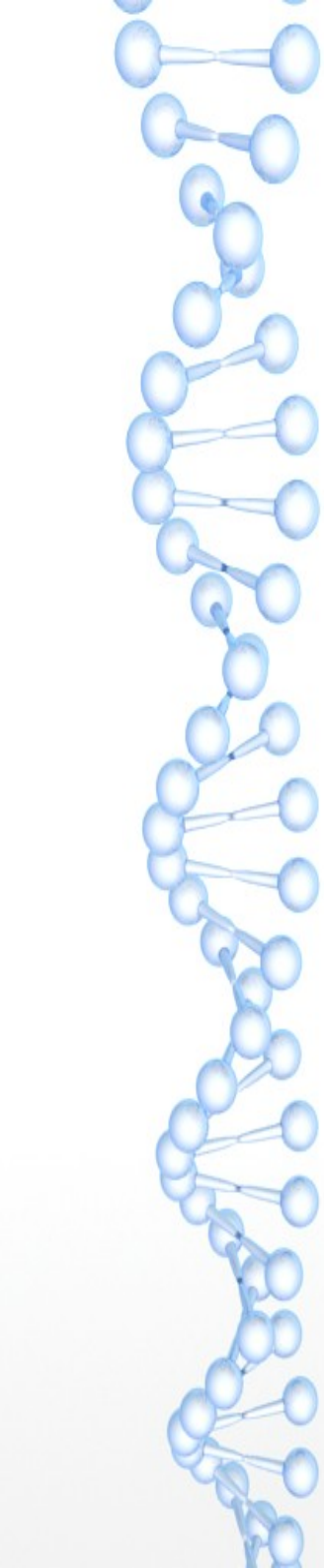
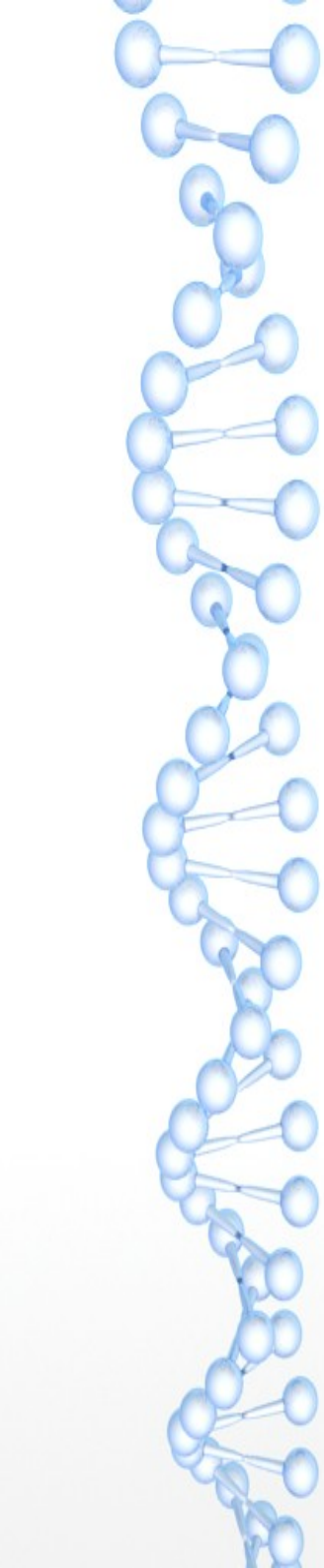


I. Structure and diversity of the genomes



I. Structure and diversity of the genomes

Definition of the term “genome”

Background question of the module

What is a genome?

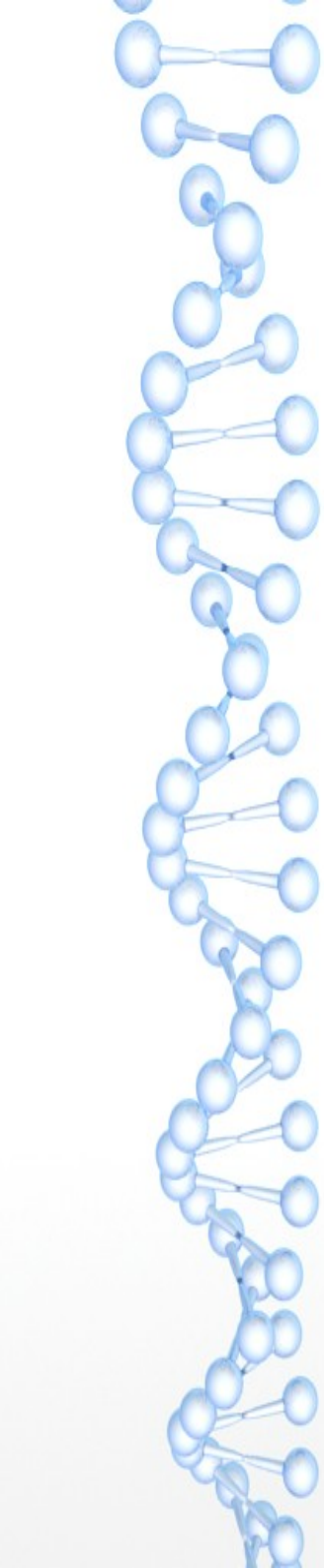
Several commonly used definitions

Genome is

- the entire set of **DNA** instructions found in a **cell**.
- genetic material of haploid set of **chromosomes**
- **information repository** of organism
- all **hereditary material** of organism

Definitions overlap significantly, but not completely!

Goldman and Landweber, 2016



I. Structure and diversity of the genomes

Genome size

Variation of genome size

How small is the smallest genome on the Earth?

How big is the biggest one?

Your suggestions?

Variation of genome size

How small is the smallest genome on the Earth?

How big is the biggest one?

Your suggestions?

Hints:

human	3.3	Gbp
drosophila	240	Mbp
baker yeasts	12	Mbp
E. coli	4.2	Mbp

Gbp = giga base pairs
Mbp = mega base pairs

Biggest genome

Paris japonica

キヌガサソウ , *Kinugasasō*

***150 Gbp** / 152.23 pg*

(Pellicer et al, 2010)

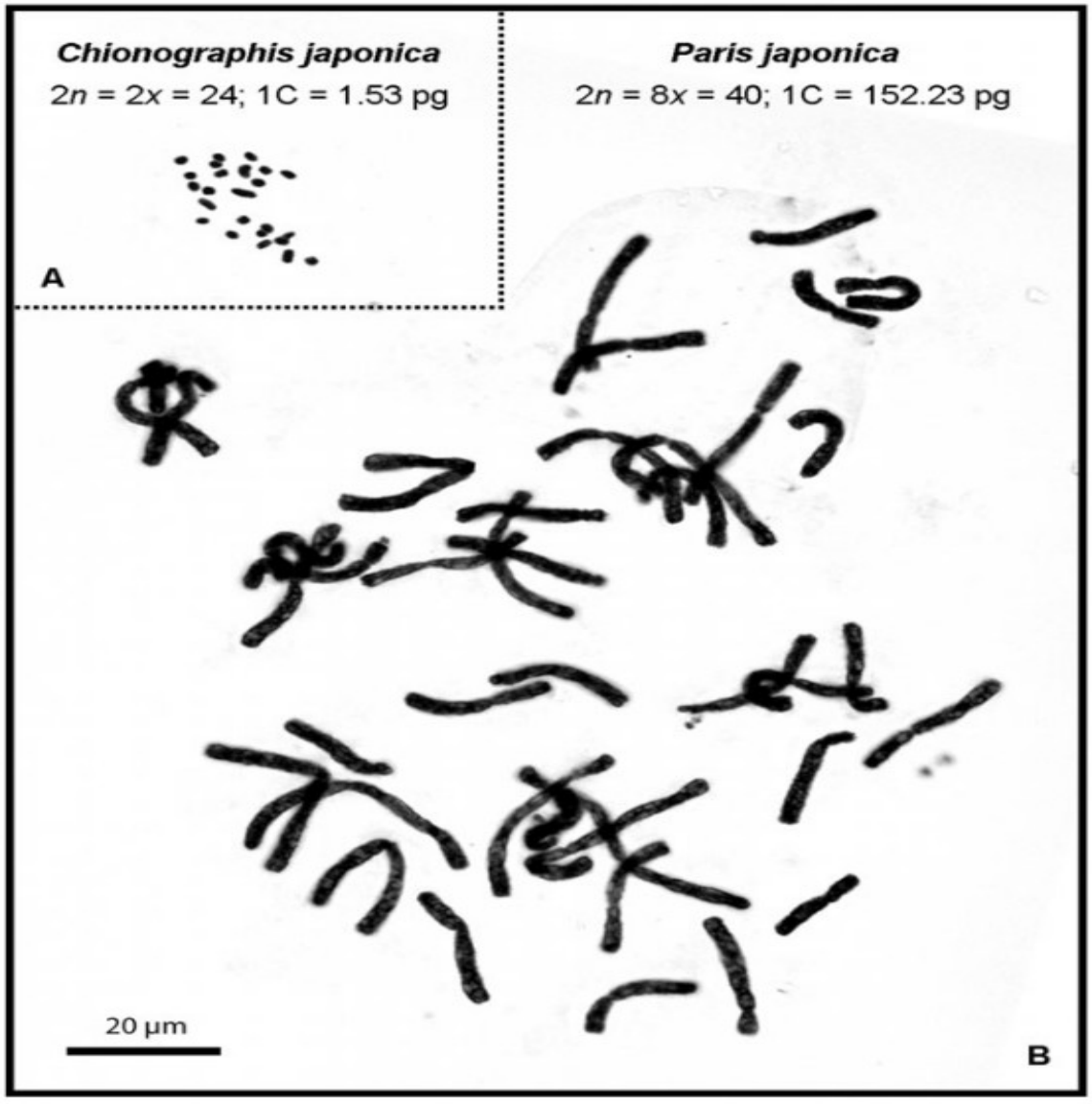
Protopterus aethiopicus

Marbled lungfish

130 Gbp / 132.83 pg

(Pedersen, 1971)

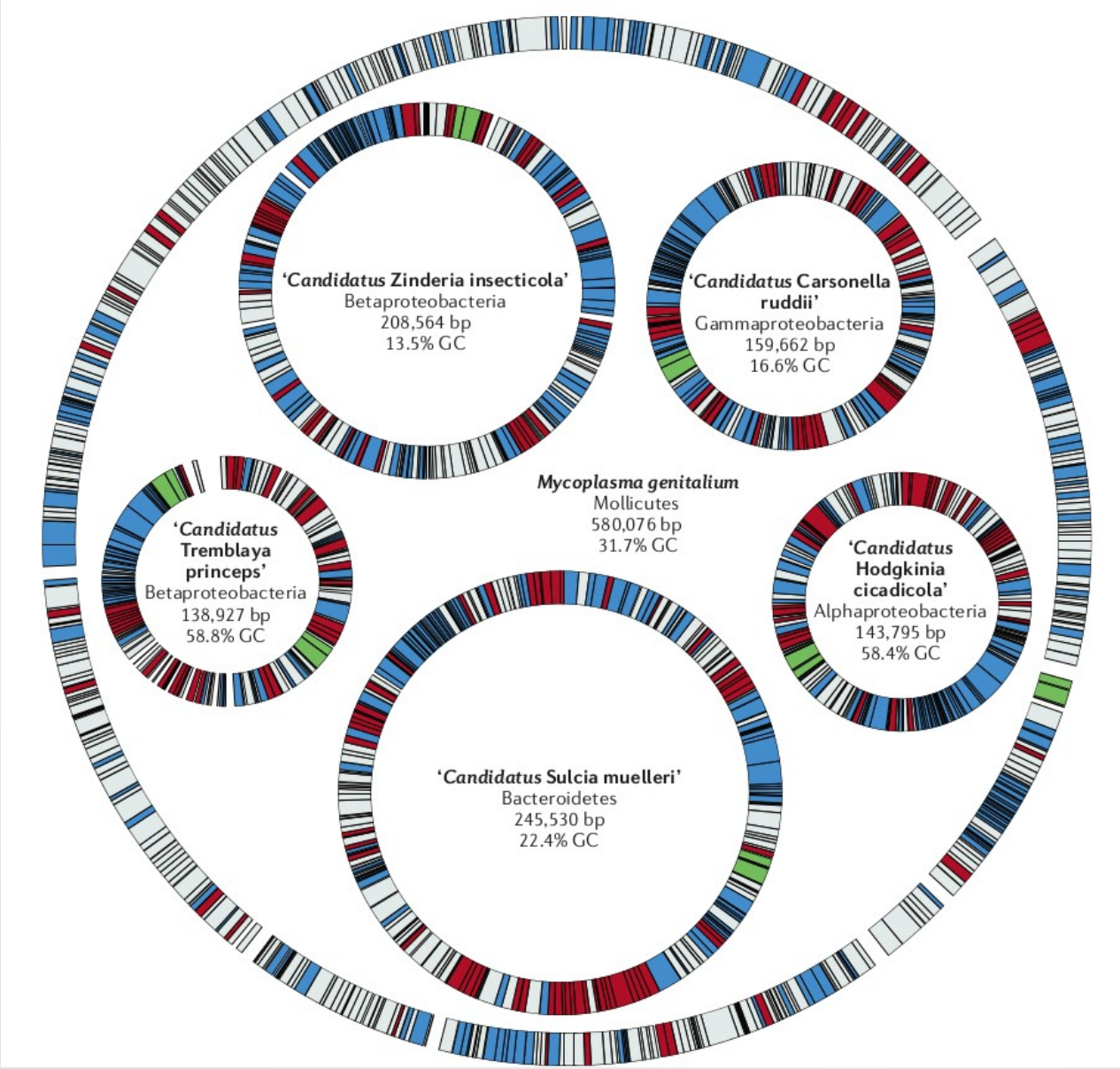
100x genome size difference in the cell



150 Gbp vs 1.5 Gbp

Pellicer et al, 2010

Small bacterial genomes



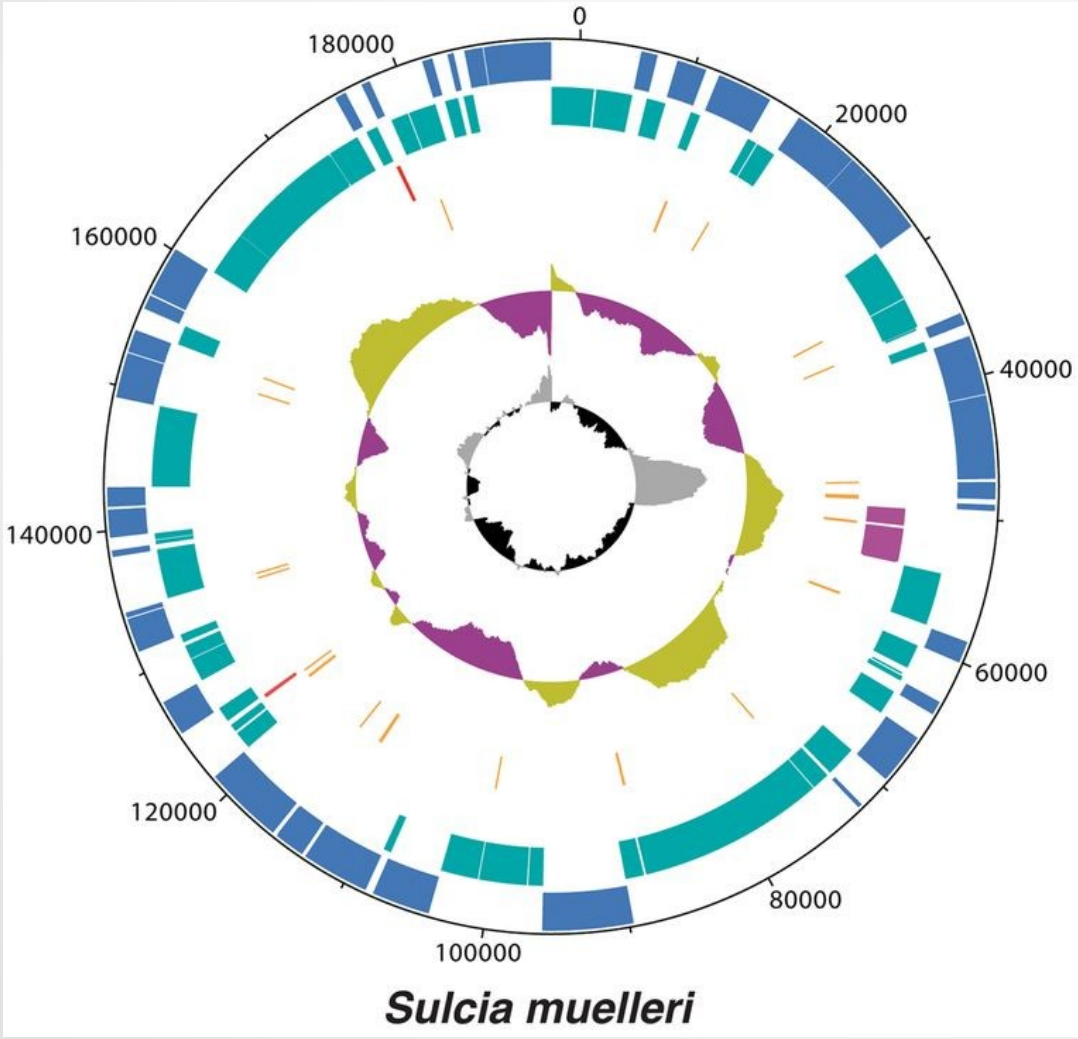
Free-living *M. genitalium*
vs
endosymbionts

580 kbp
vs
138 - 245 kbp

McCutcheon and Moran, 2012

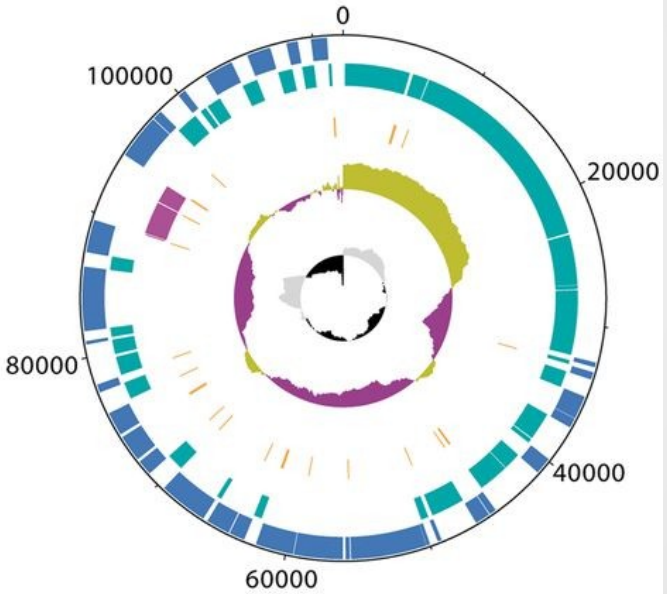
Smallest bacterial genomes

190 kbp



Sulcia muelleri

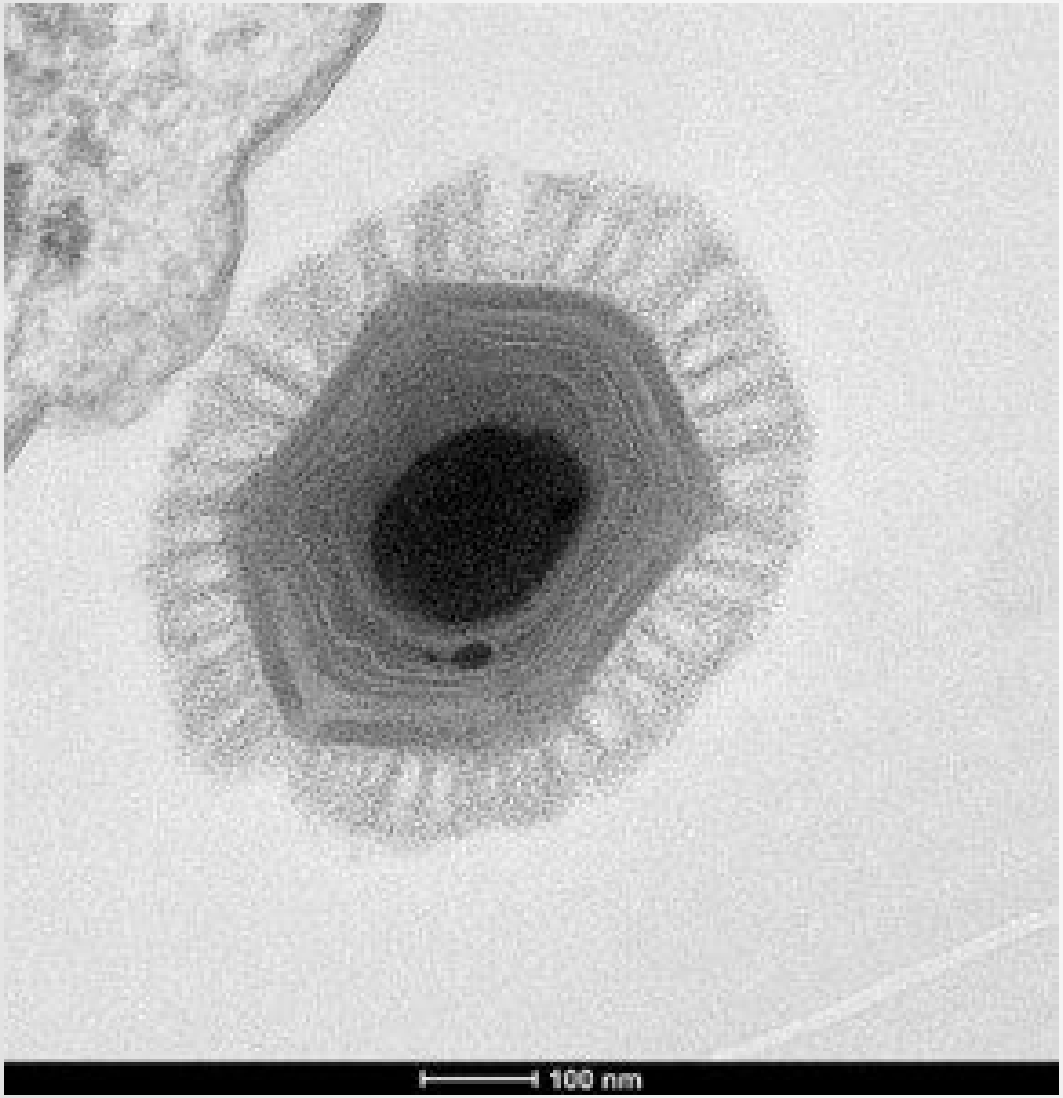
112 kbp



Nasuia deltocephalinicola

Bennett and Moran, 2013

Biggest viral genome

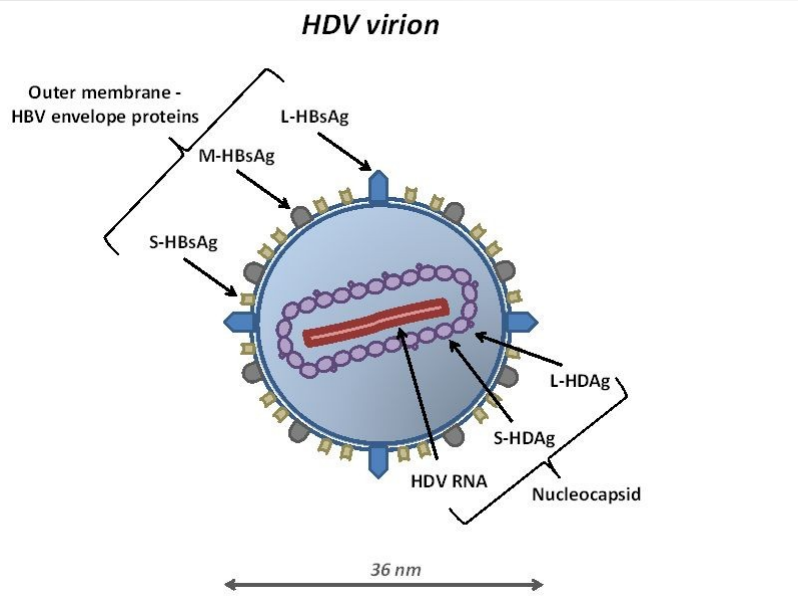


Megavirus chilensis

1.26 Mbp
1120 genes

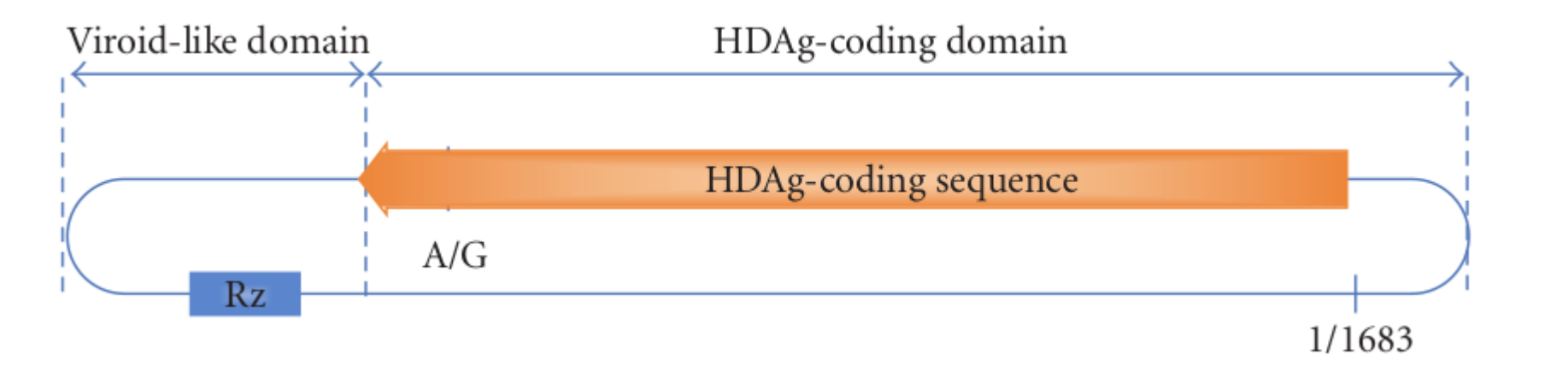
Legendre et al, 2012

Smallest viral genome



Hepatitis D virus
1683 bp

Symbiont of
Hepatitis B virus



Bennett and Moran, 2013

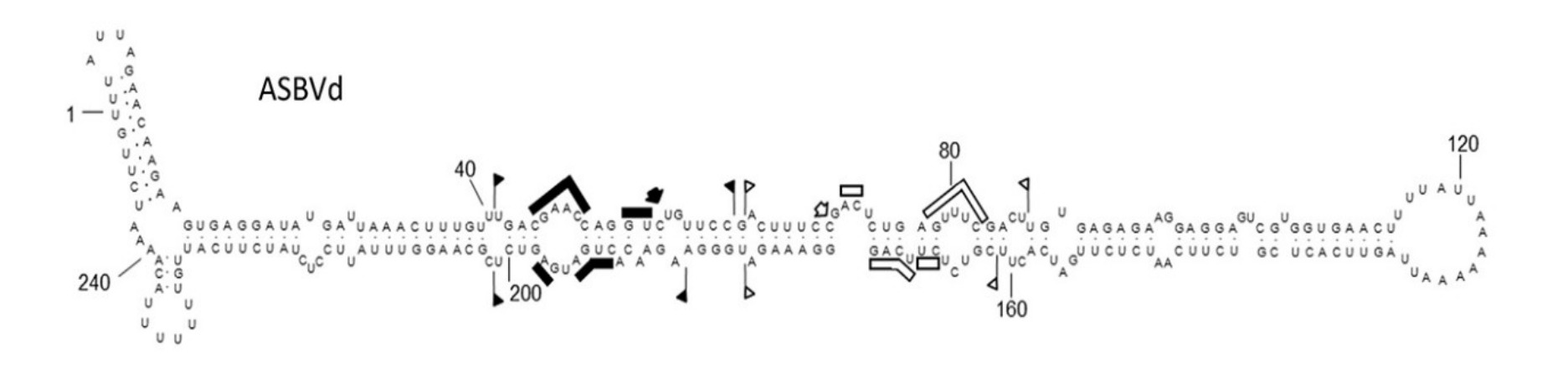
Recently was split in 8 species

Smallest genome

Avocado sunblotch viroid

248 bp of circular RNA.

No capsid. No envelop of any type.
ONLY pure naked RNA.
Nothing more.



1 molecule = 1 organism(?)

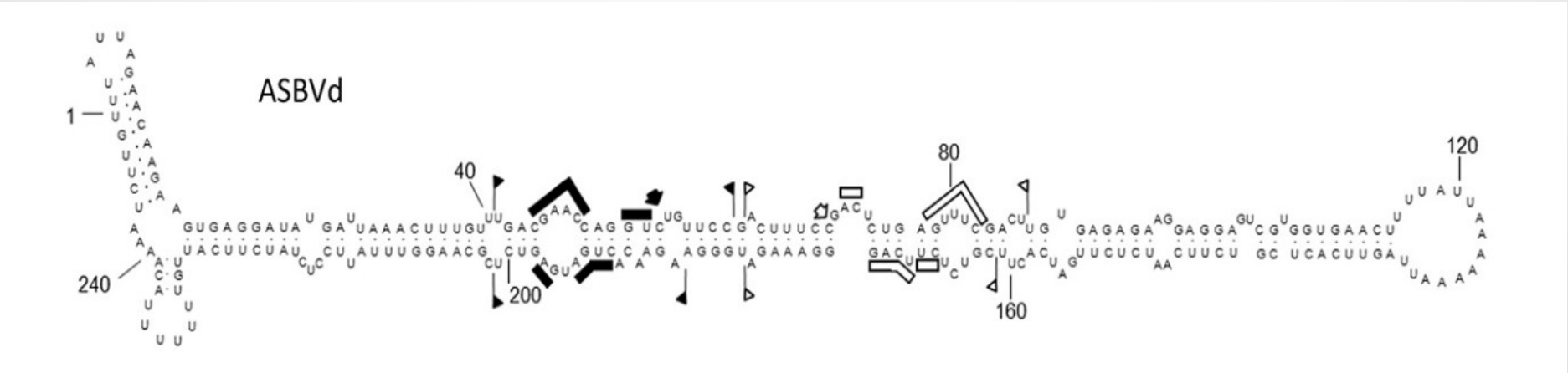
Flores et al, 2022

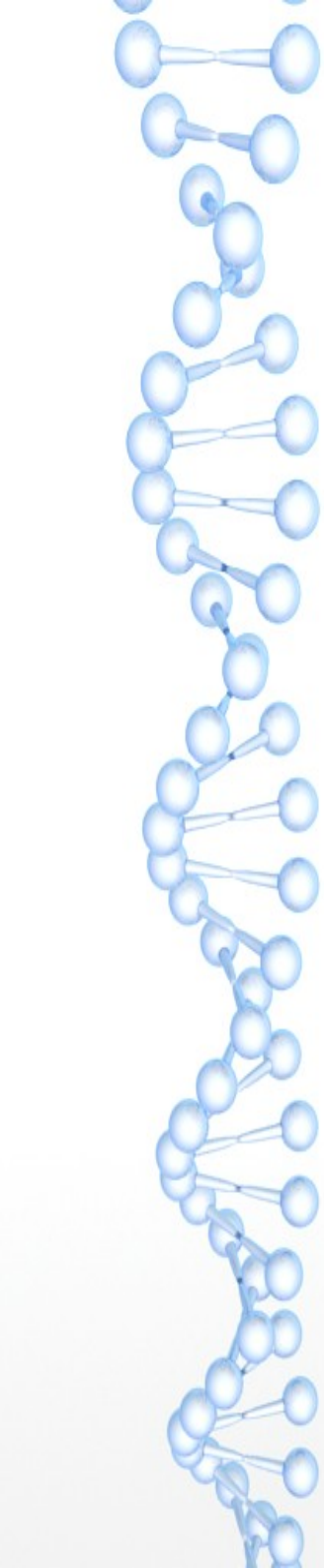
Smallest and biggest genomes

Plant vs viroid
Multicellular complex organism
vs
short circular RNA only organism

150 000 000 000 bp
vs
248 bp

vs





I. Structure and diversity of the genomes

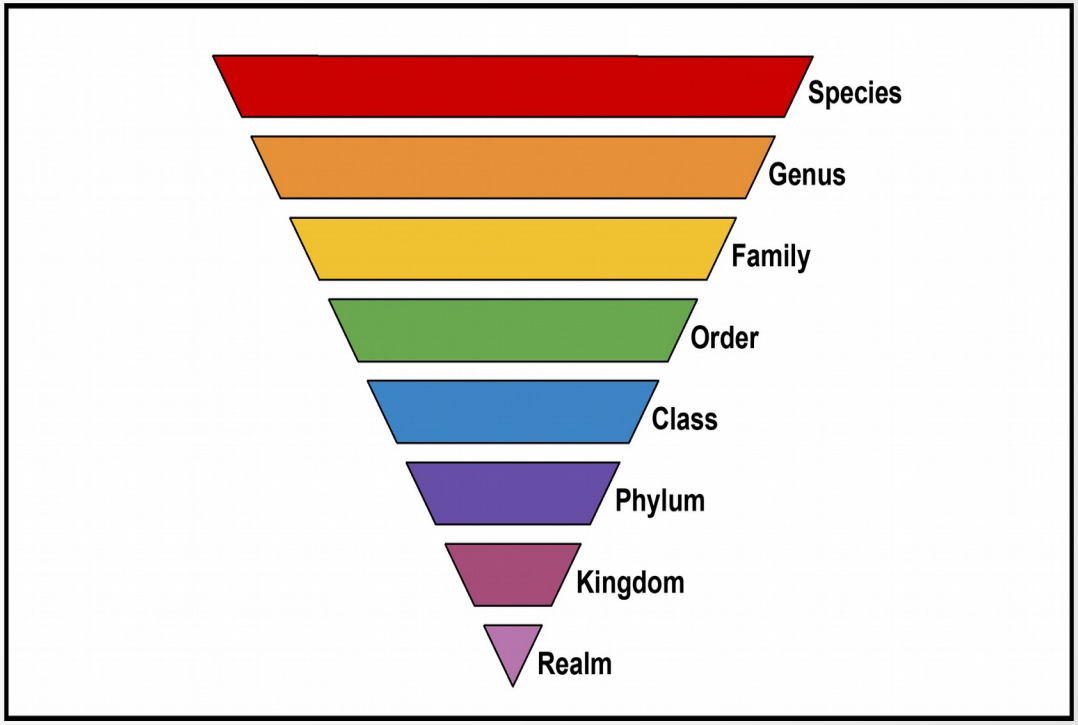
Viruses and(?) viroids

Taxonomy of viruses and(?) viroids

ICTV
International Committee on Taxonomy of Viruses

Latest Taxonomy Release #37 (2021)

- 8 realms
- 1 class without higher classification
- 19 families without higher classification
- 2 genera without higher classification

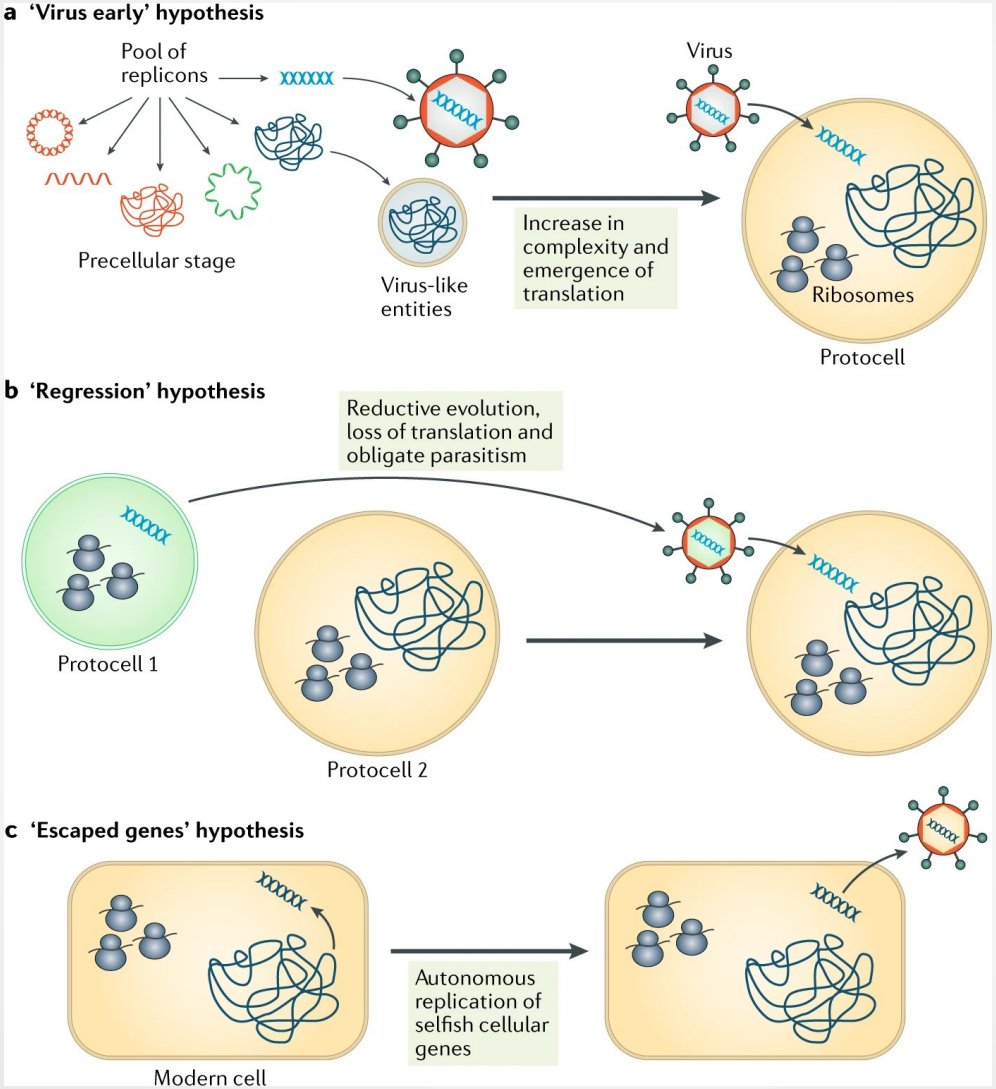


Classification of viruses is very difficult.

Some realms are not monophyletic!

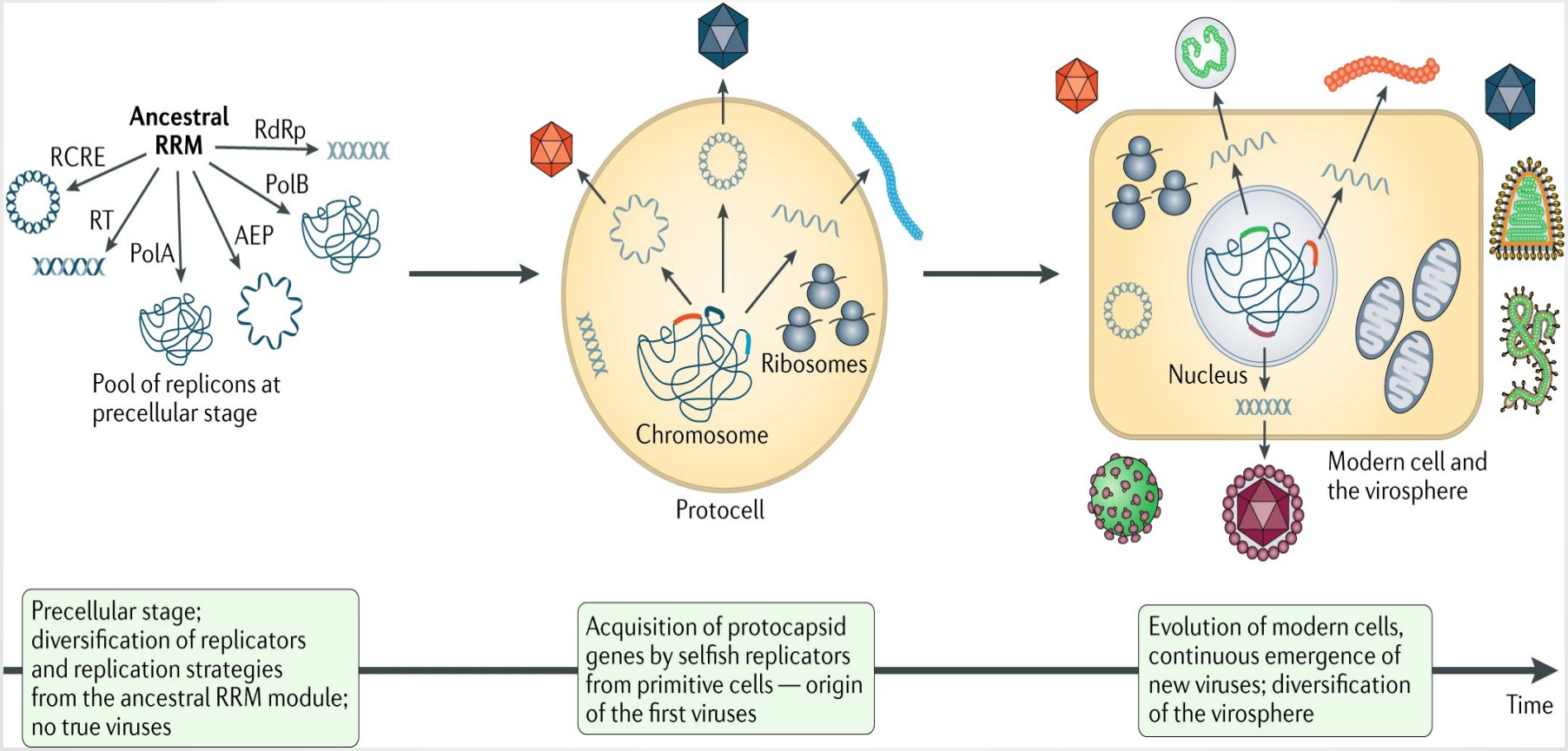
https://ictv.global

Origin of viruses: three basic scenarios



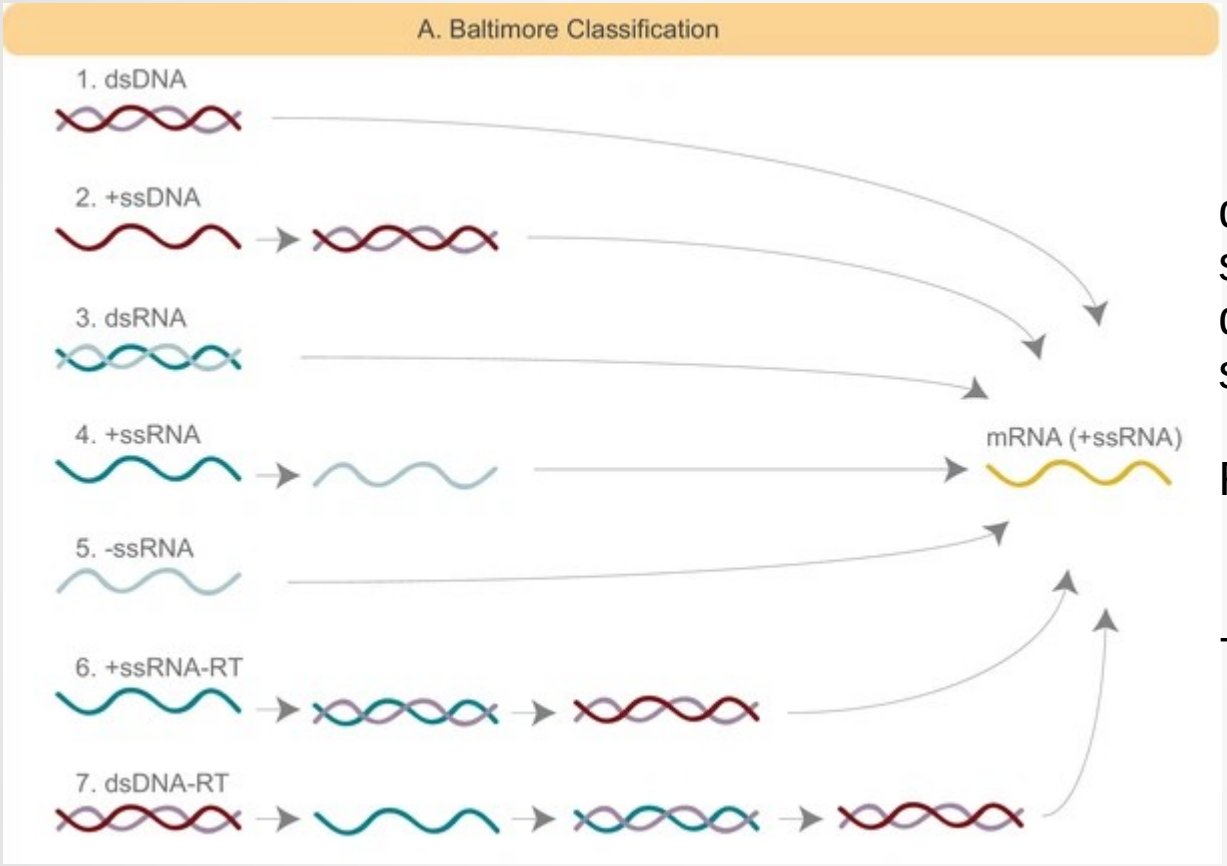
Krupovic et al, 2019

Origin of viruses: chimeric scenario



Krupovic et al, 2019

Types of viral genomes (1)



