Genome Diversity and Structure

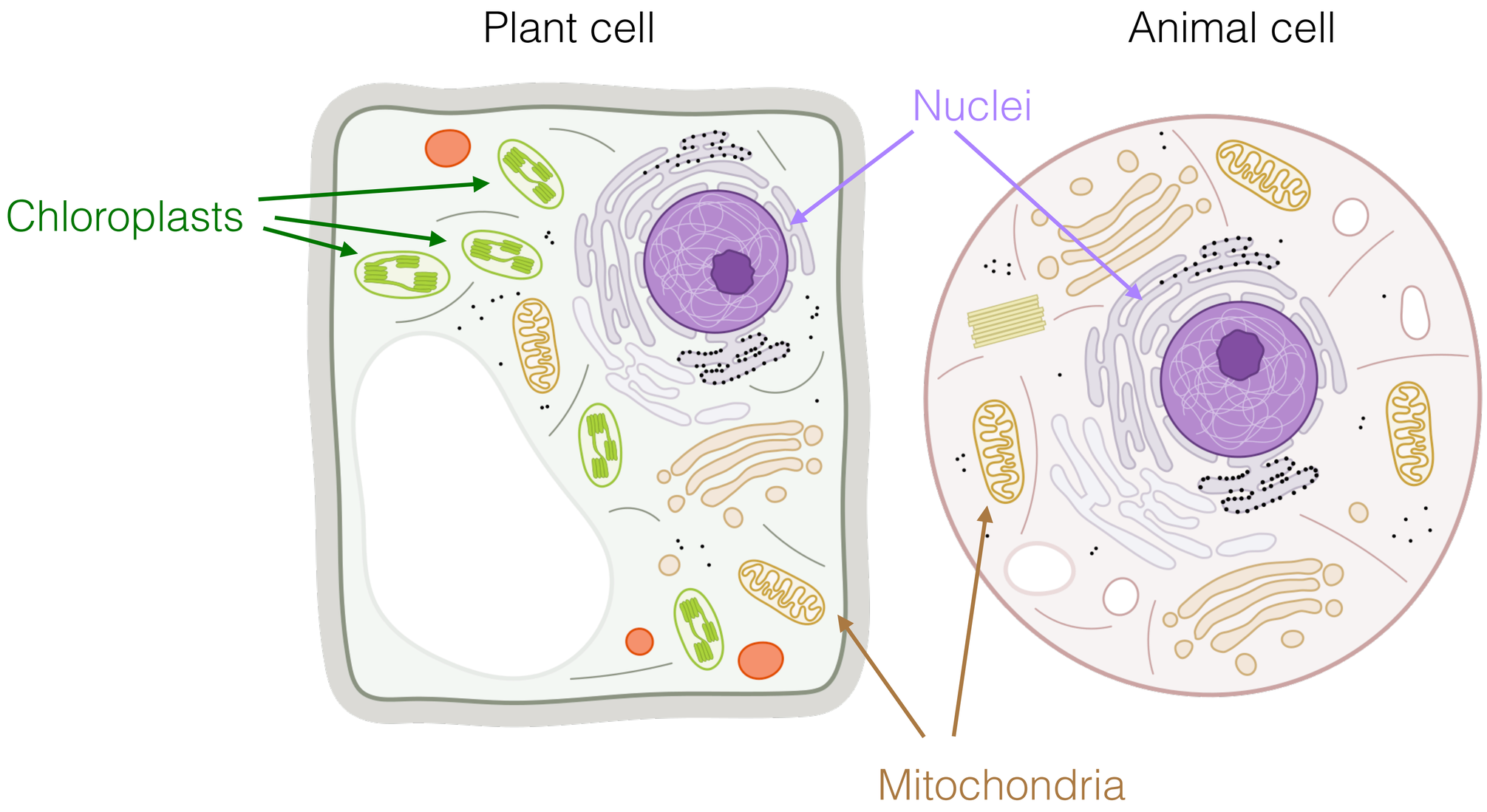
Mark Stenglein, MIP 280A4

Exercises [beginning of class]:

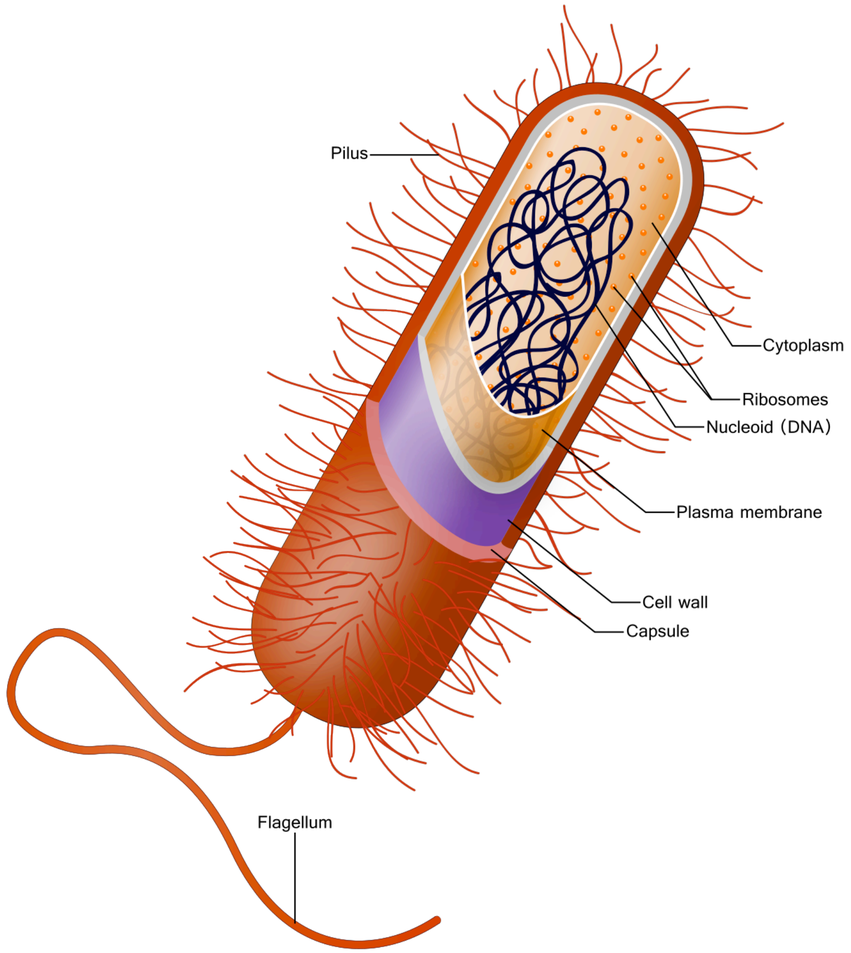
1) Briefly describe 3 ways that genomes differ from each other

2) What are 2 features shared by all genomes

Typical eukaryotic cells contain nuclear and organellar dsDNA genomes

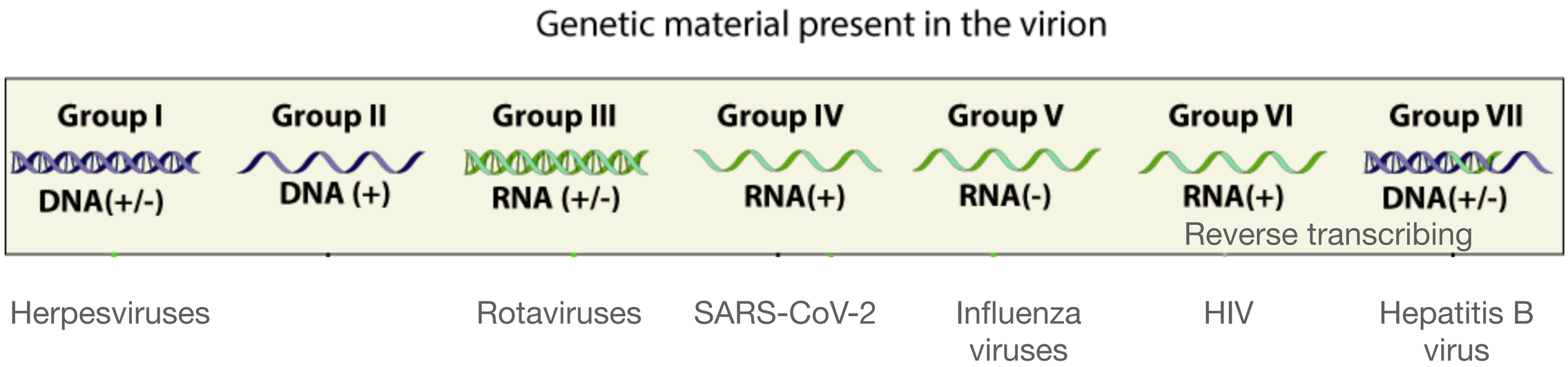


A typical prokaryotic cell contains one or more dsDNA chromosomes and often one or more plasmids.



