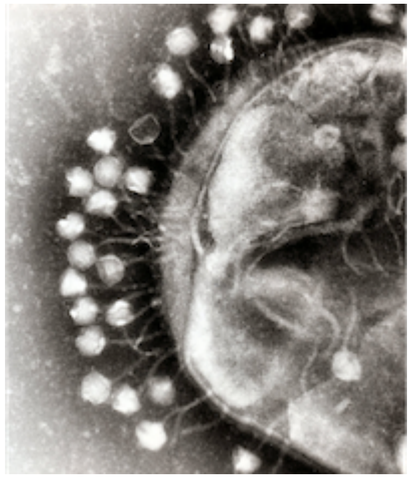
CRISPRにターゲットされるヒト腸内メタゲノム配列の網羅的データベースの構築

○杉本竜太[1]　西村瑠佳[1,2]　井ノ上 逸朗[1]

1)情報・システム研究機構国立遺伝学研究所人類遺伝研究室　2)総合研究大学院大学

Viruses

The most diverse and numerous genetic entities on Earth



About 10^{31} viruses on Earth

Number of sand grains: 10^{18}
Number of bacterial cells: 10^{30}

About 50 viral species infect one mammalian species
About 5 thousands mammalian specie are known; There are about 300 mammalian infecting viral species

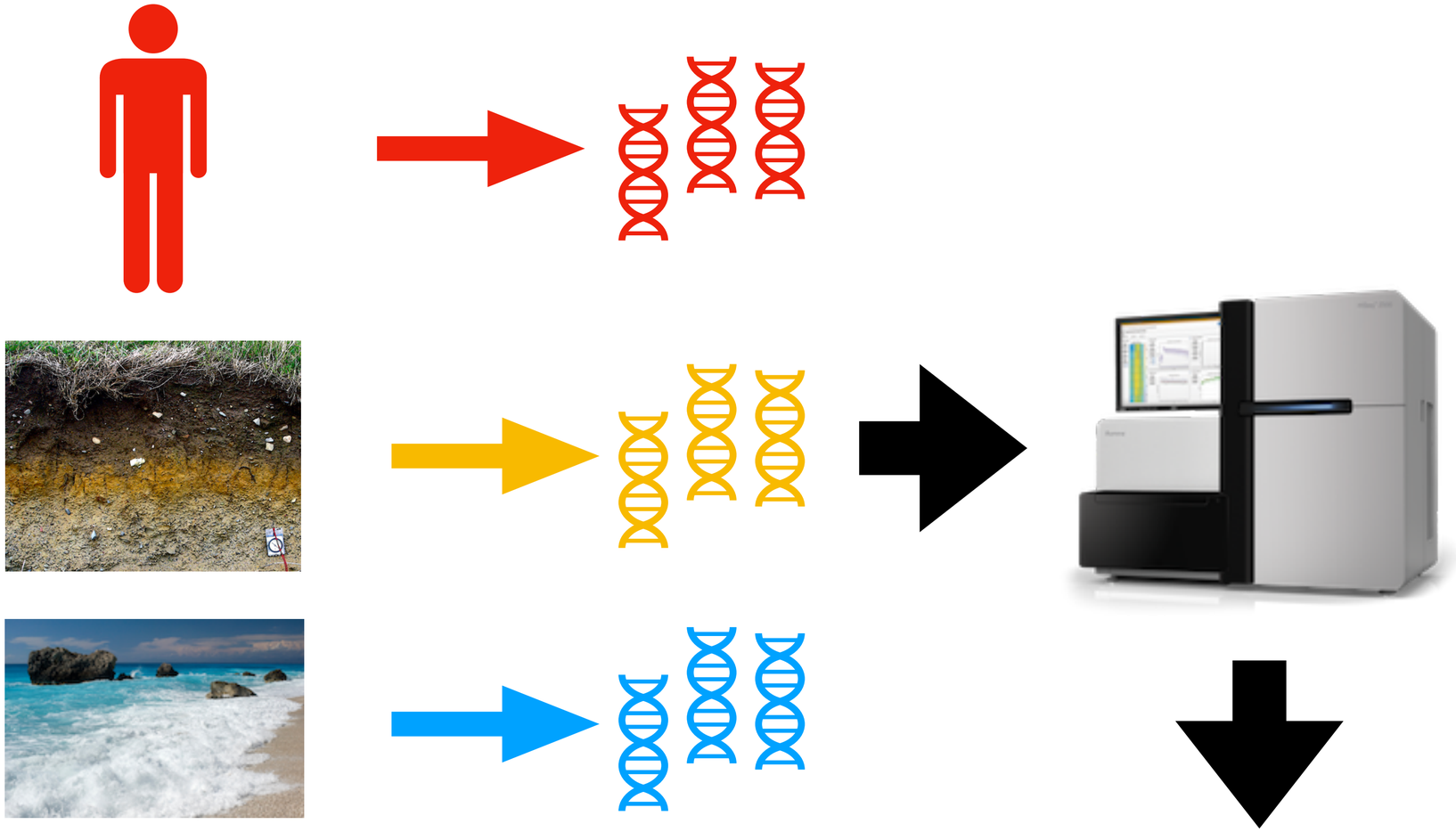
Currently, about 7 thousands viral species are recorded in ICTV
This is a portion of entire virome on Earth