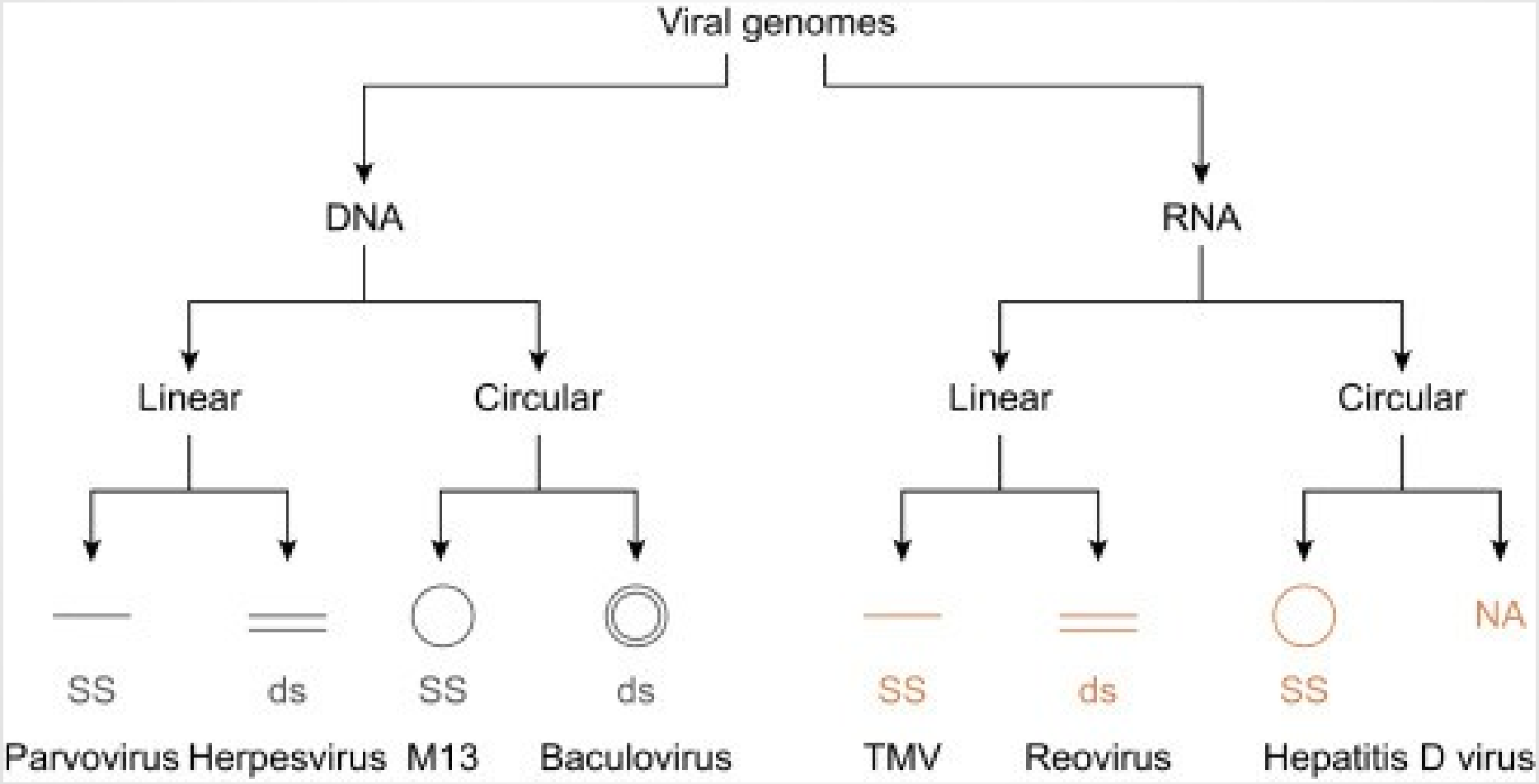
dsDNA - double stranded DNA
ssDNA - single stranded DNA
dsRNA - double stranded DNA
ssRNA - single stranded RNA

RT - reverse transcription
(RNA → DNA)

+/- - relation to expressed mRNAs

Mahmoudabadi et al, 2018

Types of viral genomes (2)



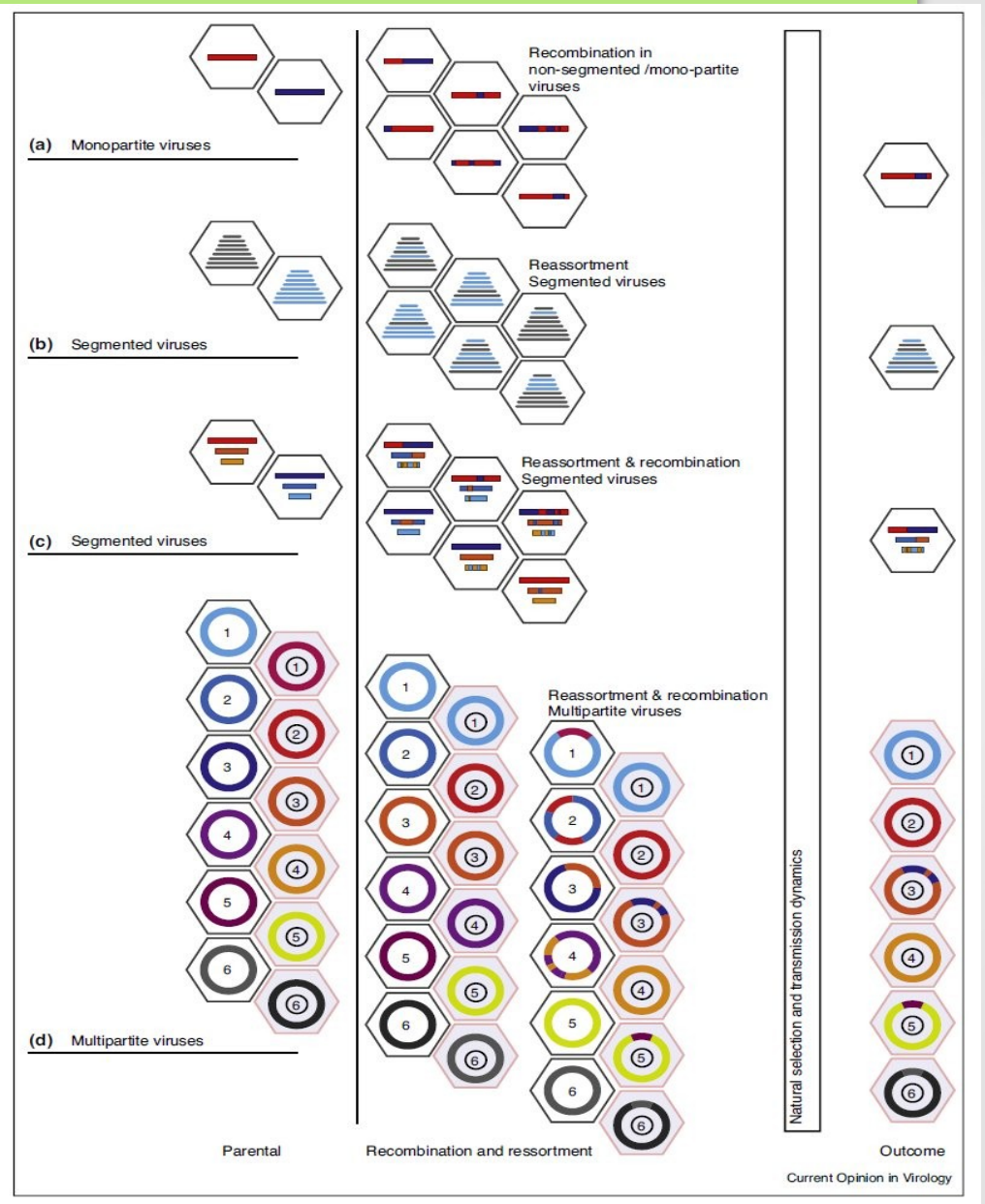
Marintcheva, 2018

Types of viral genomes (3)

nonsegmented genomes
genome = single molecule

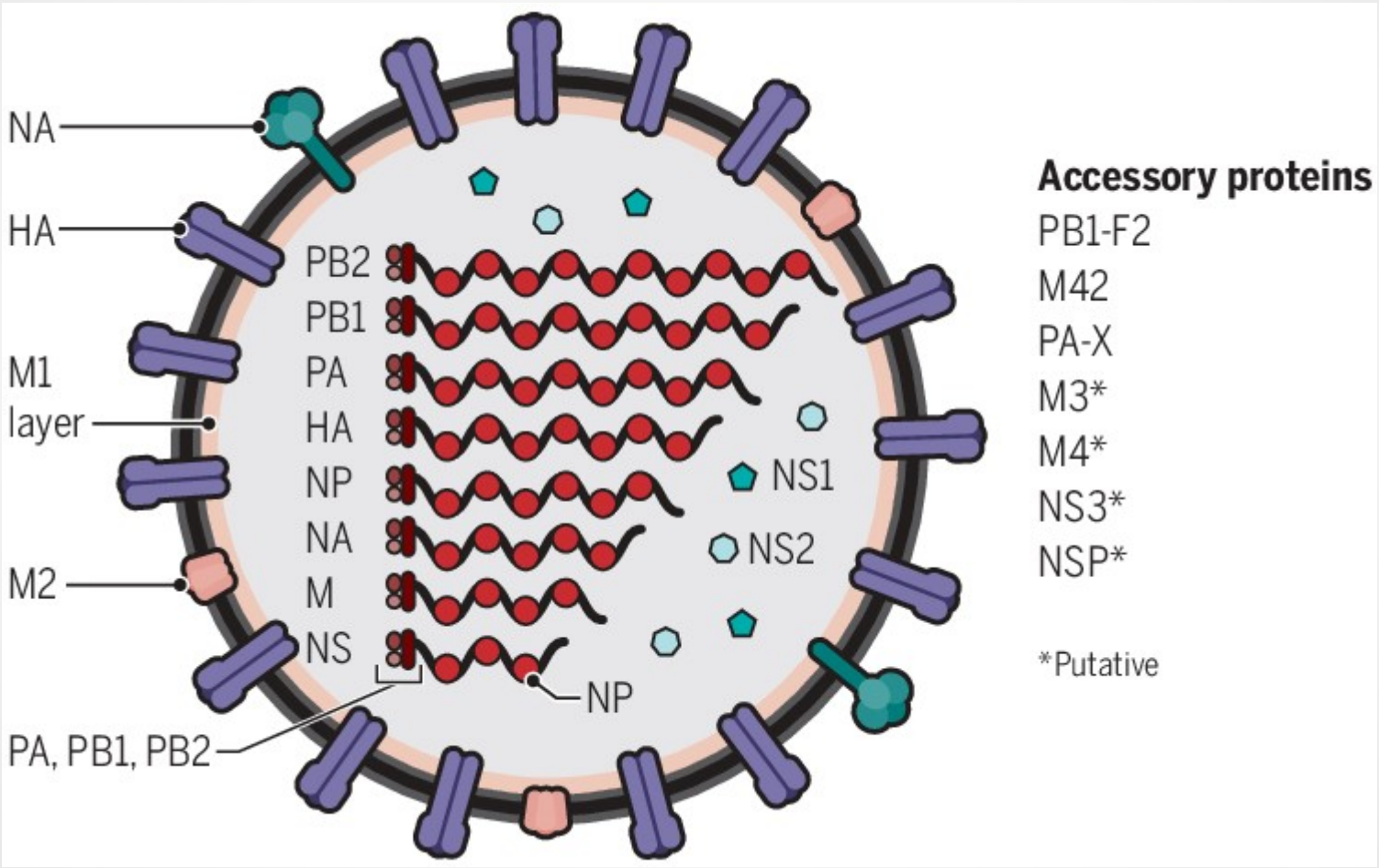
segmented genomes
genome = several molecules, must be packed in a single virion

multipartite genomes
genome = several molecules, is packed in several virions



Varsani et al, 2018

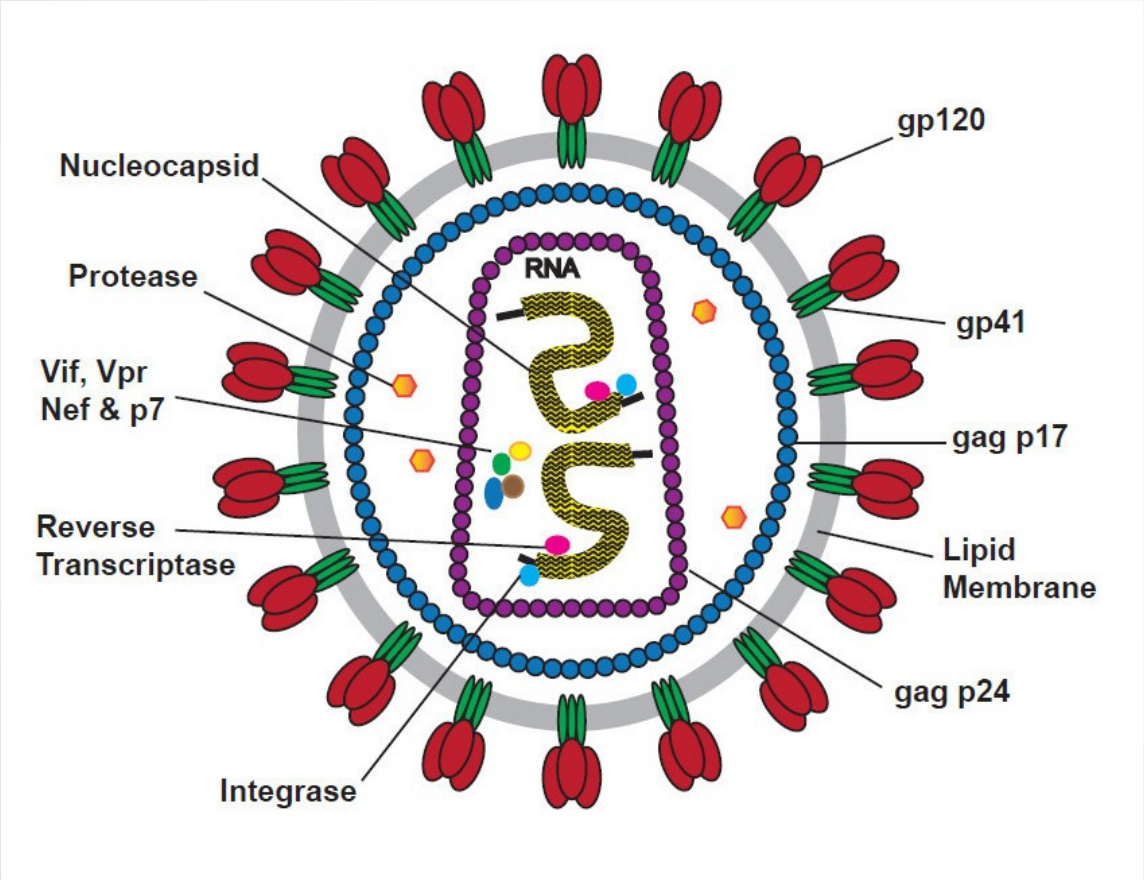
Genome of Influenza A virus



-ssRNA linear segmented genome
8 genome segments

Taubenberger et al, 2019

Human immunodeficiency virus (HIV)

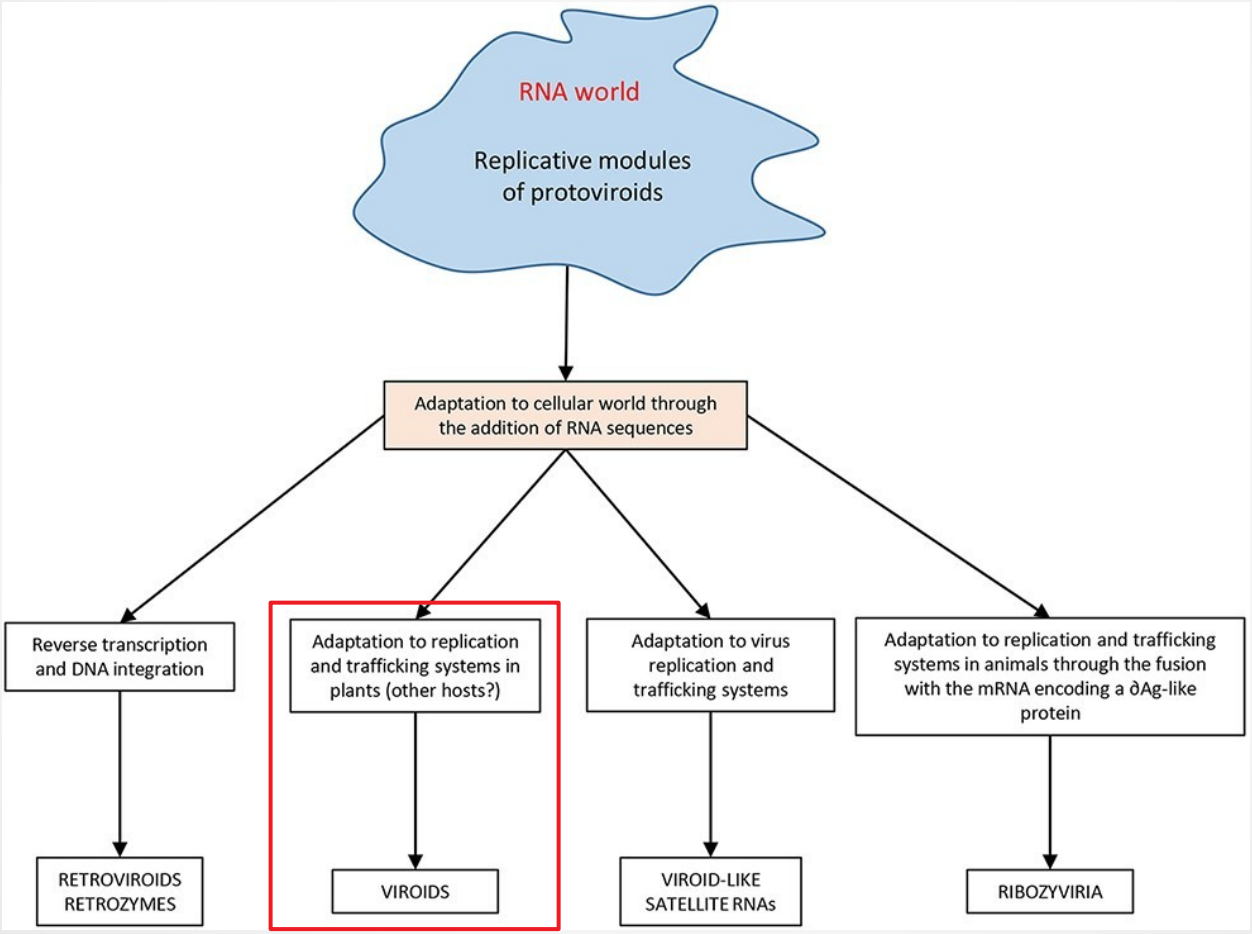


+ssRNA-RT linear
non-segmented genome

but in two copies in each virion

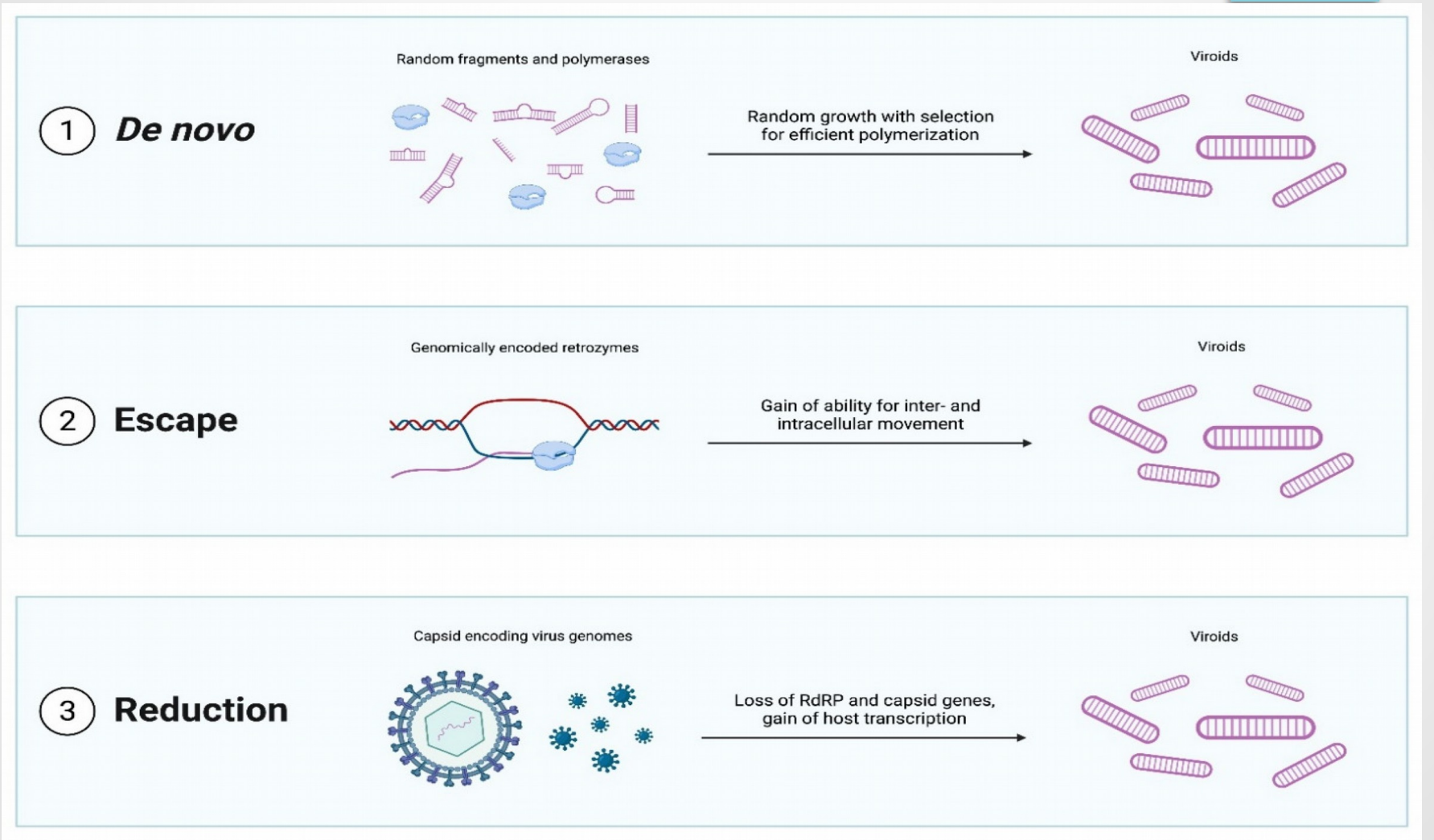
https://www.eenzyme.com/HIV-research-tools.aspx

Origin of viroids (1): remnants of RNA world?



Flores et al, 2022

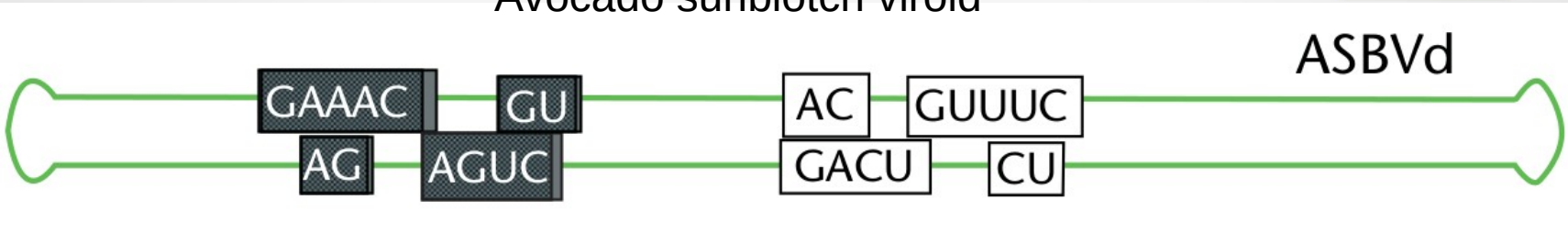
Origin of viroids (2): “modern” scenarios



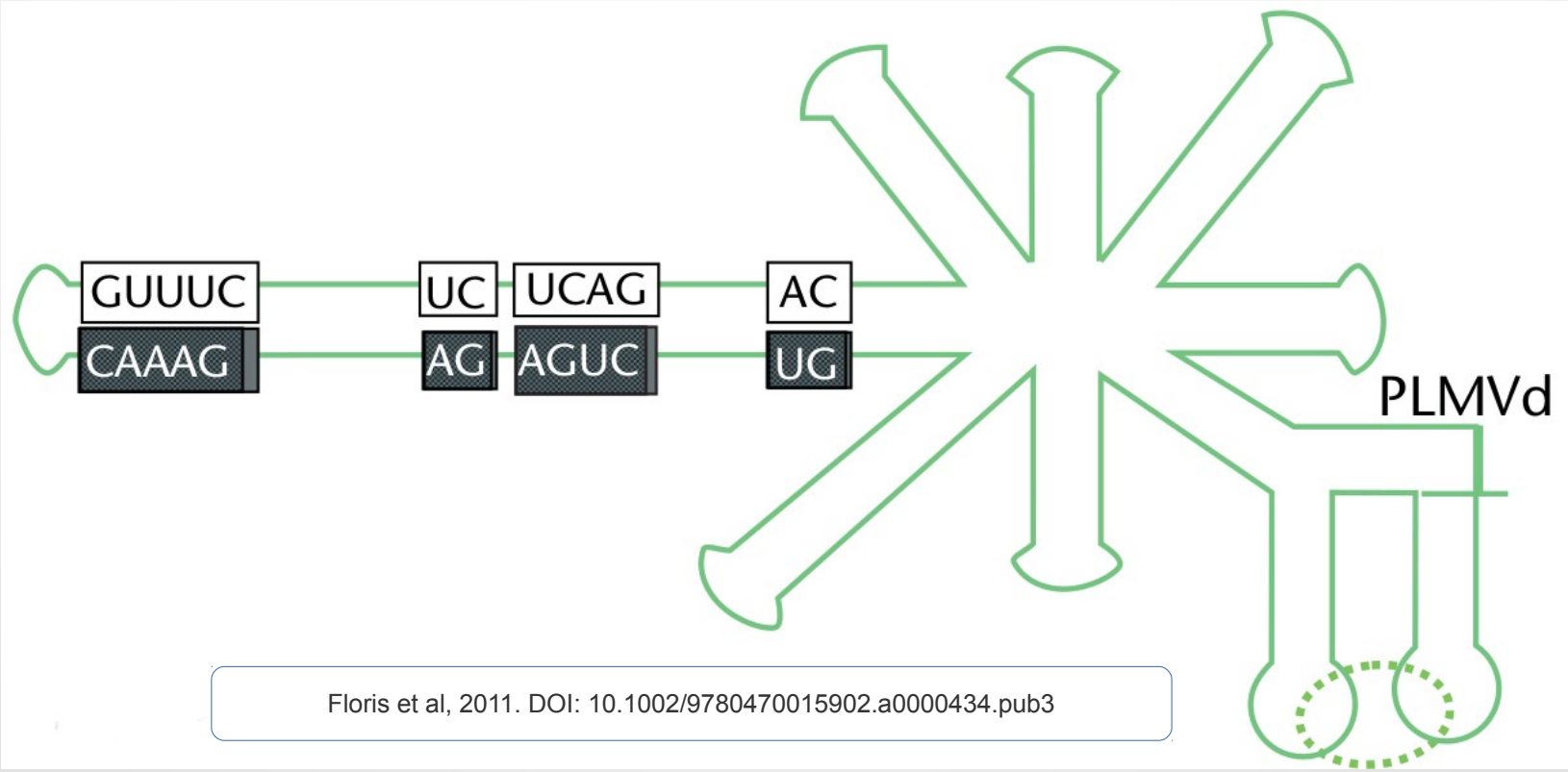
Lee and Koonin, 2022

Structure of viroids

Avocado sunblotch viroid

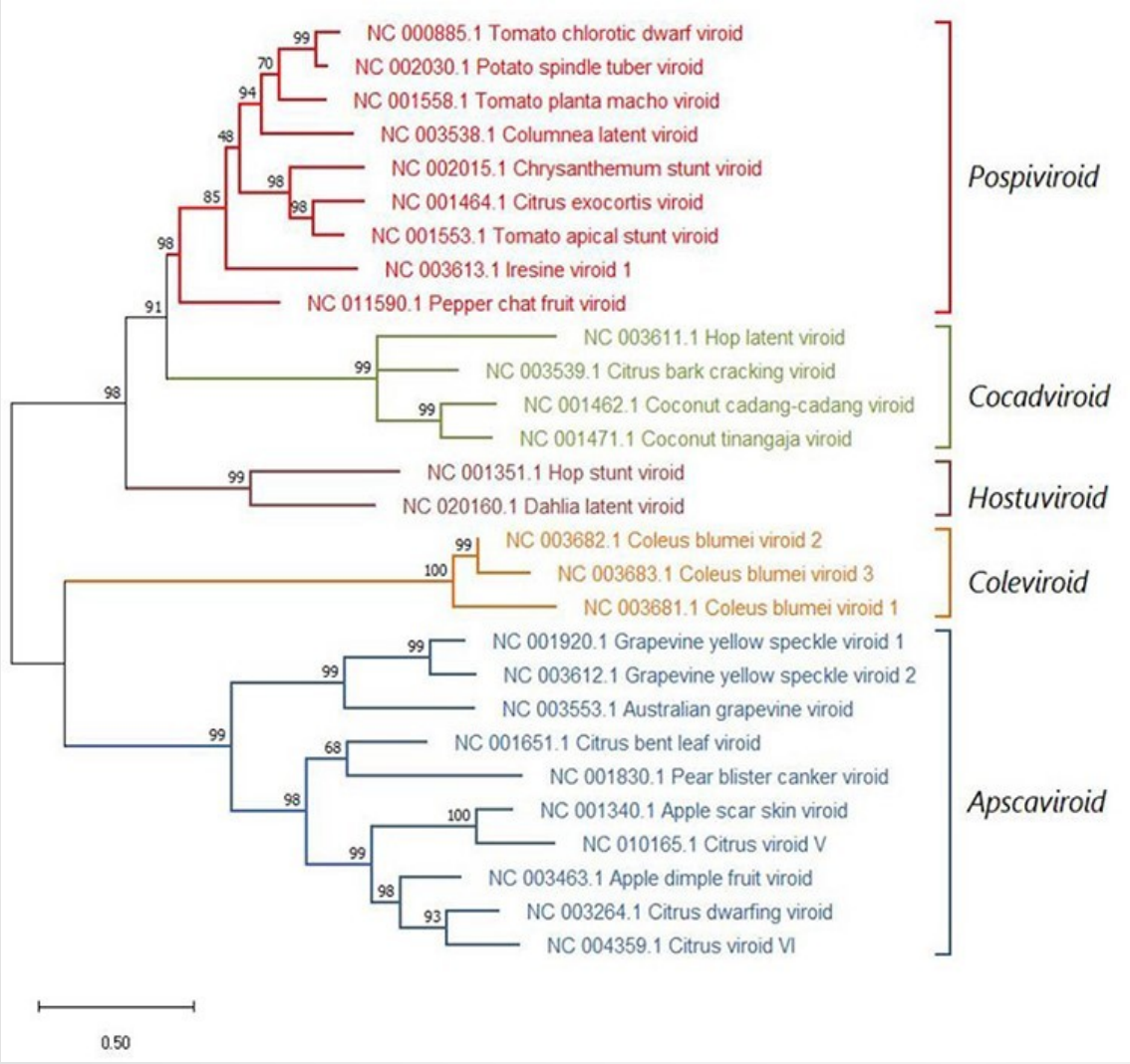


Peach latent mosaic viroid



Floris et al, 2011. DOI: 10.1002/9780470015902.a0000434.pub3

Viroid taxonomy



Flores et al, 2022

species

Family **Avsunviroidae**

Genus *Avsunviroid*	1
Genus *Pelamoviroid*	2
Genus *Elaviroid*	1

- Use plastid RNA polymerase to replicate

Family ***Pospiviroidae***

Genus *Pospiviroid*	10
Genus *Hostuviroid*	1
Genus *Cocadviroid*	4
Genus *Apscaviroid*	10
Genus *Coleviroid*	3

- Use nuclear RNA polymerase II to replicate

Summary for viral(-like) genomes

- Genomes are very simple but very diverse
- Genome size vary between 248 (1683) bp to 1.26 Mbp
- Genomes are based on RNA or DNA
- Single stranded or double stranded genomes
- Circular or linear genomes
- Nonsegmented, segmented and multipartite genomes
- Multipartite genome is a very interesting phenomena when organism could consists of several not connected parts during transmission between hosts
- horizontal gene transfer (HGT) is very common
- viruses could be vectors for (HGT) between very distant taxa

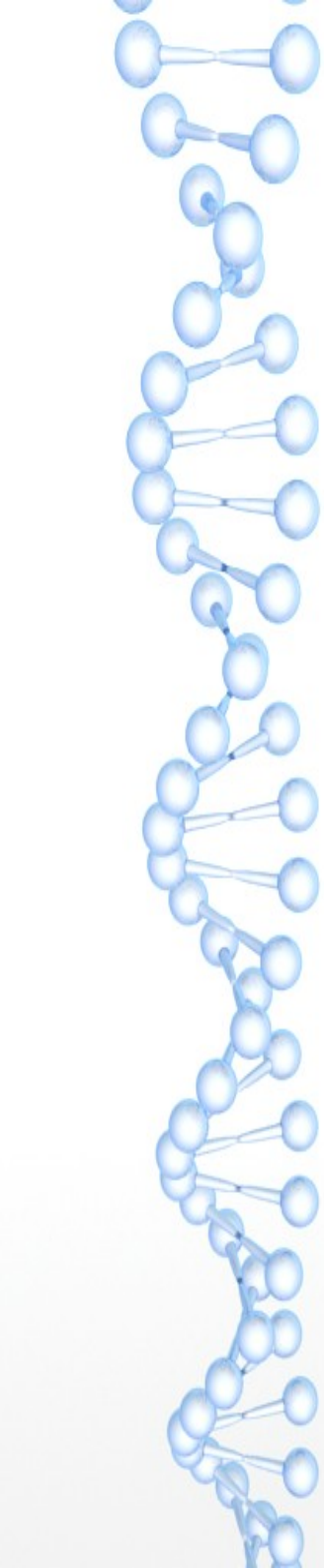
Issues with definitions of “genome” (1)

Genome is

- the entire set of ~~**DNA**~~ instructions found in a ~~**cell**~~.

RNA genomes

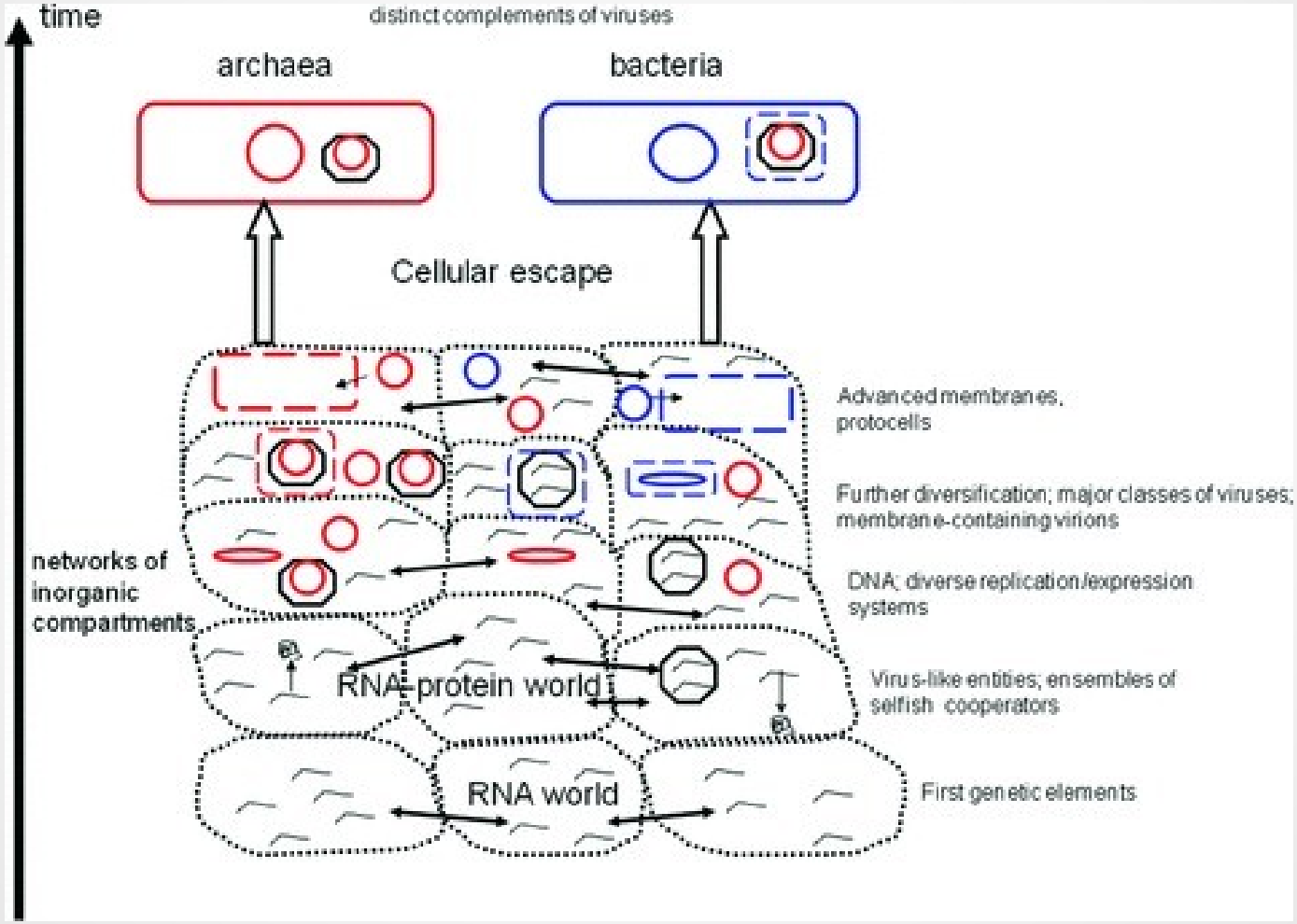
Noncellular organisms



I. Structure and diversity of the genomes

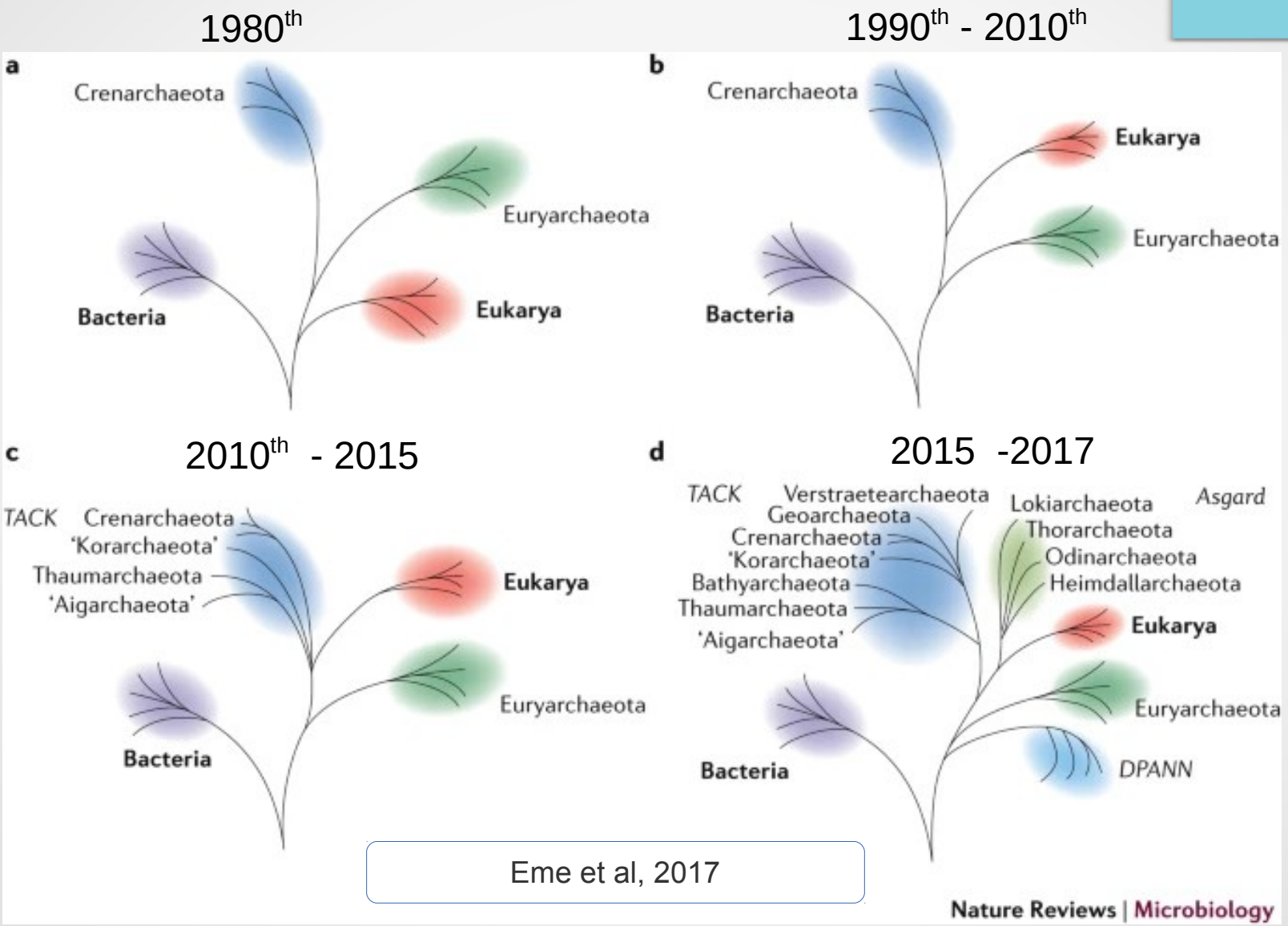
Cellular life

Possible origin of (cellular) life

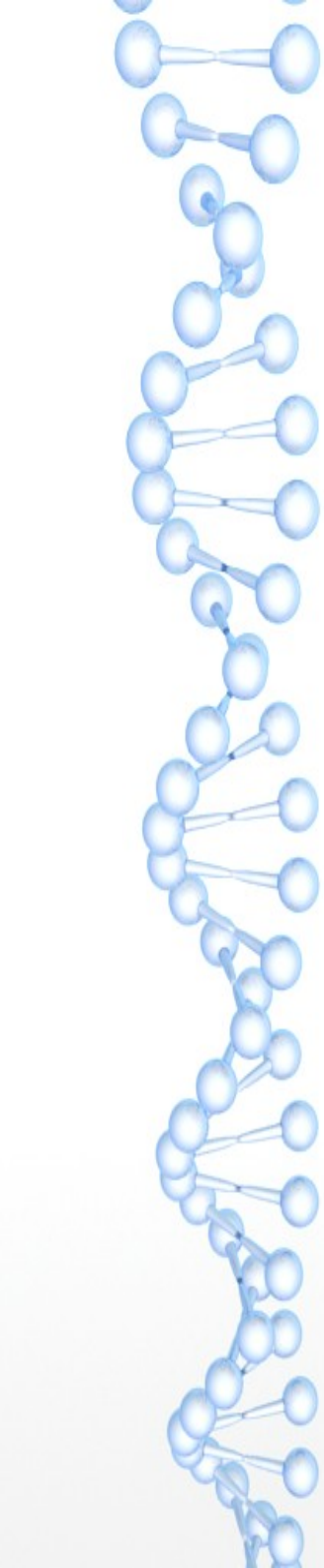


Koonin, 2009

Evolution of tree of (cellular) life



Eme et al, 2017



I. Structure and diversity of the genomes

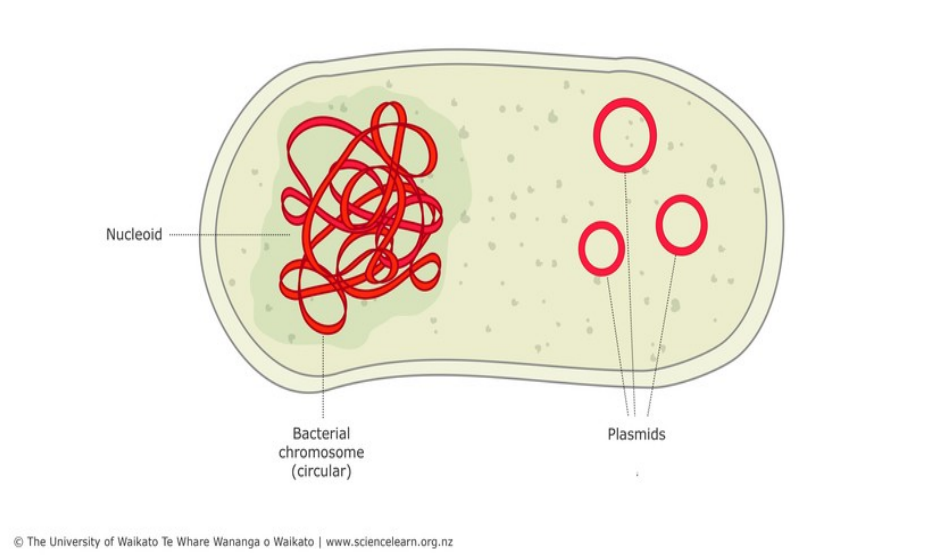
Cellular life
Prokaryotes

Structure of bacterial genome

Up to three dsDNA elements:

I. **Chromosome (obligatory)**.
- oriC/DNA replication type
- contains core genes for cell existence

II. chromids (Harrison et al, 2010).
Obligatory for some clades
- plasmid-type replication
- could contain core genes
- comparable to chromosome size
- codon usage similar to chromosome



III. plasmids
- plasmid-type replication
- doesn’t contain core genes
- could contain essential genes (resistance, conjugation)

All element usually are circular, but sometimes could be linear

Structure of archeal genome

Up to three dsDNA elements:

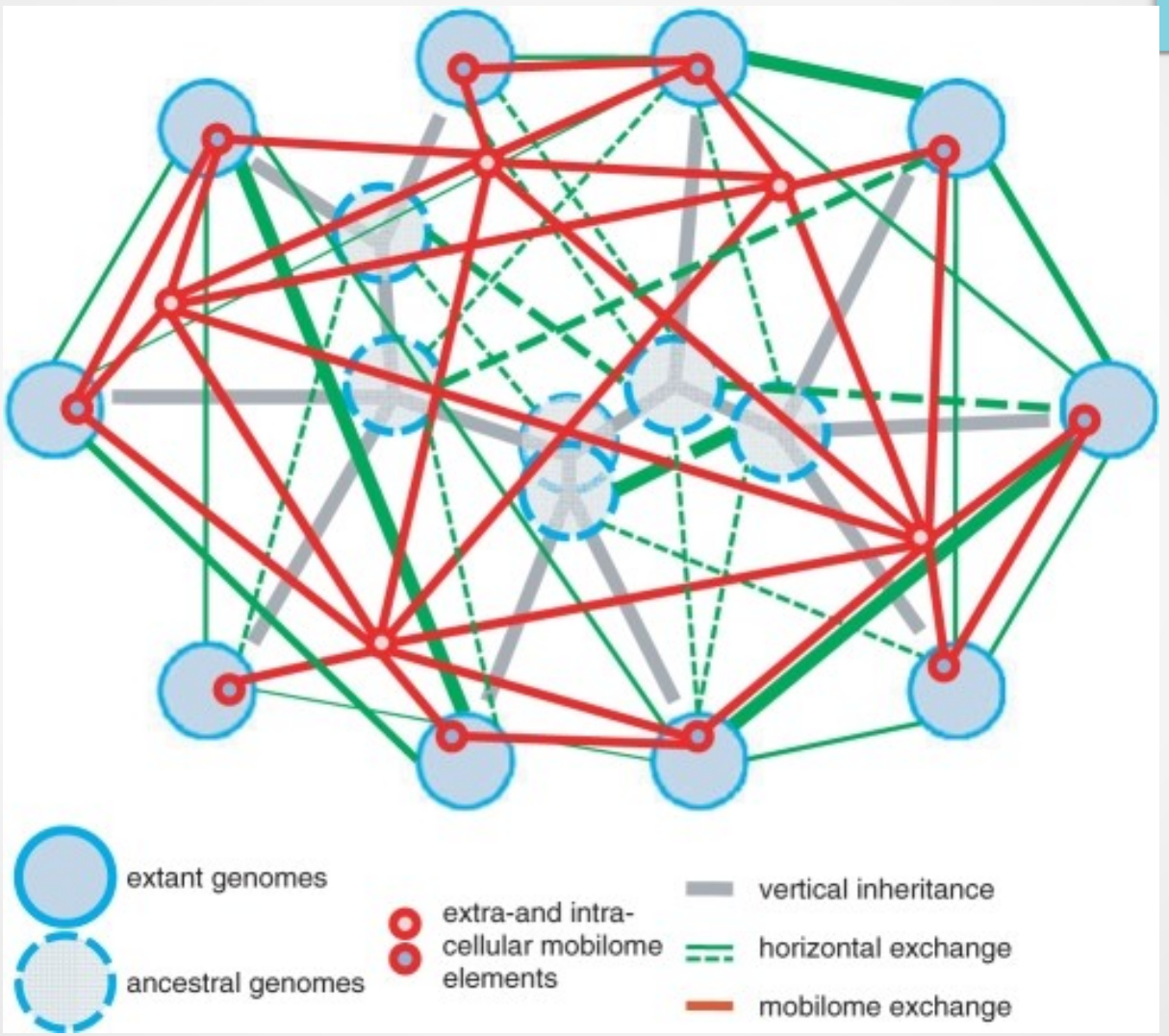
I. **Chromosome (obligatory)**.
- Orc1/Cdc6 replication type
- contains core genes for cell existence

II. mini-chromosome(s).
Obligatory for some clades
- chromosome-like replication type
- could contain core genes
- comparable to chromosome size
- codon usage similar to chromosome

III. plasmids
- plasmid-type replication
- doesn’t contain core genes
- could contain essential genes (resistance, conjugation)

All element usually are circular, but sometimes could be linear

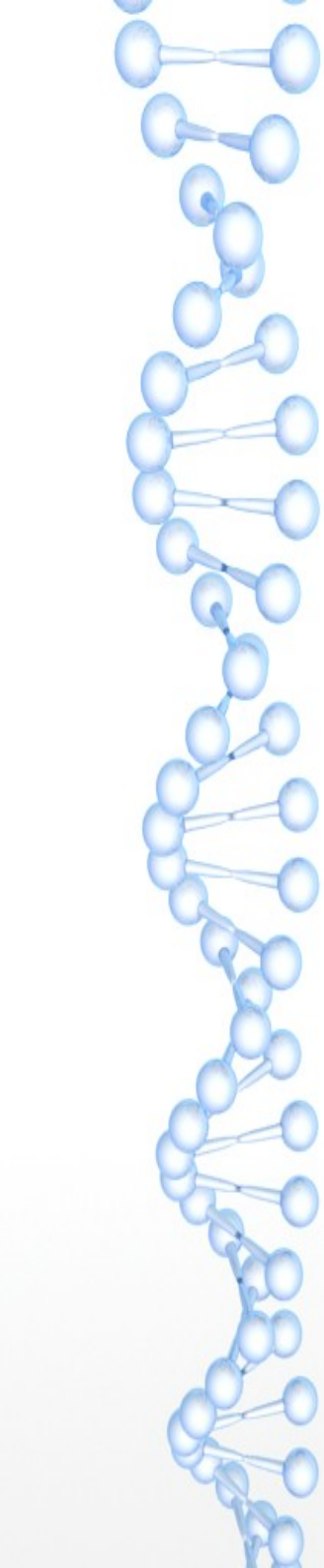
Gene flow in prokaryotic world



Koonin and Wolf, 2009

Summary for prokaryotic genomes

- Only double stranded DNA genomes
- Genome size vary between 112 kbp and 14.8 Mbp (*Sorangium cellulosum*)
- Up to three element types in the genome (chromosome + chromid/minichromosome + plasmids)
- deep difference between Bacteria and Archea in replication system, but high similarity in genome organization
- horizontal gene transfer (HGT) is very common as in viruses
- modern research supports existence of only two domains of life (Archea and Bacteria) with Eukaryotes included in Archea.