Usually but not always circular

Different viruses have almost every imaginable type of genomic nucleic acid

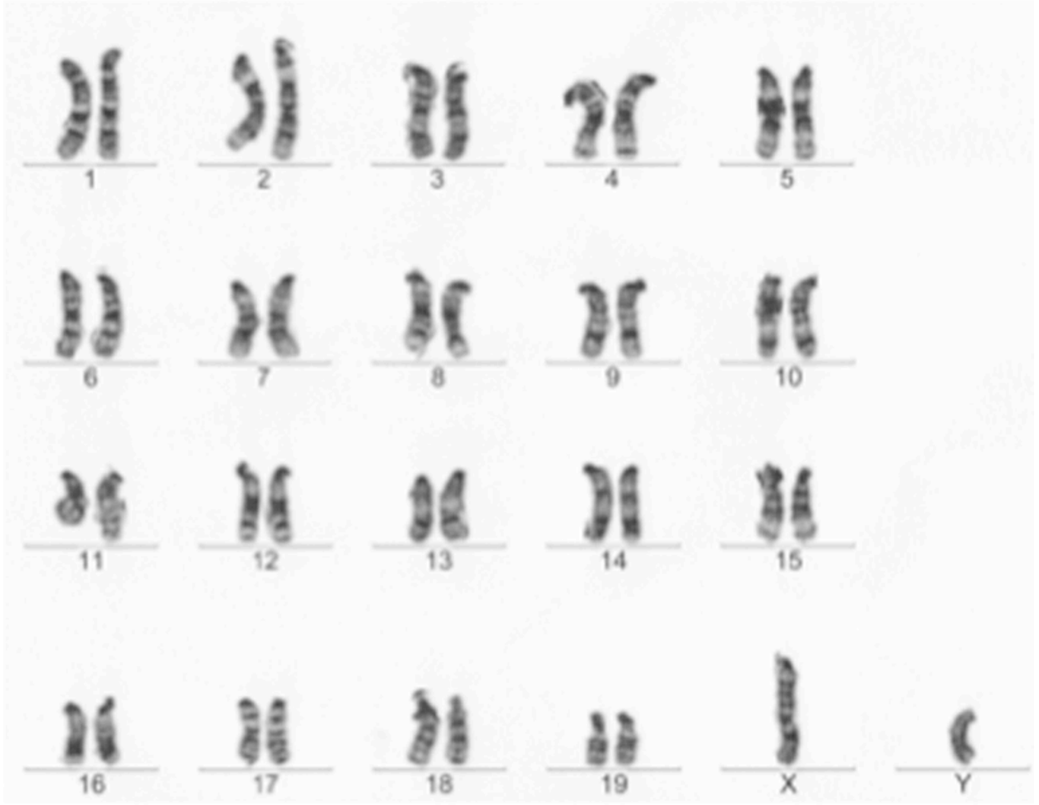


Viral genomes can be linear or circular
And composed of a single molecule or multiple molecules

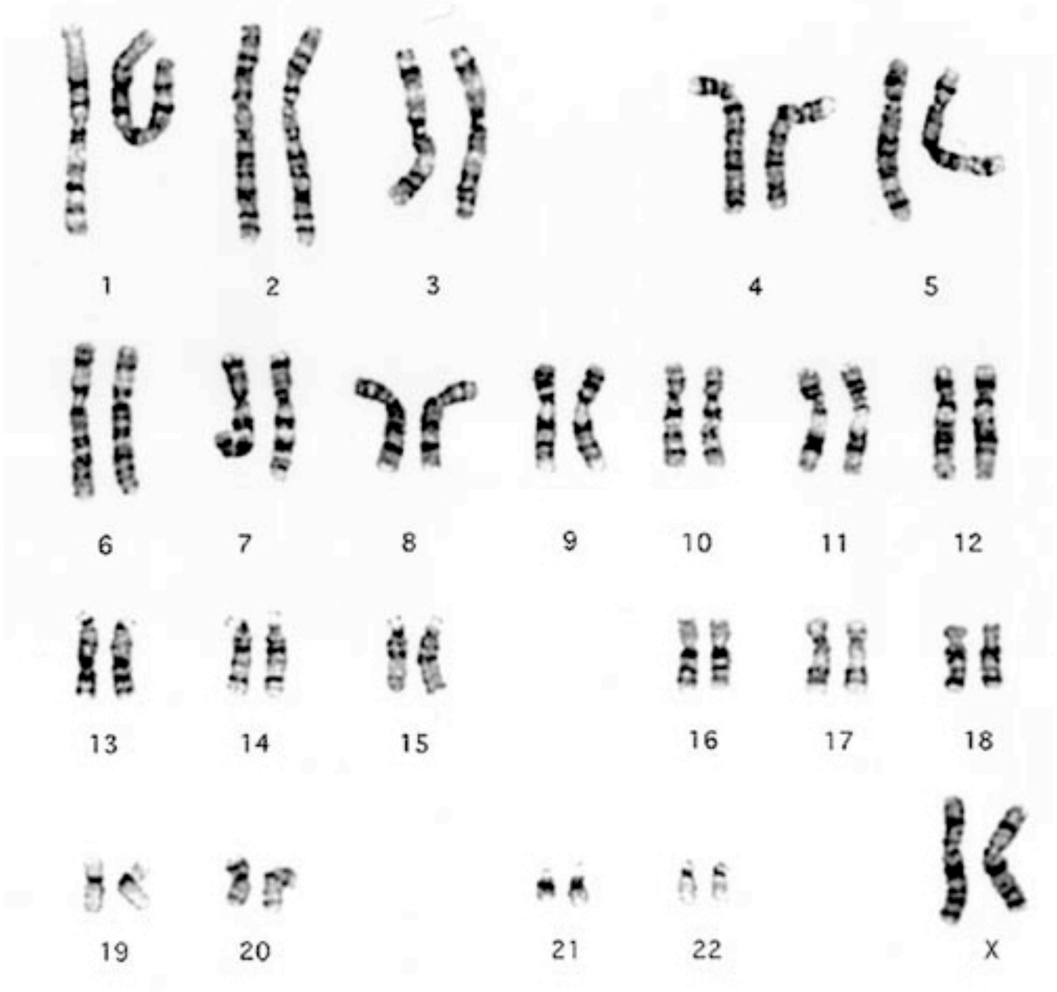
Image credit: Viral Zone Link

Different genomes are made up by different numbers of molecules

_____ mouse karyotype



_____ human karyotype

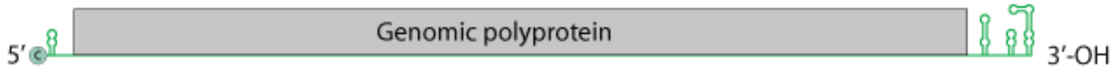


Courtesy of Dr. K. Phelan, Greenwood Genetic Center.
Noncommercial, educational use only.

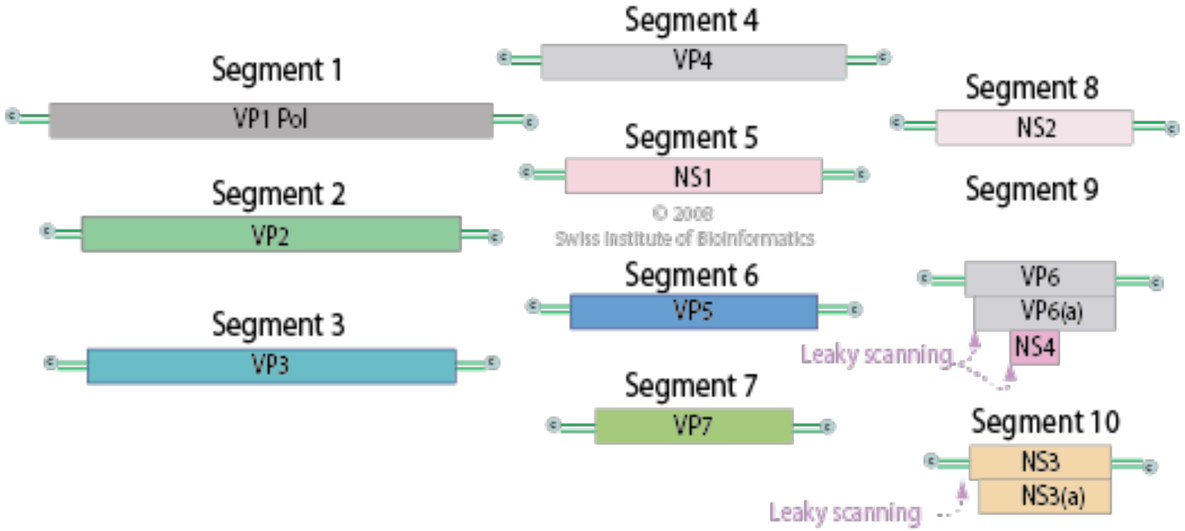
Sanchez-Partida et al (2011) PLoS One

Different viruses have different numbers of genome segments

West Nile virus
(One molecule +ssRNA)

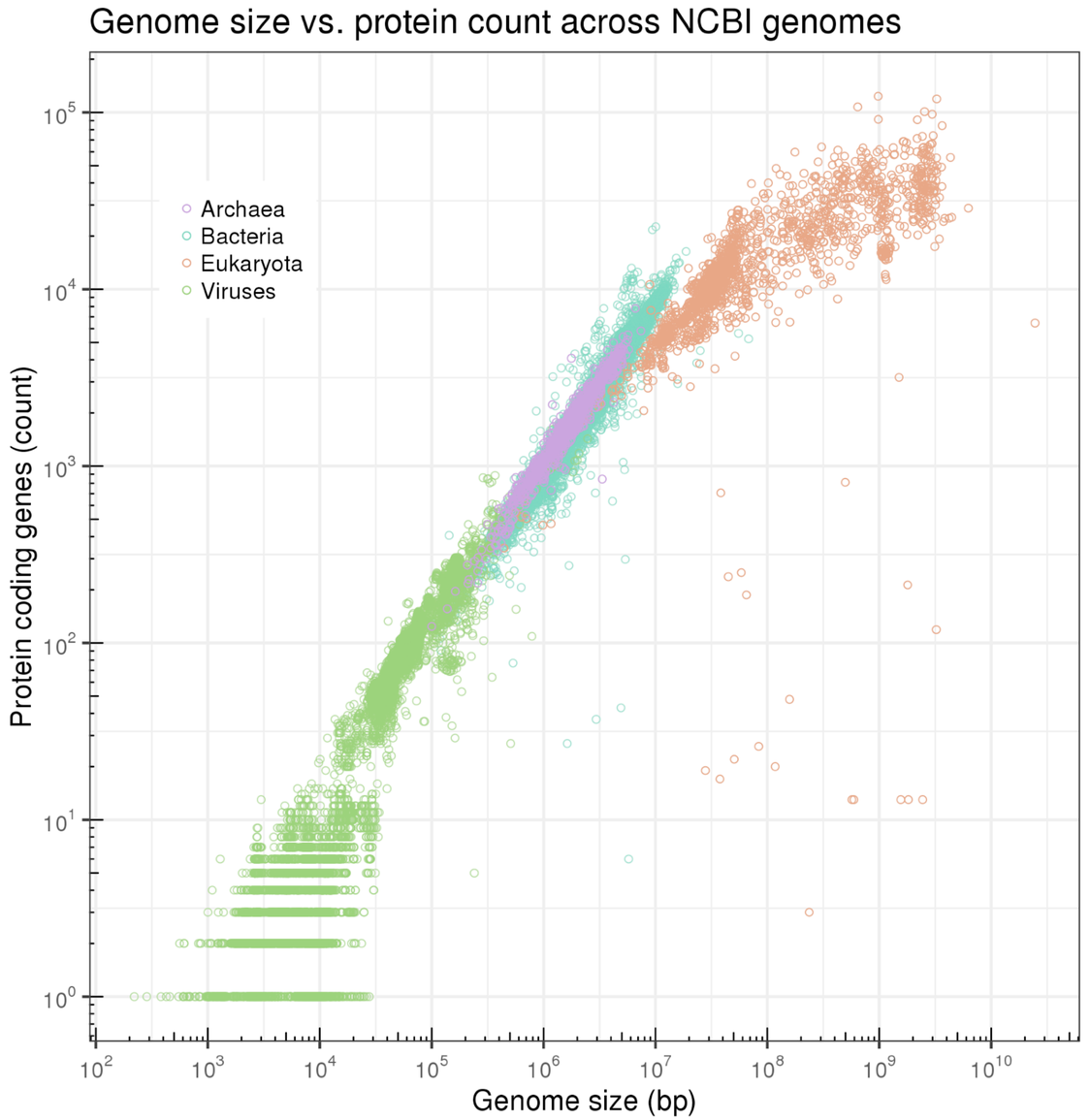


Bluetongue virus
(dsRNA 10 segments)