Virus evolutions and involvement to the cellular life

- Horizontal gene transfer
  - Viruses possibly transfer genes between species
- Evolutional arms race
  - Host evolves immune system, and viruses evolve to counter or escape it
- Intriguing origin(s) of viruses
  - Viruses likely have multiple origins, and they are very old

To understand the evolution of viruses, we need diverse viral genomes

Metagenome contains viral genetic informations



AATACCTGCCTGTCACGCAGGGGGTCGCGGGTTCGAGTCC...

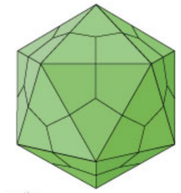
CAGTGTAATGTCAAGATAAACTTGGGACAGCTCATTA...

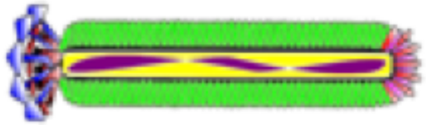
TTATCGAATCCCTCAGCACGGGCCAGGGAAA...

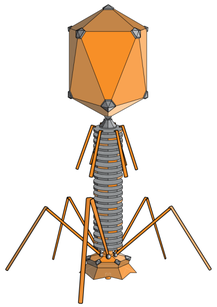
Metagenome is a mixture of all organisms and viruses in the sample

How do we identify viral genomes from metagenome?

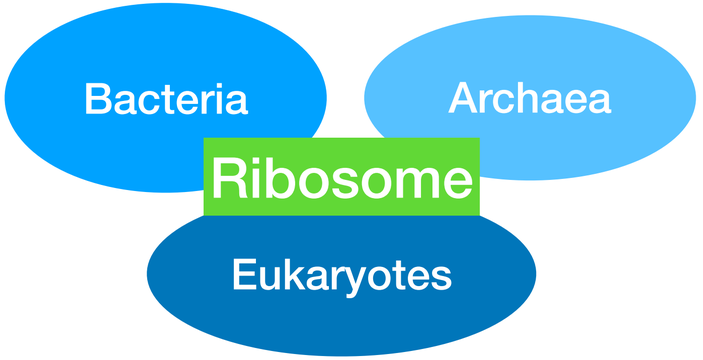
Difficulties to infer viral genomes







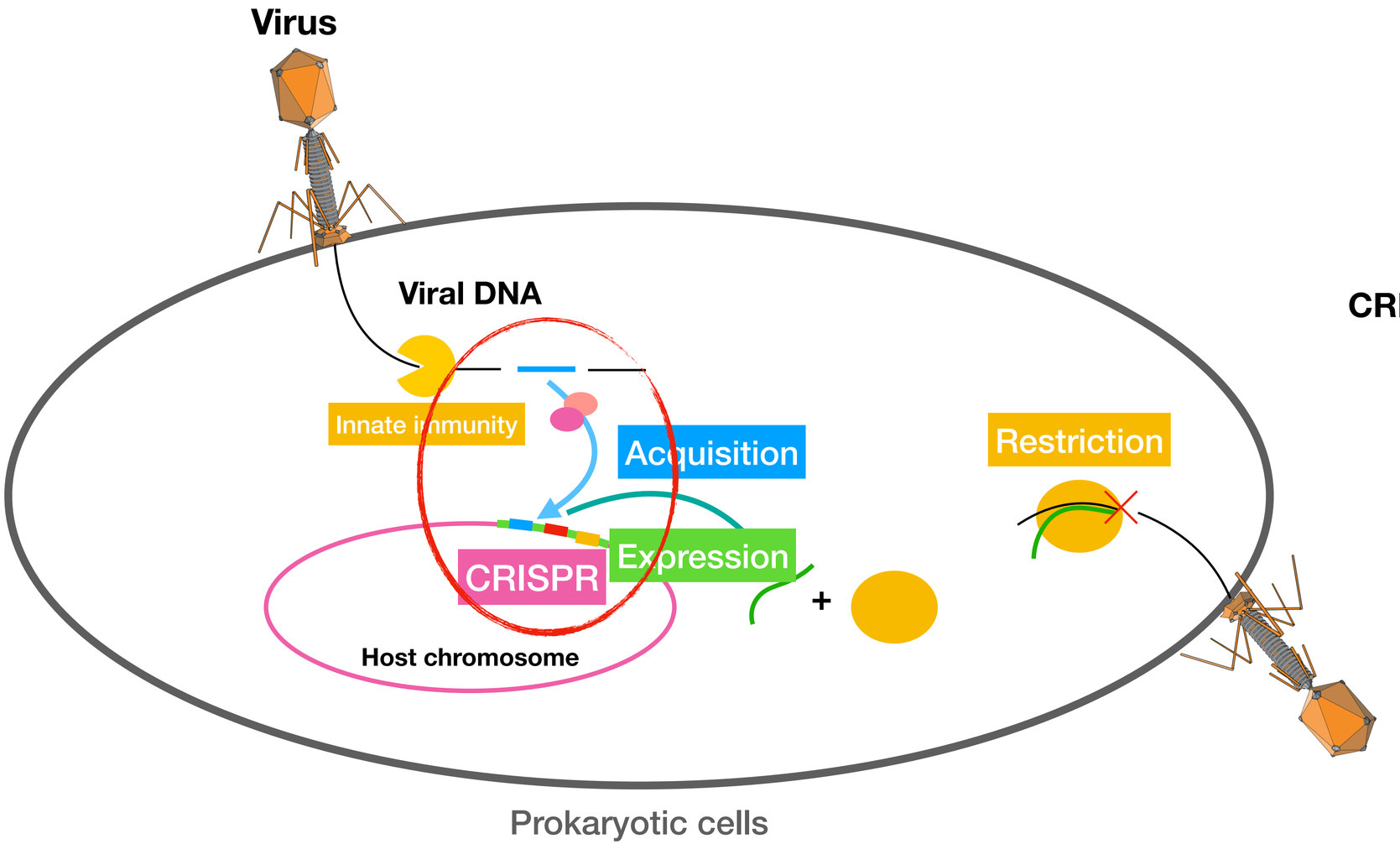
There is no gene shared across all viral lineages