I. Structure and diversity of the genomes

Cellular life
Eukaryotes

Eukaryotic genomes

I. Basic type:
Nuclear genome
+ mitochondrial genome (many copies)

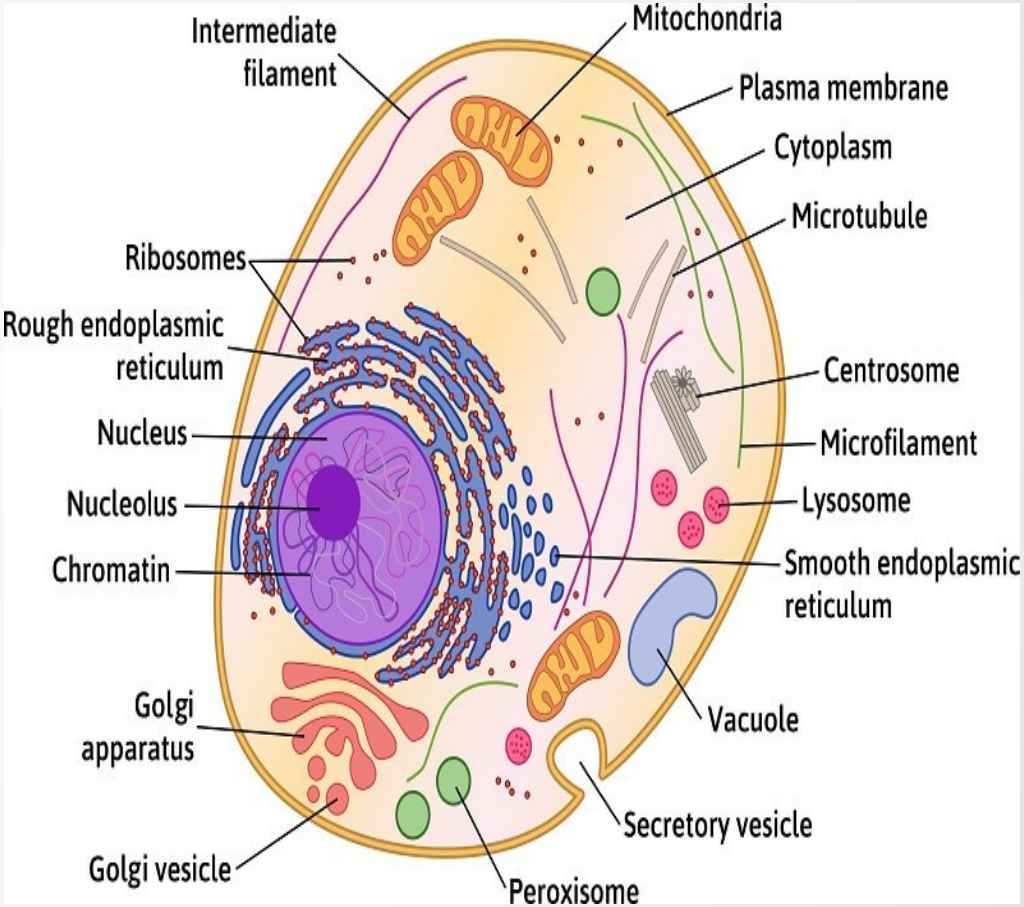
Ia. Basic type **with lost mitochondria:**
Nuclear genome only

One protist is known
Monocercomonoides sp

II. Plant type:
Nuclear genome
+ mitochondrial genome (many copies)
+ chloroplast genome (many copies)

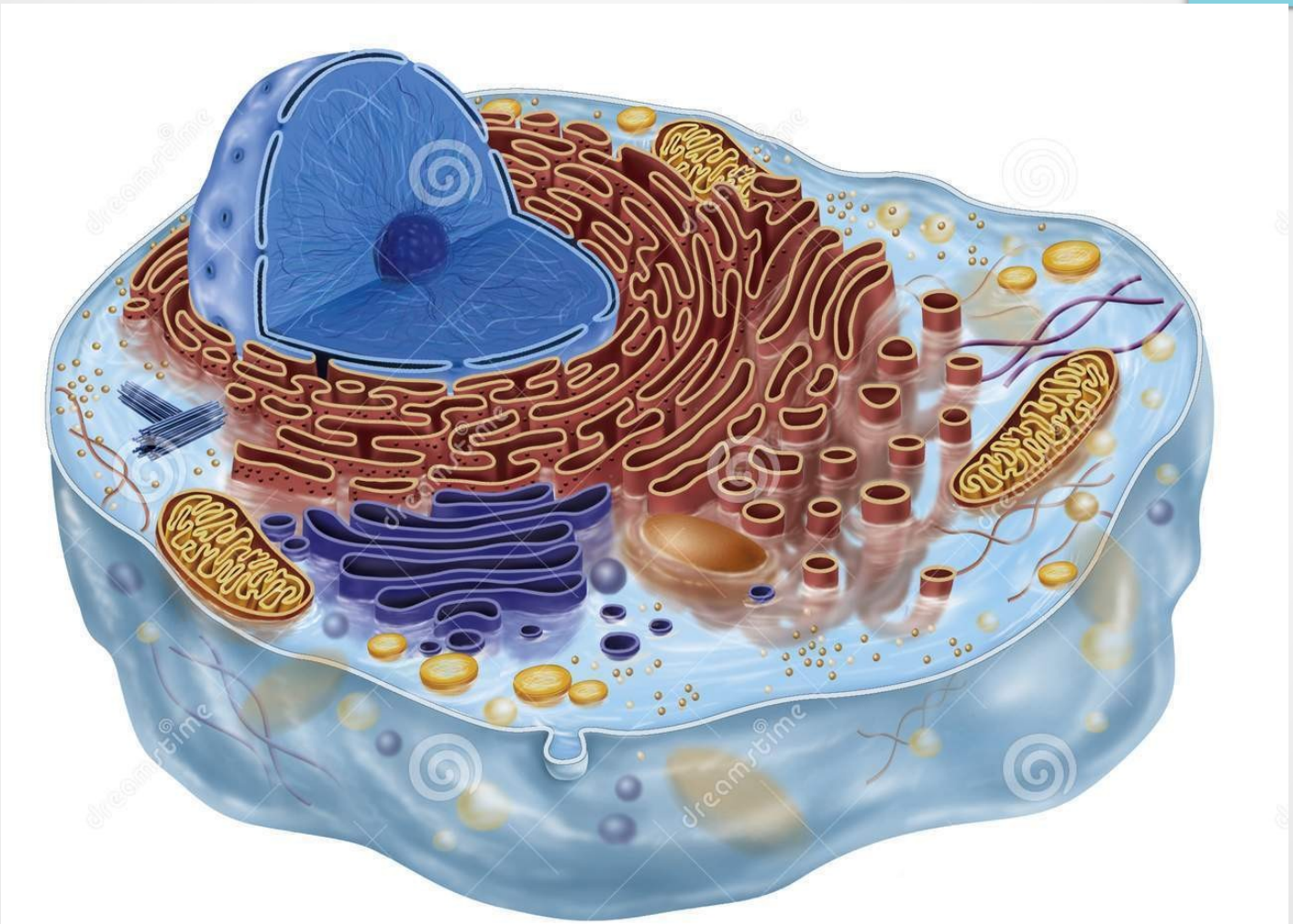
IIa. Plant type with lost chroroplasts:
Nuclear genome
+ mitochondrial genome (many copies)
One (?) genus *Rafflesia*

Animal cell

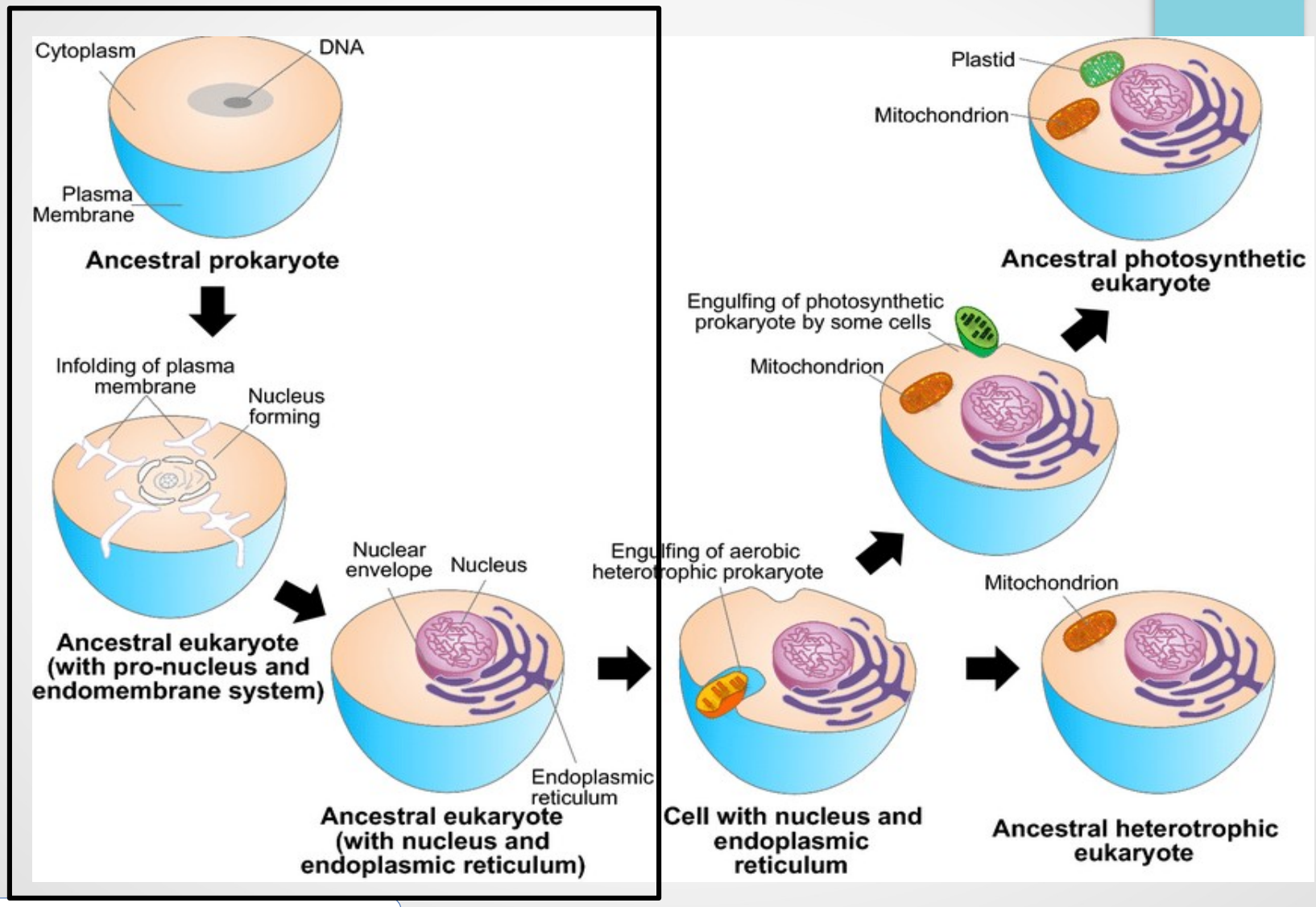


https://biologydictionary.net/animal-cell/

Animal cell


Download from
Dreamstime.com
This watermarked comp image is for previewing purposes only.
ID 37889236
Frank Liu | Dreamstime.com

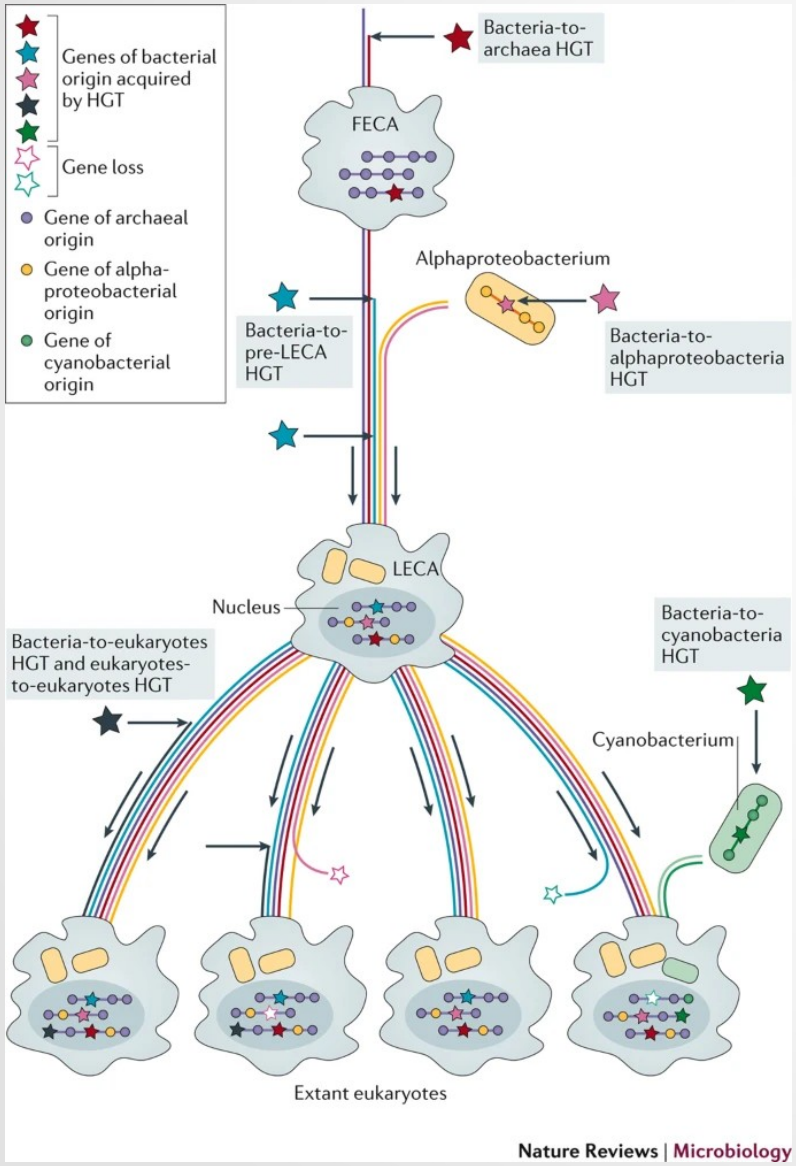
Origin of nucleus* and organells



Almedia and Amaral, 2020

hypothesis of membrane invagination-based origin

Chimeric origin of Eukaryotes



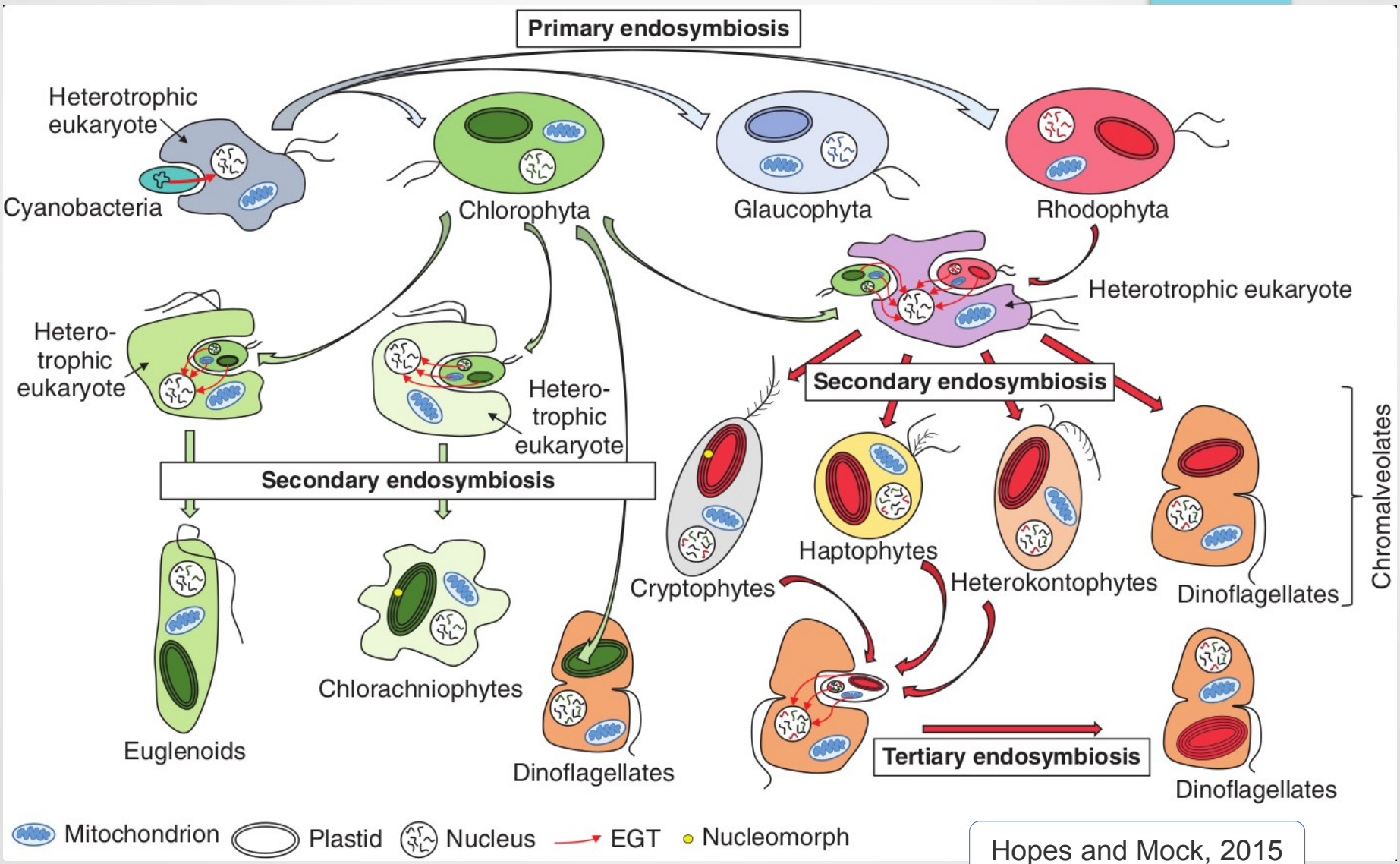
FECA - first eukaryotic common ancestor

LECA - last eukaryotic common ancestor

HGT - horizontal gene transfer

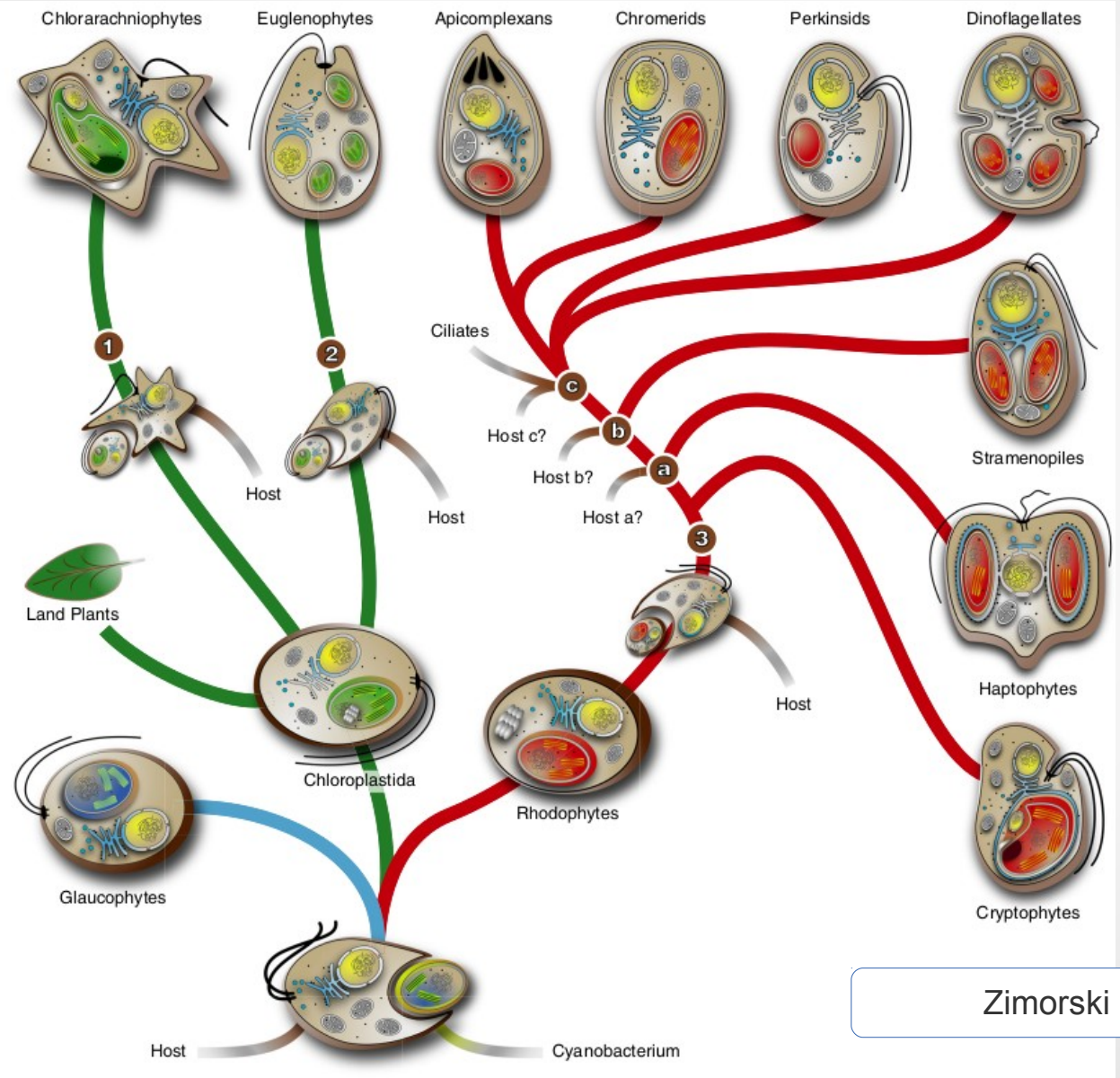
Eme et al, 2017

Origin of plastids (1)

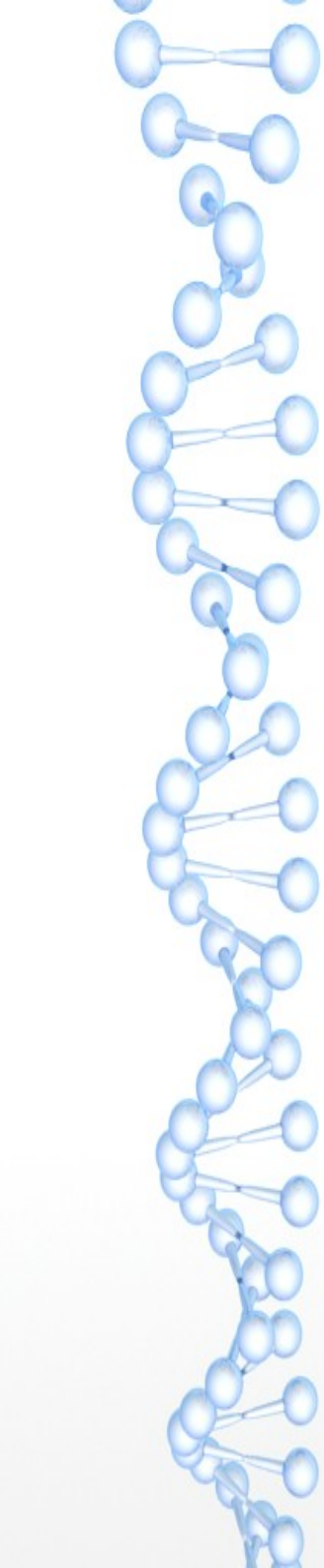


Hopes and Mock, 2015

Origin of chloroplasts (2)



Zimorski et al, 2014



I. Structure and diversity of the genomes

Genome in vivo
Karyotypes and chromosomes

Definitions

Chromosome

DNA and protein-containing structures in cells of eukaryotes, microscopically visible as a rod-shaped body during cell division metaphase

Karyotype

complete set of metaphase chromosomes in cells of an organism of a particular species.

Chromatine

DNA-protein complex. Material of chromosomes

Euchromatine

lightly packed and transcriptional active chromatine. Enriched by genes.

Heterochromatine

highly packed chromatine. Low transcriptional activity**.** Few genes

Lewin et al, 2019

Issues with definitions of “genome” (2)

- **Genome** is

  genetic material of haploid set of **chromosomes**

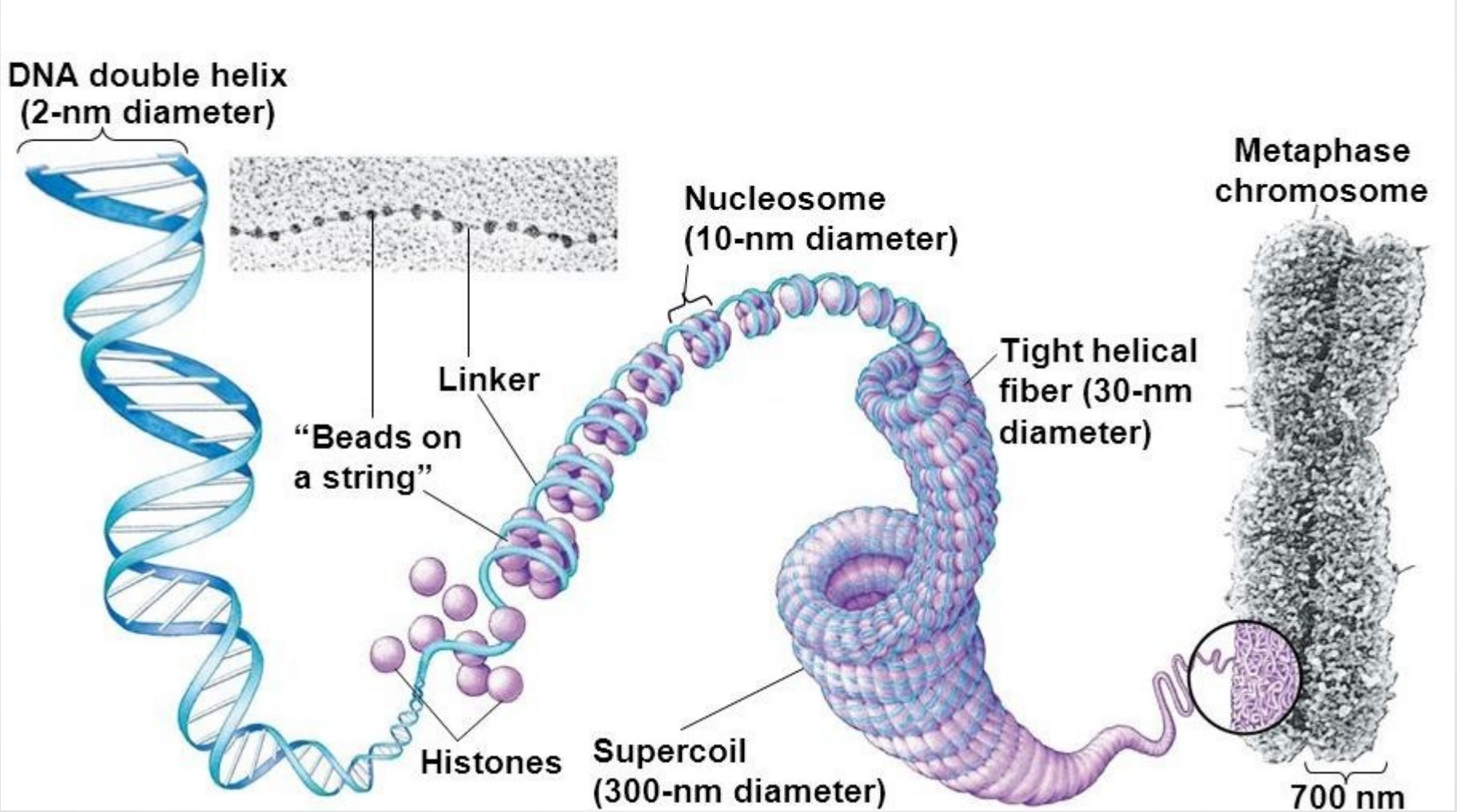
Chromosome

DNA and protein-containing structure in cells of eukaryotes, microscopically visible as a rod-shaped body during cell division metaphase

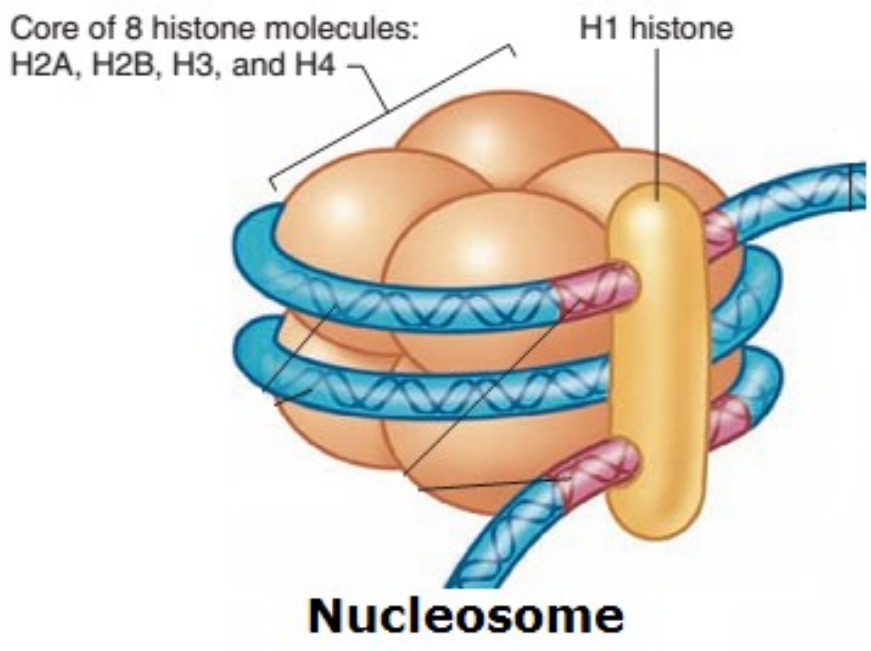
Bacteries? Viruses? mtDNA? chloroplast DNA?
Why not included?

Definition (this) of chromosome was proposed by cytogeneticists of “high” eukaryotes (plants and animals)