Images: ViralZone.expasy.org

Genome sizes can differ by as much as 10,000,000x



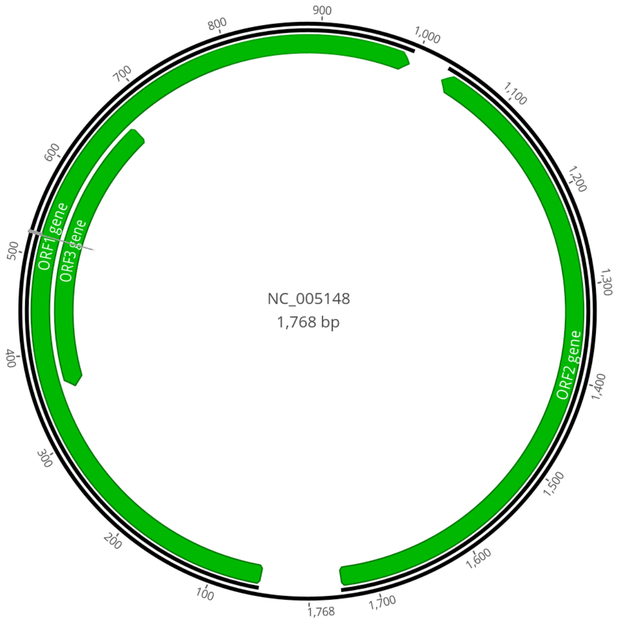
1000 → 10 billion

10 million fold difference in genome size

Image credit: smsaladi CC BY-SA 4.0 Link

ssDNA viruses have the smallest known genomes

One of the smallest genomes:
porcine circovirus



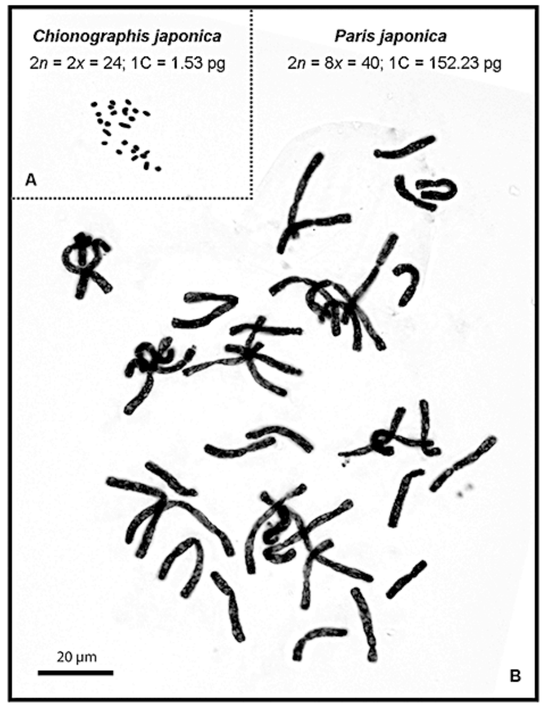
1700 nt ssDNA
3 proteins encoded

PCV makes pigs sick

Image credit: Dr. Joaquim Segalés (Universitat Autònoma de Barcelona)

Plants have the largest known genomes

Current largest estimated genome:
Paris Japonica (canopy plant)



~150 billion bp dsDNA

Current largest *sequenced* genome:
Pinus lambertiana (sugar pine)

27 billion bp dsDNA

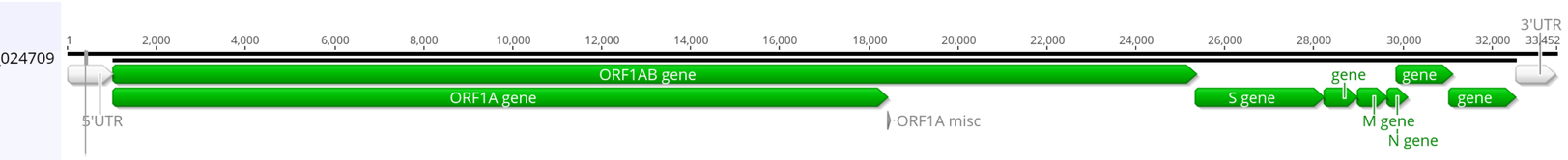
Pellicer et al (2010) J. Linn. Soc., Bot.

Image credit: Alpsdake, Public Domain Link

Image credit: S. Rae, CC BY-SA 2.0 Link

Fun fact: For a few years a virus I discovered held the record for the longest *RNA* genome

Ball python nidovirus



33,500 nt +ssRNA

Stenglein et al (2014) MBio

The current longest known RNA genome is that of a planarian virus

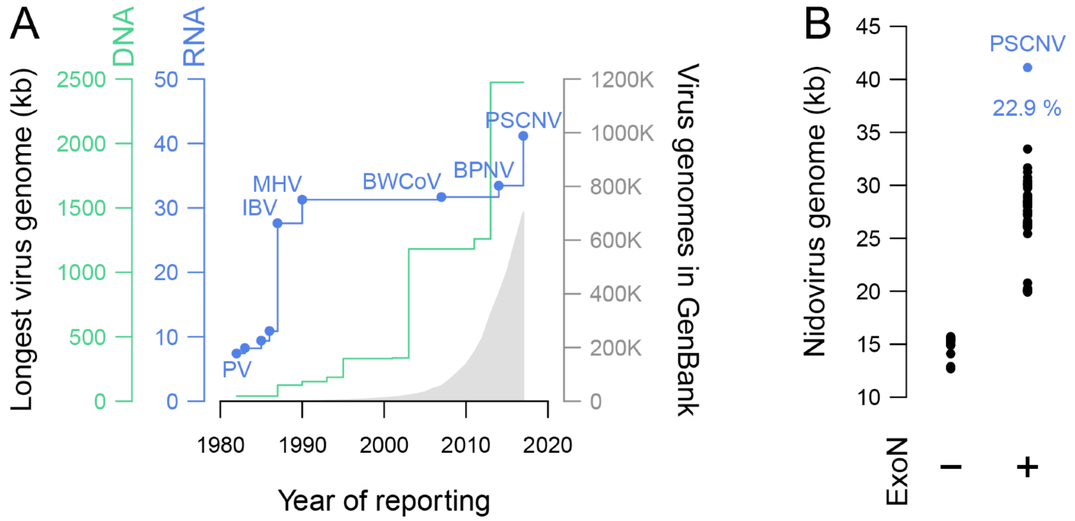
The planarian *Schmidtea mediterranea*

Can be infected by a nidovirus with a 41,100 nt ssRNA genome

RESEARCH ARTICLE

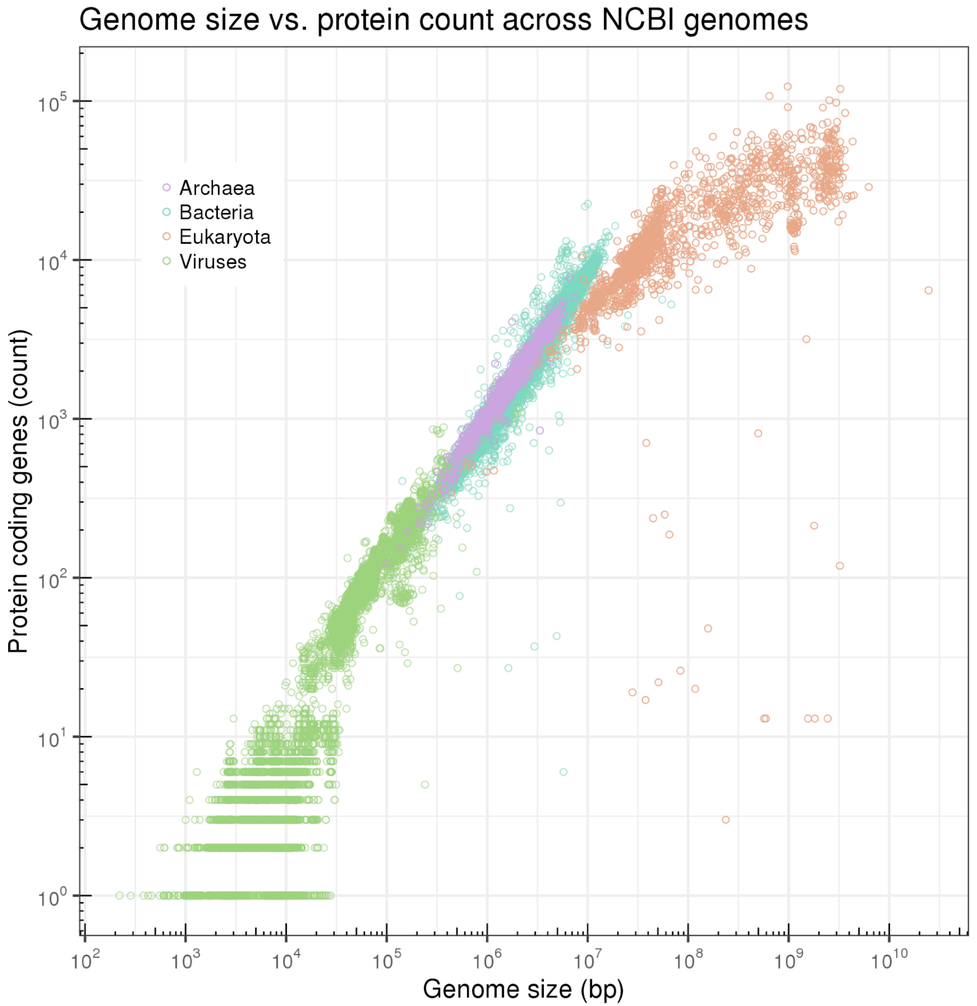
A planarian nidovirus expands the limits of RNA genome size