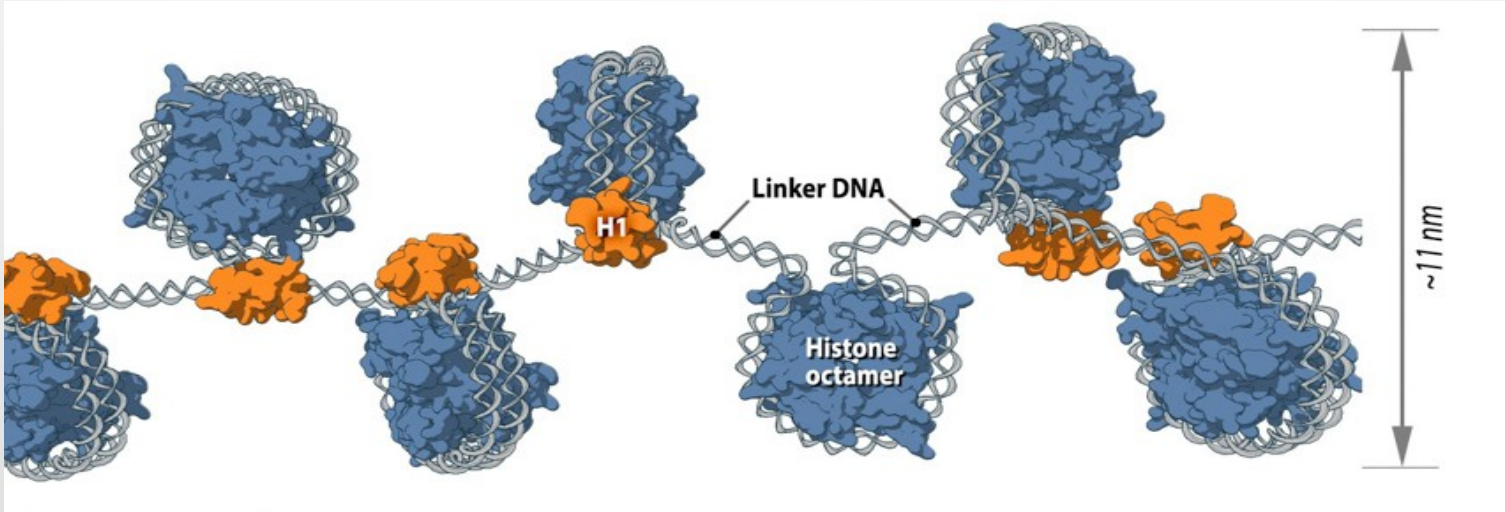
From DNA to chromosome



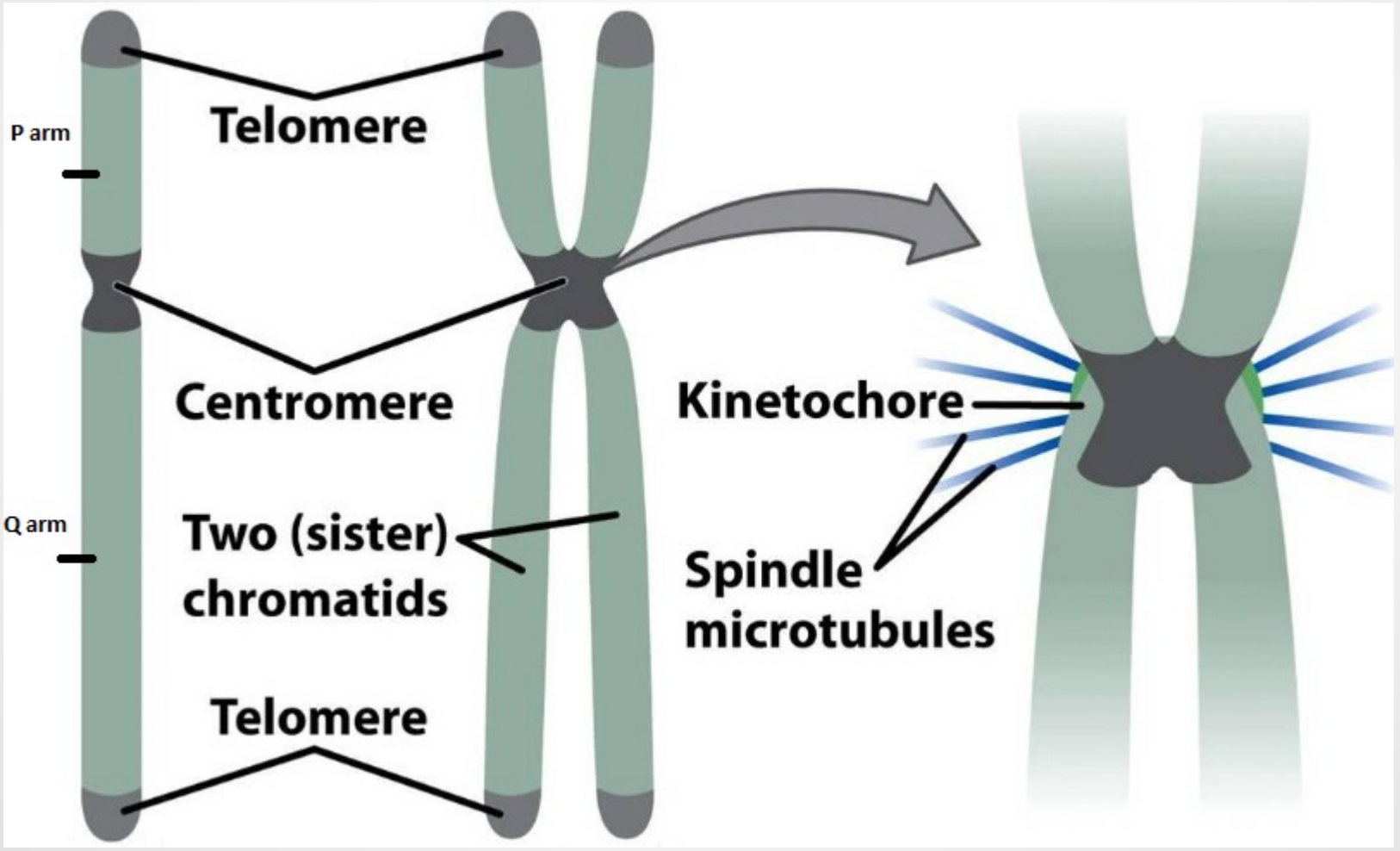
Nucleosome



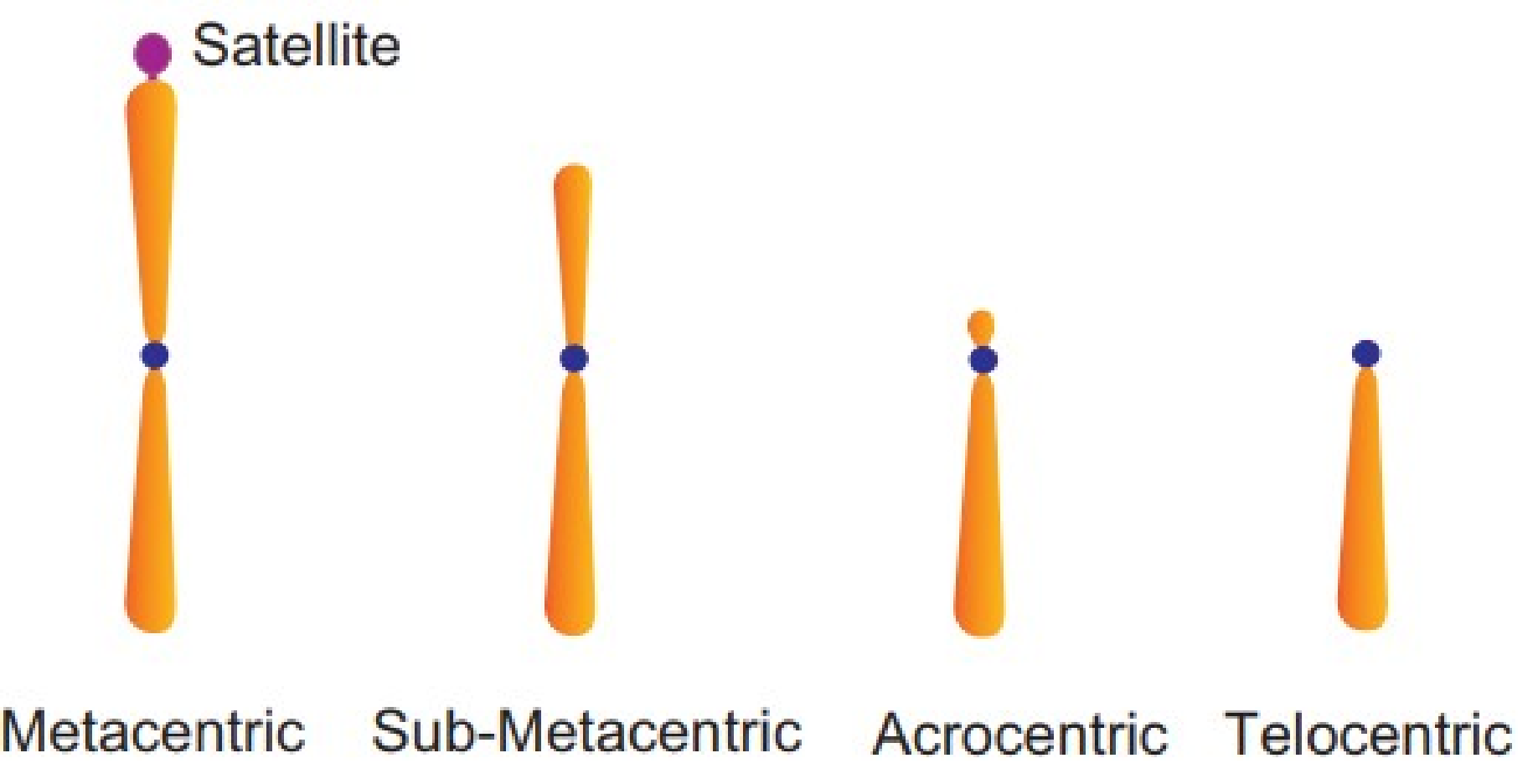
147 bp per nucleosome



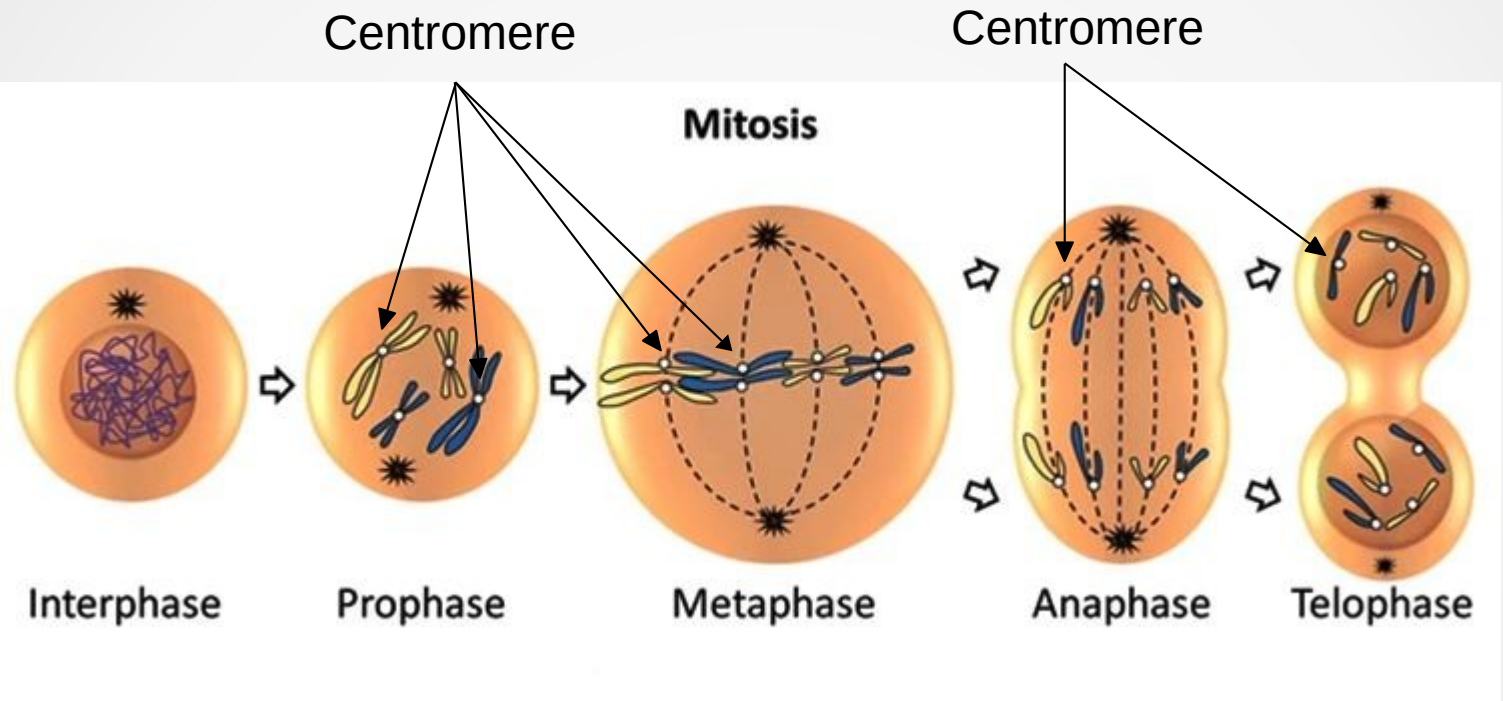
Chromosome structure



Chromosome types

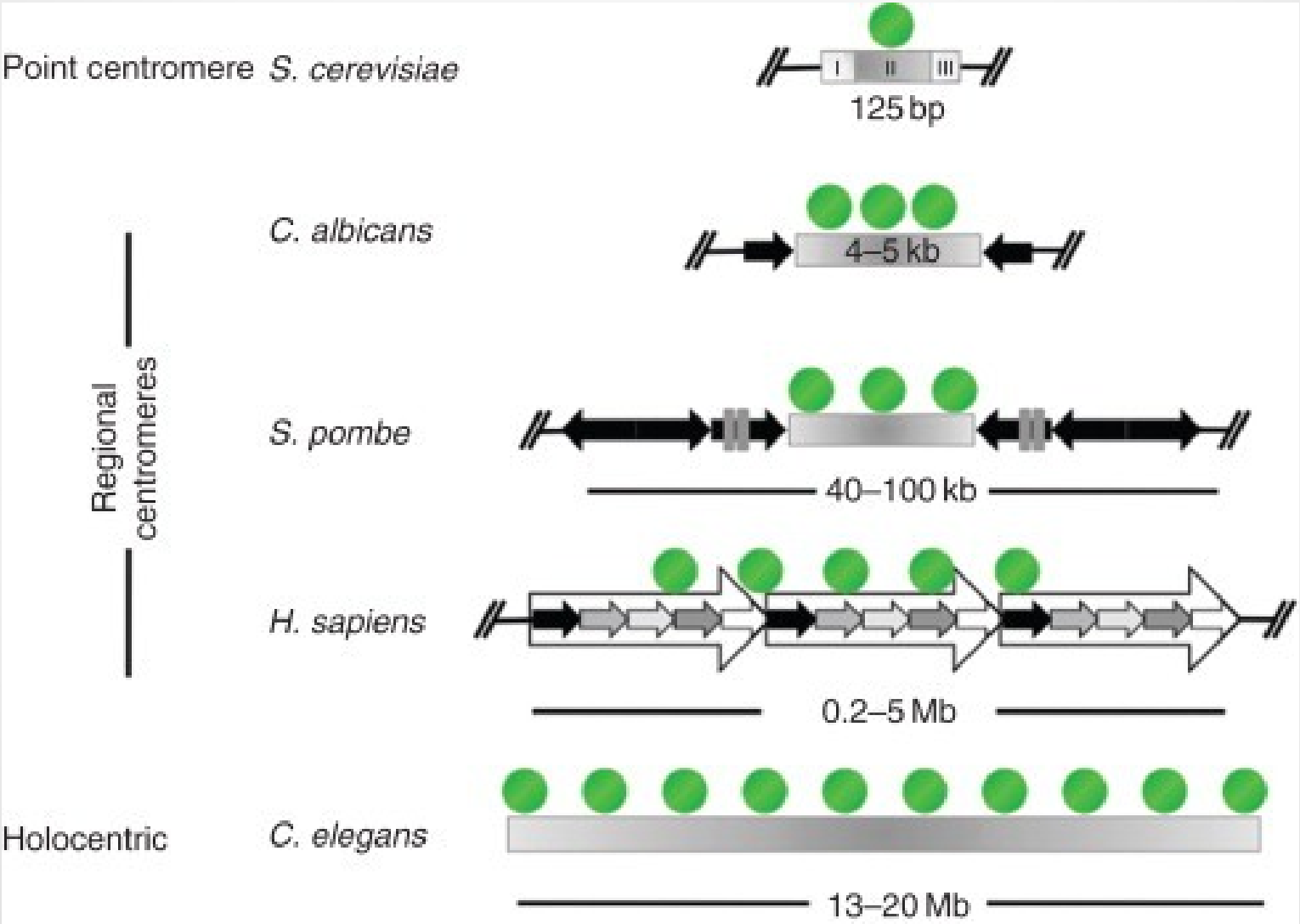


Role of centromere

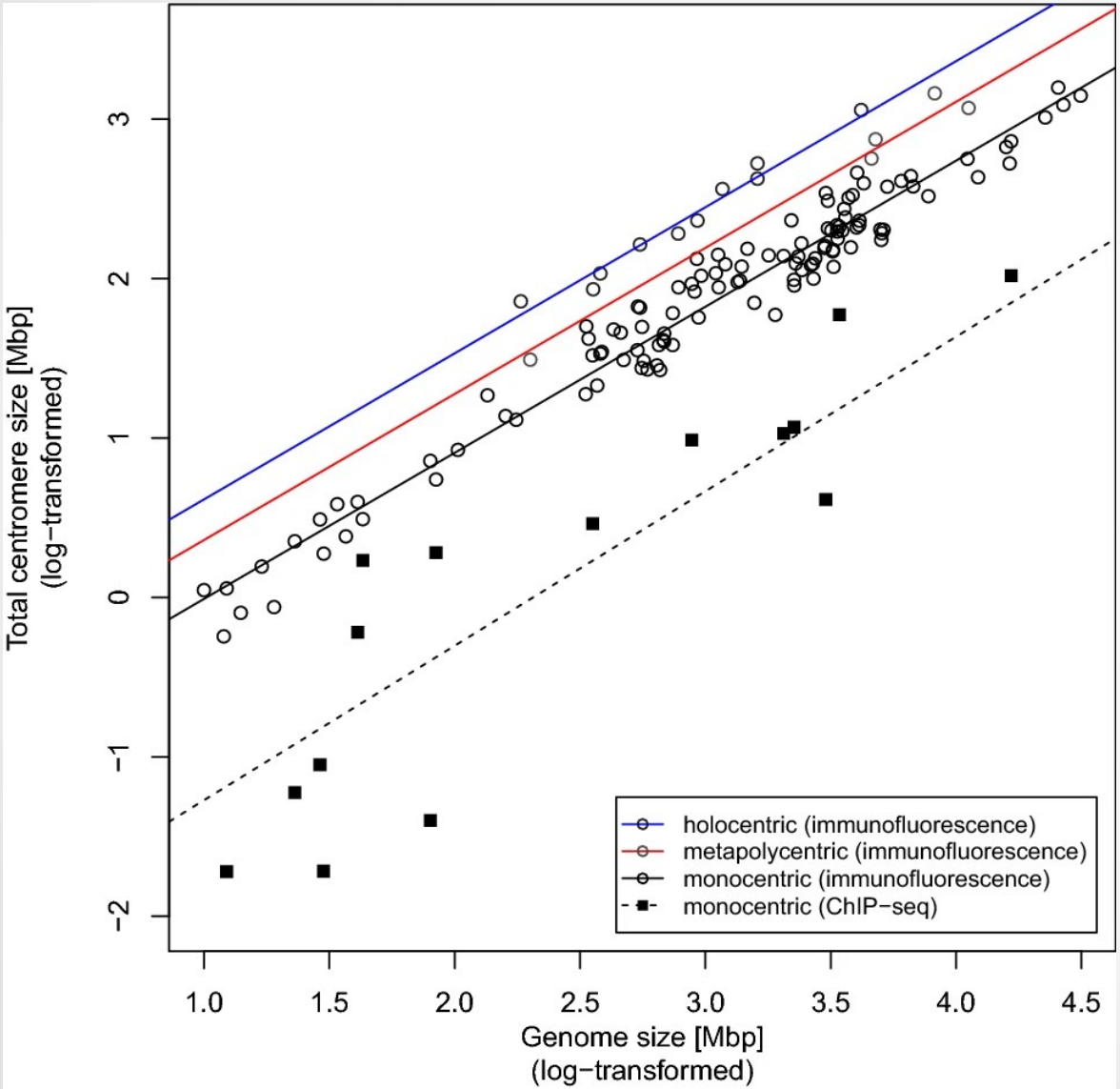


Centromeres are crucial for proper separation of chromosomes between daughter cells

Centromere structure and length



Centromere length and genome size



Total centromere length is strongly correlated with genome size

R2 = 0.964

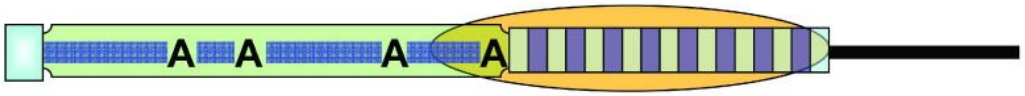
Plačková et al, 2021

Telomers are:

- located at the end of linear chromosomes
- highly repetitive regions
- protection system against end-replication losses
- protection system against interchromosomal fussions
- stabilizators of genome
- marker of cell age
- maintained by telomerase (or by similar mechanism) but not in all cells

Variants of telomere structure

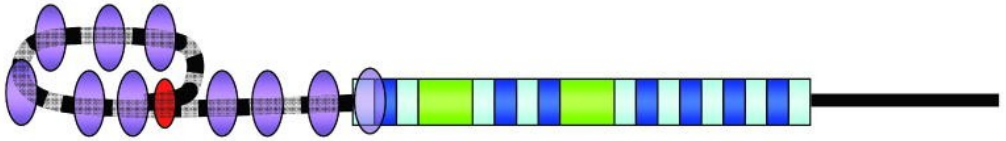
Drosophila



Cap	*HeT-A, TART, TAHRE* (HTT) array	Telomere associated sequence	Euchromatin
HOAP HP1	H3Me3K9 H3Me3K4 Z4 JIL-1 HP1 PROD	PC E(Z) PH PSC SCM H3Me3K27	

Array of mobile elements

Mammalian

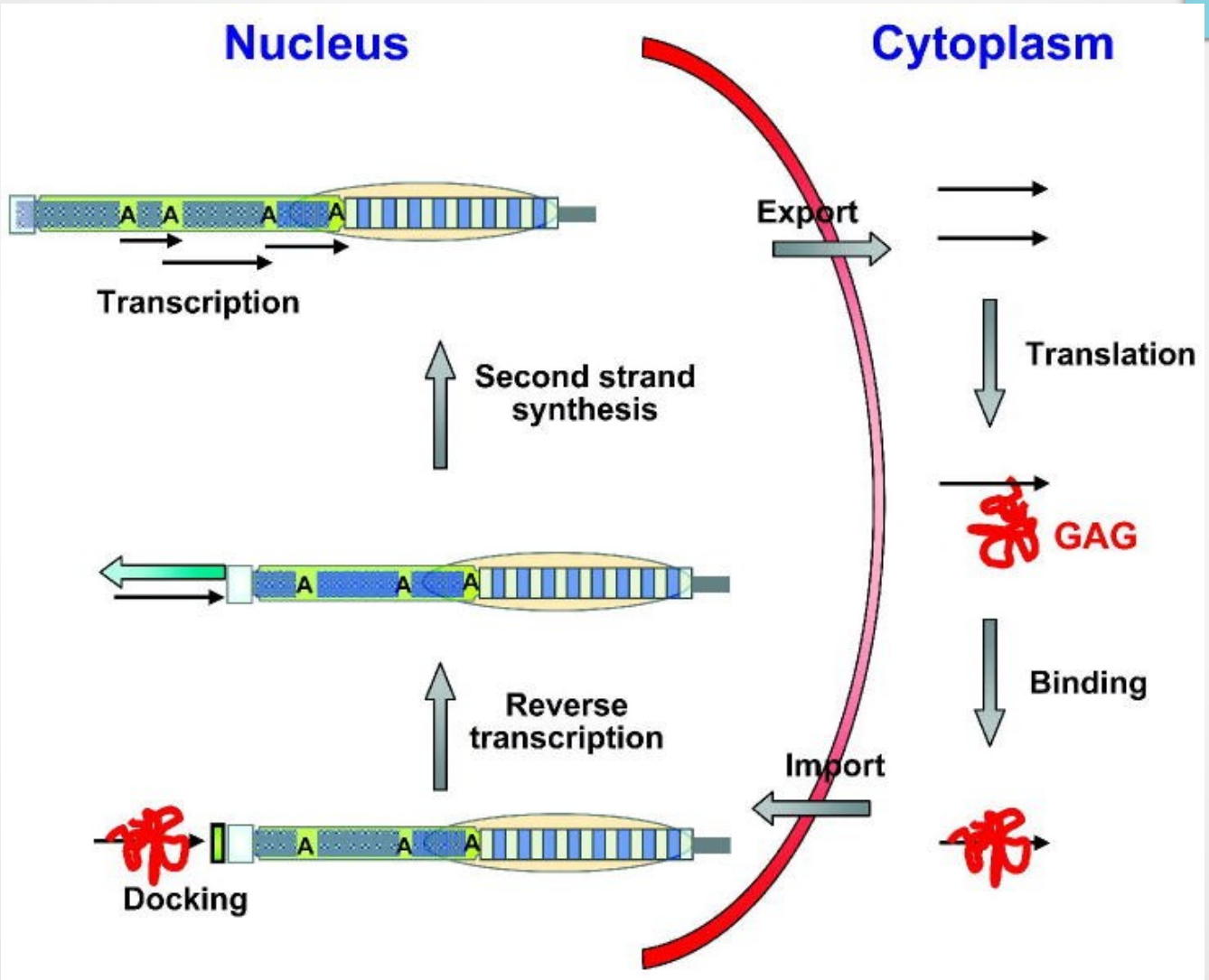


Shelterin		Telomere associated sequence	Euchromatin
TRF1 TRF2 RAP1 TIN2 POT1 TPP1	HP1a HP1b H3Me3K9 H4Me3K20 Ku70/Ku80?	HP1a HP1b H3Me3K9 H4Me3K20	

Tandem telomeric repeats

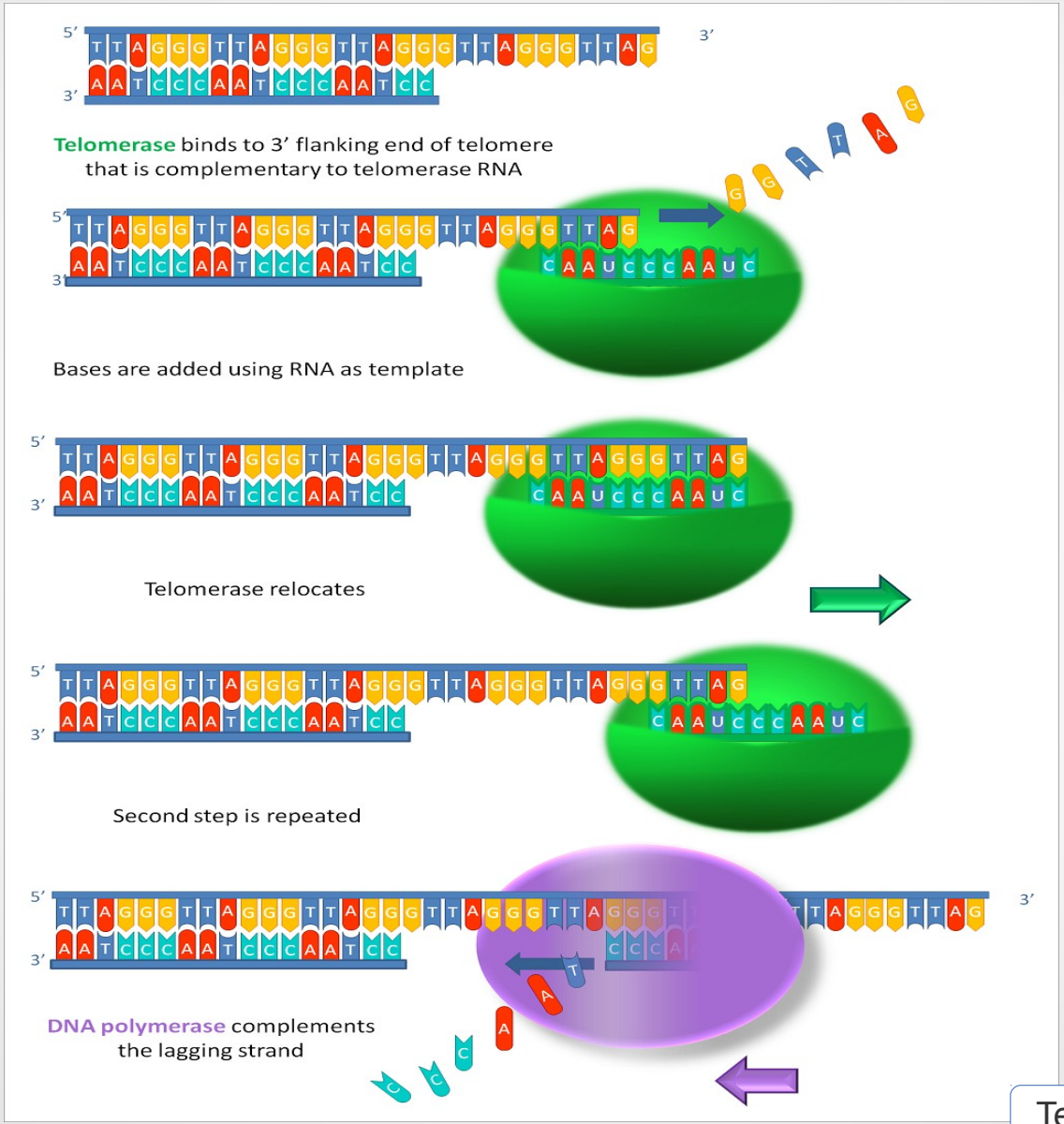
Mason et al, 2008

Elongation of Drosophila telomers



Mason et al, 2008

Telomere elongation

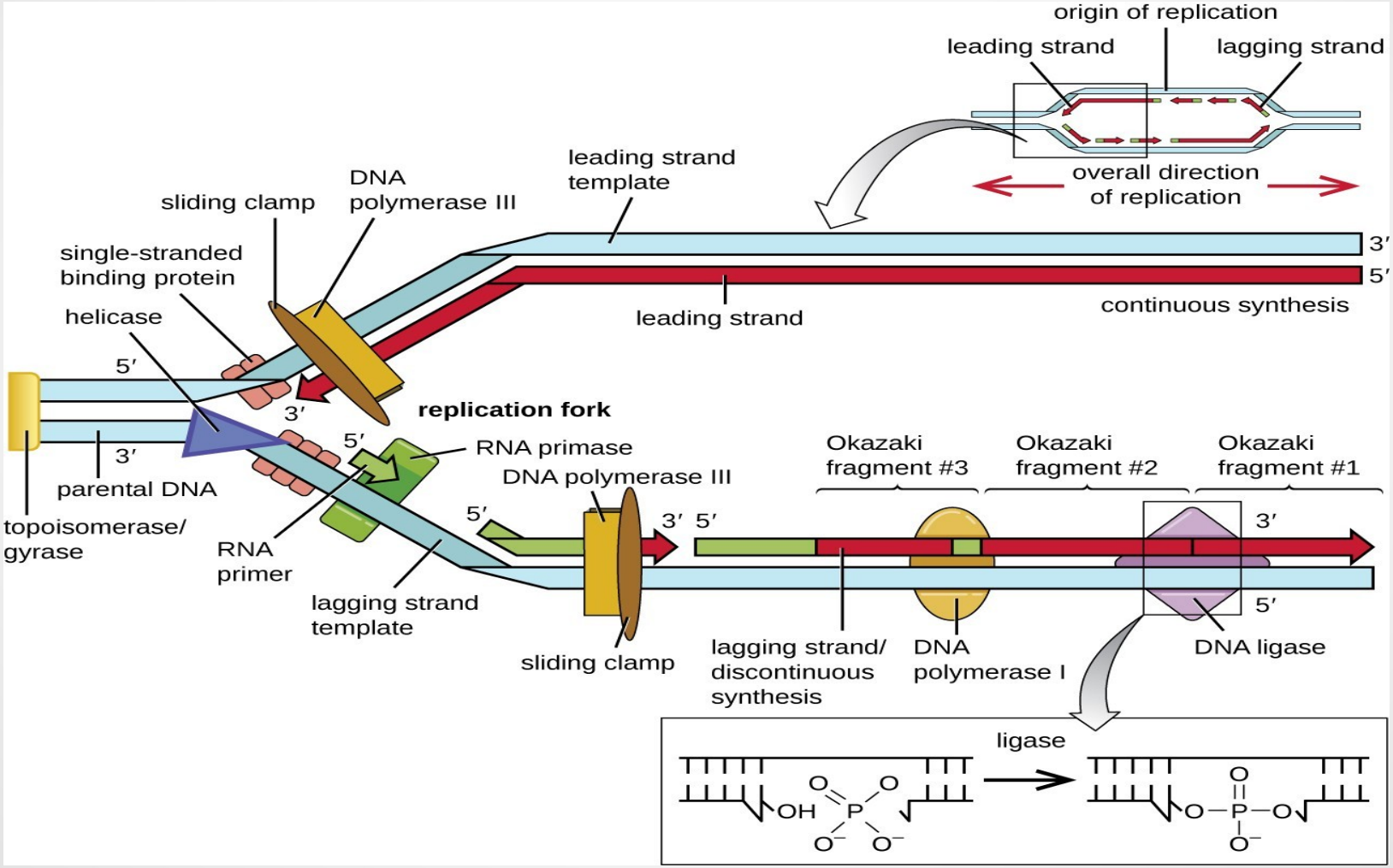


Telomerase. Wikipedia.

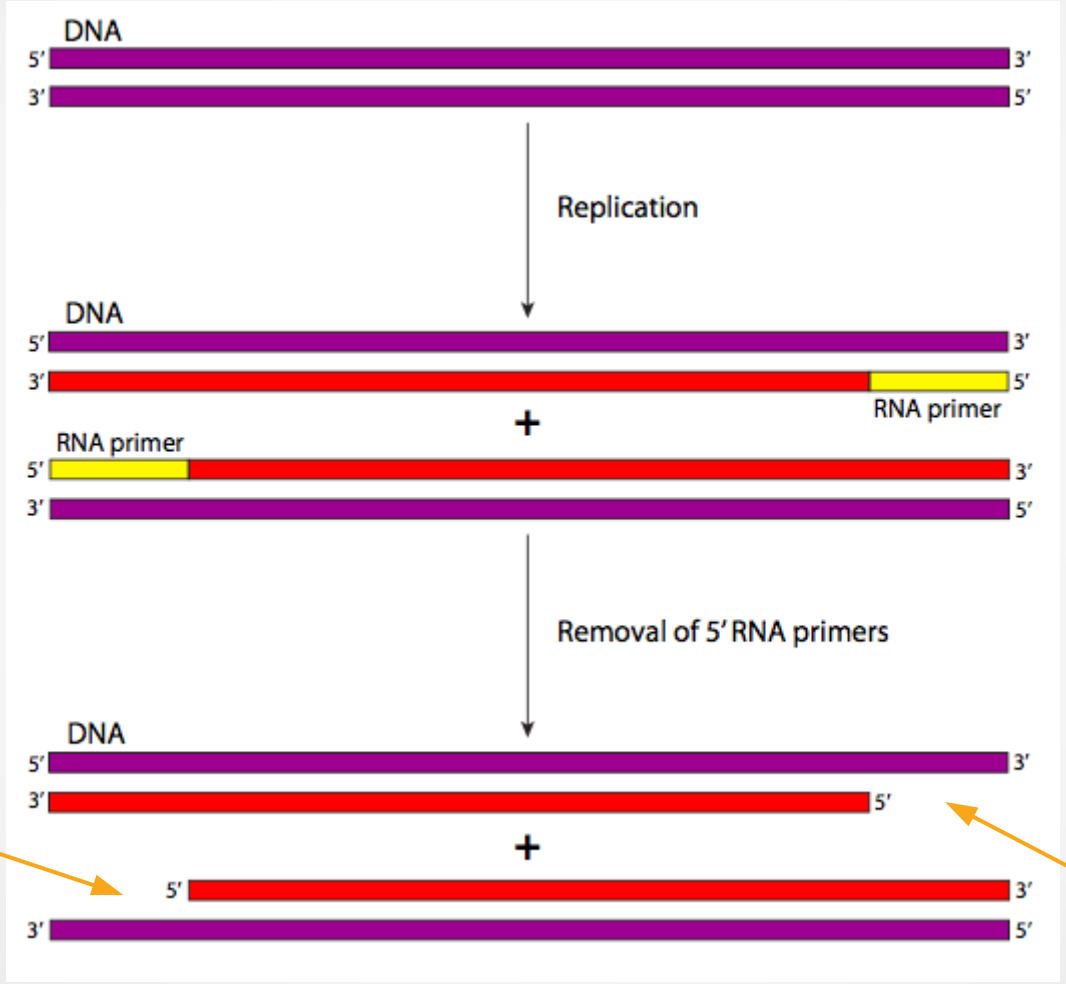
Telomeric repeats

Group	Organism	Repeat monomer
Vertebrates	Human, mouse, *Xenopus*	TTAGGG
Filamentous fungii	*Neurospora crassa*	TTAGGG
Slime moulds	*Physarum*, *Didymium*	TTAGGG
	Dictyostelium	AG(1-8)
Kinetoplastid protozoa	*Trypanosoma*, *Crithidia*	TTAGGG
Ciliate protozoa	*Tetrahymena*, *Glaucoma*	TTGGGG
	Oxytricha, *Stylonychia*, *Euplotes*	TTTTGGGG
Apicomplexan protozoa	*Plasmodium*	TTAGGG(T/C)
Higher plants	*Arabidopsis thaliana*	TTTAGGG
Green algae	*Chlamydomonas*	TTTTAGGG

Replication



End replication problem

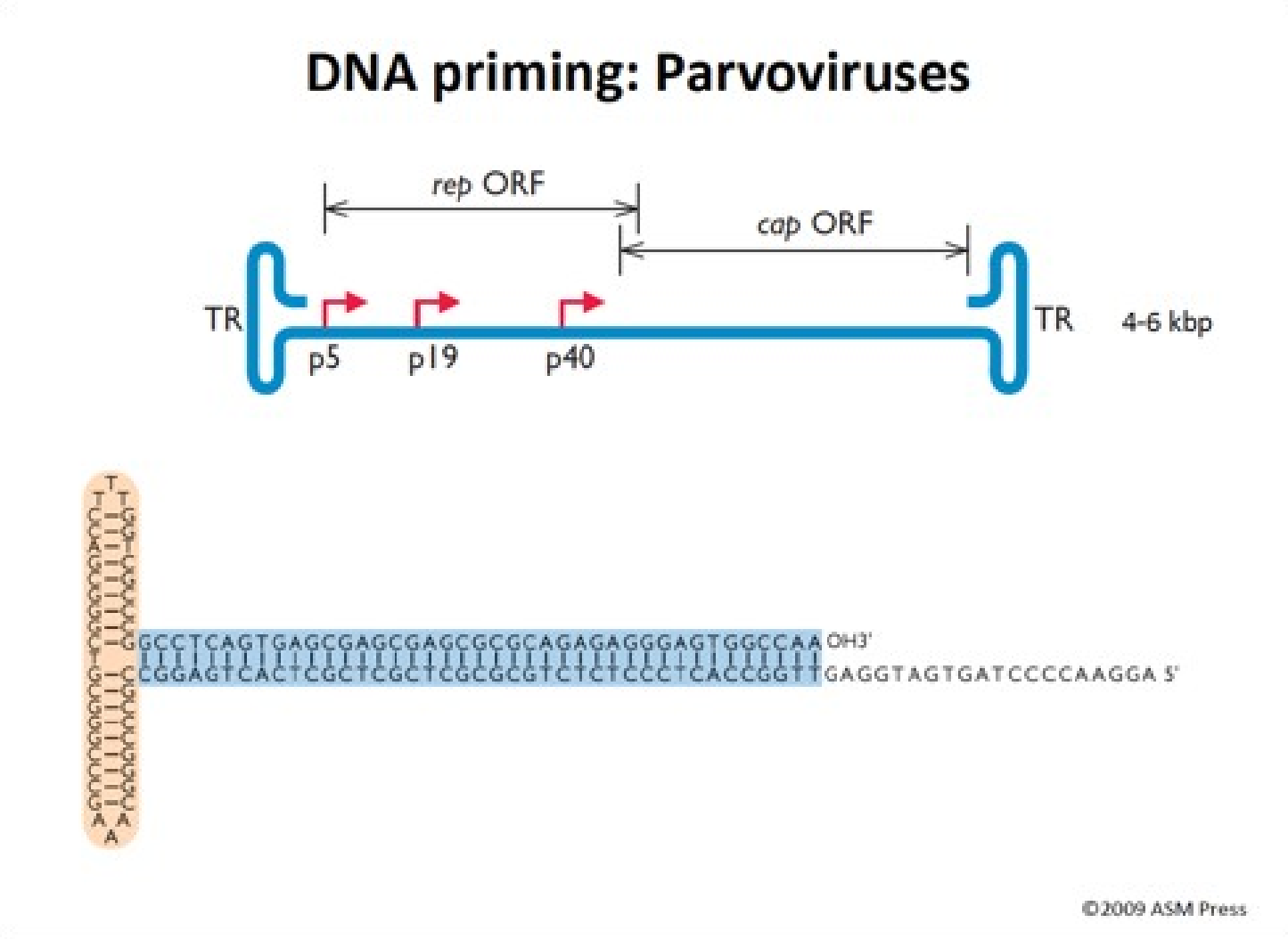


loss

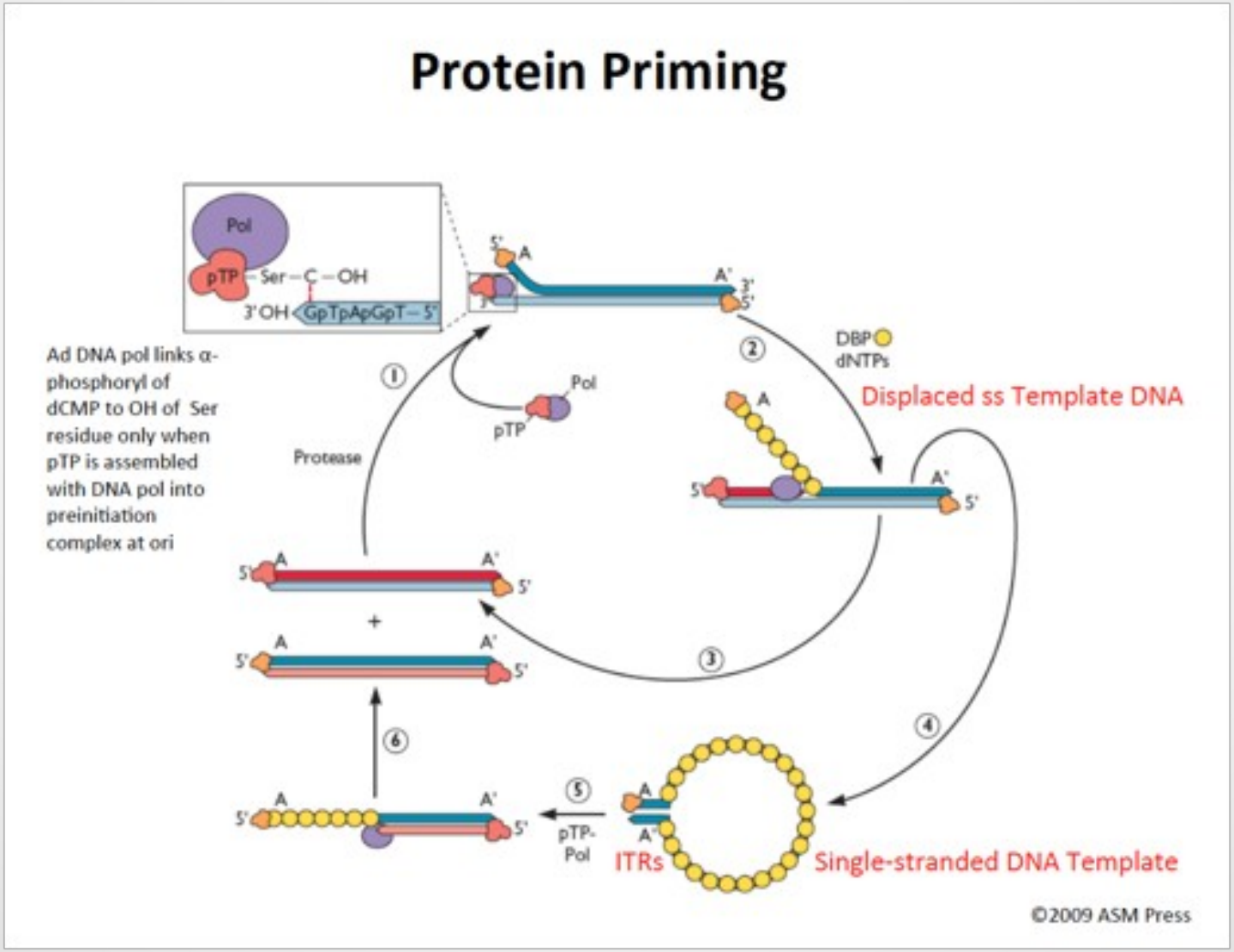
loss

Specific problem of linear chromosome replication

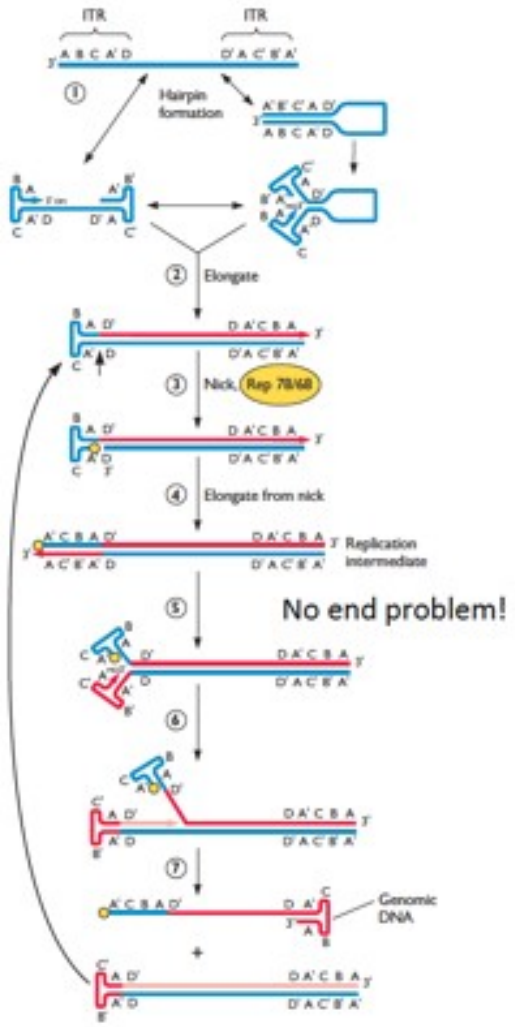
Viral solutions of end problem: parvoviruses (1)



Viral solution of end problem: adenoviruses

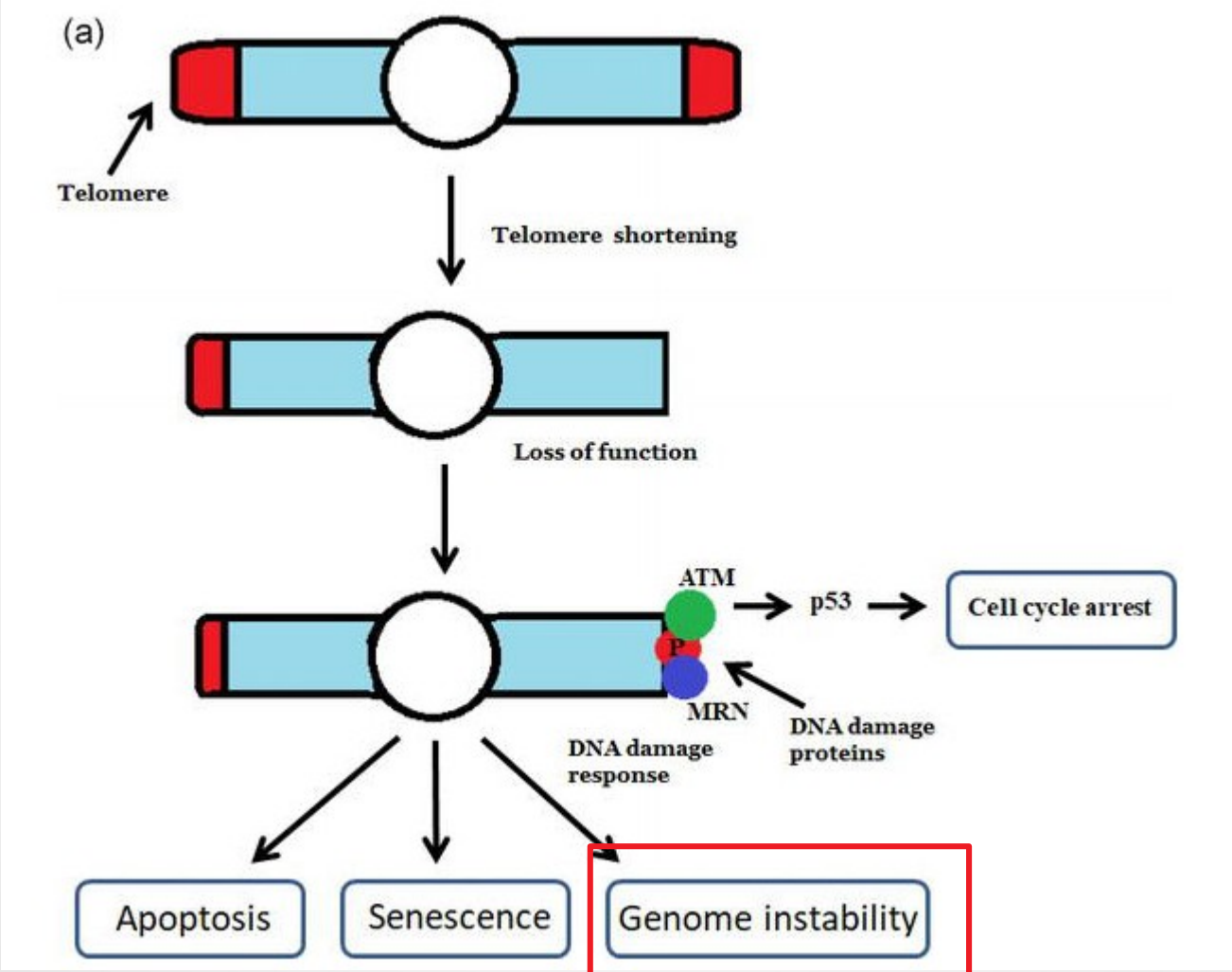


Viral solution of end problem: parvoviruses (2)



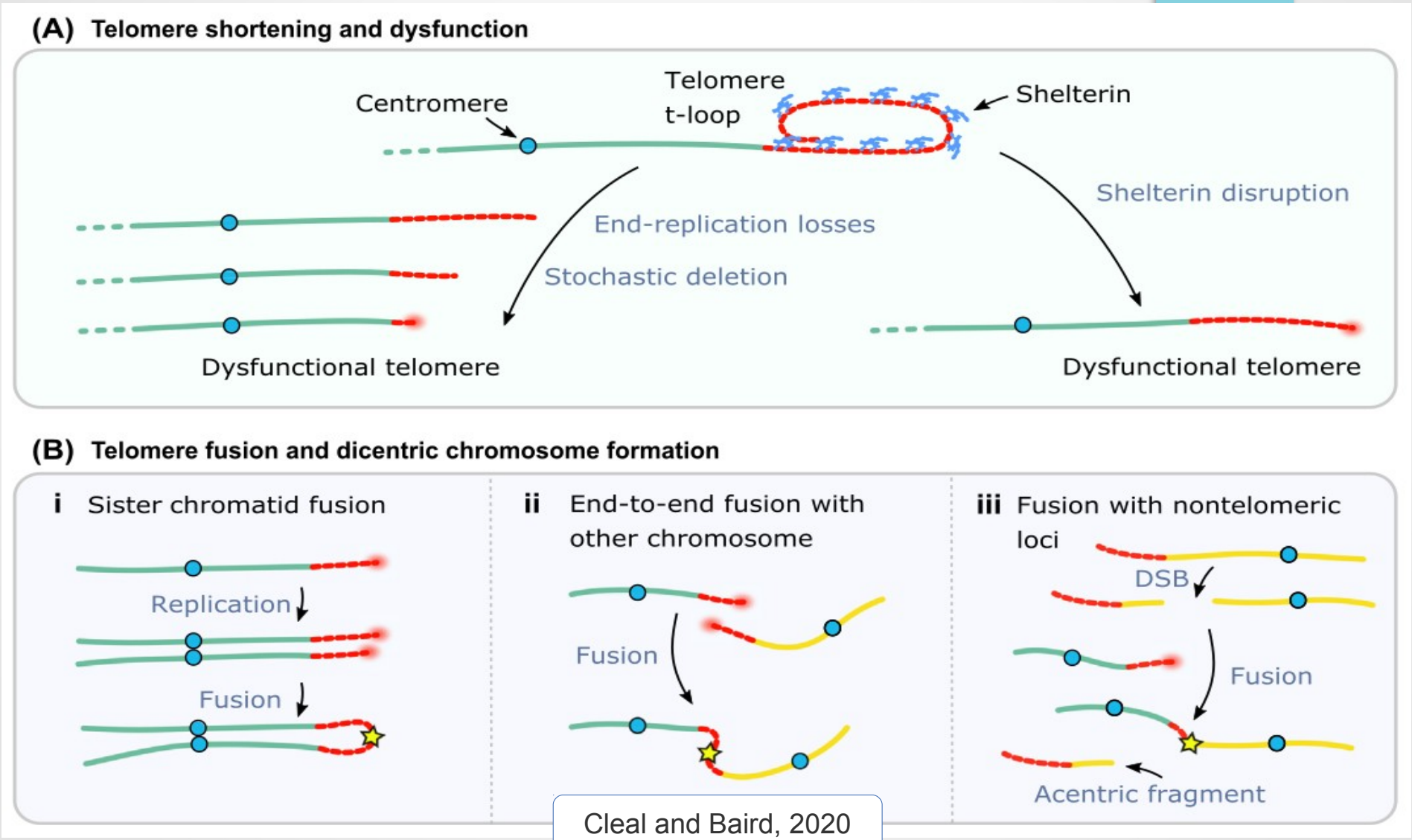
- Replication is continuous
- No pol α, uses ITR to self-prime
- Requires pol δ, Rf-C and PCNA
- Rep78/68 proteins are required for initiation and resolution: endonuclease, helicase, binds 5' terminus
- No replication fork, strand displacement

Results of telomere shortening (1)



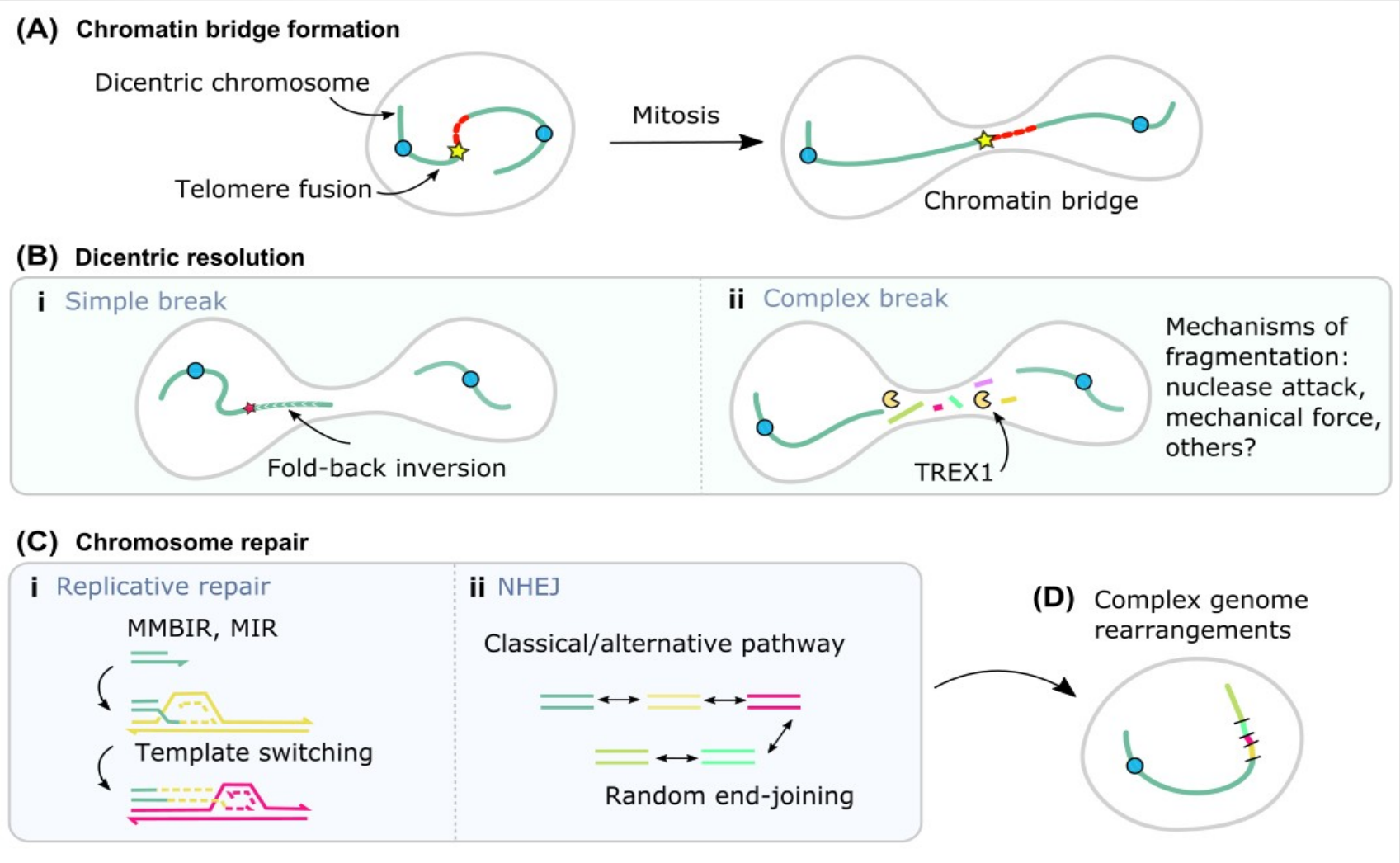
Nag, 2019

Results of telomere shortening (2)



Cleal and Baird, 2020

Results of telomere shortening (3)



Cleal and Baird, 2020

Can we see the chromosomes?

Yes, in microscope, but we need… to stain them first!

Staining type	Light bands	Dark bands
G-Banding	GC-rich	AT-rich
R-Banding	AT-rich	GC-rich
C-banding	euchromatine	constitutive heterochromatine

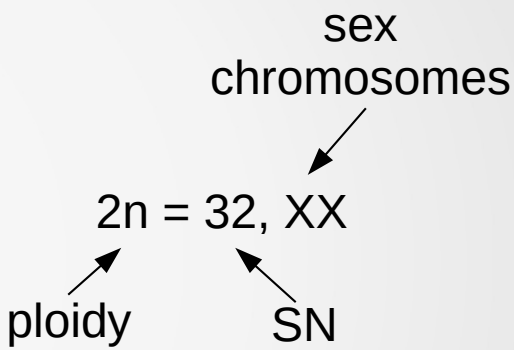
G-banded chromosome (right)
and
its ideogram (left)

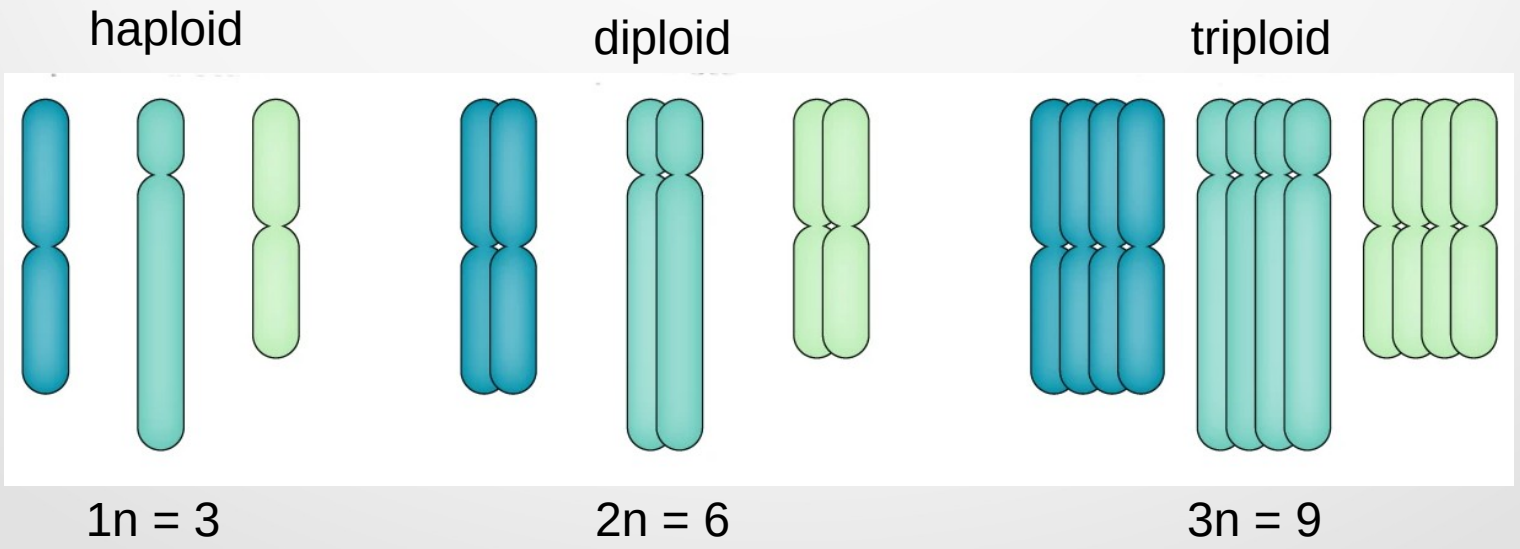
Karyotype description

Karyotyping is
process of visualizing and evaluating the karyotype, i.e. all set of chromosomes from the cell

Ploidy is
the number of complete sets of chromosomes in a cell.

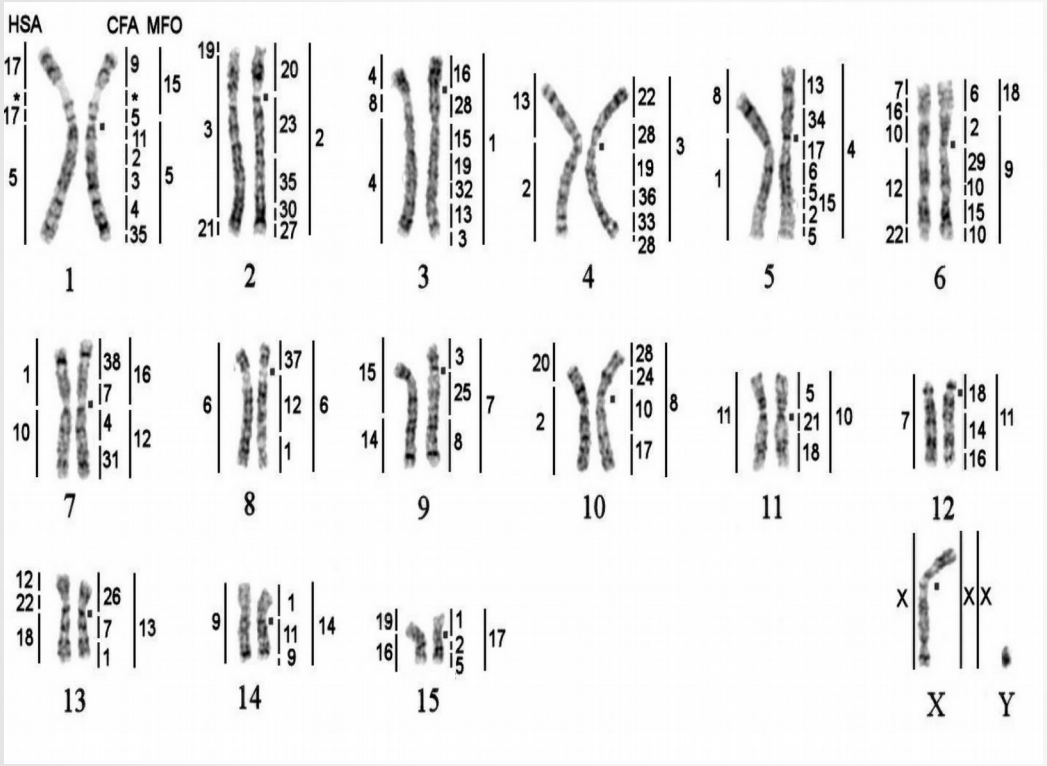
Somatic number (SN) is
number of all chromosomes in the cell





Examples of karyotypes (1)

G-bands



C-bands

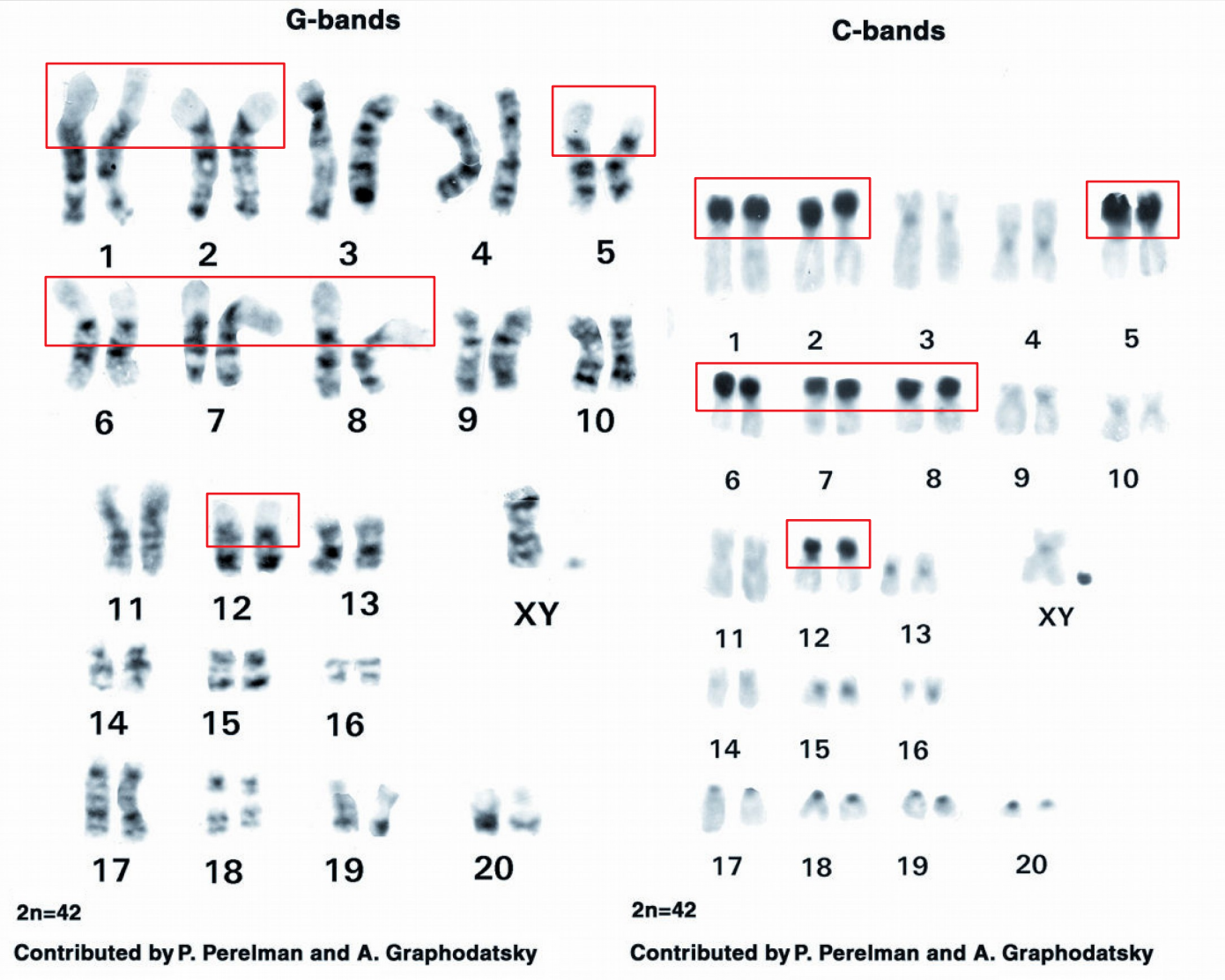
Beklemisheva et al, 2016

Beklemisheva et al, 2020

Pusa sibirica
Baikal seal

~ 2.2-2.3 Gbp

Examples of karyotypes (2)



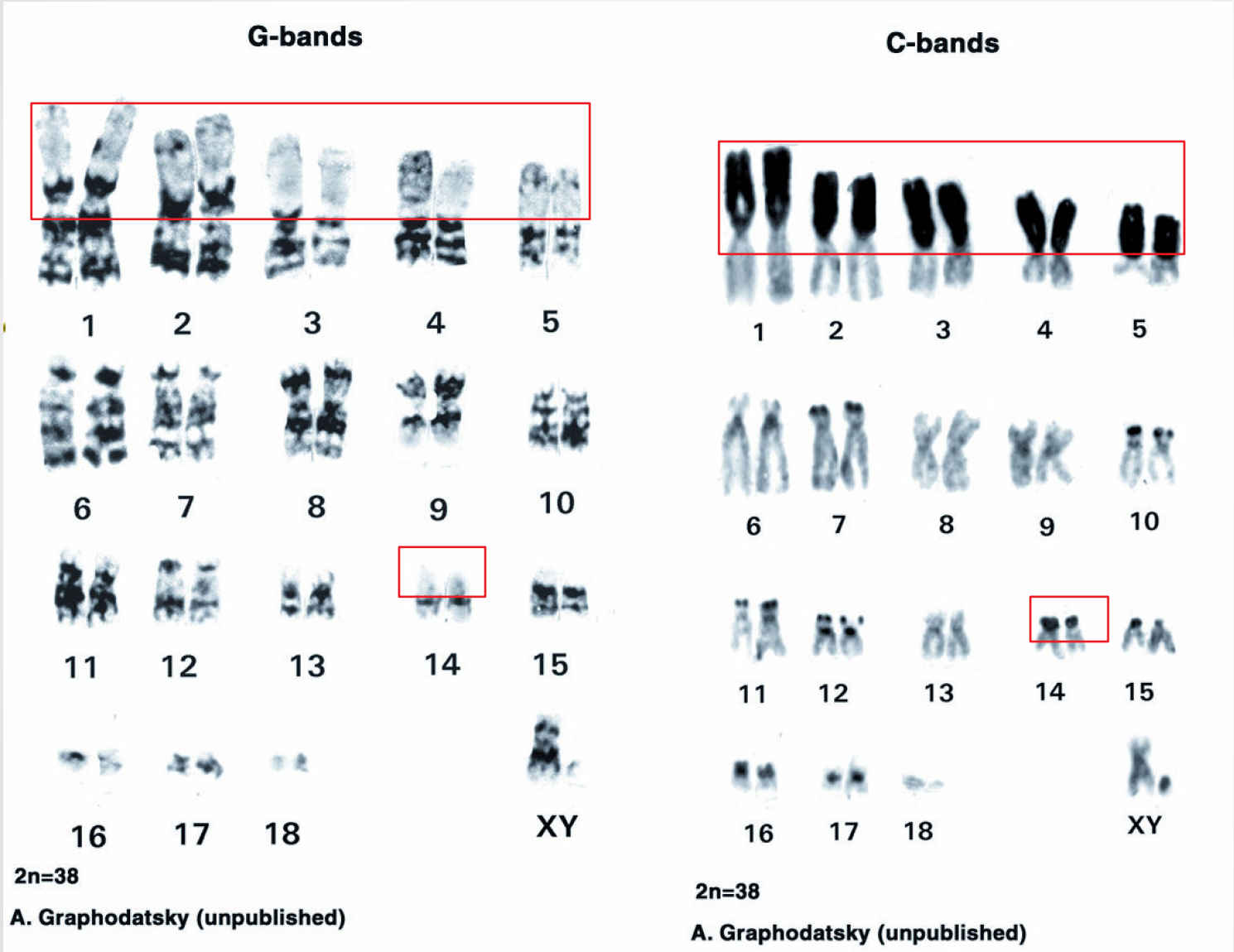
Graphodatsky et al, 2020

Mustela nivalis

the least weasel

~ 3.0 - 3.2 Gbp

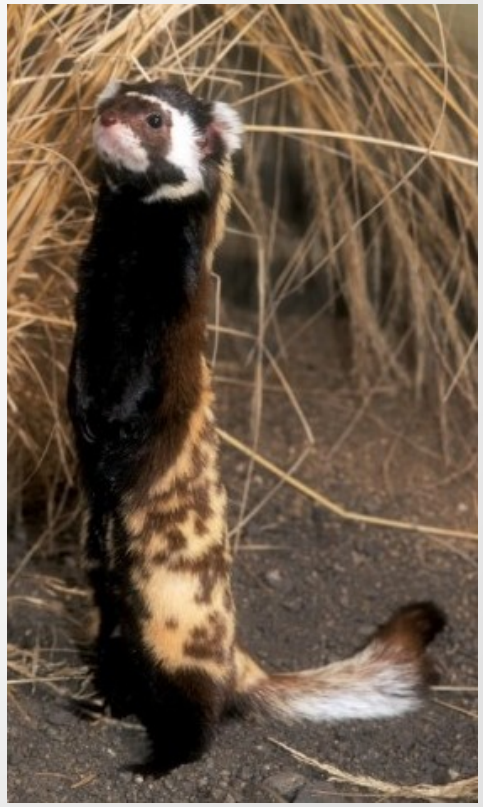
Examples of karyotypes (3)



Graphodatsky et al, 2020

Vormela peregusna

Marbled polecat

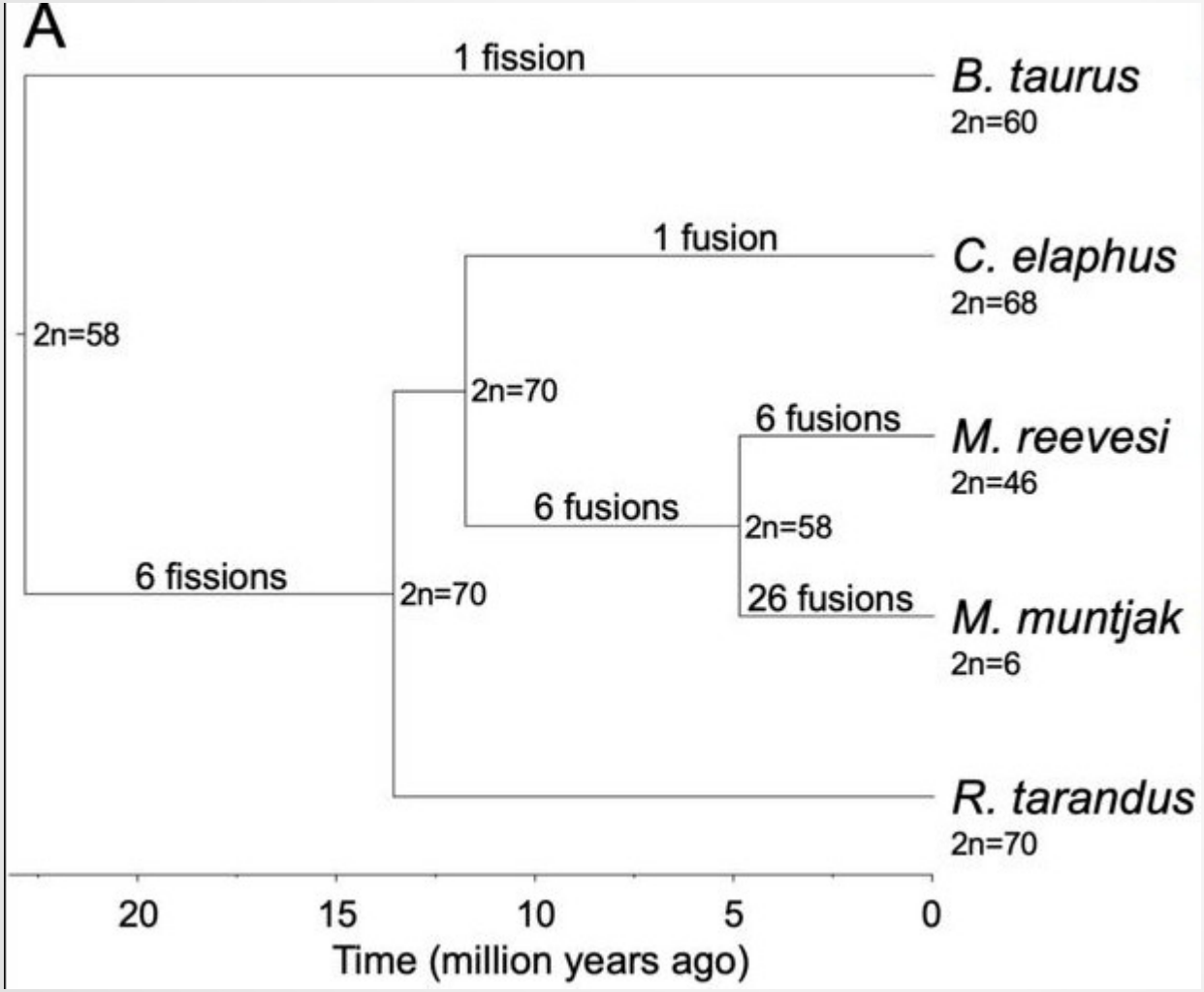


~ 6.4 Gbp (!)

How stable are karyotypes? Muntjac case

Indian muntjac (*Muntiacus muntjac*)

How stable are karyotypes? Muntjac case (1)



Cow

Red deer

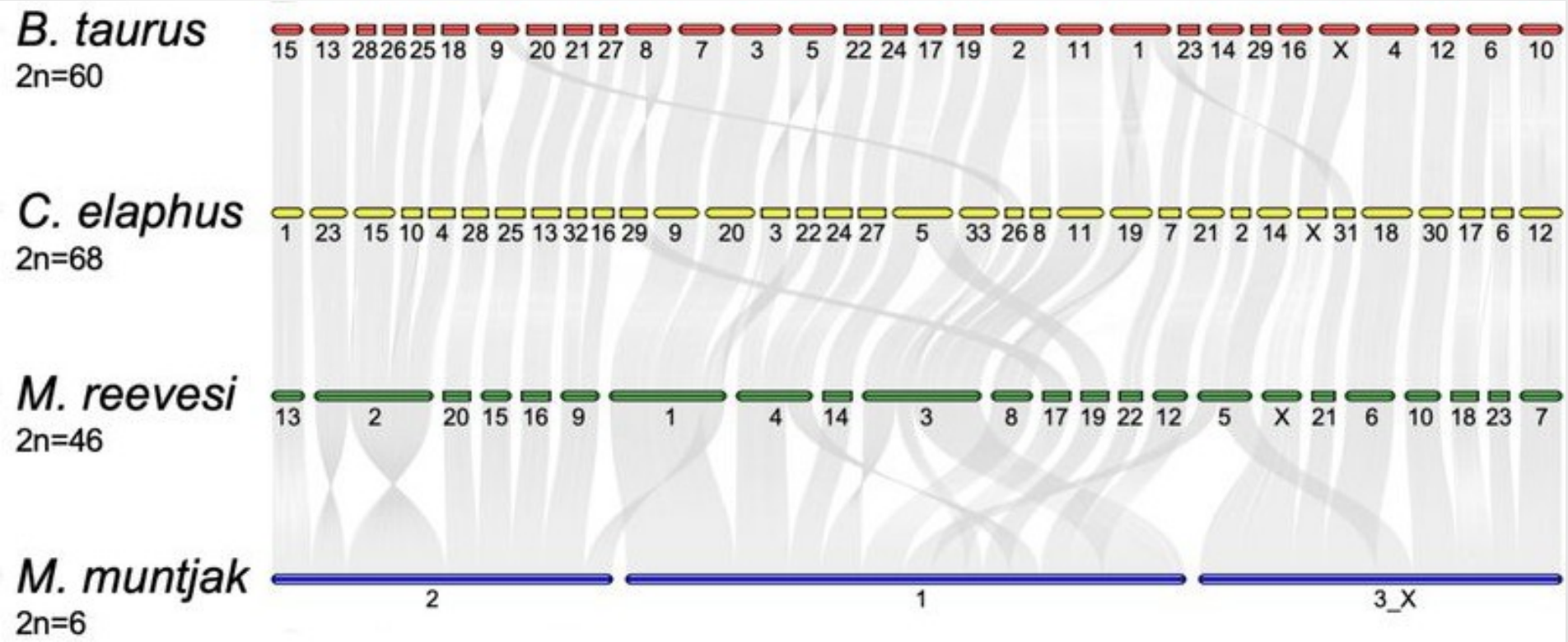
Chinese muntjac

Indian muntjac

Reindeer

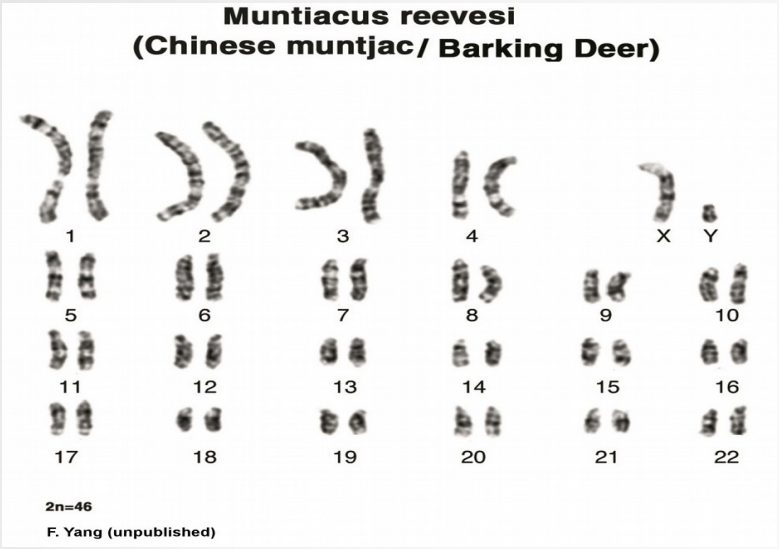
Mudd et al, 2020

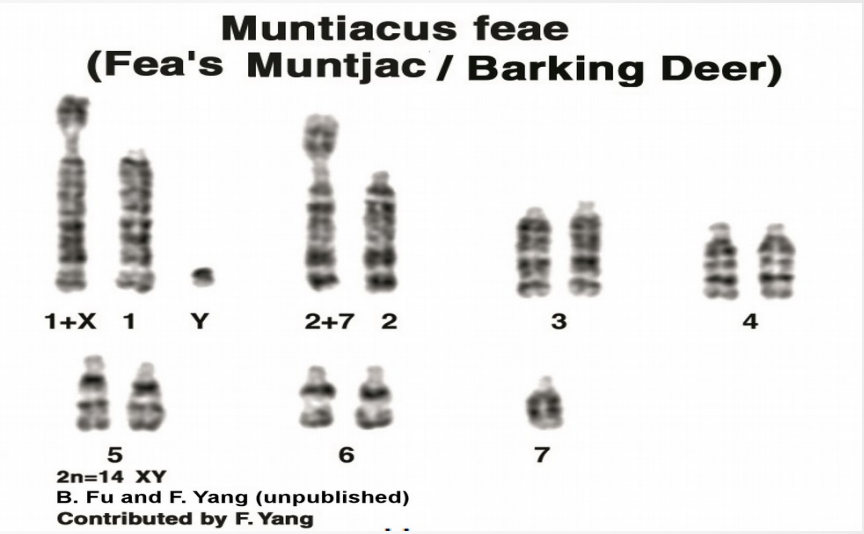
How stable are karyotypes? Muntjac case (2)

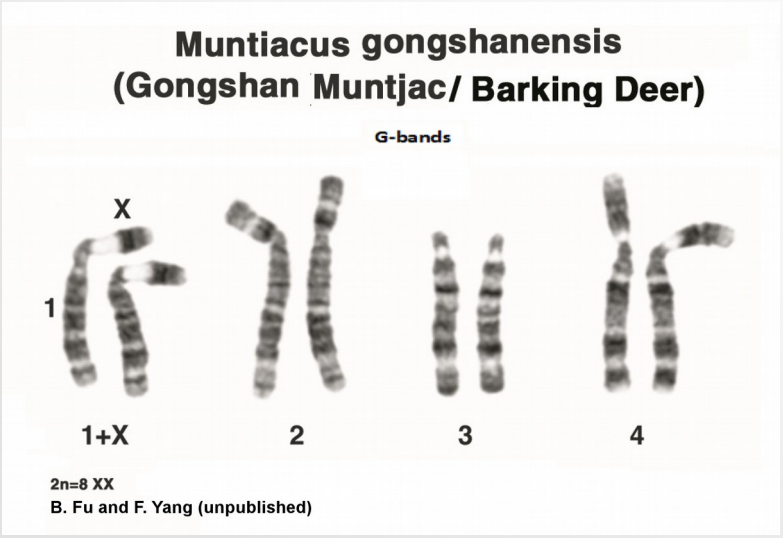


Mudd et al, 2020

How stable are karyotypes? Muntjac case (3)





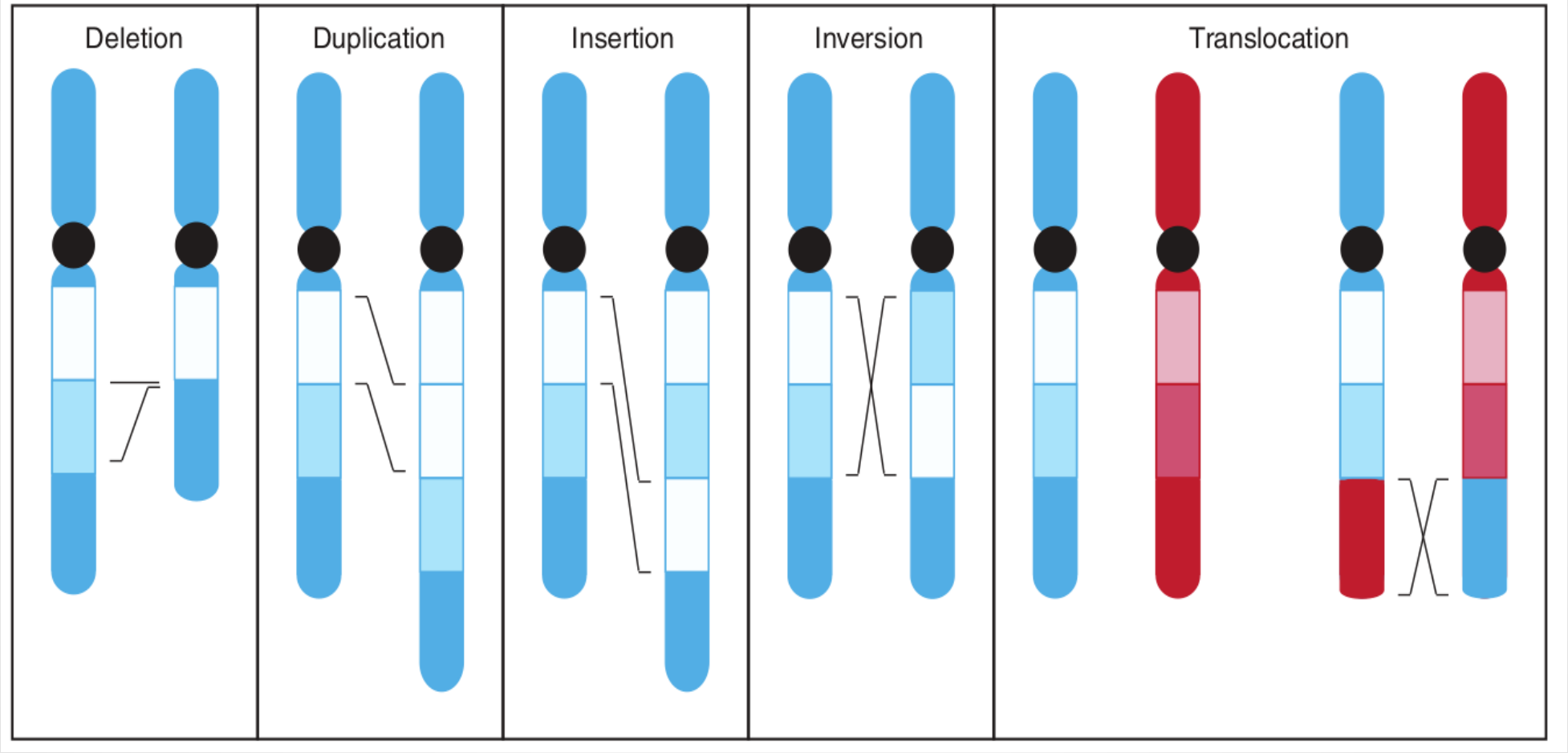


Chinese muntjak	2n=46
Fea's muntjak	2n=14
Gongshan munjak	2n=8
Indian muntjak	2n=6

Same genus, but very different number of chromosome

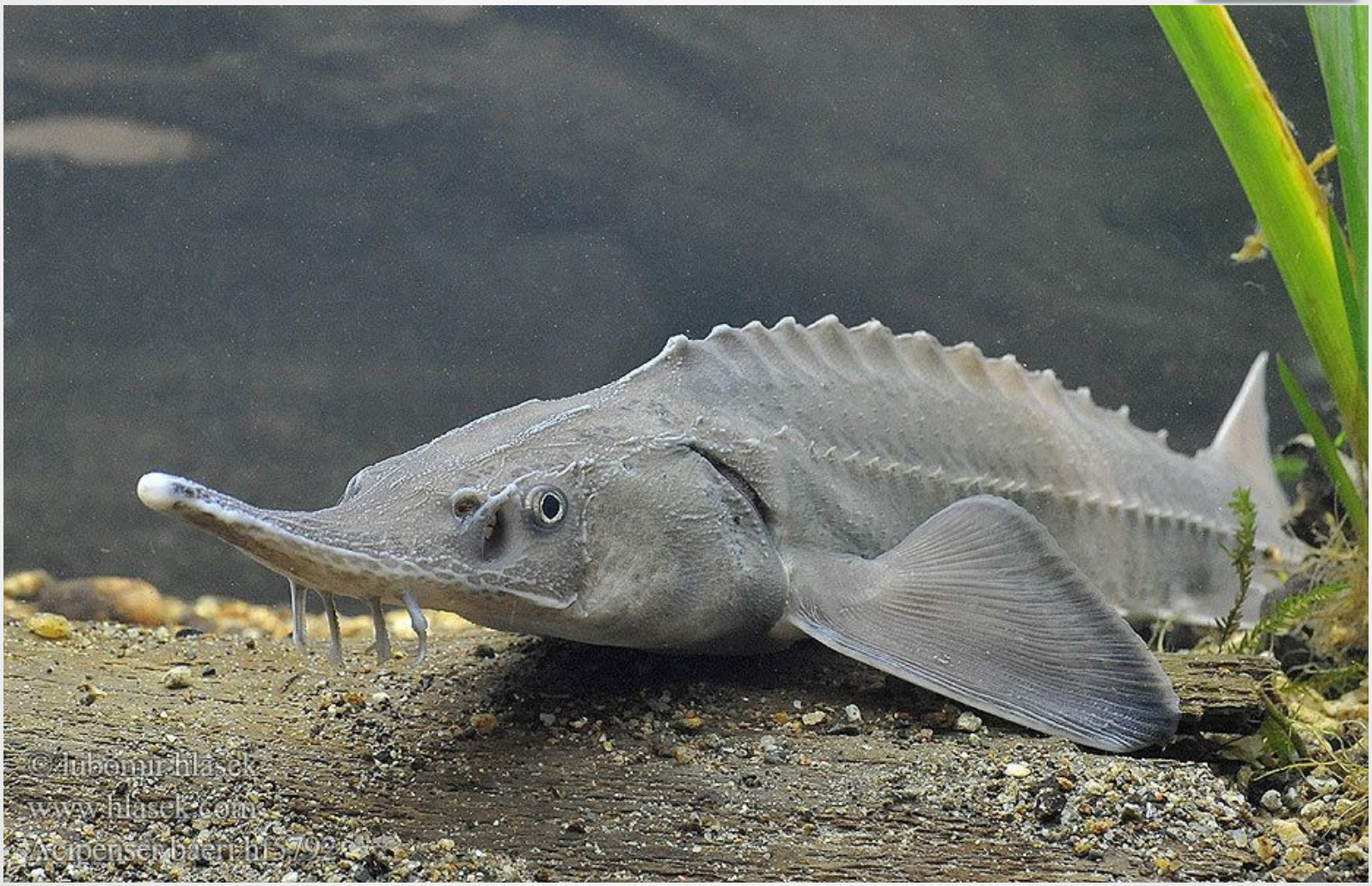
Graphodatsky et al, 2020

Basic types of genome rearrangements



Wijchers and de Laat, 2010

How stable are genomes? Case of sturgeons



Siberian sturgeon (*Acipenser baerii*)

Ploidy of different sturgeon species

Species	Chromosome number	Ploidy level		Reference
		"evolutionary scale"	"recent scale"	
Polyodon spathula (Walbaum)	120	4	2	Dingerkus and Howell, 1976
Scaphirhynchus platorynchus (Rafinesque)	~112	4	2	Ohno et al., 1969
Acipenser sturio Linnaeus	116 ± 4	4	2	Fontana and Colombo, 1974
A. nudiventris Lovetsky	118 ± 2	4	2	Arefjev, 1983; Sokolov and Vasil'ev, 1989a
A. ruthenus Linnaeus	118 ± 2	4	2	Fontana et al., 1975; Vasil'ev, 1985; Birstein and Vasil'ev, 1987
	118 ± 4	4	2	Ráb, 1986; Fontana, 1994
A. stellatus Pallas	118 ± 2	4	2	Vasil'ev, 1985; Birstein and Vasil'ev, 1987
A. oxyrinchus Mitchill	121 ± 3	4	2	Fontana et al., 2008b
A. huso Linnaeus	116 ± 4	4	2	Fontana and Colombo, 1974; Vasil'ev, 1985
A. gueldenstaedtii Brandt and Ratzeburg	250 ± 8	8	4	Vasil'ev, 1985; Vlasenko et al., 1989
A. persicus Borodin	~258	8	4	Nowruzfashkhami et al., 2000
A. baerii Brandt	249 ± 5	8	4	Vasil'ev et al., 1980; Sokolov and Vasil'ev, 1989b
A. naccarii Bonaparte	239 ± 7	8	4	Fontana and Colombo, 1974
A. brevirostrum Lesueur	~372	12	6	Kim et al., 2005
	372 ± 6	12	6	Fontana et al., 2008a
A. transmontanus Richardson	248 ± 8	8	4	Fontana, 1994
	~271	8	4	Van Eenennaam et al., 1998
A. sinensis Gray	264 ± 4	8	4	Yu et al., 1987
A. fulvescens Rafinesque	262 ± 6	8	4	Fonatana et al., 2004
A. mikadoi Hilgendorf	262 ± 4	8	4	Vasil'ev et al., 2008, 2009; this study
A. medirostris Ayres	249 ± 8	8	4	Van Eenennaam et al., 1999
A. schrenskii Brandt	238 ± 8	8	4	Song et al., 1997
	266 ± 4	8	4	This study
A. dauricus Georgi	268 ± 4	8	4	Vasil'ev et al., 2008, 2009; this study

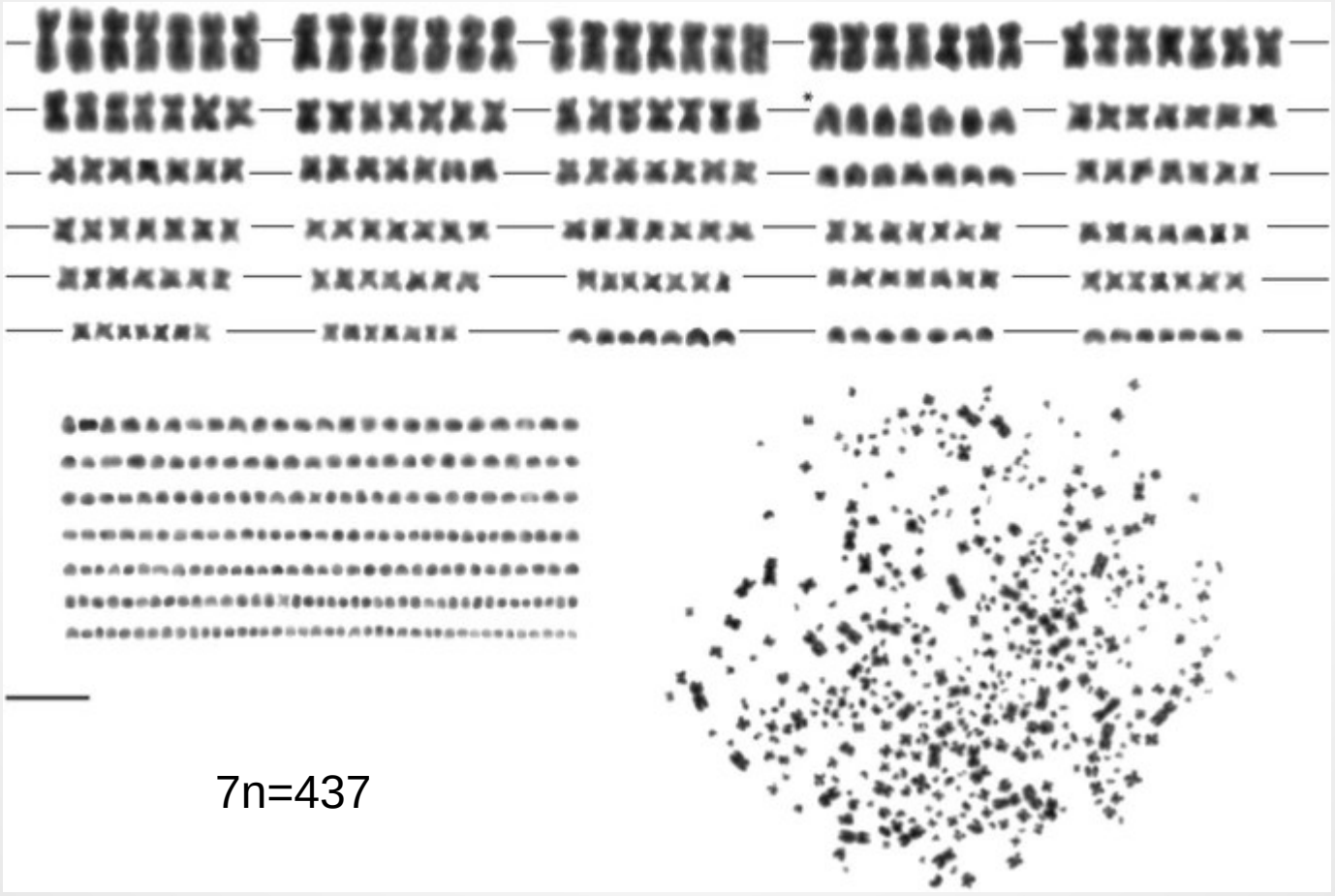
Polyploidization is an increase of ploidy.
Most common cases are
2n → 4n
2n → 3n

Diploidization is a conversion of polyploid species to diploid.
Some chromosomes might be lost, the rest diverge.