Amir Saberi[1☯¤a], Anastasia A. Gulyaeva[2☯], John L. Brubacher[3], Phillip A. Newmark[1¤b]*, Alexander E. Gorbalenya[2,4]*



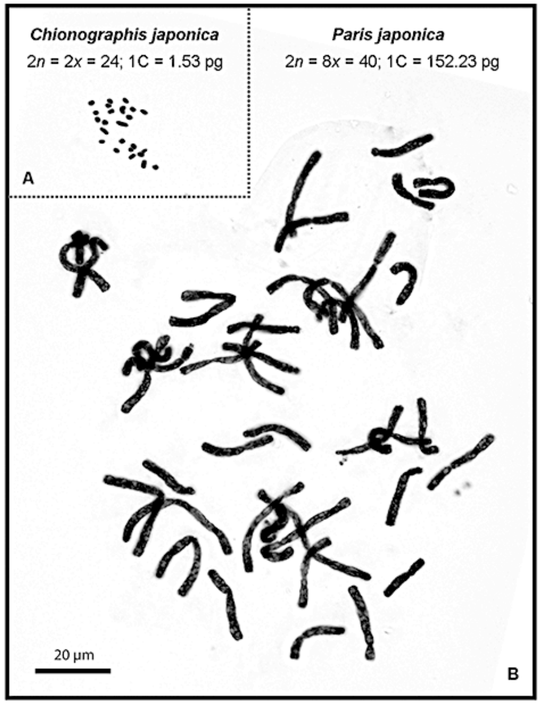
The size of RNA genomes is hypothesized to be limited by the high error rate of RNA virus RNA polymerases and the lack of post-replication error correction

Even within related groups of organisms genome size varies widely



There is overlap between the genome sizes of viruses and bacteria, and bacteria and eukaryotes

Plants in the same family with 1000x difference in genome size



Chionographis japonica

Paris japonica

Image credit: smsaladi CC BY-SA 4.0 Link

Pellicer et al (2010) J. Linn. Soc., Bot.

Image credit: Alpsdake, Public Domain Link

Exercise: Order the following genome sizes from largest (#1) to smallest (#6)

SI prefixes used to describe genome size

Name	Symbol	Base 10	Decimal	English Name
giga	G	10^9	1,000,000,000	billion
mega	M	10^6	1,000,000	million
Kilo	k	10^3	1,000	thousand

15 kbp

1.5 Gbp

15 Mbp

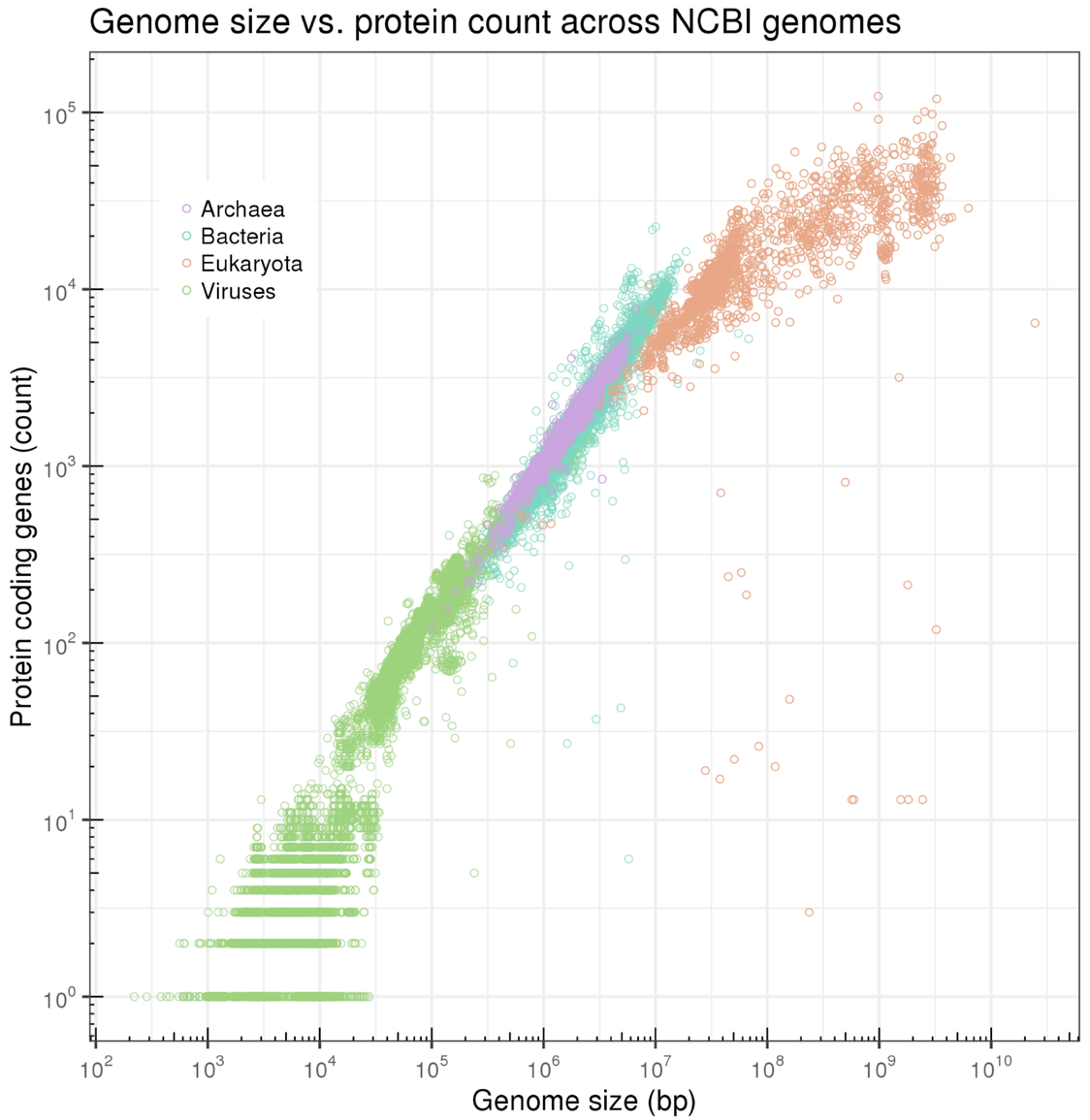
1.5 x10^6 bp

1500 bp

150,000 bp

Gene number varies less than genome size

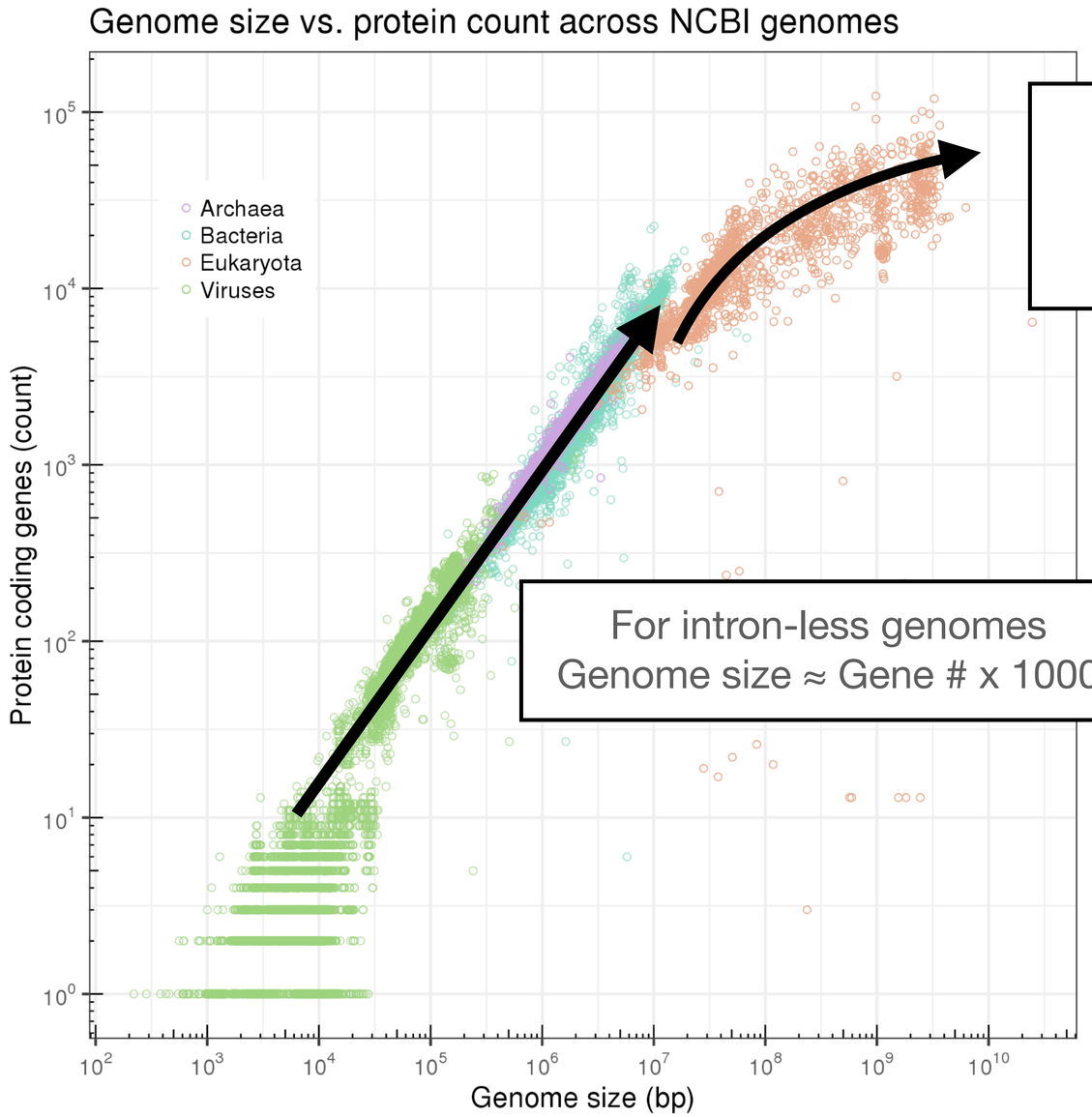
100,000 fold difference
In number of protein-coding genes