Vasi'ev et al, 2009

Spontaneous polyploidization

… is common in sturgeons!

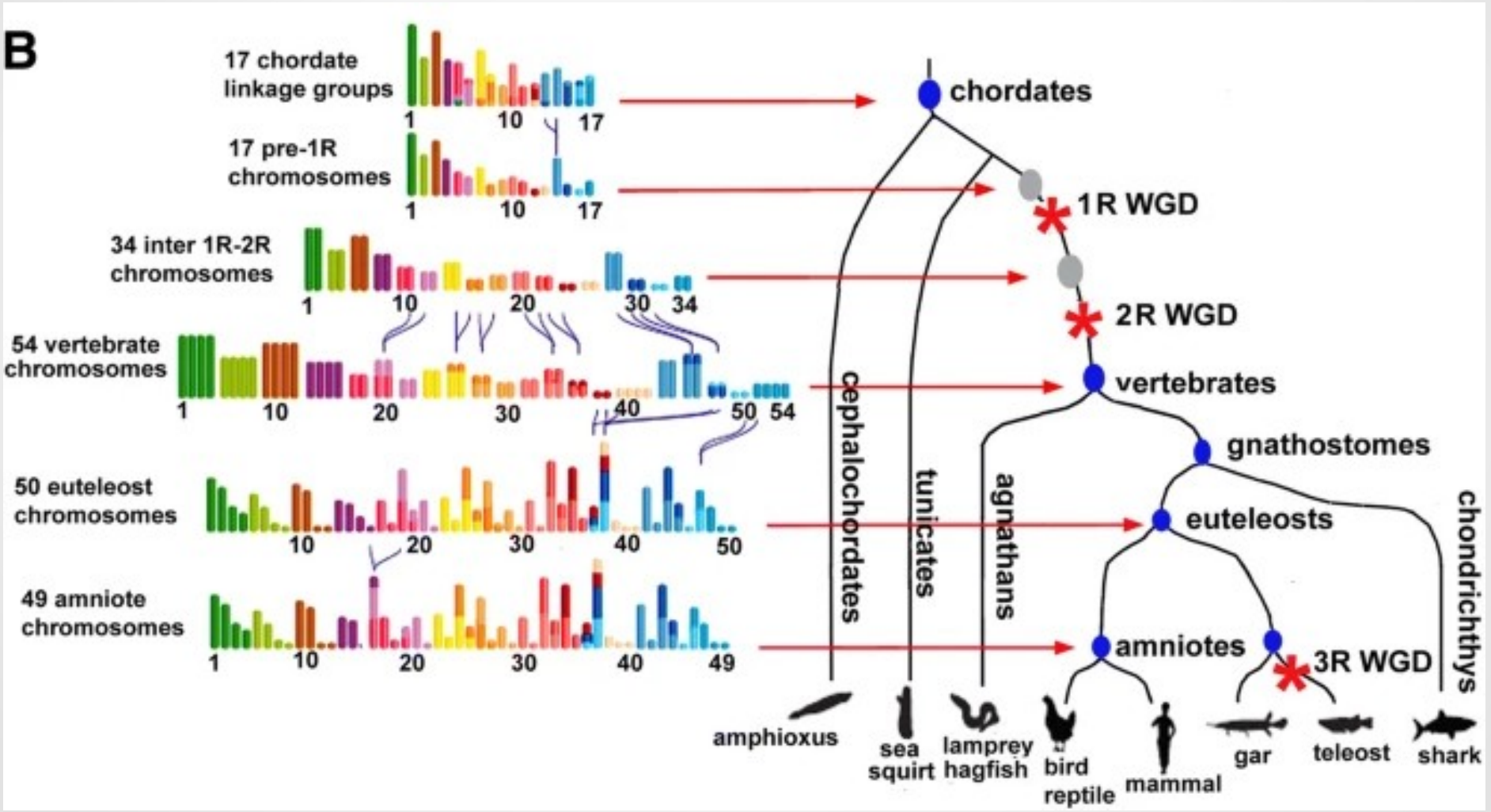


7n=437

Karyotype of heptaploid Siberian Sturgeon individual

Usual karyotype is 4n=250!

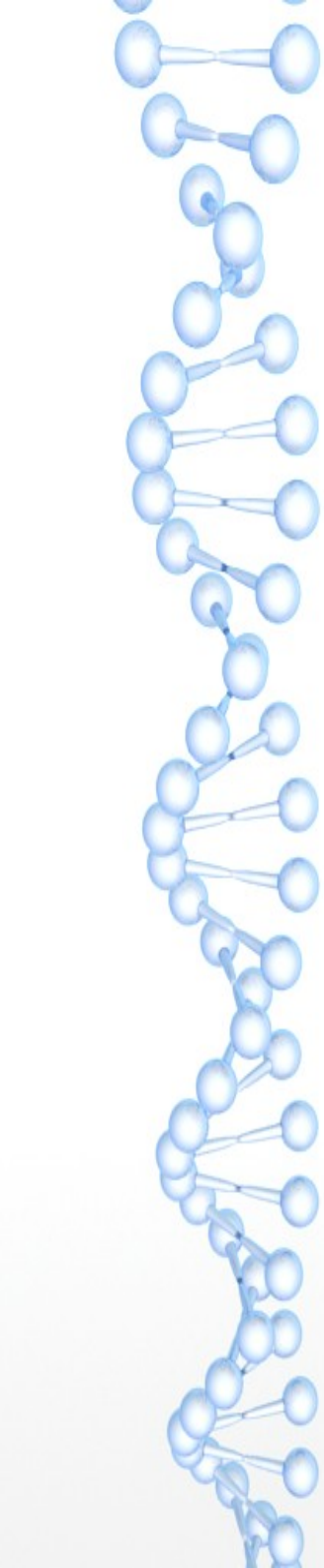
Whole genome duplications in evolution



Holland et al, 2018

Summary

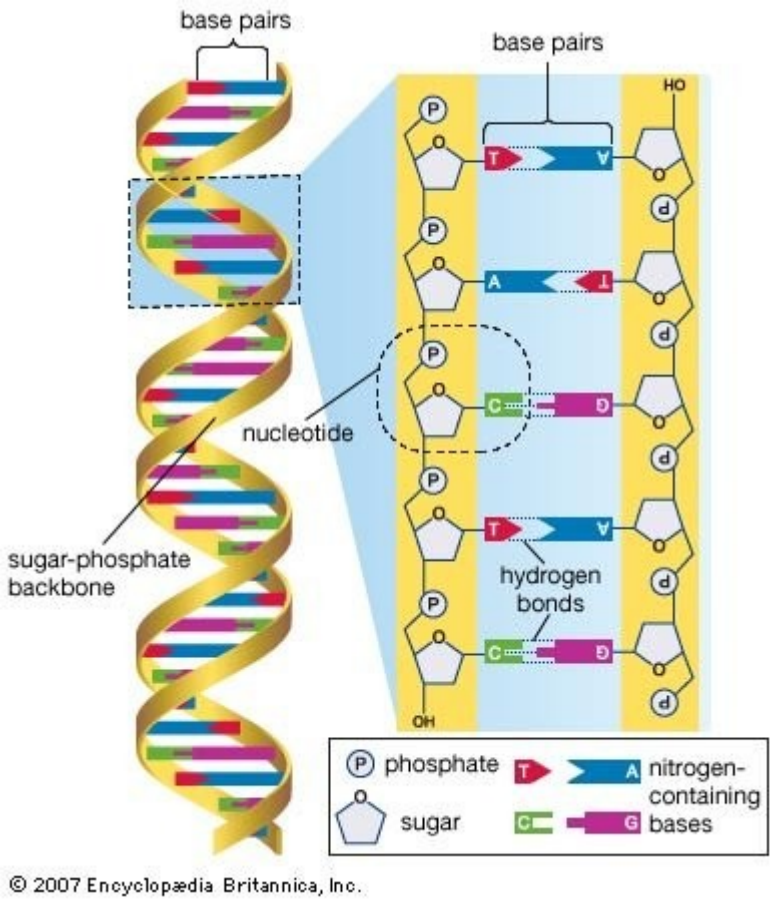
- DNA is not naked in cell. It is associated with proteins and with them forms chromosomes
- chromosomes have structural elements important for its stability and inheritance: telomers and centromers
- linear chromosomes have a specific problem of end replication, which is solved in eukaryotes by telomers and telomerase
- Karyotyping and differential stainning provides important information about genome structure: presence of absence of heterochromatine, chromosome number, ploidy and etc
- Genome size, ploidy and number of chromosomes can vary even in closely related species

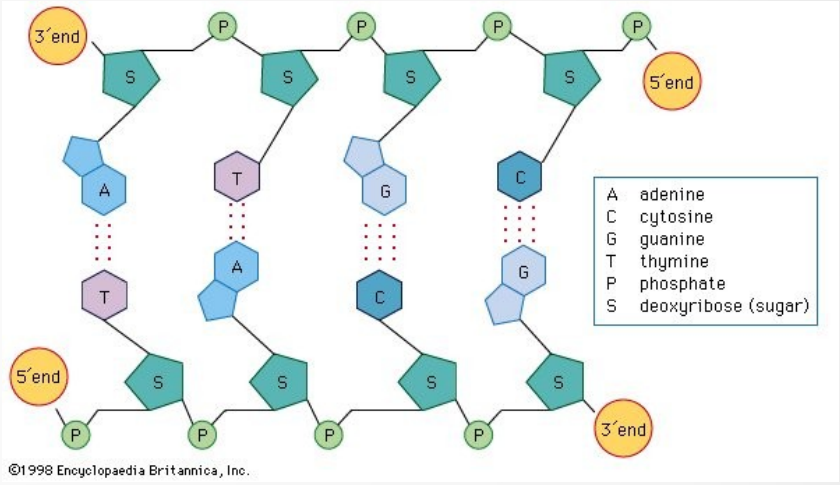


I. Structure and diversity of the genomes

Genome *in silico*
Assemblies and databases

DNA structure



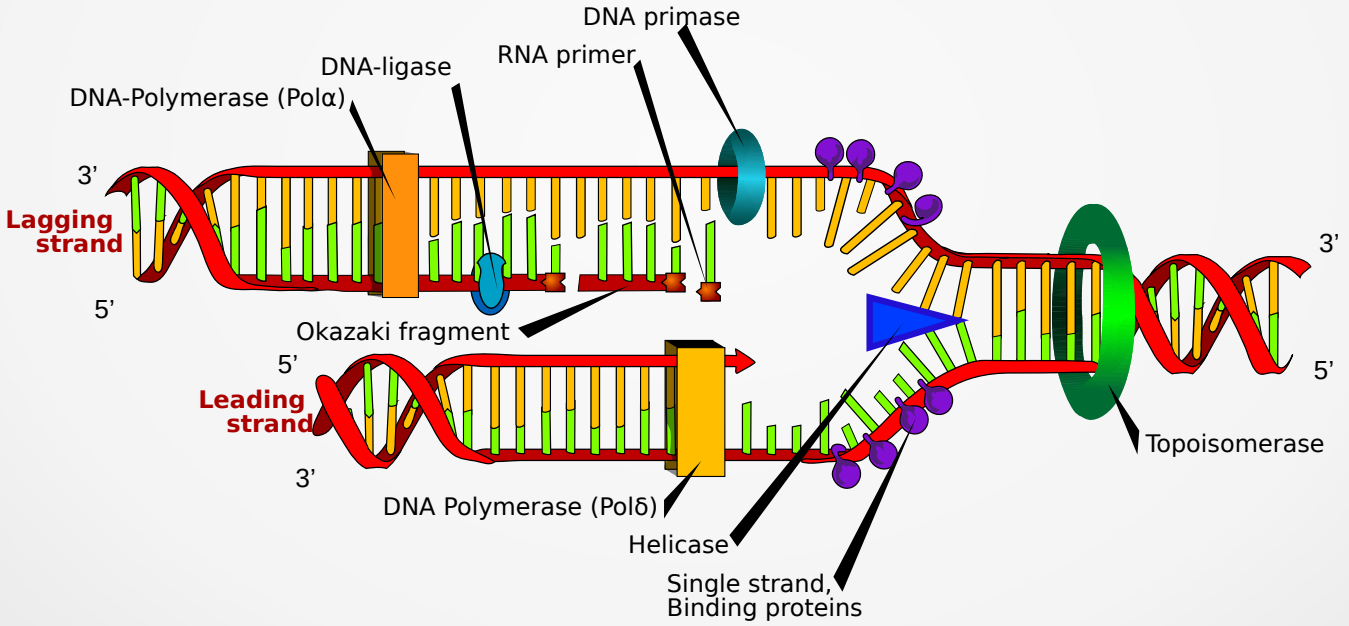


forward strand 5’ - TACG - 3’
reverse strand 3’ - ATGC - 5’

In databases all DNA/RNA sequences are stored as **one strand in 5’ → 3’ direction**

Why 5’ → 3’ orientation?

All DNA/RNA matrix processes in cell happen in 5’→3’ direction



During replication DNA polymerase synthesize complementary strand in 5’→ 3’

Nucleotides in databases

Symbol	Meaning	Description Origin
G	G	**G**uanine
A	A	**A**denine
T	T	**Thymine**
C	C	**C**ytosine
R	G or A	pu**R**ine
Y	T or C	p**Y**rimidine
M	A or C	a**M**ino
K	G or T	**K**etone
S	G or C	**S**trong interaction
W	A or T	**W**eak interaction
H	A or C or T	**H** follows G in alphabet
B	G or T or C	**B** follows A in alphabet
V	G or C or A	**V** follows U in alphabet
D	G or A or T	**D** follows C in alphabet
N	G or A or T or C	a**N**y

unambiguous nucleotides

ambiguous nucleotides

Sequence formats

FASTA

GENBANK

NEXUS

Fasta format

header1

sequence1

header2

sequence2

header3

sequence3

```
>sequence1 Description1
gtaacatgcagttcttccggacttttctattatagactcgtt
ctgctattgtcactaagtcctgcaagctcttctctcccggtct
ctctactctttgcaatttcttttaatgttcagaacGGCCTGg
ctaacaaaagccatgataattgcggcctttgtttccggggctt
ctgggttcataggggtgtactgcctaaaagcctctattgctcc
ctgtaagaaggctgctggactctcatccttaccctgtctcaca
tcatagaccttggccagattggtagtcttgcacgtggcagcct
>sequence2 Description2
ggagacctgccattagagtctggcggtggacccggagtcgctc
cttaccttcaggcgagttgtgtgctttggggagtaaaaacggg
ttttagccaaggaggggaattctcgatcatgtcctgacaggcc
aggatgttgggcatctggtcagggtggccatctggtttgtccc
tgaagttcacggccttaacctgtaagataataggtaggtgaaa
>sequence3 Description3
tcacggccttaacctgtaagataataggtaggtgaaa
```

Extensions

- most common .fa, .fasta, .fna, .faa
- often used .cds, .pep

Multiple sequence in file

- each sequence is preceded by header
- sequence could be one line or multiline (prefered for long sequences)

Header format

- `>sequence2 Description2`

sequence id

description

space

Genbank format

Extensions
most common .gb, .genbank

Contains metadata

Contains annotations

Multiple sequence in file
sequence records are separated by //
followed and preceded by newline symbol

Most of fields are optional
It even could contain only annotation

```
LOCUS       MZ571765                406 bp    DNA     linear   INV 10-OCT-2021
DEFINITION  Haemoproteus tinnunculi isolate R3 cytochrome b (cytb) gene,
            partial cds; mitochondrial.
ACCESSION   MZ571765
VERSION     MZ571765.1
KEYWORDS    .
SOURCE      mitochondrion Haemoproteus tinnunculi
  ORGANISM  Haemoproteus tinnunculi
            Eukaryota; Sar; Alveolata; Apicomplexa; Aconoidasida; Haemosporida;
            Haemoproteidae; Haemoproteus.
REFERENCE   1  (bases 1 to 406)
  AUTHORS   Alrefaei,A.F.
  TITLE     Molecular diversity of falcon blood parasites in Saudi Arabia:
            Cytochrome b lineages of the genera Haemoproteus (Haemosporida)
            from Riyadh
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 406)
  AUTHORS   Alrefaei,A.F.
  TITLE     Direct Submission
  JOURNAL   Submitted (15-JUL-2021) Zoology Department, King Saud University,
            King Khalid Road, Riyadh, Riyadh 2455, KSA
COMMENT     ##Assembly-Data-START##
            Sequencing Technology :: Sanger dideoxy sequencing
            ##Assembly-Data-END##
FEATURES             Location/Qualifiers
     source          1..406
                     /organism="Haemoproteus tinnunculi"
                     /organelle="mitochondrion"
                     /mol_type="genomic DNA"
                     /isolate="R3"
                     /db_xref="taxon:2588639"
     gene            <1..>406
                     /gene="cytb"
     CDS             <1..>406
                     /gene="cytb"
                     /codon_start=3
                     /transl_table=4
                     /product="cytochrome b"
                     /protein_id="UBY00644.1"
                     /translation="HILRGLNYSYVYLPLSWITGLVIFLISIVTAFMGYVLPWGQMSF
                     WGATVITNLLYFIPGLVSWICGGYTISDPTLKRFFVLHFIFPFIALCIVFIHIFFLHL
                     QGSSNPLGYDTALKIPFYPSLLCLDVKGFNNV"
ORIGIN
        1 tacatattct aagaggatta aattattcat atgtatattt acctttatca tggataactg
       61 gattagttat attcctaatc tctatagtca ctgcttttat gggttatgtt ctaccttggg
      121 gtcaaatgag tttctggggt gcaacagtta ttactaattt attatatttc ataccaggac
      181 tagtatcatg gatttgtggt ggatatacta ttagtgaccc tactttaaaa agattctttg
      241 tattacactt tatattccca tttatagccc tatgtatcgt atttatacat atattcttct
      301 tacatttaca aggtagctct aacccttag gatatgatac agctttaaaa ataccttct
      361 atccaagtct attatgttta gatgttaaag gatttaataa tgtatt
//
```

metadata

annotations

sequence

Nexus format

Extensions
most common .nex, .nexus

Could contain nearly everything. Even commands for tools

Multiple sequence in file

Most of fields are optional

Mostly used in phylogenetics and phylogeography

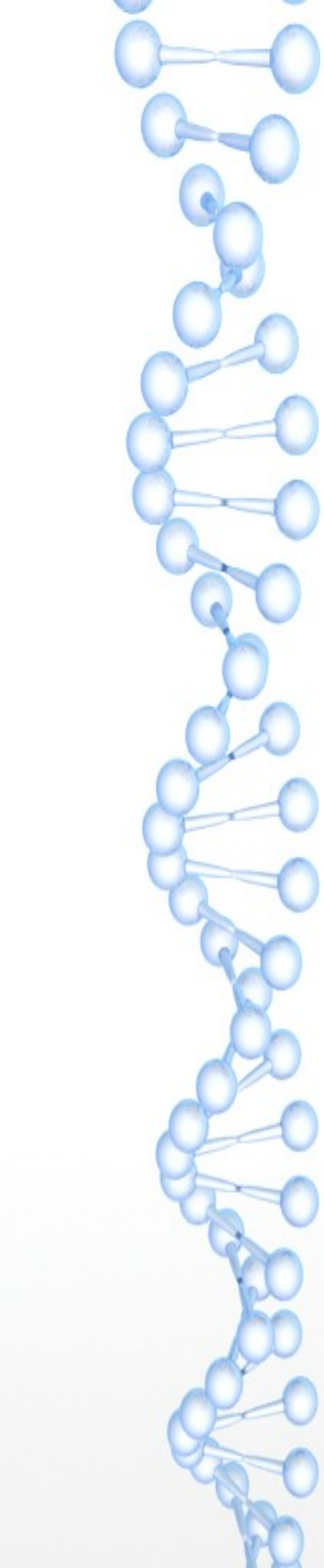
```
#NEXUS
Begin TAXA;
  Dimensions ntax=4;
  TaxLabels SpaceDog SpaceCat SpaceOrc SpaceElf
End;

Begin data;
  Dimensions nchar=15;
  Format datatype=dna missing=? gap=- matchchar=.;
  Matrix
    SpaceDog   atgctagctagctcg
    SpaceCat   ......??...-.a.
    SpaceOrc   ...t.......-.g.
    SpaceElf   ...t.......-.a.
  ;
End;

BEGIN TREES;
  Tree tree1 =
(((SpaceDog,SpaceCat),SpaceOrc,SpaceElf));
END;
```

Available genome databases

- **NCBI Genome** https://www.ncbi.nlm.nih.gov/genome
  - huge whale of bioinformatics
- **Ensembl** https://www.ensembl.org
  - just an elephant
- **GenomeArk** https://vgp.github.io/genomeark/
  - contains genomes assembled according VGP standards
- **DNAzoo** https://dnazoo.org
  - contains genomes assembled by DNAzoo team to chromosome level

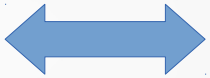


I. Structure and diversity of the genomes

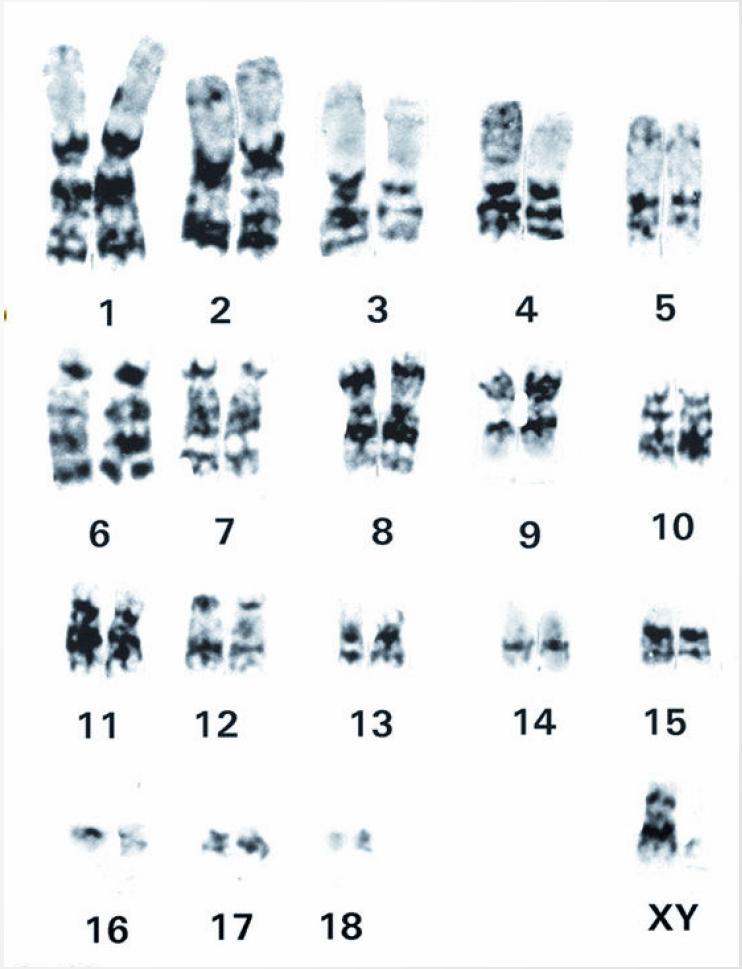
Genome *in silico*
Connection between assembly and karyotype

How to assign chromosome names to C-scaffolds?

```
>HiC_scaffold_1
gtaacatgcagttcttccggacttttttctattatagactcgtt
ctgctattgtcactaagtcctgcaagctcttctctcccggtct
ctctactctttgcaatttctttttaatgttcagaacGGCCTGg
ctaacaaaagccatgataattgcggcctttgtttccggggctt
ctgggttcataggggtgtactgcctaaaagcctctattgctcc
ctgtaagaaggctgctggactctcatccttaccctgtctcaca
tcatagaccttggccagattggtagtcttgcacgtggcagcct
…
…
…
tttaatgttcagaacGGCCTGgctaacaaaagccatgataatt
>HiC_scaffold_2
ggagacctgccattagagtctggcggtggacccggagtcgctc
cttaccttcaggcgagttgtgtgctttggggagtaaaaacggg
ttttagccaaggaggggaattctcgatcatgtcctgacaggcc
aggatgttgggcatctggtcagggtggccatctggtttgtccc
tgaagttcacggccttaacctgtaagataataggtaggtgaaa
…
…
```

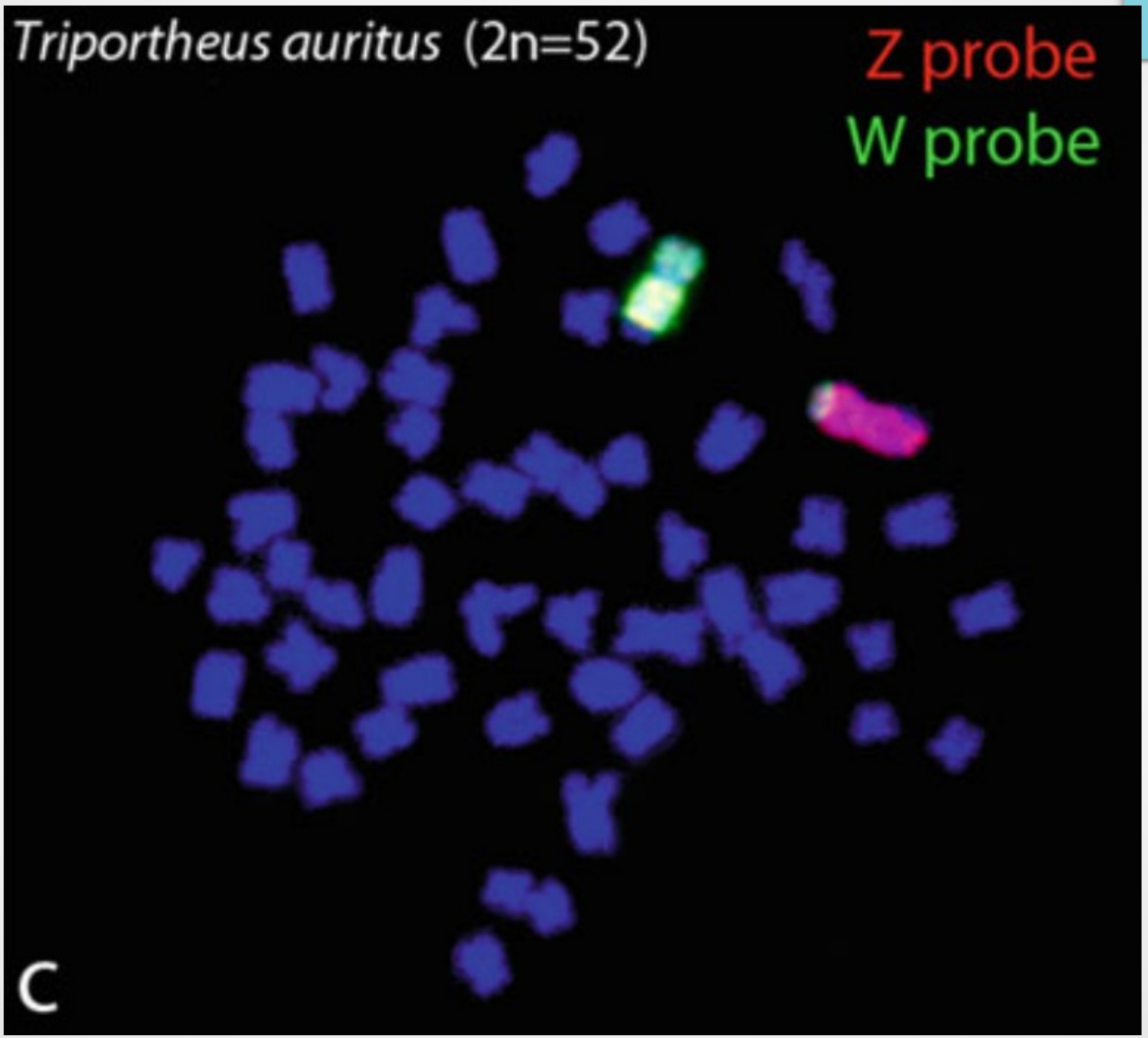


?



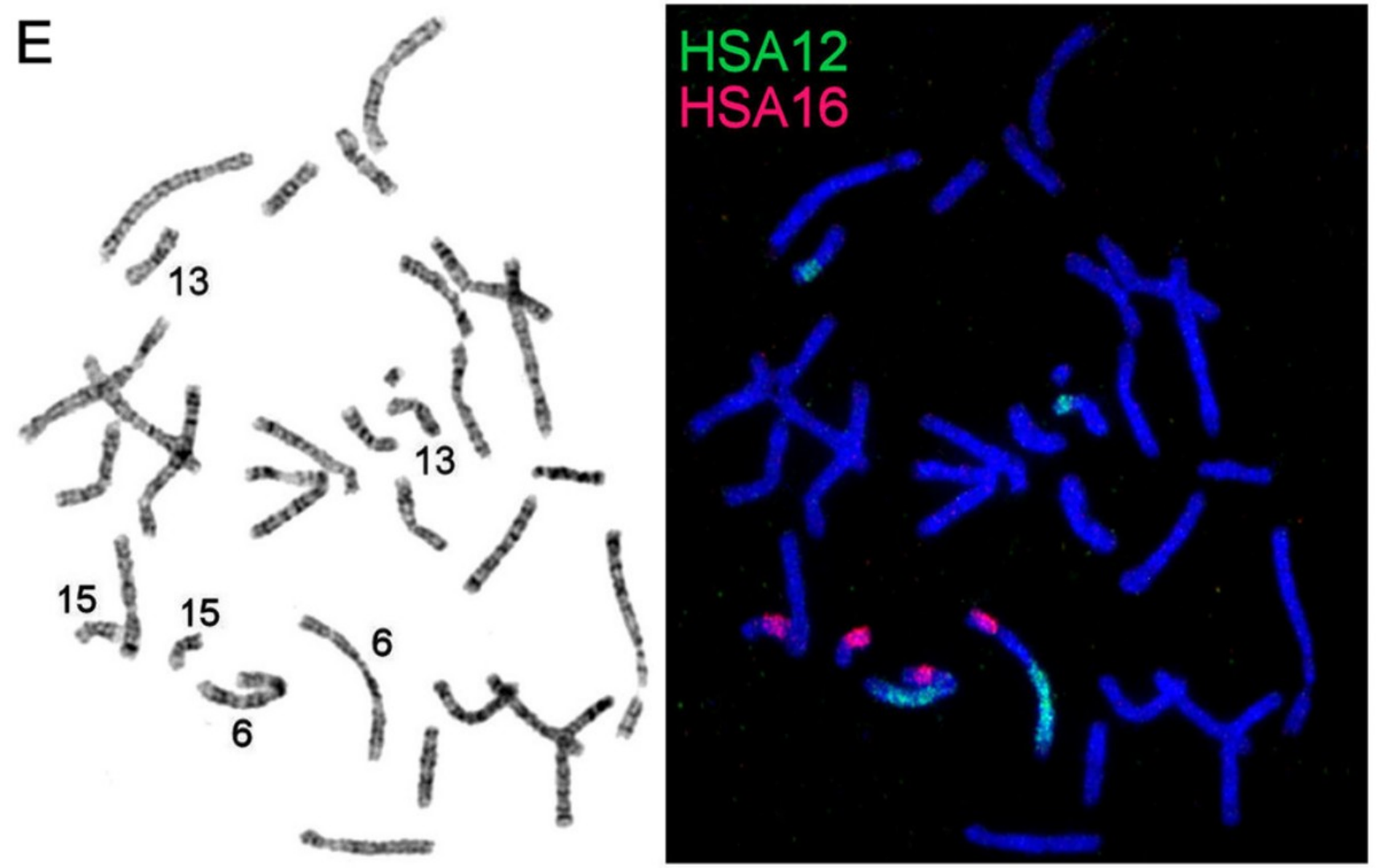
We need markers which we could detect both in the assembly and karyotype.
What can we use as marker? How many of them do we need?

Cross-species chromosome painting (ZooFISH)



Yano et al, 2017

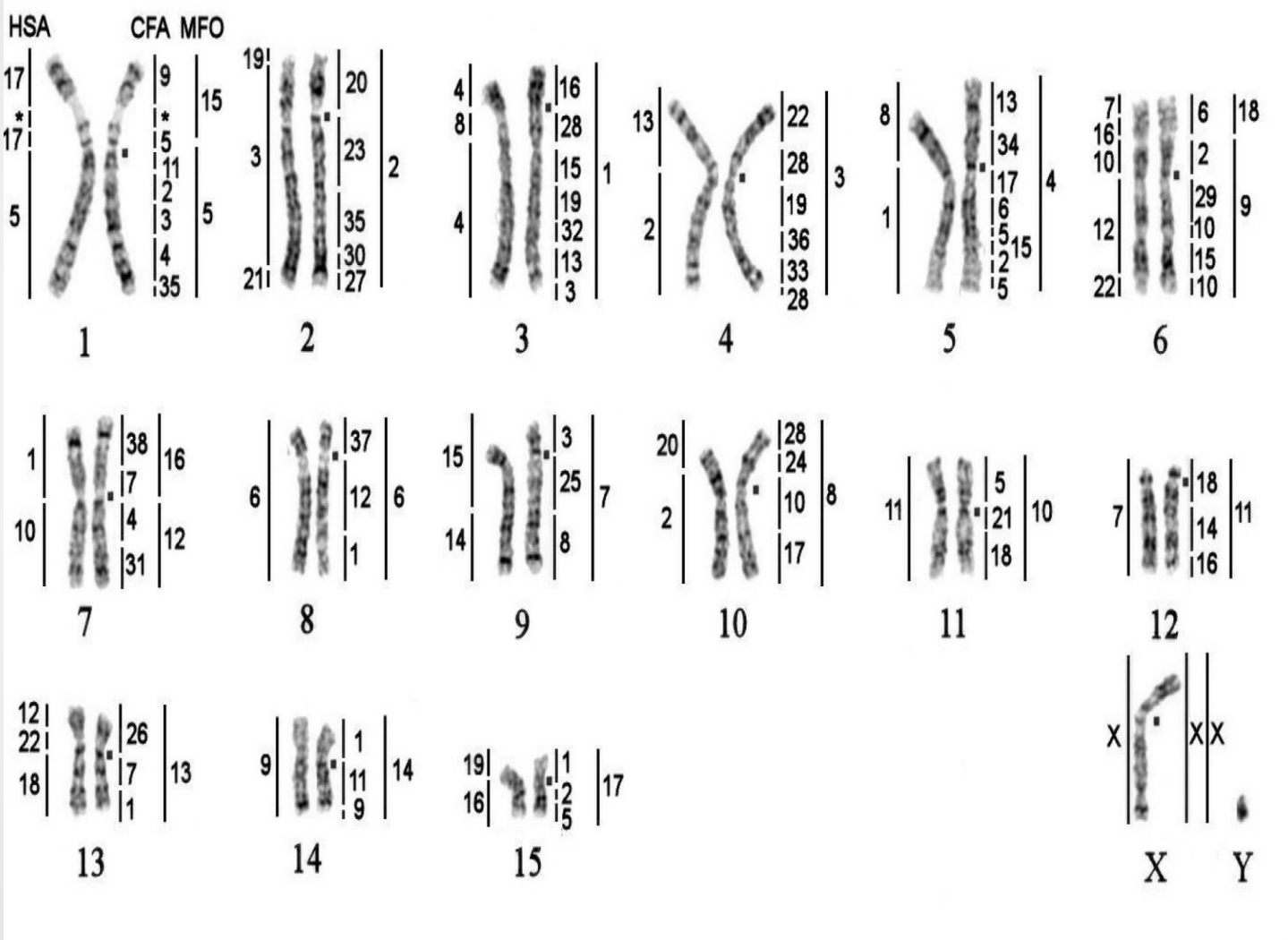
Cross-species chromosome painting (ZooFISH)



Beklemisheva et al, 2016

human probes hybridized with chromosomes of Baikal seal

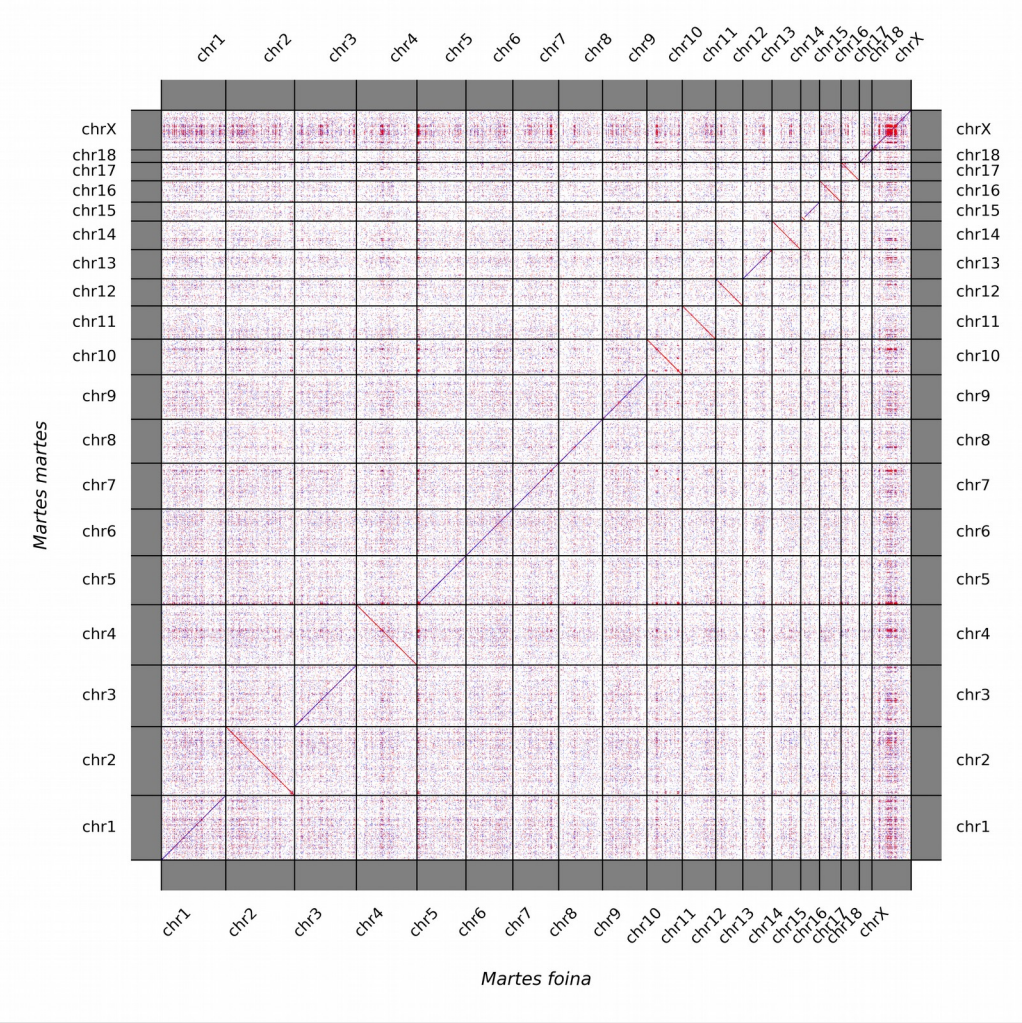
FISH maps

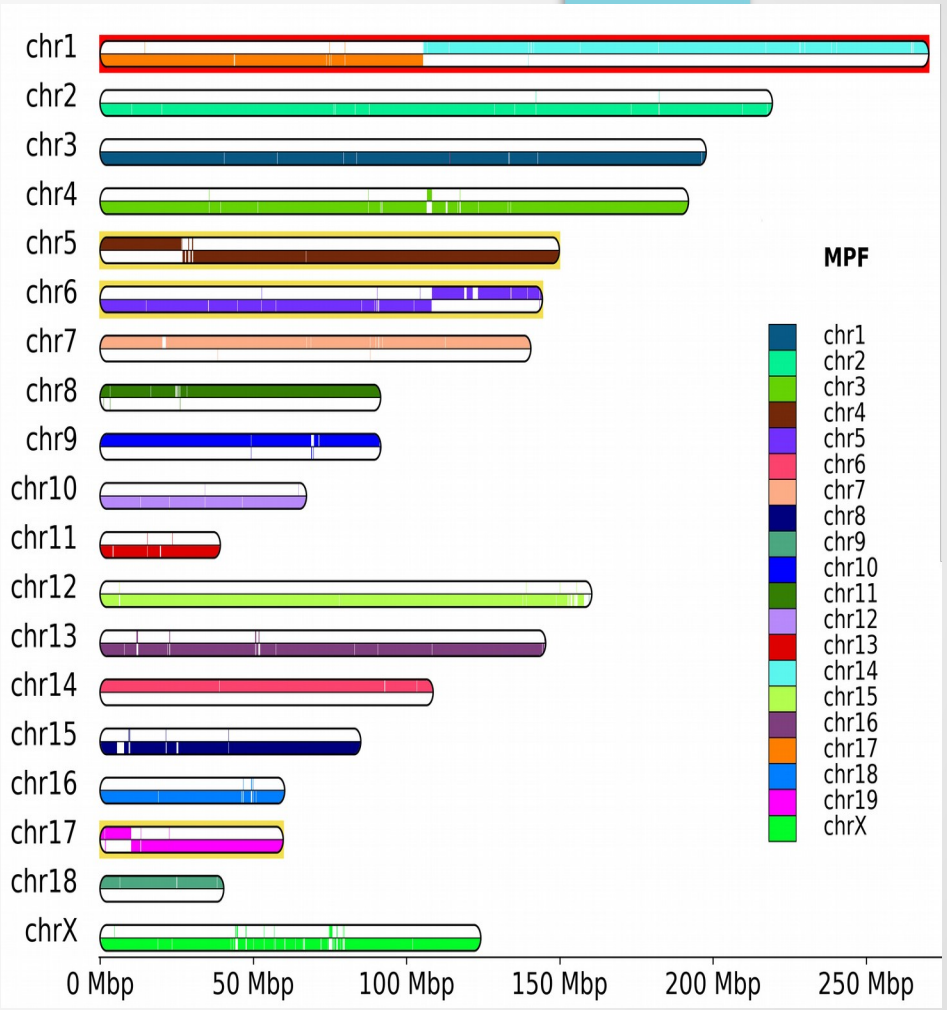


Beklemisheva et al, 2016

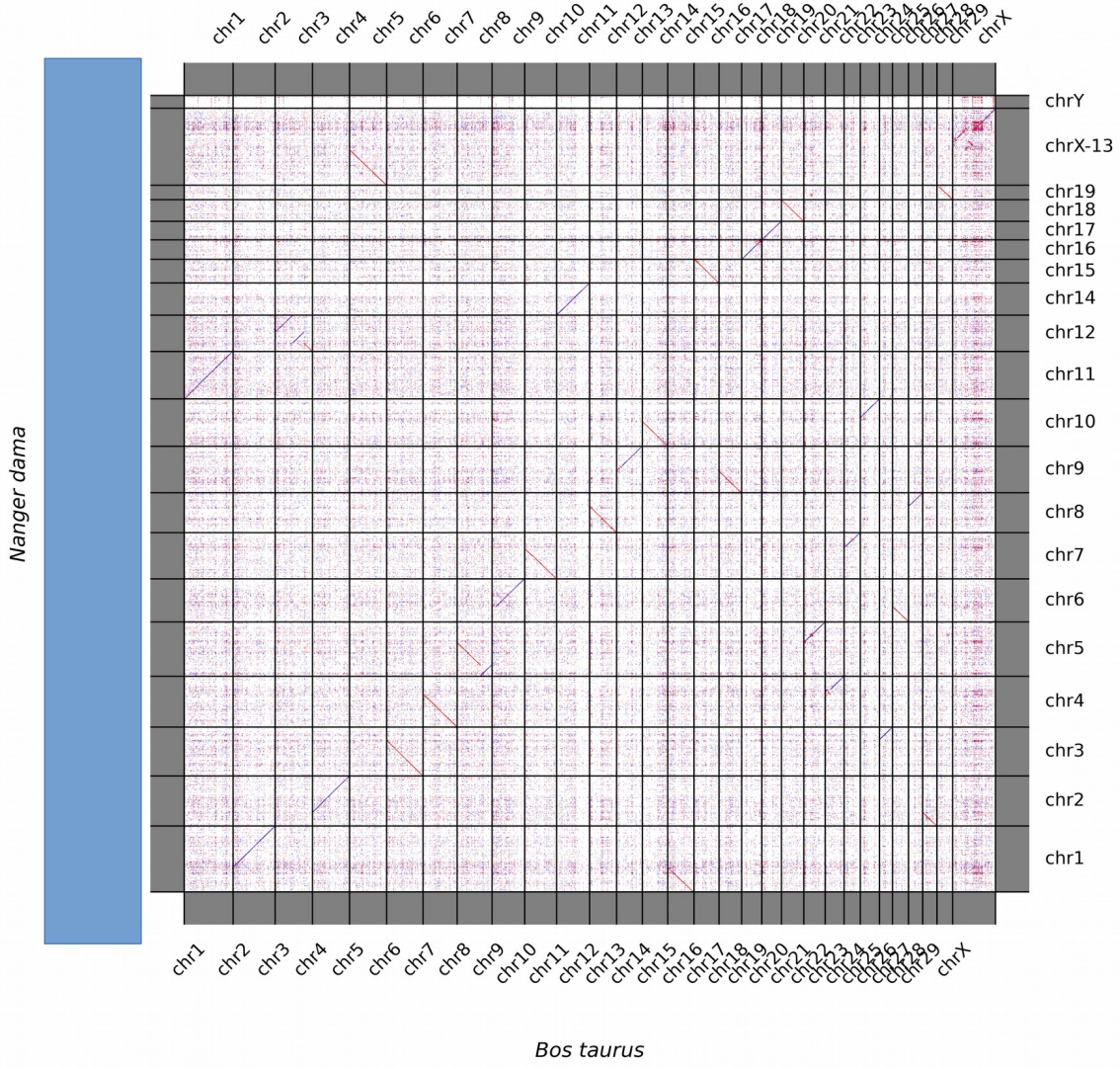
FISH maps of Baikal seal based on human (HSA), dog (CFA) and stone marten (MFO) probes