1000 → 10 billion

10 million fold difference in genome size

Gene number varies less than genome size



For intron-containing eukaryotic genomes Genome size grows faster than the number of genes

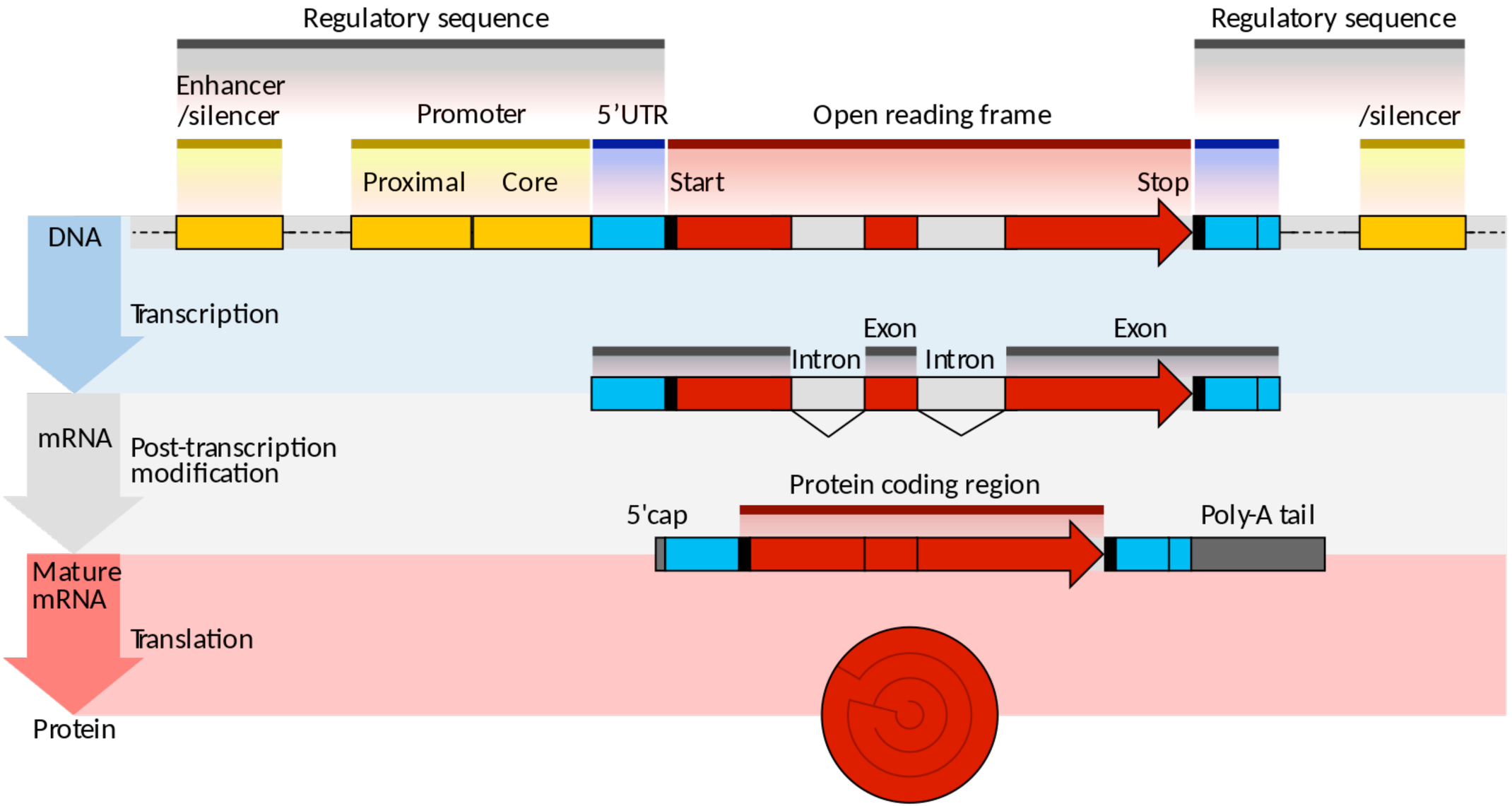
For intron-less genomes Genome size ≈ Gene # x 1000