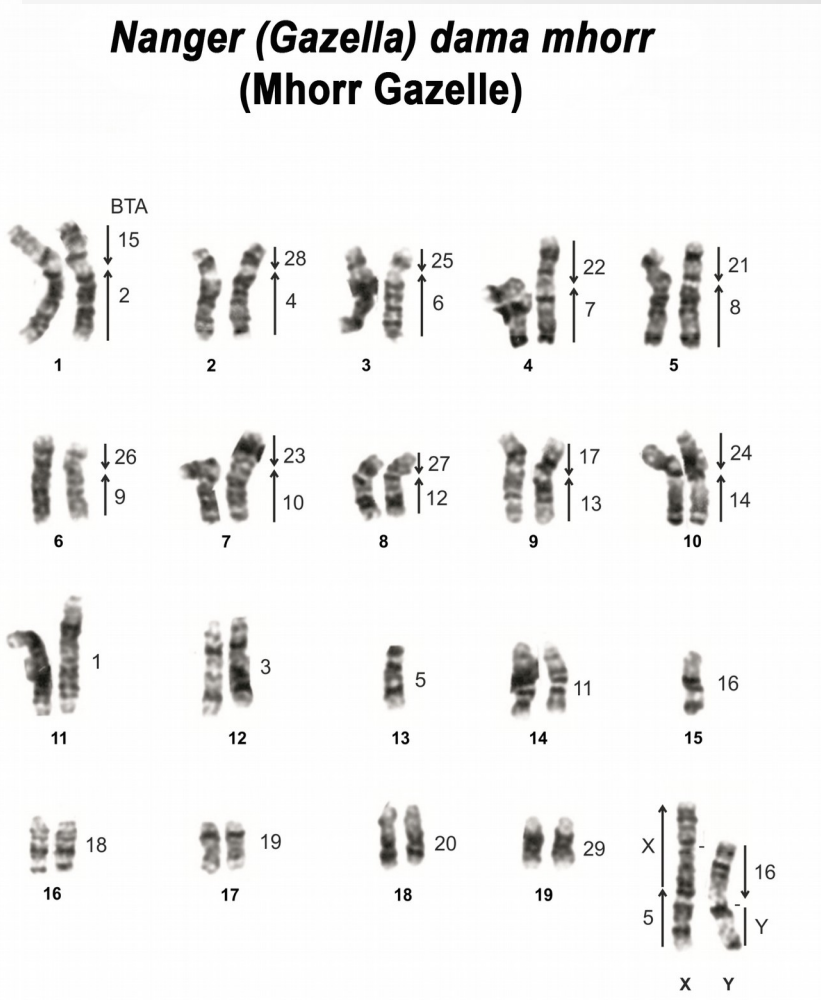
Whole genome alignment





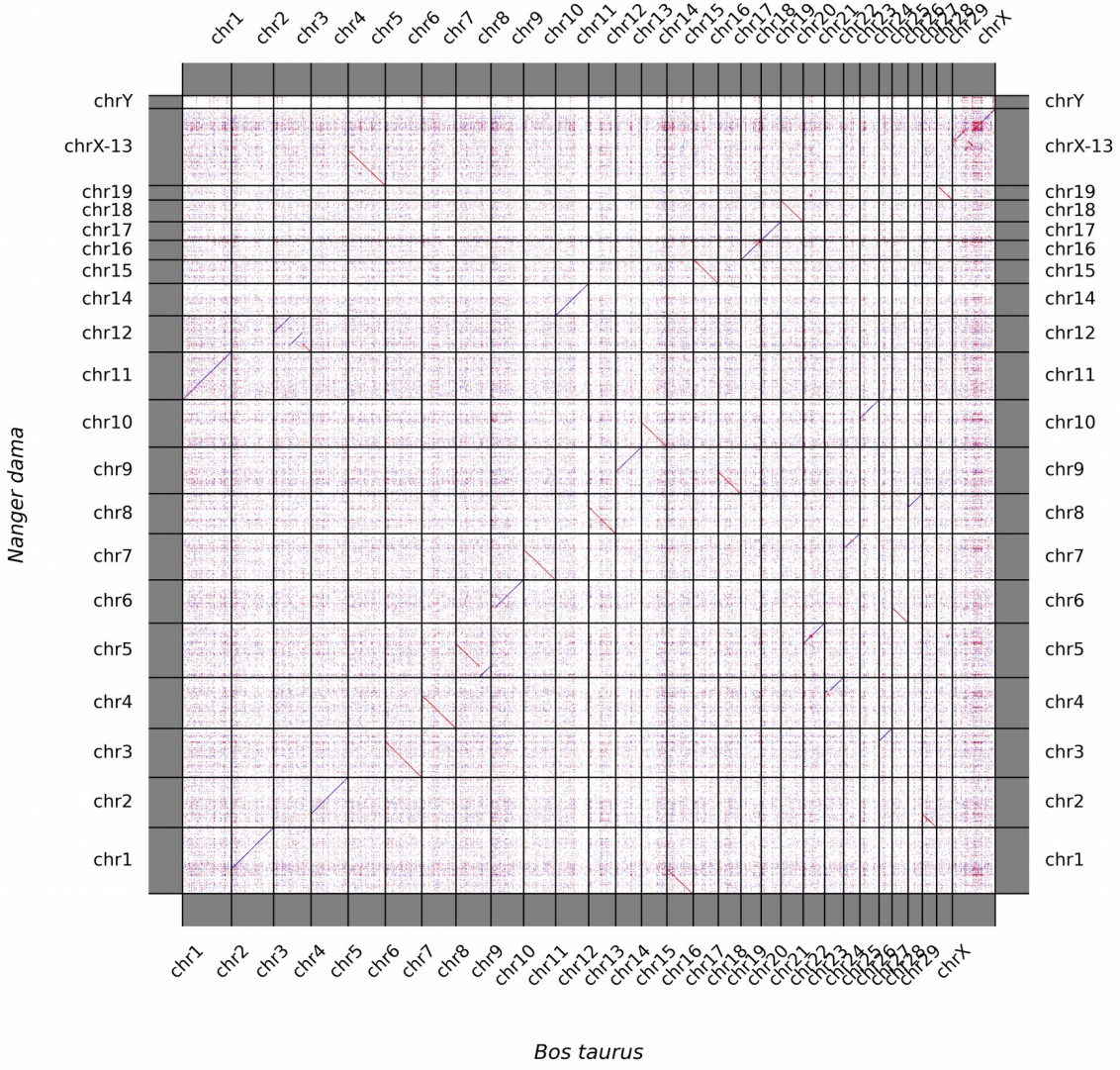
Correspondence between WGA and FISH map (1)

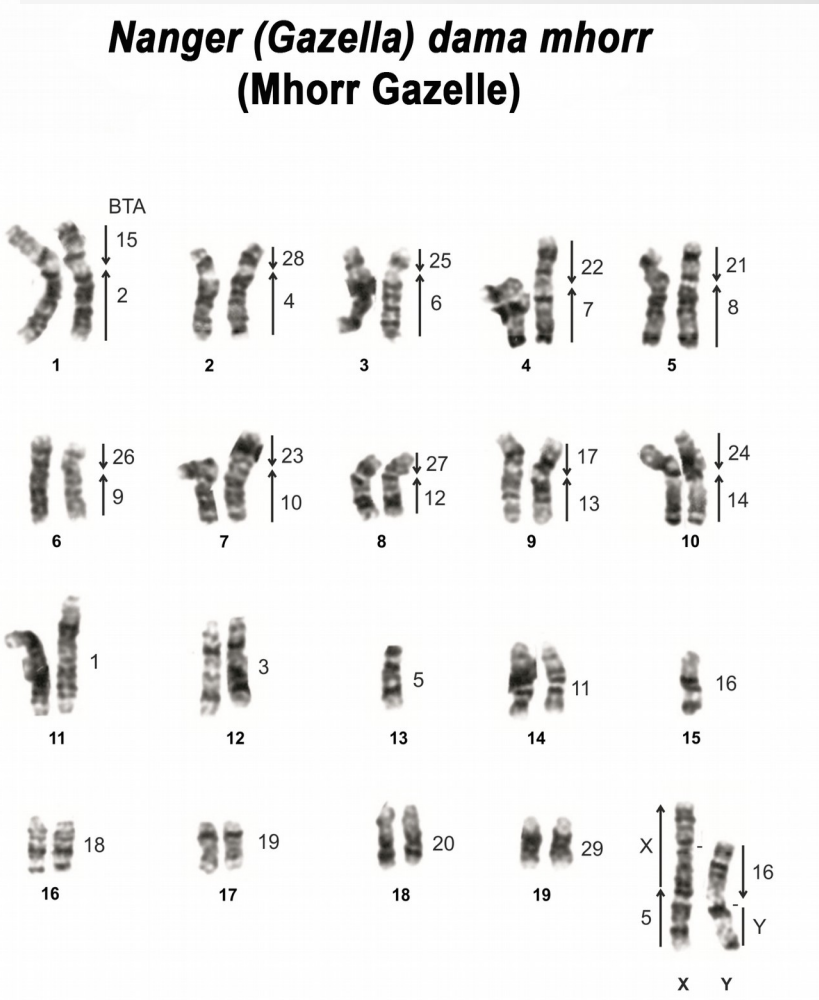




Chernohorska et al, 2012

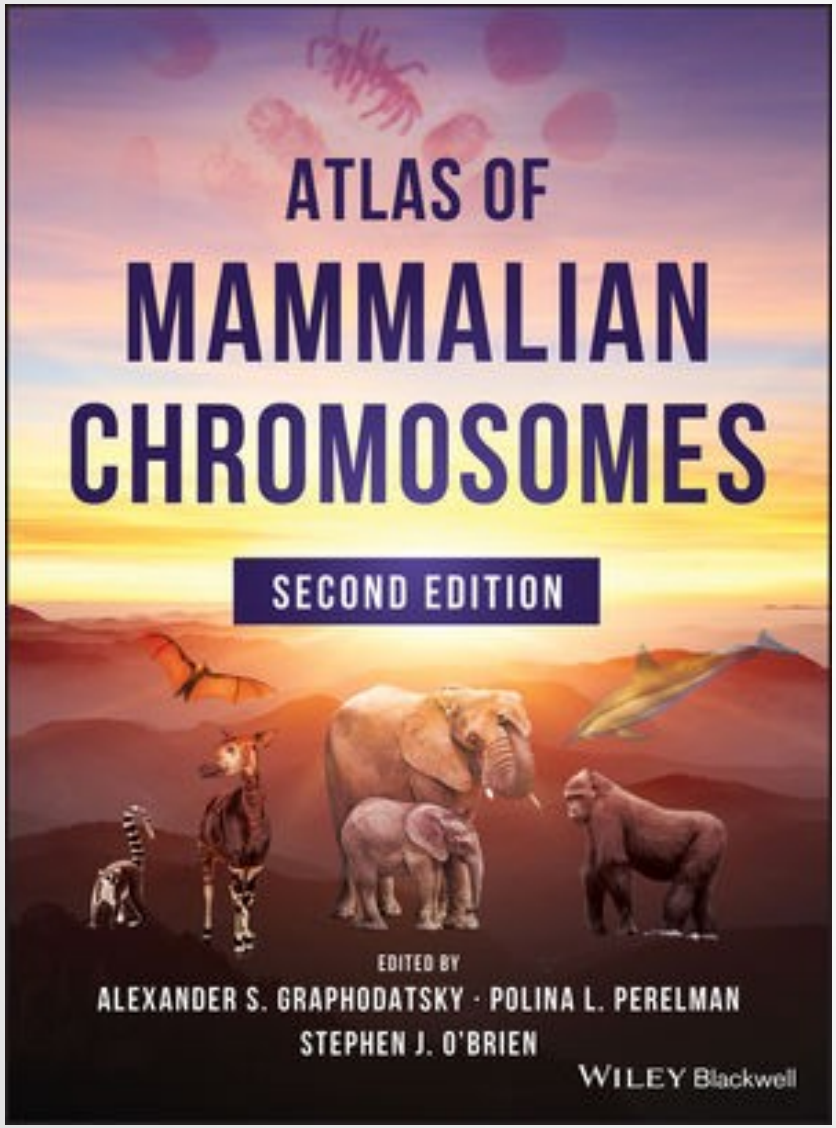
Correspondence between WGA and FISH map (2)





Chernohorska et al, 2012

Atlas of mammalian chromosomes



Big book contains
hundrends karyotypes and FISH maps

Very useful for *de novo* assembly

Rules for ordering and orientation of chromosomes

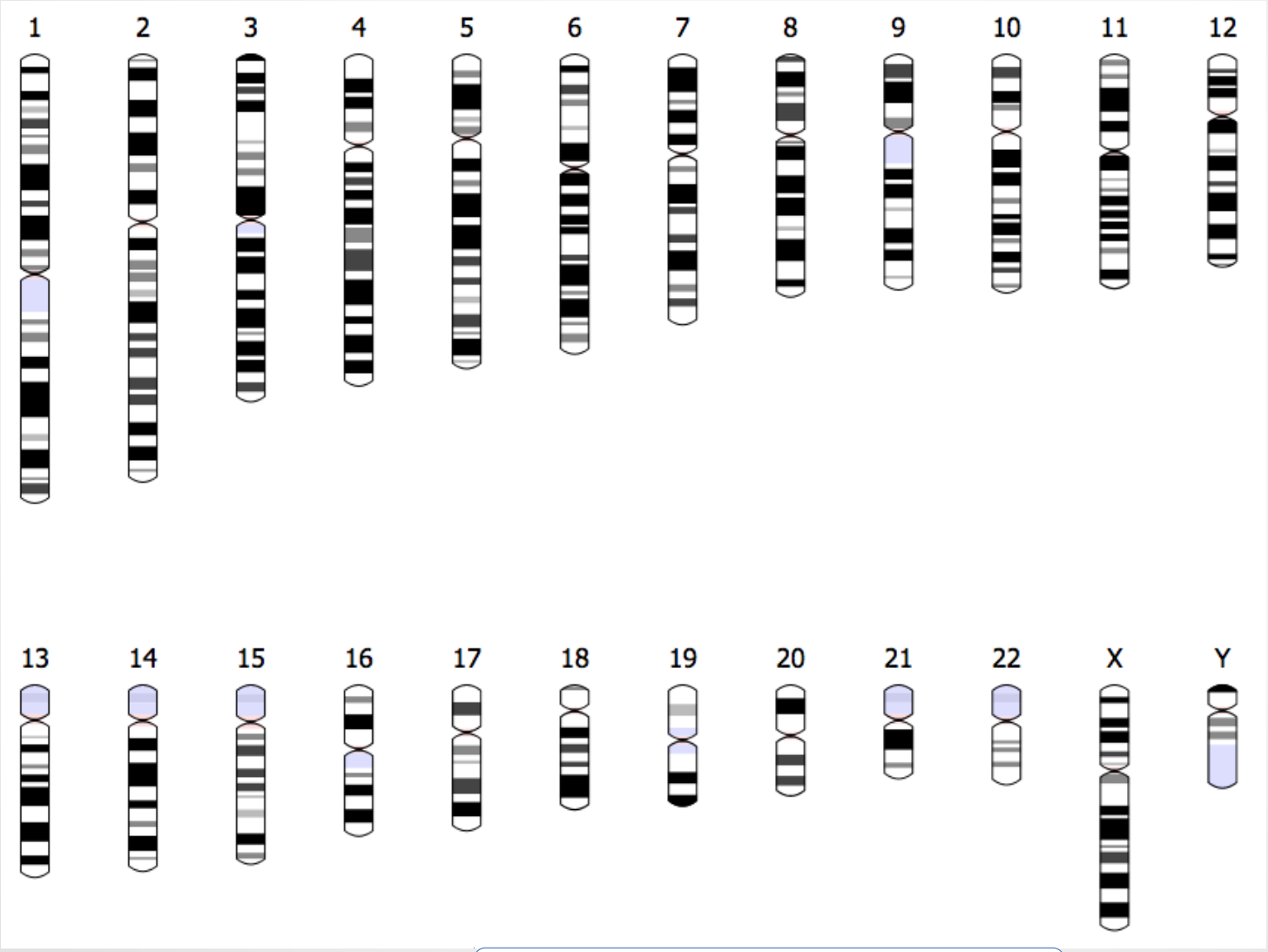
Rules for orientation and ordering of chromosomes and C-scaffolds:

- Autosomes and C-scaffolds should be ordered by length, from longest to shortest
- Each chromosome and C-scaffold should start from p-arm (short arm)

Issues:

- Length of C-scaffold might significantly differ from length of corresponding chromosome
- Chromosomes might have a very similar length
- For proper orientation of C-scaffolds we need to know the location of centromere in each C-scaffold to understand where is p-arm and q-arm.
- for nearly ideal metacentric (p- nd q- arm are nearly equal in length) chromosomes we might not be able to distinguish p- and q-arm
- Old nomenclature, contradicting to rules, might exist

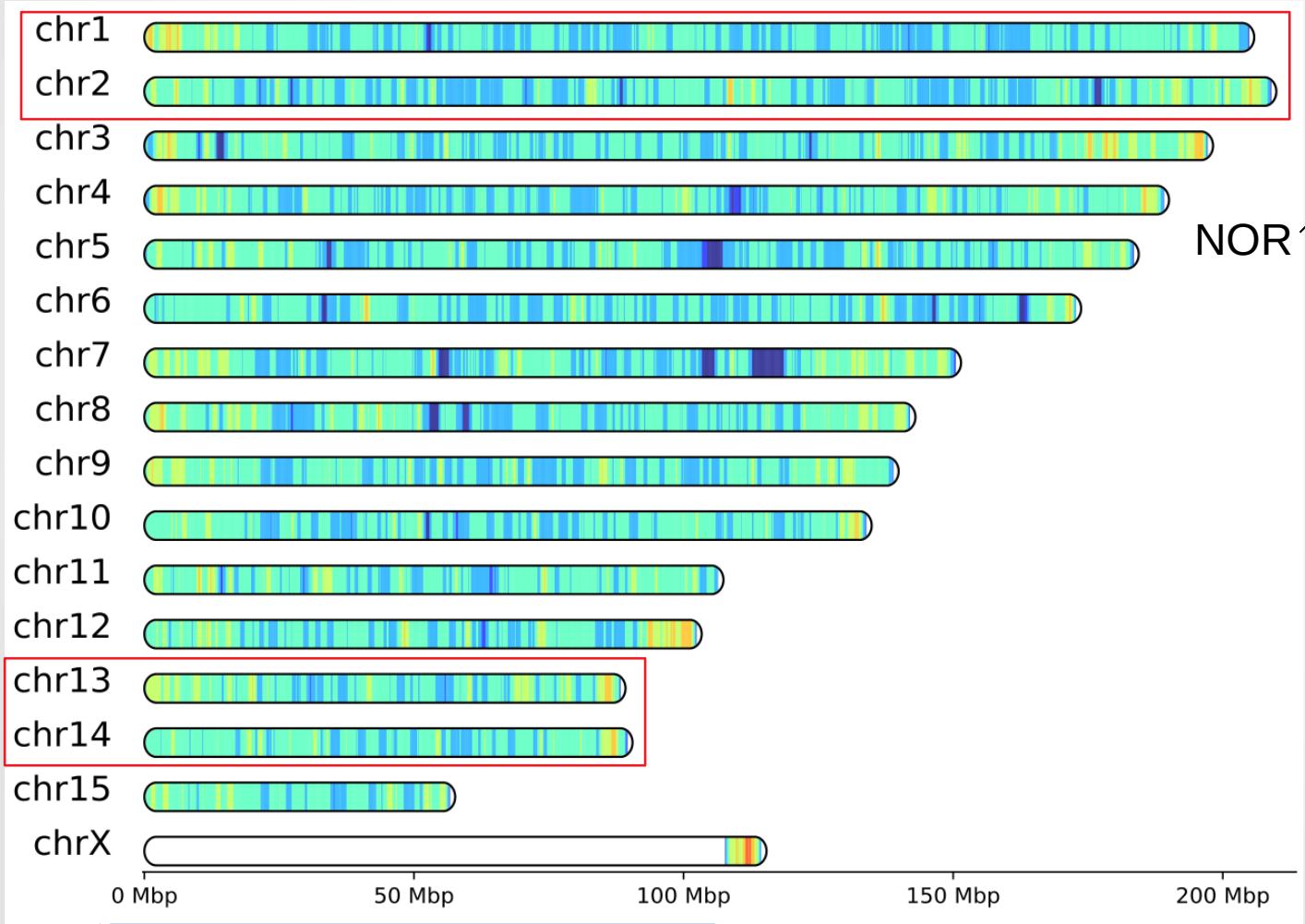
Orientation of human chromosomes



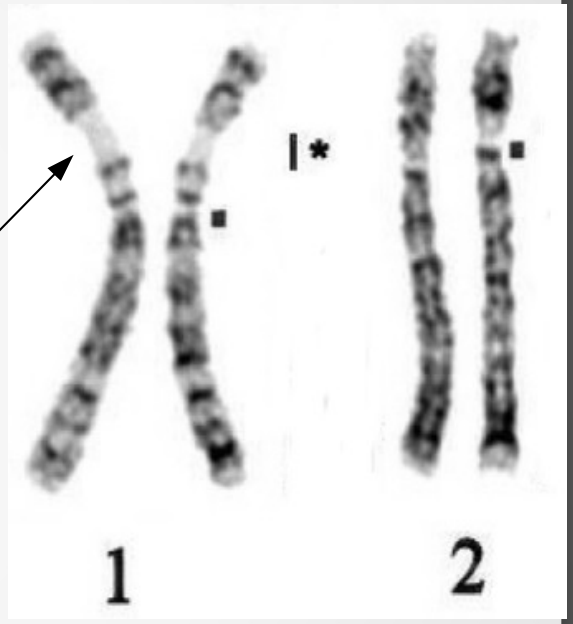
https://eweitz.github.io/ideogram/

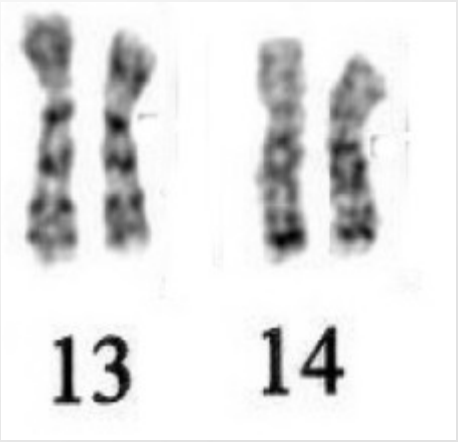
Difference between length of C-scaffolds and chromosomes: case of baikal seal

Assembly



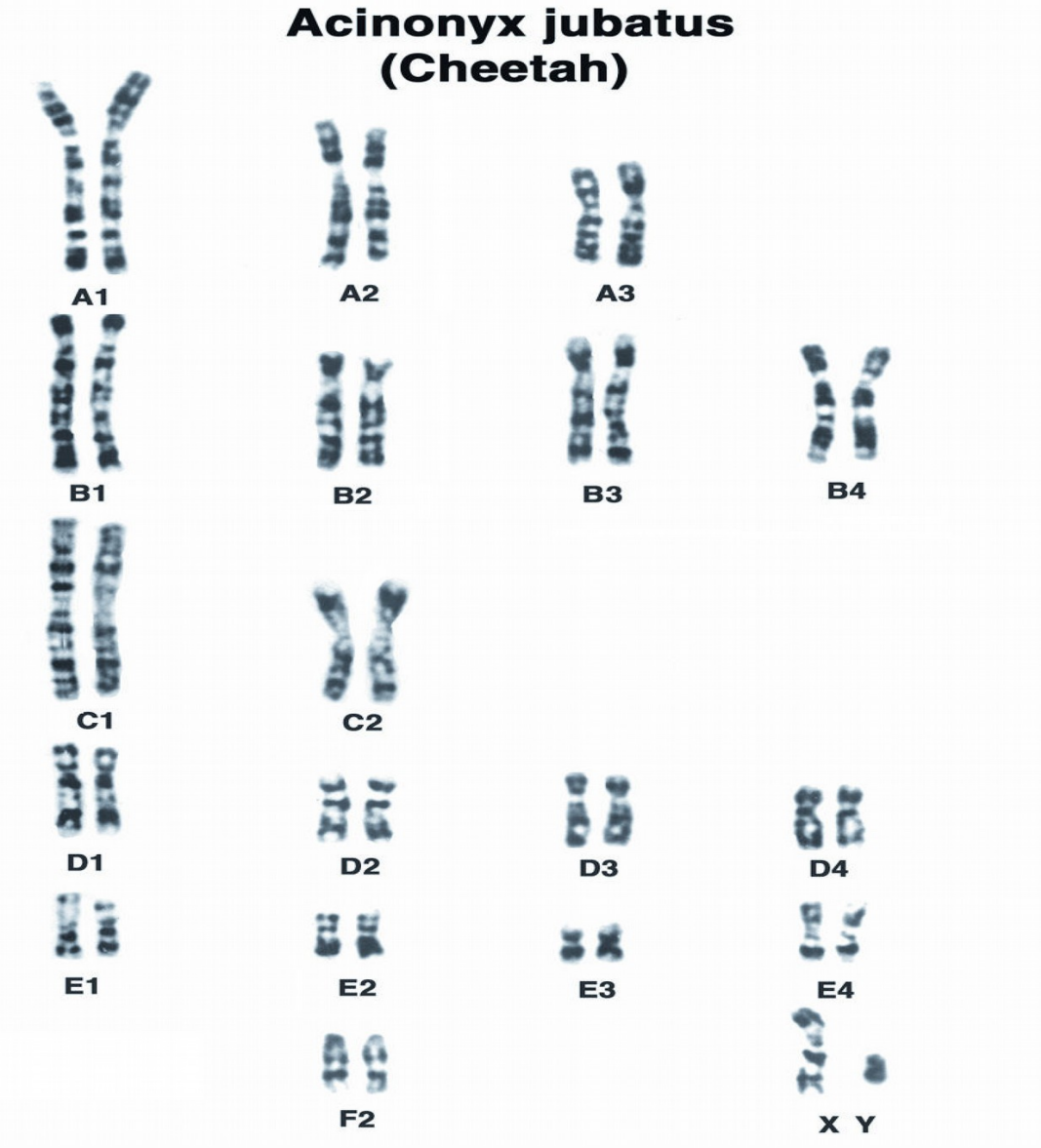
Karyotype





Yakupova et al, 2022

Old nomenclature contradicting rules: cat case

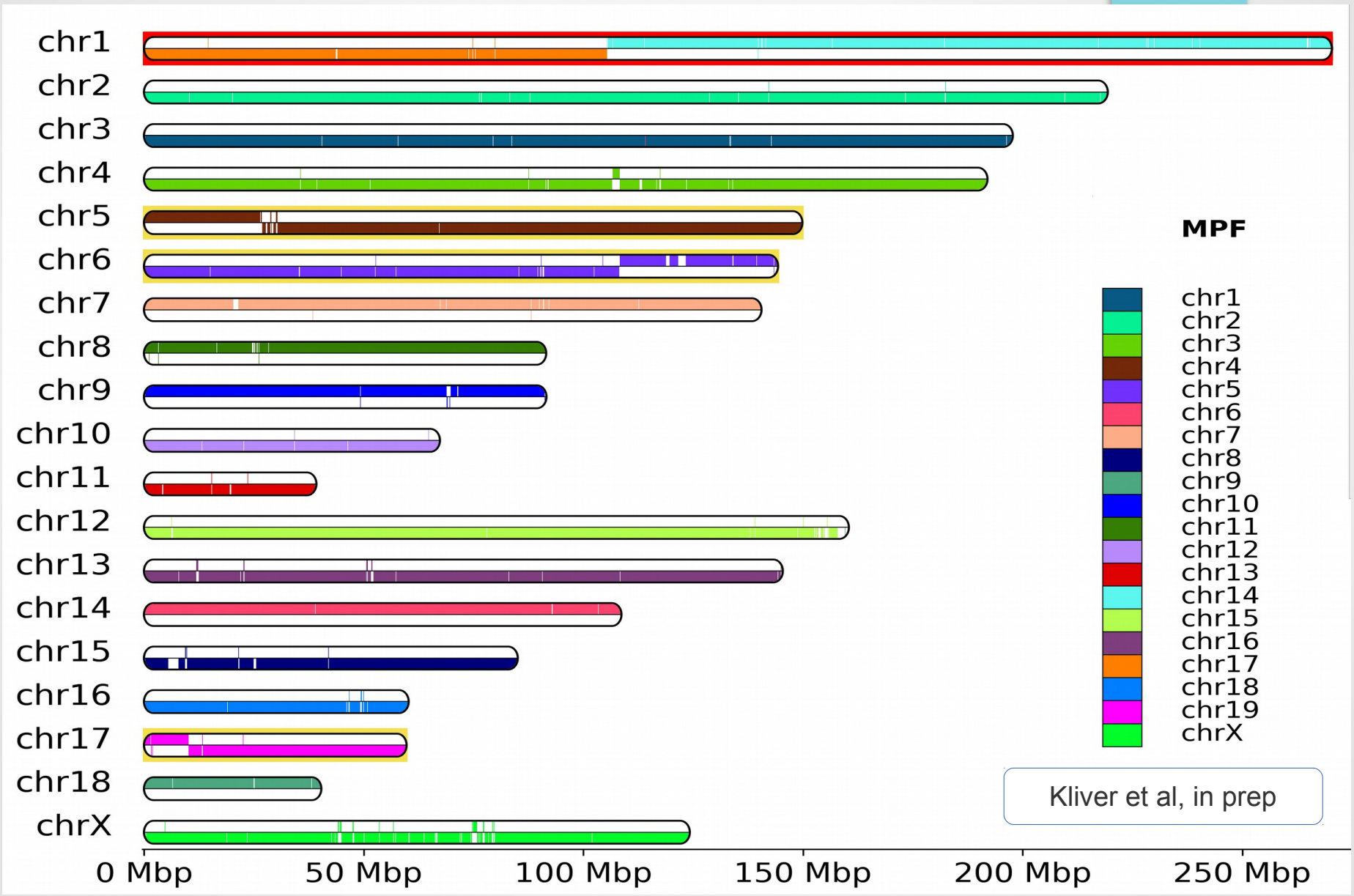


A, B, C, D, E, F groups are used for genomes of all cats.

Inherited from … 1965

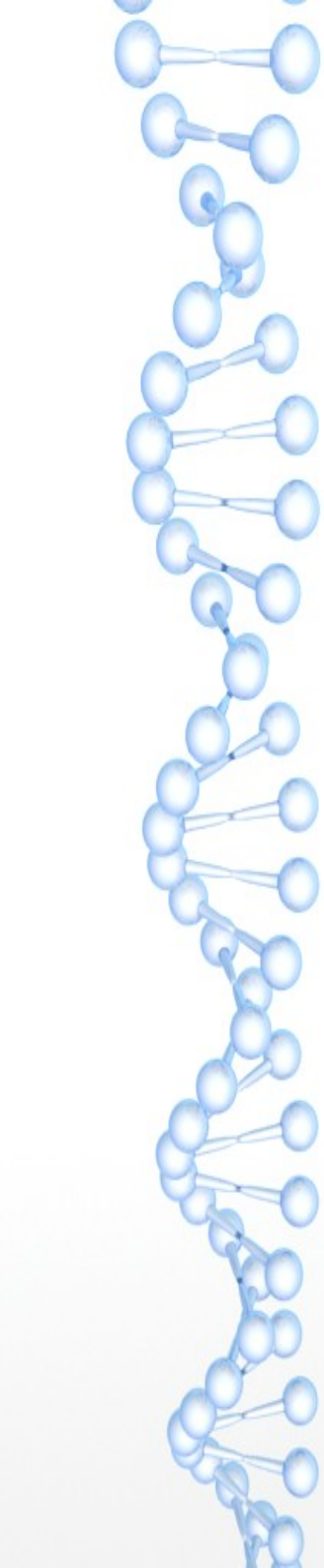
Graphodatsky et al, 2020

Example of different C-scaffold orientation in related species: black-footed ferret and domestic ferret



Summary

- Rules of chromosome nomenclature and ordering are very often impossible to follow. You should consider them as recommendation
- Always check if elder nomenclature exists for your species. Follow it if it exists.
- Do not forget that even closely related species with similar chromosome numbers and karyotypes could have different orientation of C-scaffolds(chromosomes) in the assembly. Assembler/scaffolder choose orientation of sequences randomly
- To assign chromosome names to C-scaffolds of your new chromosome-level genome assembly you need to find a reference species for which it was already done and for chromosomes of which you already have FISH map. Then just compare FISH map and WGA between your ad reference species. Sometimes you to do it multiple times if required FISH map is unavailable.
- If no cytogenetic data is available directly of indirectly, for mammals you could identify X-chromosome. And order autosomes by length.



I. Structure and diversity of the genomes

Mutations

Classification of mutations

- genomic mutations - change of chromosome number
  - whole genome duplication
  - aneuploidy (loss or gain individual chromosomes)
- chromosomal mutations
  - fusions and fisions (Robertson translocations)
  - big insertions, inversions, deletions, translocations
  - small translocations
- gene mutations
  - small insertions, inversions and deletions
  - single nucleotide substitutions

Large-scale mutations

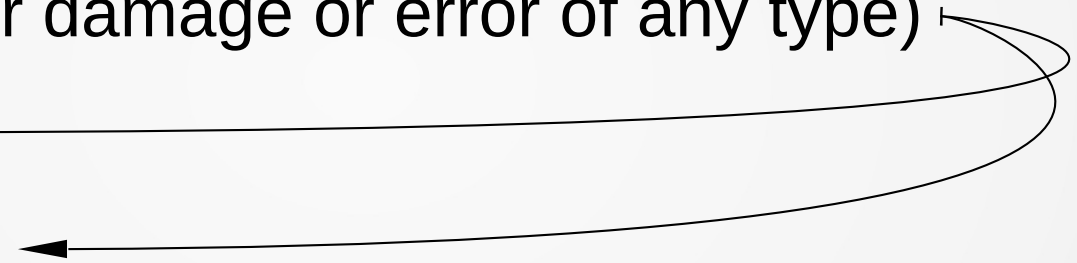
Small-scale mutations

This commonly used classification is ambiguous !

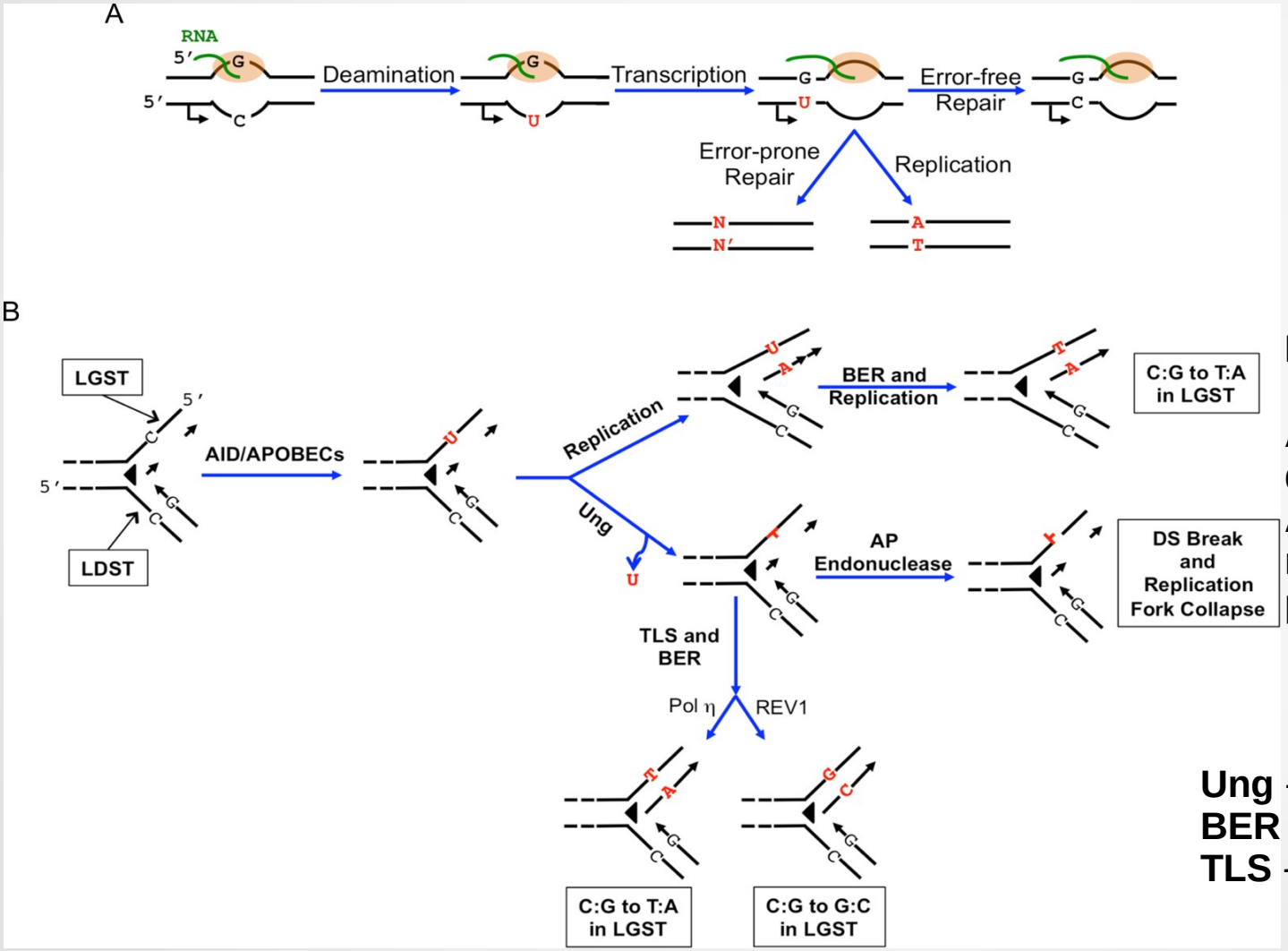
Sources of mutations

Mutations are **results** of *errors* of:

- Reparation (after damage or error of any type)
- Replication
- Recombination
- Separation of chromosomes between daughter cells



Difference between DNA damage and mutation: deaminase case



Siriwardena et al, 2016

deaminases convert cytosine to uracile in ssDNA

Examples of deaminases:

AID - Activation-induced deaminase
APOBEC - apolipoprotein B mRNA-editing catalytic polypeptide-like

Ung - uracil−DNA glycosylase
BER - base-excision repair
TLS - translesion synthesi

LDST - leading strand template
LGST - lagging strand template

DNA polymerases δ and ε

- eukaryotic polymerases
- multi subunit protein complex
- participate in genome replication
- three types of activities:
  - 5’ → 3’ polymerase - strand elongation
  - 5’ → 3’ exonuclease - excision of obstacles
  - 3’ → 5’ exonuclease - self-proofreading

Difference between error and mutation. Precision of replication

Error rate of polymerase reaction

$\sim 10^{-4} - 10^{-5}$

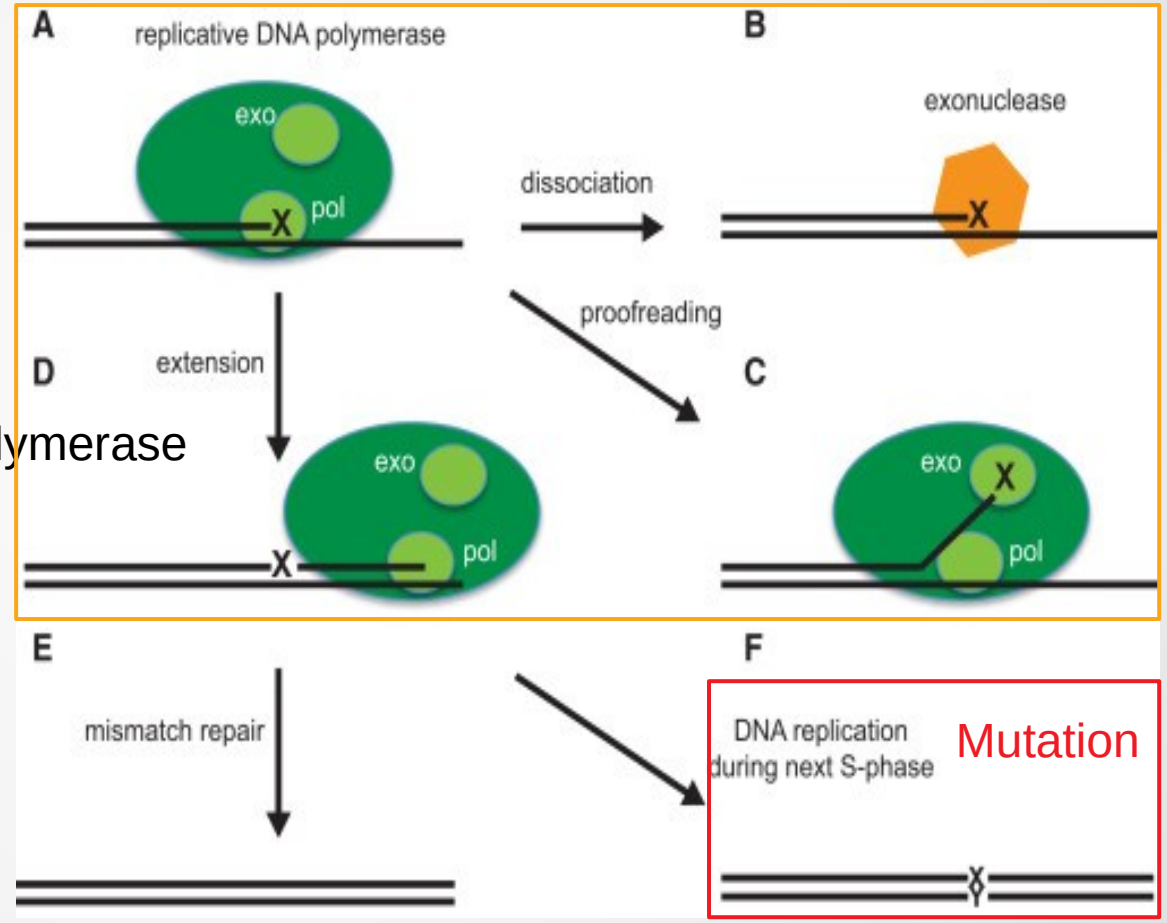
Error rate after proofreading by polymerase

$\sim 10^{-7} - 10^{-8}$

Ganai and Johansson et al, 2016

X - mismatching nucleotide

Error



Error rate after final proofreading (metilation-dependent)

$\sim 10^{-10}$

Mutation problem in big multicellular organisms

Replication error rate: 10^{-10} per base per division

Human genome size: $3.3 * 10^{9}$ bp

Cells in human body:

- human:
  - all: $3.0 \cdot 10^{13}$ cells
  - without unnuclear cells: $0.3 \cdot 10^{13}$ cells
- bacterial: $3.8 \cdot 10^{13}$ cells

Sender et al, 2016

How to reduce a number of inherited mutations?

Mutation problem in big multicellular organisms

Replication error rate: 10^{-10} per base per division

Human genome size: $3.3 * 10^{9}$ bp

Cells in human body:

- human:
  - all: $3.0 \cdot 10^{13}$ cells
  - without unnuclear cells: $0.3 \cdot 10^{13}$ cells
- bacterial: $3.8 \cdot 10^{13}$ cells

Sender et al, 2016

How to reduce a number of inherited mutations?

Reduce number of divisions.

How? Less divisions - less cells.

Mutation problem in big multicellular organisms

Replication error rate: 10^{-10} per base per division

Human genome size: $3.3 * 10^{9}$ bp

Cells in human body:

- human:
  - all: $3.0 \cdot 10^{13}$ cells
  - without unnuclear cells: $0.3 \cdot 10^{13}$ cells
- bacterial: $3.8 \cdot 10^{13}$ cells

Sender et al, 2016

How to reduce a number of inherited mutations?

Reduce number of divisions

How? Less divisions - less cells

Separate germline and somatic cells as earlier as possible

Somatic and germline cells

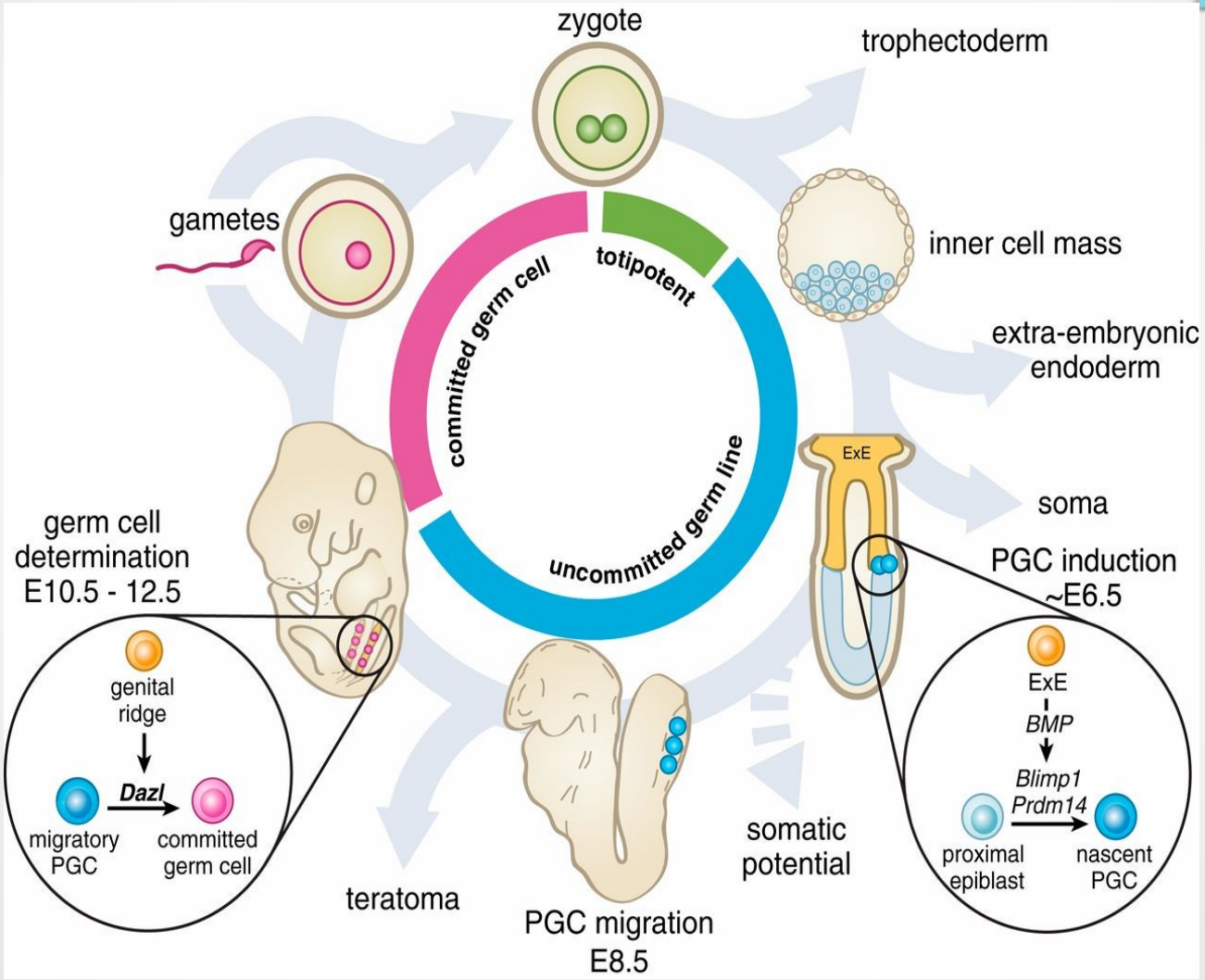
Germline (germ) cells
gametes (eggcells, spermatozoa, etc) and their ancestor cells

Somatic cells
all other cells

Primordial germ cells (PGCs)
embryonic precursors of germline cells

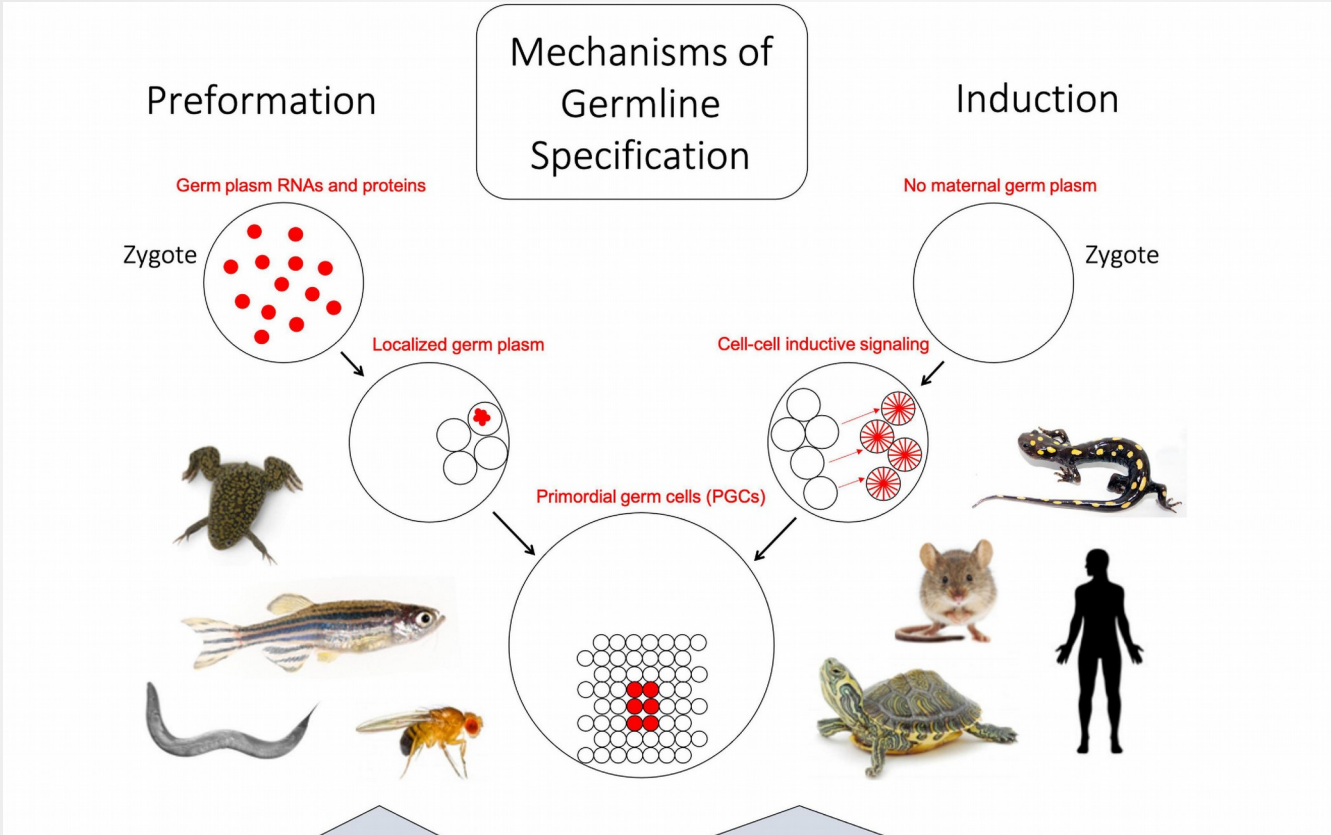
Divergence of somatic and germline cells happens very early in embryogenesis !

Germline cells in development of mammals



Nicholls et al, 2019

Germline cells in development of animals



"that is to say, the fate of the cells is determined by forces situated within them, and not by external influences"
– Weismann 1892

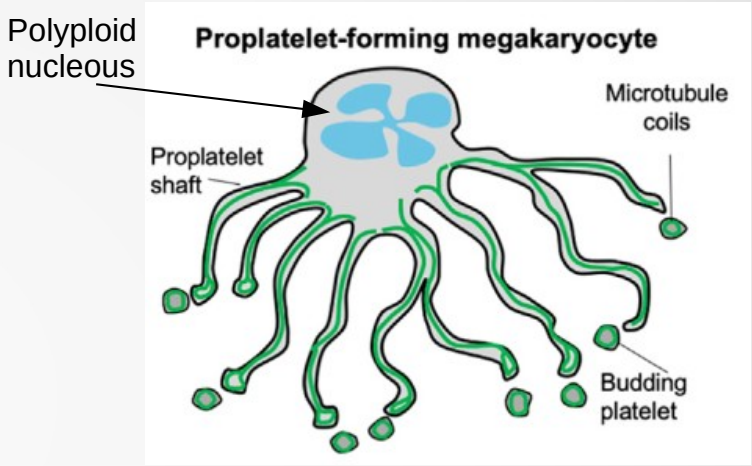
"there are forces external to the cells that bid them assume the individual characters appropriate to their individual relations to the whole; the determining forces are not within the cells, as the doctrine of determinants supposed" – Hertwig 1900

Hansen and Pelegri, 2021

Variation of genome in somatic cells

Polyploid cells in human:

- trophoblast giant cells, critical for pregnancy.
- megakaryocytes, produce platelets (thrombocytes)
- hepatocytes, liver cells
- etc

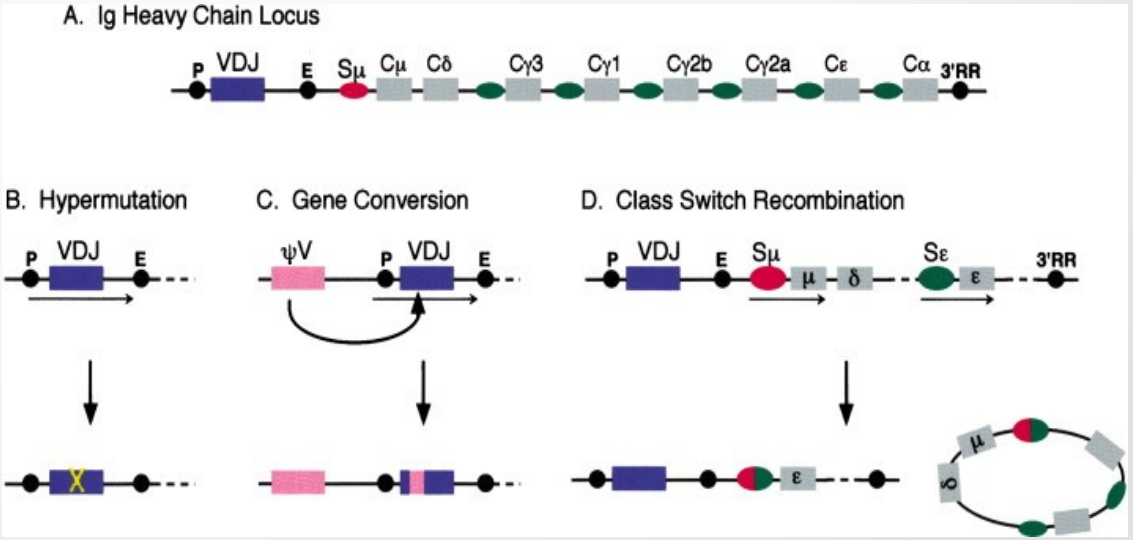


Somatic hypermutation

loci of immunoglobulin genes

in B cells

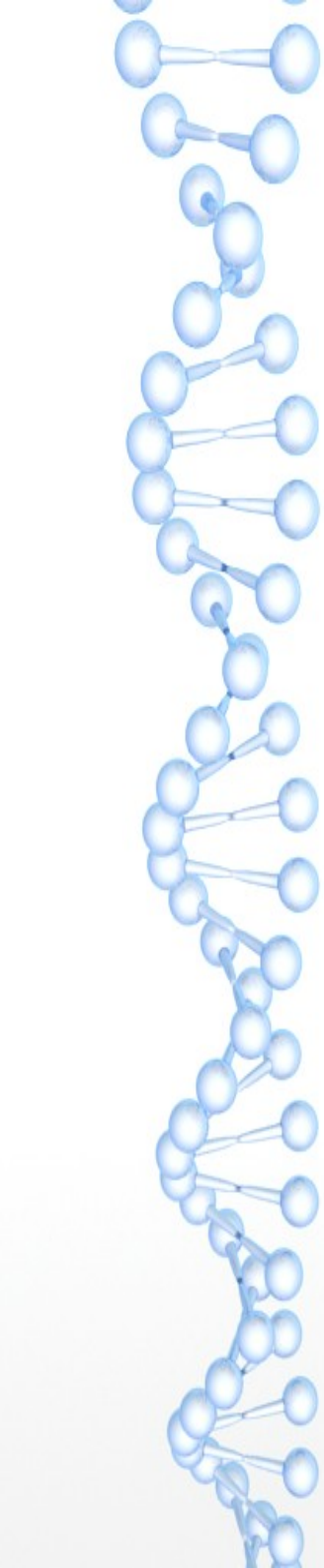
- point hypermutagenesis
- gene conversion
- deletions



Heib et al, 2021; Papavasiliou and Schatz, 2002

Summary

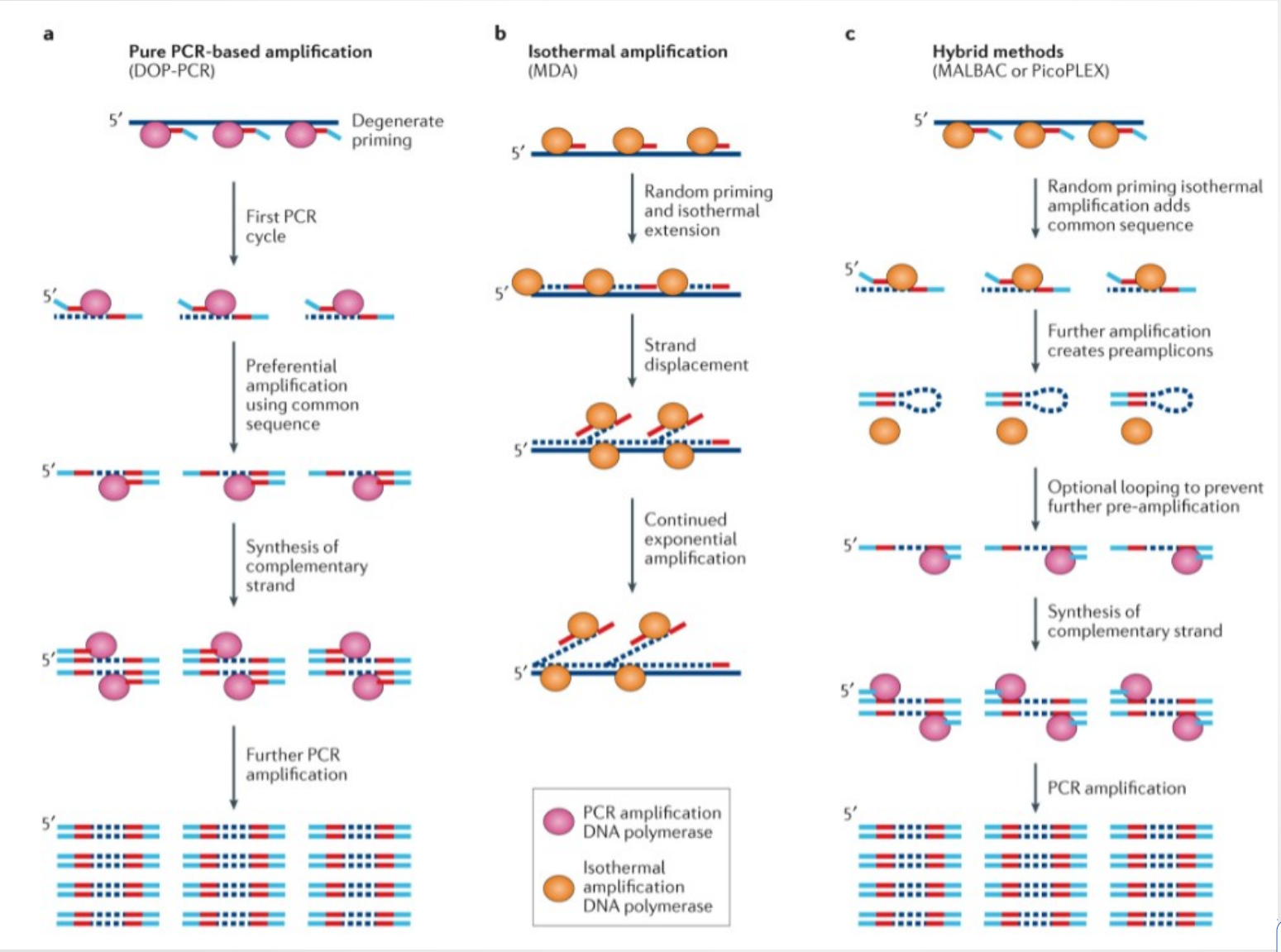
- Mutation classification is ambiguous
- There is a difference between damage or error of polymerase and mutation. Mutation is a DNA change that was fixated in next replication.
- There several proofreading systems that reduce mutation rate during replication to 10^{-10} per base per division
- methylation, and therefore epigenome, is important for error correction
- separation of somatic and germline cells is one more mechanism reducing mutation rate by lowering the number of replication cycles. It happens very earlier in embryogenesis.
- isolation of somatic cells from inheritance reduces requirements for genome stability and integrity
- Mutagenesis (even with prefix super-) might be a part of normal functioning of organism



I. Structure and diversity of the genomes

Genome “levels”

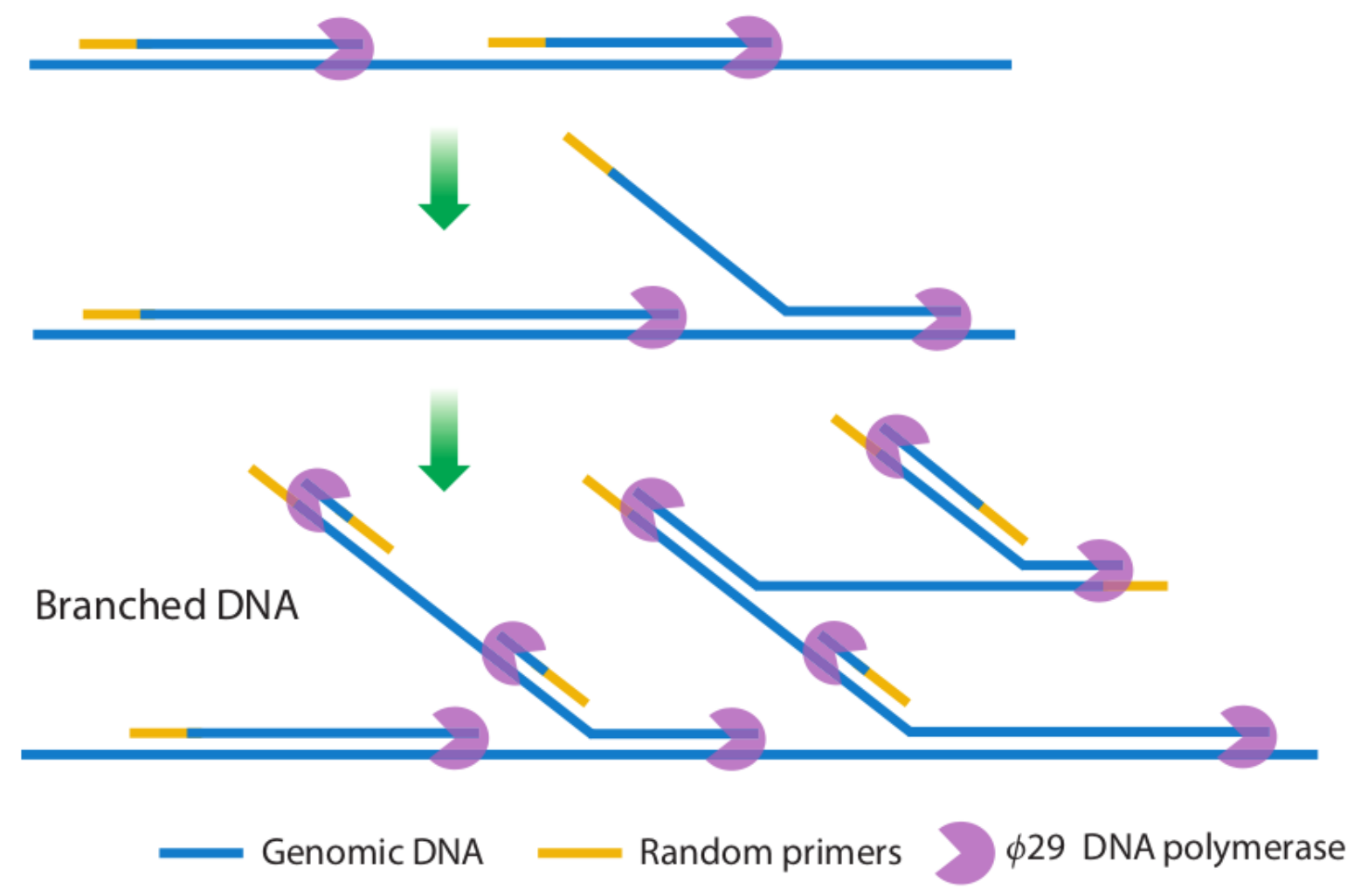
Could we assemble a **genome** from a single cell?



In general yes, but we will need to do a **whole genome amplification (WGA)** before sequencing.

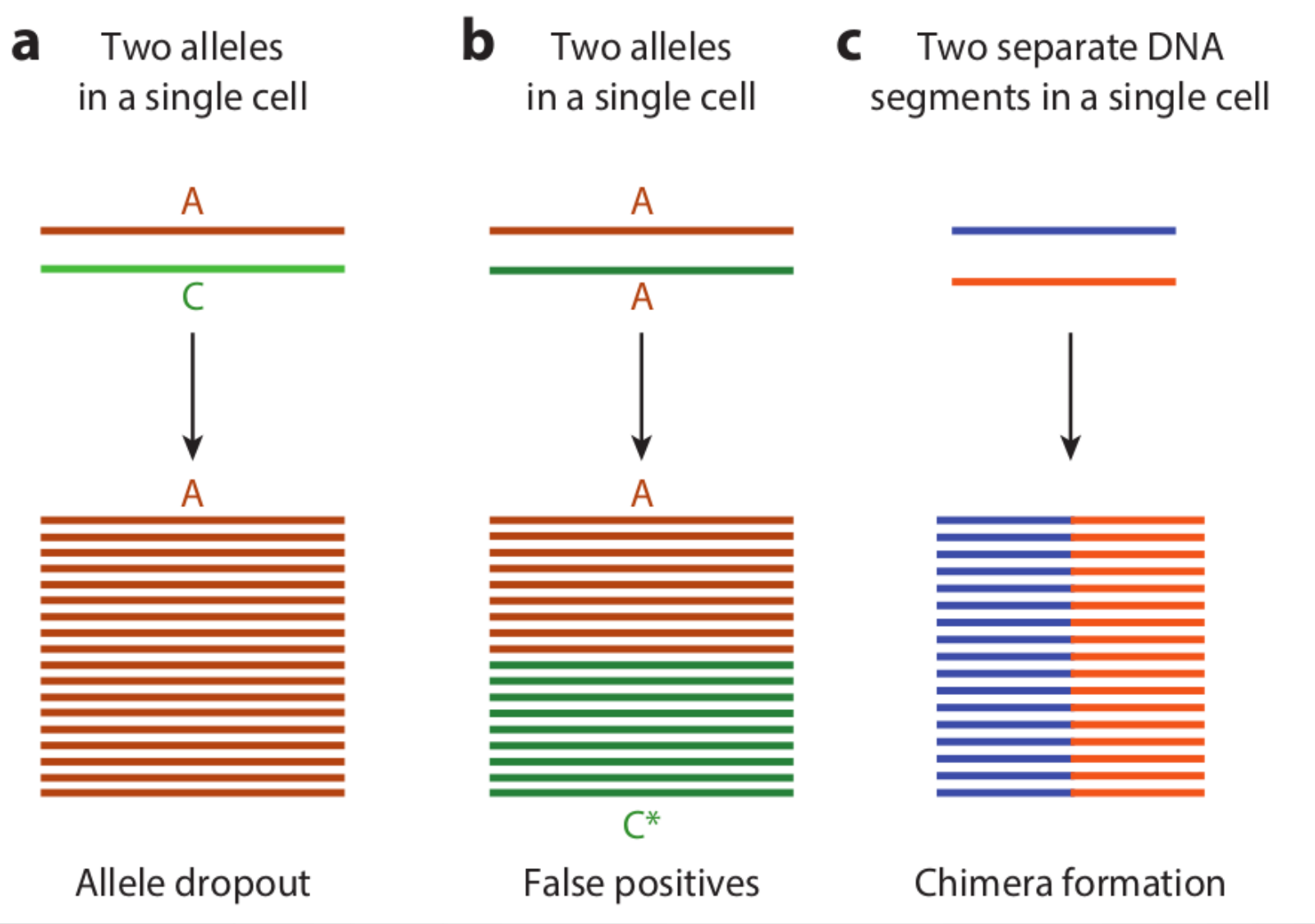
Gawad et al, 2016

Multiple displacement amplification (MDA)



Huang et al, 2016

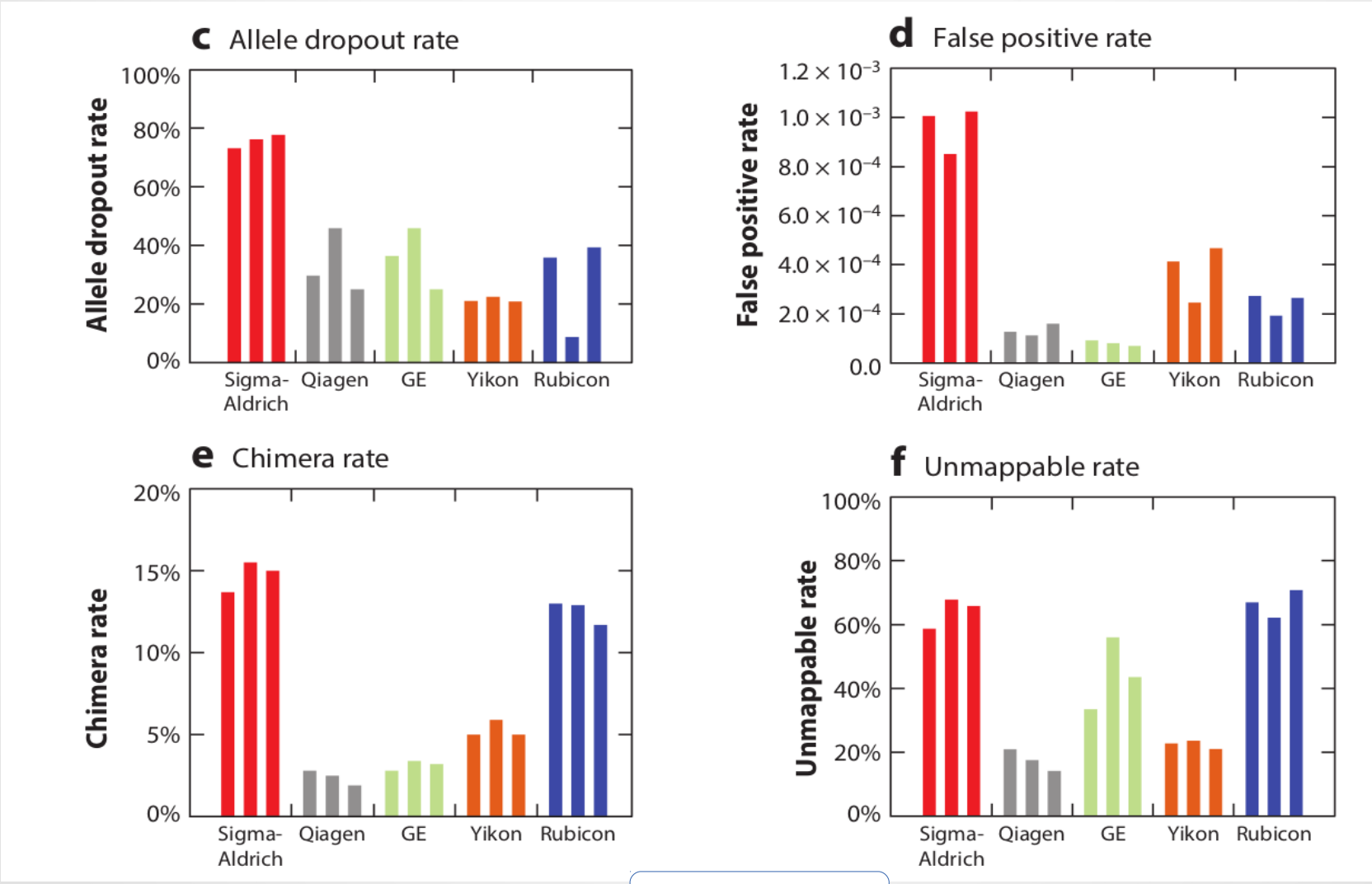
Typical errors of whole genome amplification



Some regions could just … **not amplify**

Huang et al, 2016

Efficiency of five commercial kits for WGA

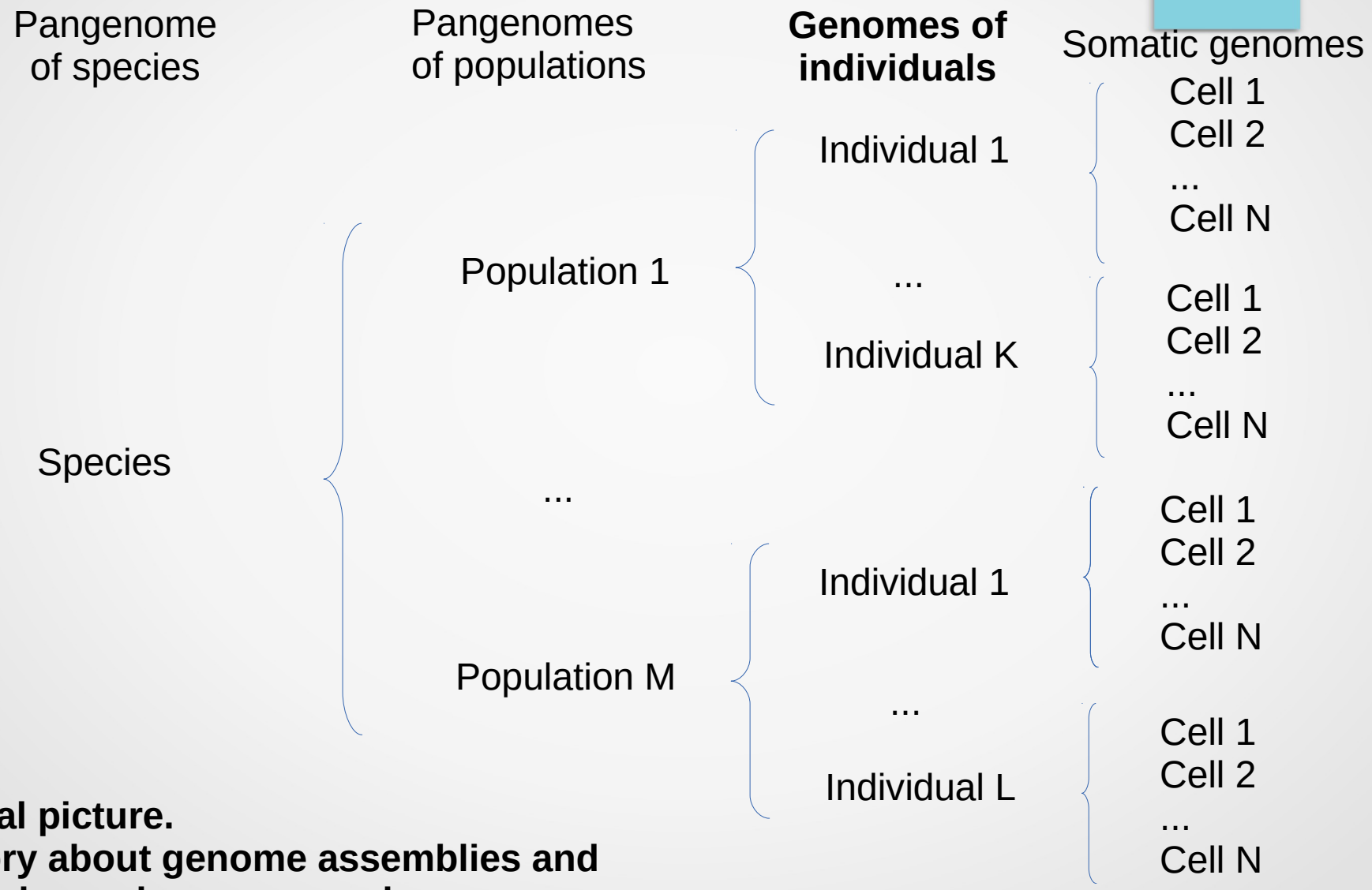


Huang et al, 2016

Could we assemble a **reference genome** from a single cell?

No.

Genome and pangenome



Ideal picture.
Story about genome assemblies and databases is more complex.

Important features of genome assemblies

- source for DNA extraction - somatic cells:
  - blood or muscles
  - primary cell lines (usually fibroblasts)
- we try to approach original genome of zygote by simultaneous sequencing genomes of multiple somatic cells and assembling the consensus.
- in ideal case all sequencing should be done from the same individual by often it is impossible. So often, very often, genome assembly is chimera created from multiple individuals.
- Most of genome assemblies are haploid, i.e. their sequences are haploid consensus between maternal and paternal haplomes

Human reference genome GRCh38

GRC38 (latest release GRCh38.p14) includes:

- C**-scaffolds(chromosomes) and mtDNA**
- **unlocalized scaffolds (unknown exact position, but knowm chromosome)**
- **unplaced scaffolds (completely unknown position)**
- alternative scaffolds
- Epstein-Barr virus sequence and additional decoy sequences

GRC38 is something intermediate between haploid assembly and graph assembly

Pangenome projects

Human pangenome

https://humanpangenome.org

Maize pangenome

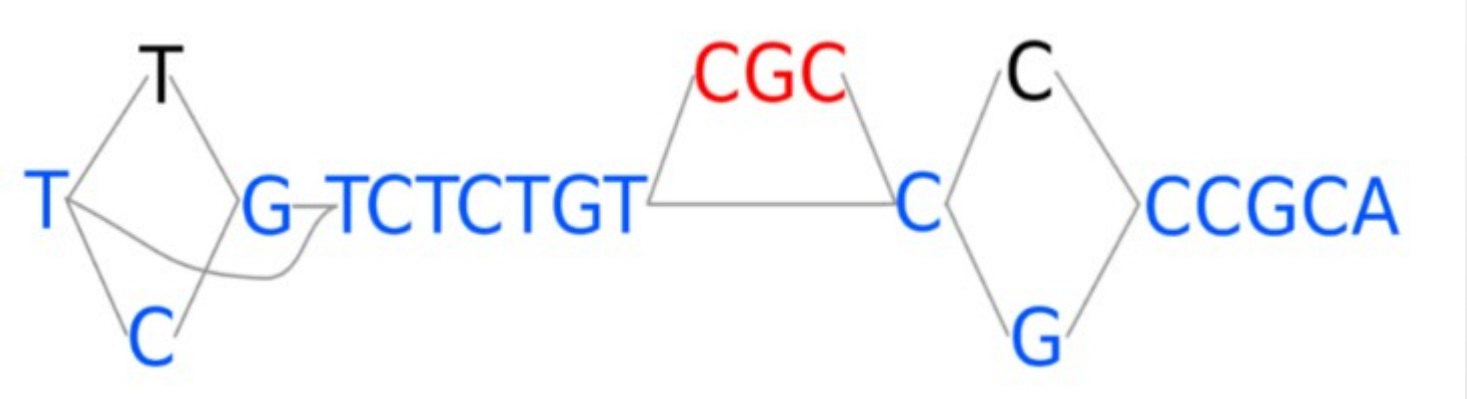
https://maize-pangenome.gramene.org

Rice pangenome

https://cgm.sjtu.edu.cn/3kricedb/

Tomato pangenome

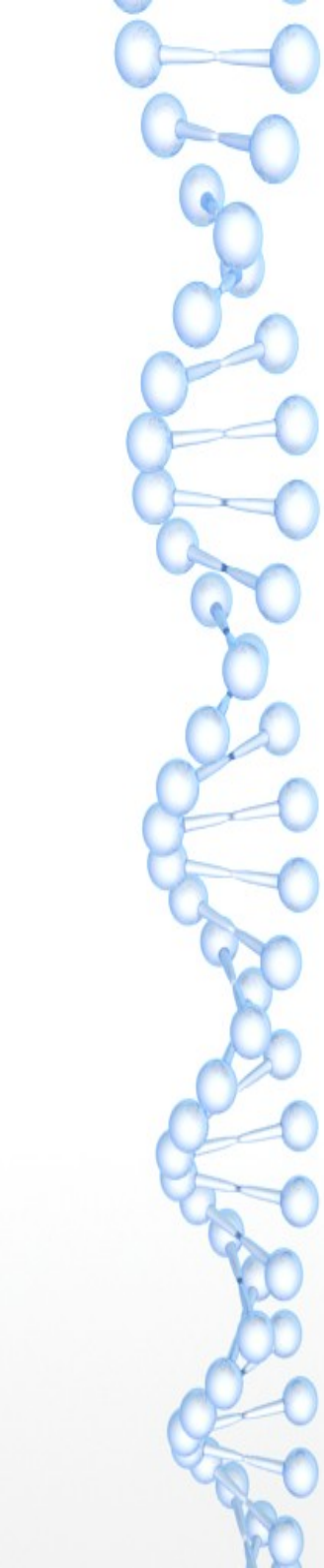
https://solgenomics.net/projects/tgg



scheme of graph-based references

Summary

- genome assembly of big multicellular organism nearly always is an approximation to the genome of the zygote (from which organism was developed) based on sequencing of multiple somatic cells and assembling the consensus. In some cases assembly might be even a consensus derived from multiple individuals.
- Under genome of species we usually mean a genome of a single reference individual. You should remember about this bias.
- Some assemblies like human GRCh38 could include alternative contigs to represent differences between diffirent populations. Also additional sequences (absent in genome) might be present to improve read mapping.
- There is a trend of switching from haploid flat assemblies to diploid and graph-based assemblies



I. Structure and diversity of the genomes

Definition of the term “genome”

Several commonly used definitions

Genome is

- the entire set of **DNA** instructions found in a **cell**.
- genetic material of haploid set of **chromosomes**

- **information repository** of organism
- all **hereditary material** of organism

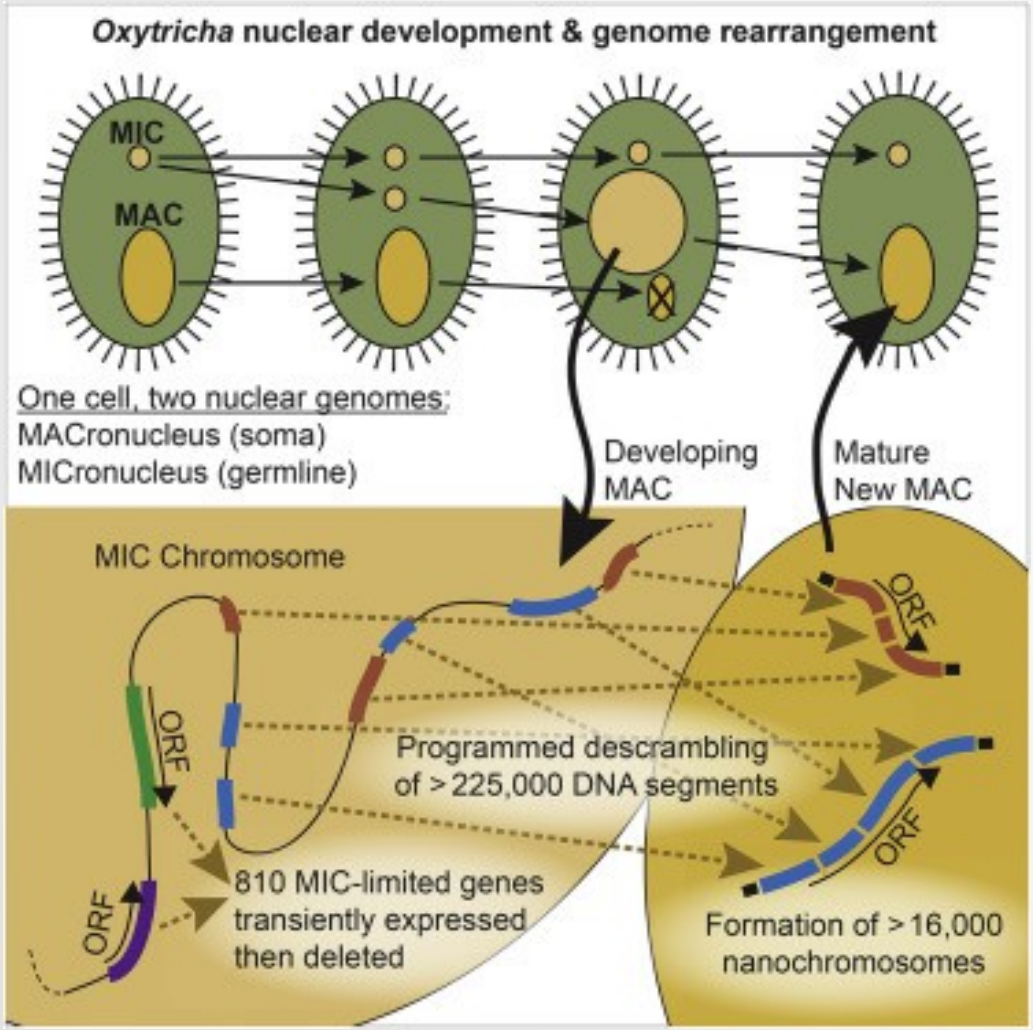
include epigenome (methylation and other modifications)

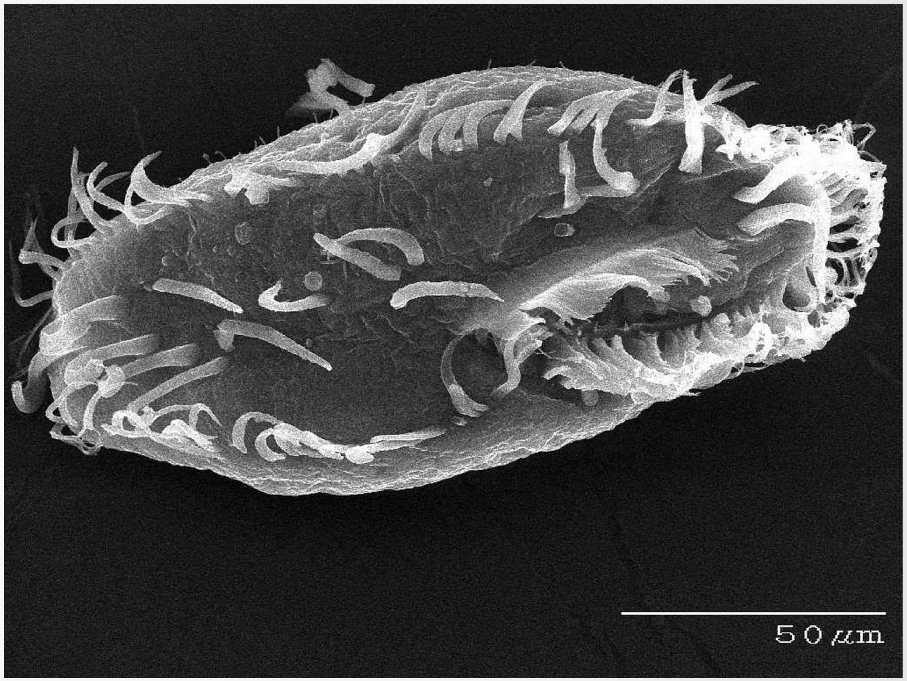
Definitions overlap significantly, but not completely!

Goldman and Landweber, 2016

RNAs and genome: case of Cilliates

Sterkiella histriomuscorum (or *Oxytricha trifallax*)





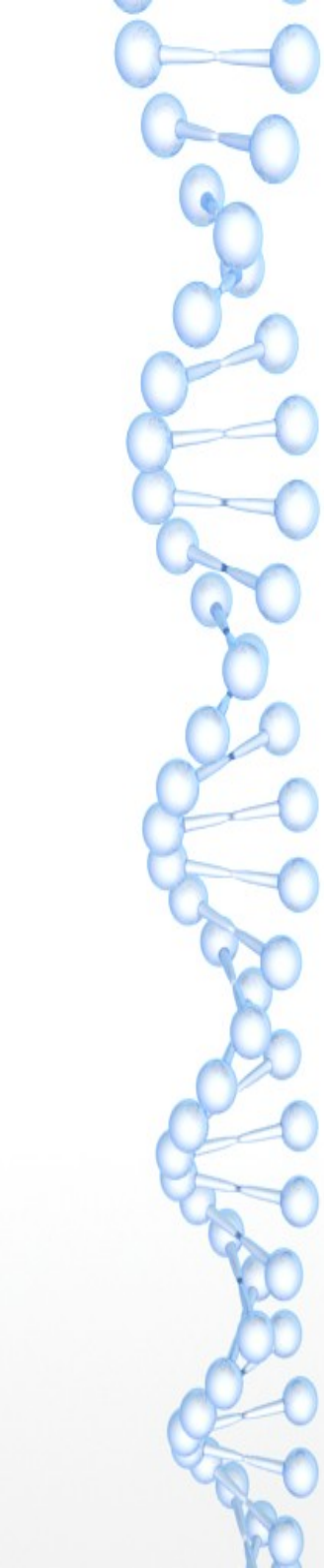
Chen et al, 2014

Information about the order of segments in scrambled genes is stored in RNA copies of nanochromosomes during at least one stage

Definitions of term genome

- **information repository** of organism
- all **hereditary material** of organism

Two definitions pretending to encompass all the variety of genomes.
But still there are issues.



End of Module I