1000 → 10 billion

10 million fold difference in genome size

Image credit: smsaladi CC BY-SA 4.0 Link

Eukaryotic coding sequences are typically on exons separated by introns



Prokaryotic coding sequences are typically continuous

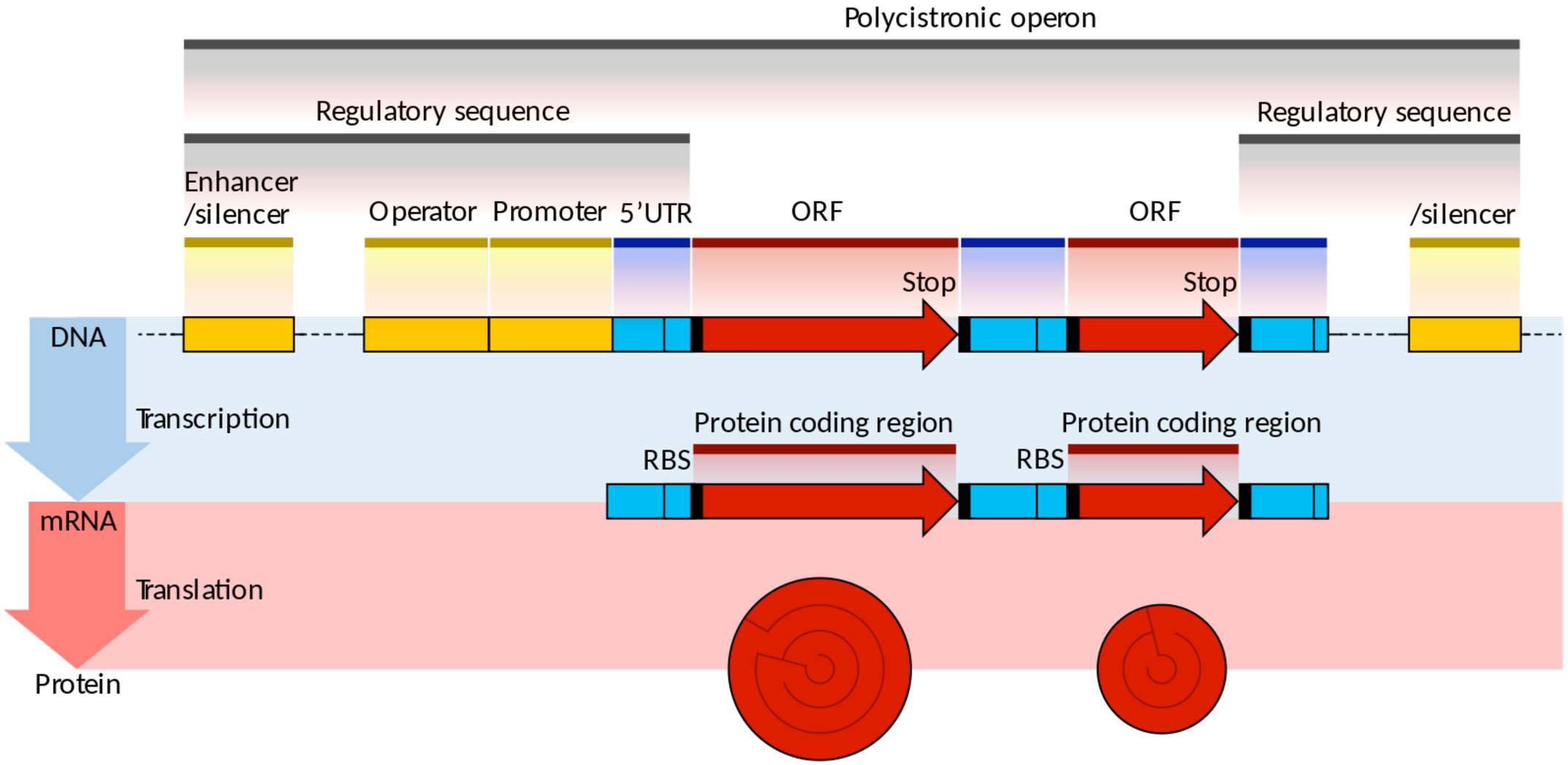


Image credit: Thomas Shafee CC BY 4.0, Link

Virus genomes are generally compact regardless of other molecular details

HIV uses splicing and ribosomal frame shifting

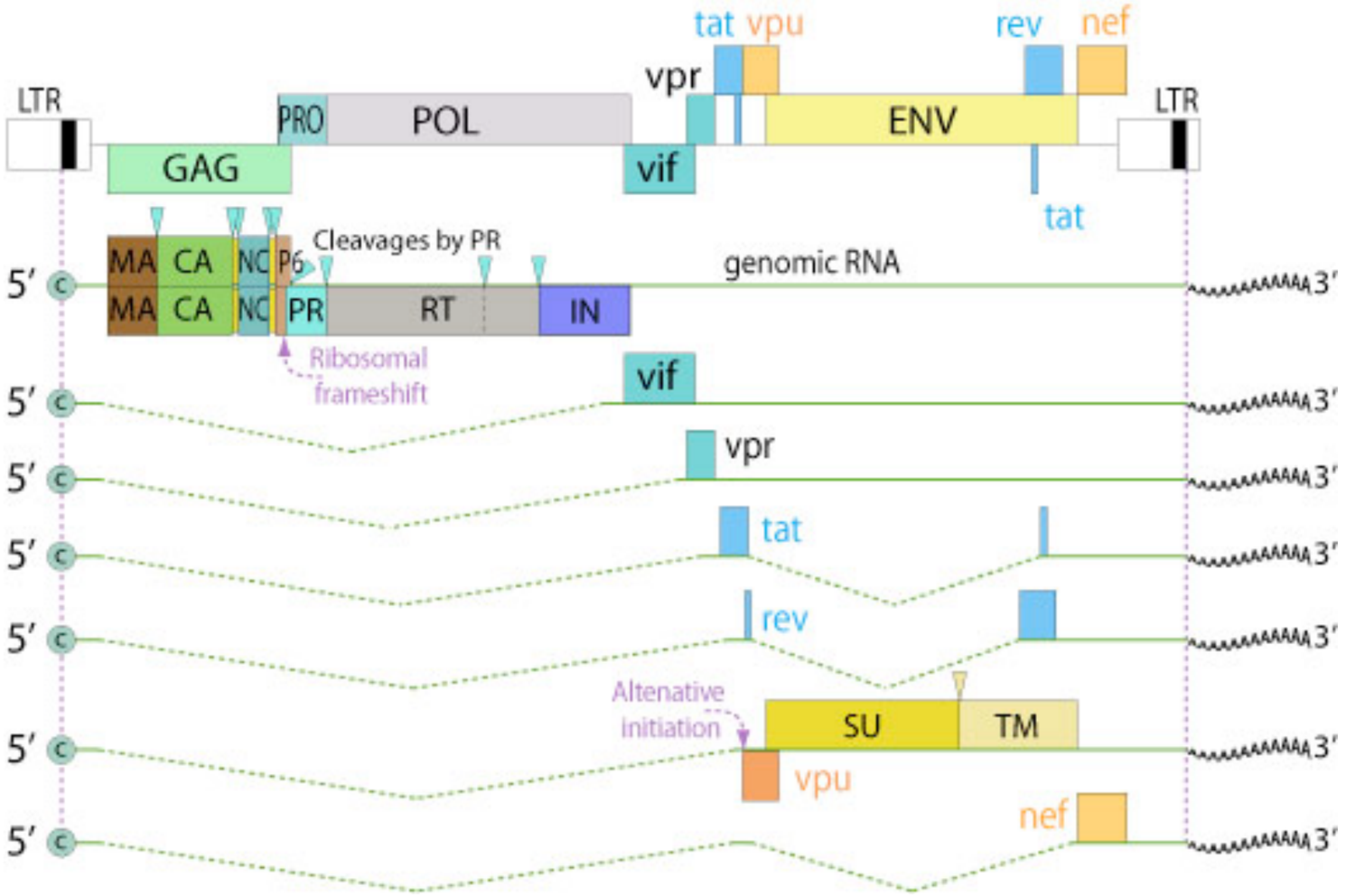


Image credit: Viralzone

Genomes differ in the fraction of genome occupied by repetitive content

The human genome is 50% repetitive

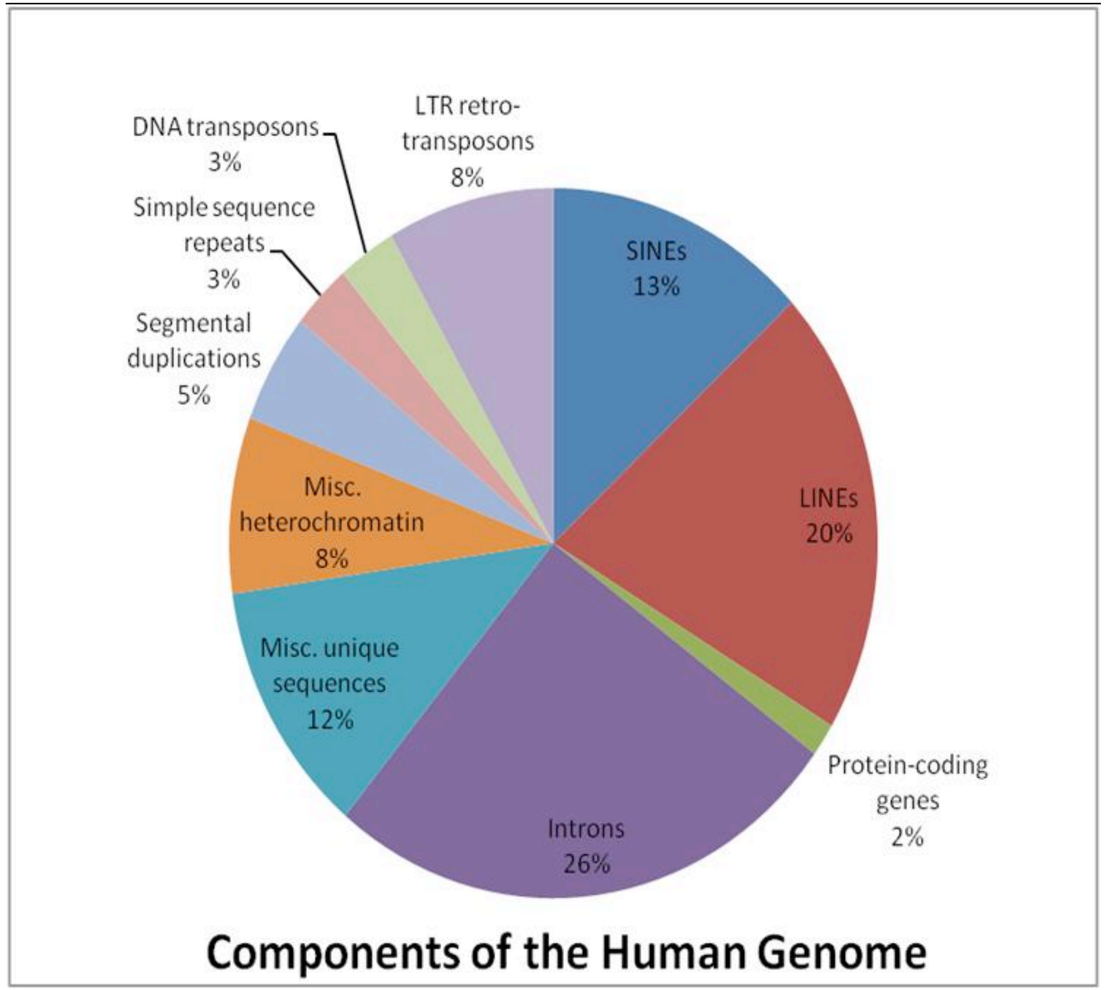
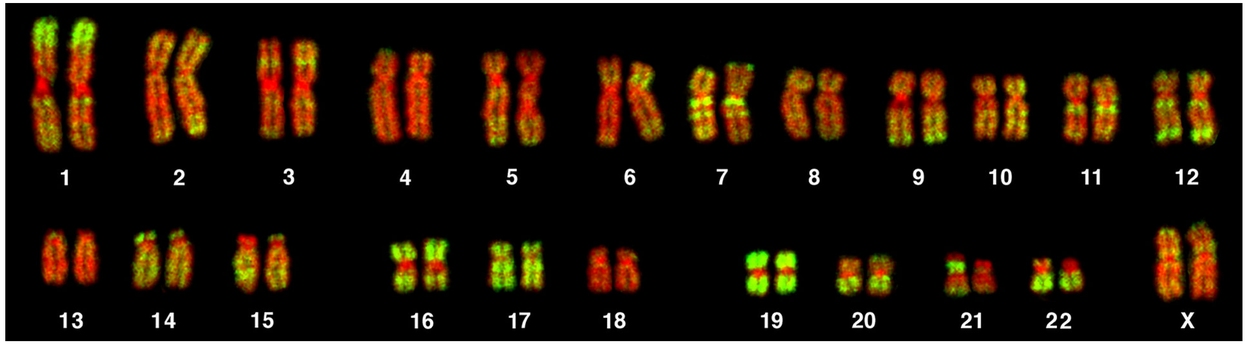


Image credit: Alglascock CC BY-SA 3.0 Link

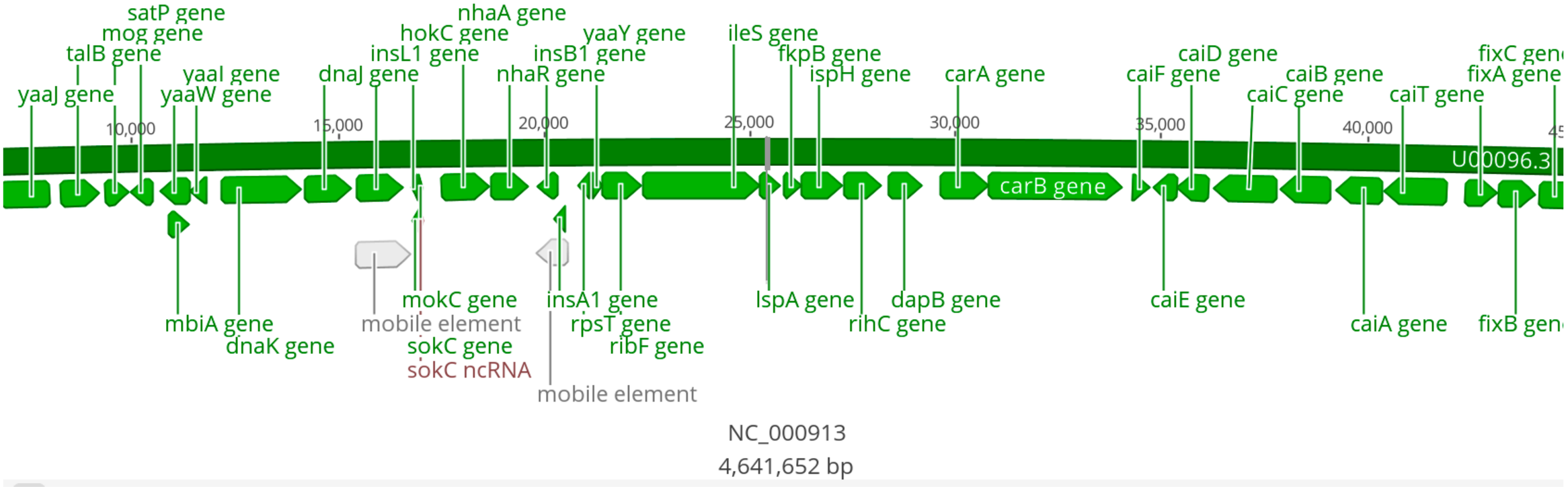
Alu elements alone compose ~10% of the human genom



Bolzer et al (2005) PLoS Biol

Prokaryotic and viral genomes are generally streamlined

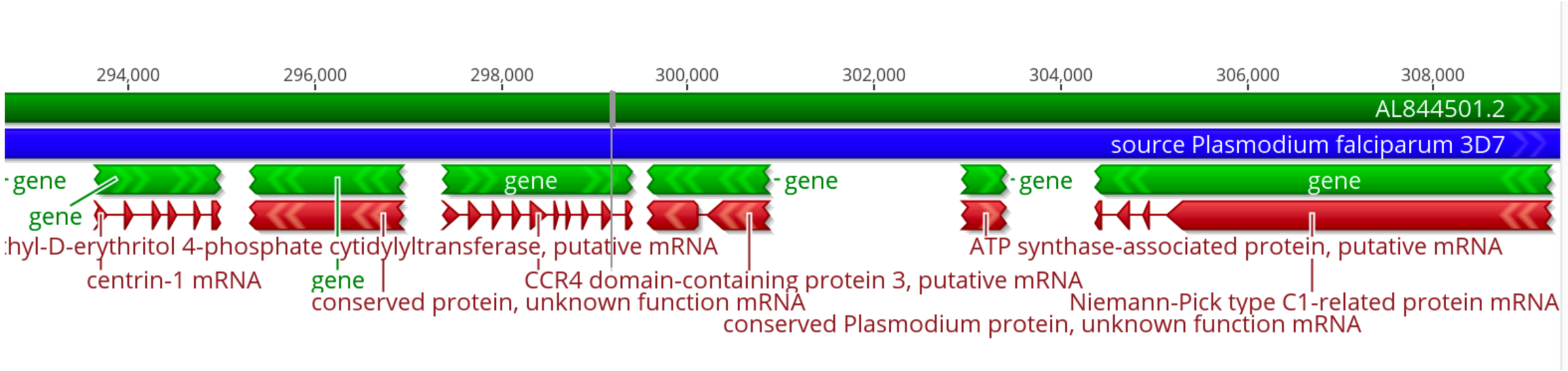
A small region of the E. coli K12 genome

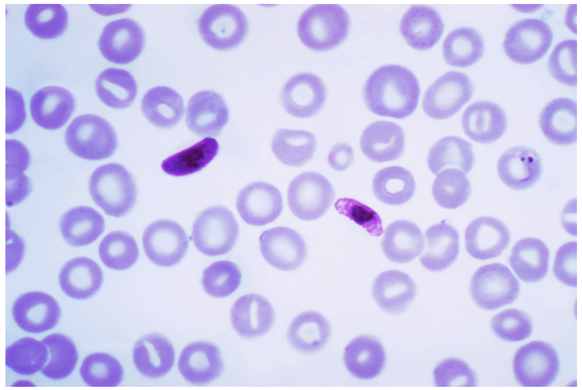


Note that there *are* repetitive elements in bacterial genomes but they constitute a minor fraction of the genomes

Some eukaryotic genomes are relatively compact

Plasmodium falciparum genome: 14 chromosomes, 23.5 Mbp, 5929 protein-coding genes



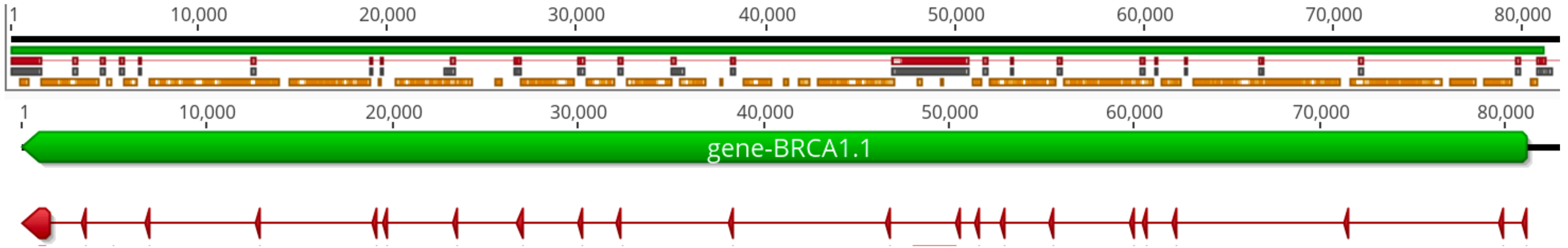


6 genes spread over ~16 kbp of the gnome
There are introns but they are short
Much less repetitive content than larger eukaryotic genomes

Image credit: Mae Melvin/CDC, Public Domain Link

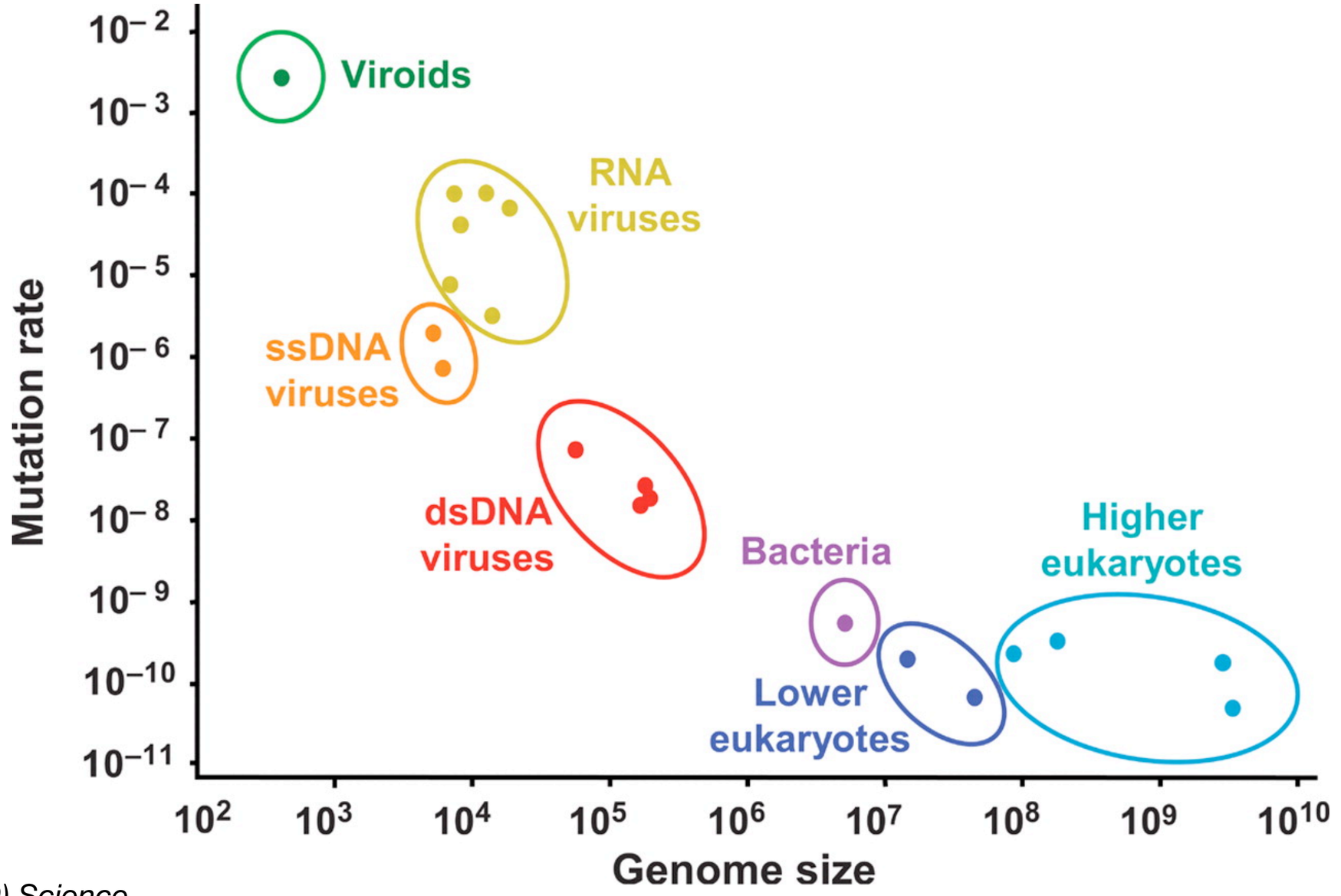
A typical large, repeat filled eukaryotic genome

Homo sapiens genome: 23 chromosomes, ~3 Gbp, 20,000 protein-coding genes



1 gene (BRCA1) spread over 80 kb of the human genome
Long introns
Tons of repeat elements

Mutation rates are generally ≈ 1 / genome size



Gago et al (2009) Science

Between species variation

ring-tailed cat

kinkajou

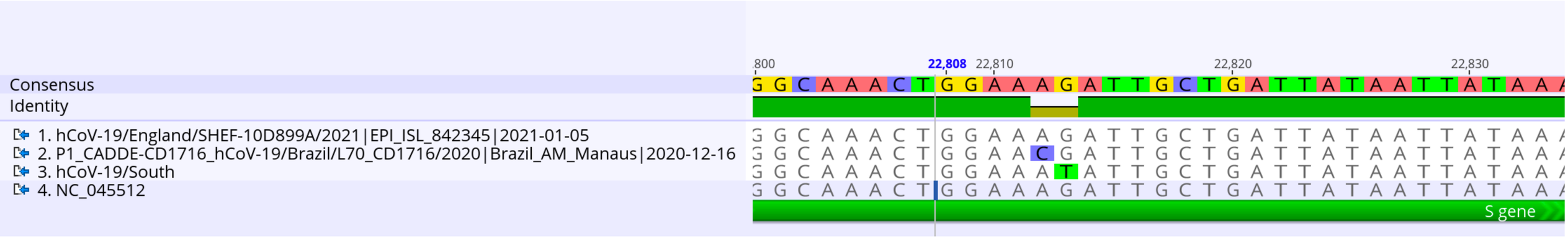
raccoon

```
ringtail  GAGGTCACACGCATGGTCATCATCATGGTCATTGCATTCCTGATCTGCTGGGTGCCCT
raccoon   GAGGTCACGCGCATGGTCATCATCATGGTCATTGCATTCCTGATCTGCTGGGTGCCCT
kinkajou  GAGGTCACACGCATGGTCGTCATCATGGTCATCGCATTCCTGATTTGCTGGTTGCCCT
```

Procyonid images: wikipedia, San Diego Zoo

Within species variation

SARS-CoV-2 spike gene sequences from viruses from 4 lineages



Structural variation occurs within and between species

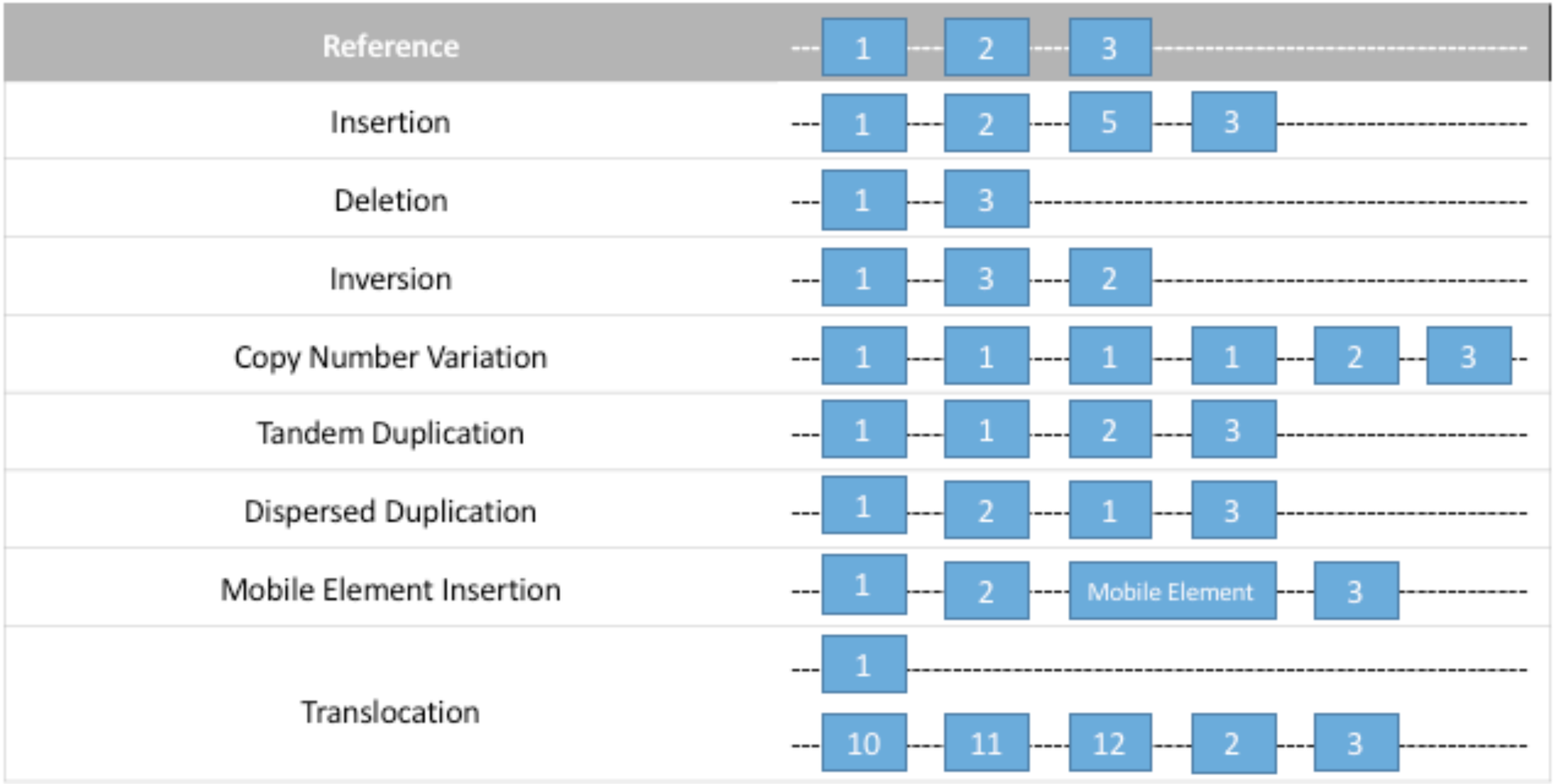
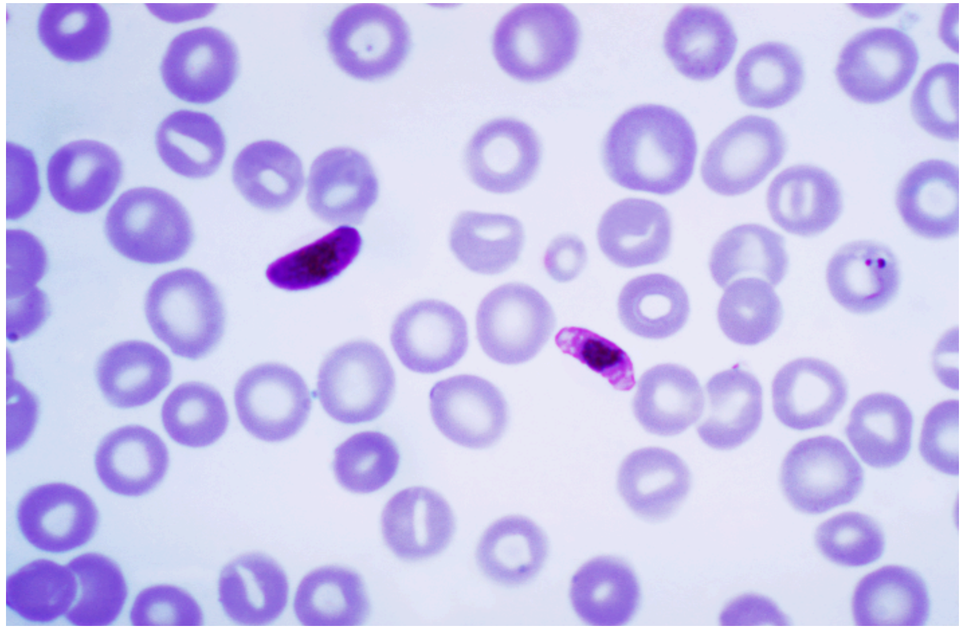
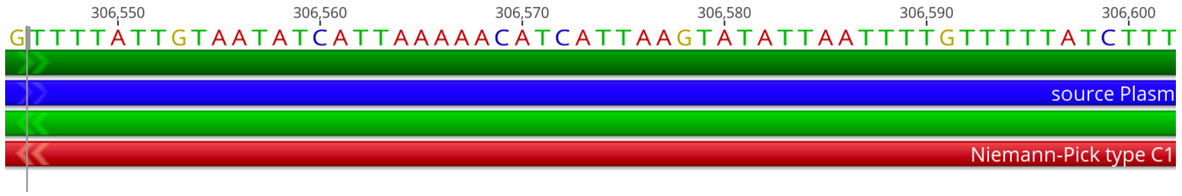


Figure 1: Depiction of different types of structural variants compared to the reference genome. Each different number represents a different gene.

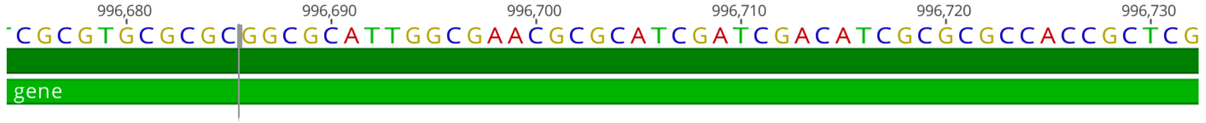
Genomes vary to a surprising degree in their average GC content

Plasmodium falciparum
Average GC content: 20%





Burkholderia pseudomallei
Average GC content: 68%



Ways that genomes differ from each other

1. Nucleic acid type (DNA, RNA, ss, ds, +/- sense)
2. Nucleic acid topology (circular, linear)
3. Number of molecules that make up the genome (chromosomes, viral genome segments)
4. Size of genome
5. Number of genes
6. Repeat content
    1. % repetetive
    2. Types of repeat elements
7. Structure of genes
    1. Introns or no introns
8. Copy # / ploidy
9. GC content
10. Mutation rate / substitution rate
11. Genetic code used for translation
12. Within and between species variation
    1. Single nucleotide variation
    2. Structural variation

Exercise: predict the organism that these genomes belong to

- NC_016072
- 1.3 Mbp
- Linear dsDNA
- 1120 protein-coding genes
- No introns

- NC_012920
- 16.6 kbp
- Circular dsDNA
- 13 protein-coding genes
- No introns

- NC_045512
- 29.9 kbp
- Linear ssRNA
- 12 protein-coding genes
- No introns

- NC_018414
- 162.6 kbp
- Circular dsDNA
- 190 protein-coding genes
- No introns

- NC_027779
- 7.3 kbp
- Circular dsDNA
- 6 protein-coding genes
- No introns

- GRCh38.p13
- 2.8 Gbp
- Linear dsDNAs
- ~20,000 protein-coding genes
- Introns

- NC_005148
- 1.7 kbp
- Circular ssDNA
- 3 protein-coding genes
- No introns

- NC_035963
- 446.3 kbp
- Circular dsDNAs
- 39 protein-coding genes
- No introns

- GCF_000854445
- 19.1 kbp
- Linear dsRNAs
- 10 protein-coding genes
- No introns

- GCA_001447015
- 27.6 Gbp
- Linear dsDNAs
- >20,000 protein-coding genes
- Introns

- NC_000913
- 92.7 kbp
- Circular dsDNA
- 85 protein-coding genes
- No introns

- NC_000962
- 4.4 Mbp
- Circular dsDNA
- 3906 protein-coding genes
- No introns

Exercises [end of class]:

1) Briefly describe 3 additional ways that genomes differ from each other