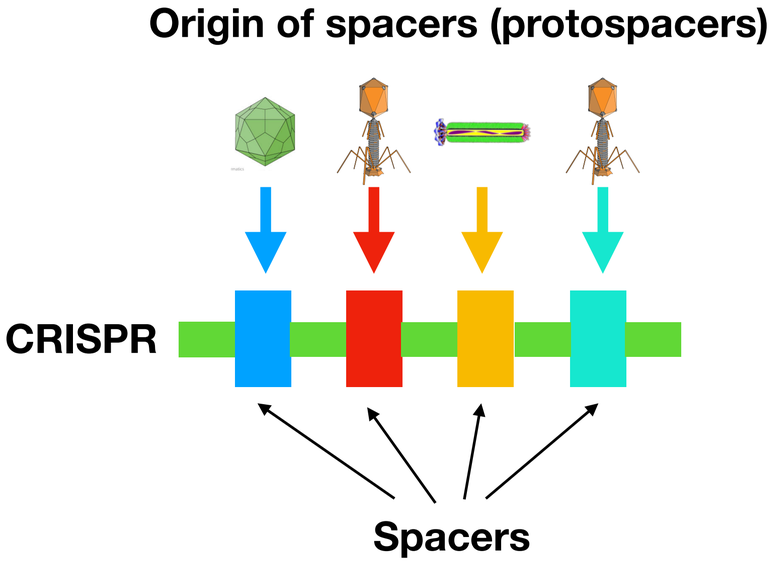


Cellular life share genes such as ribosome

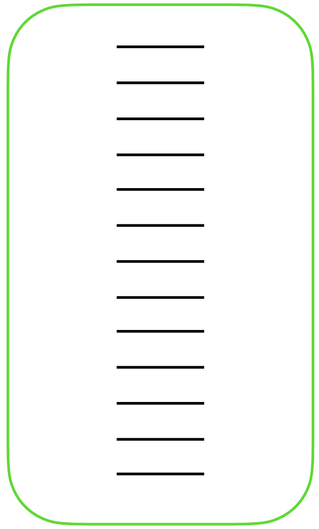
Unlike cellular life, we cannot define a marker gene for viral genomes

CRISPR is a prokaryotic adaptive immunity

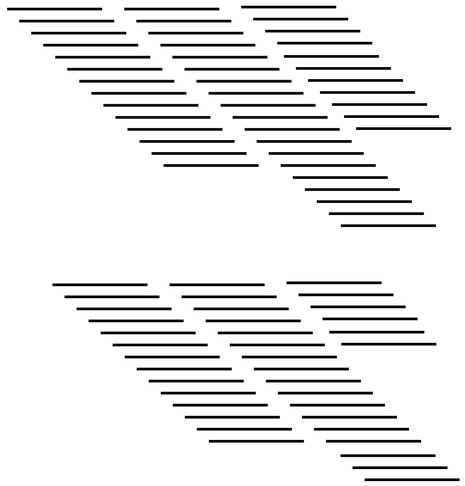




Using CRISPR spacers, we can identify viral genomes from metagenome

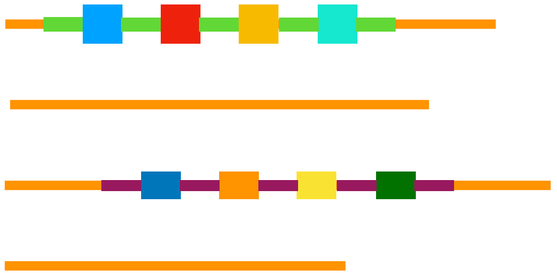


Metagenome reads



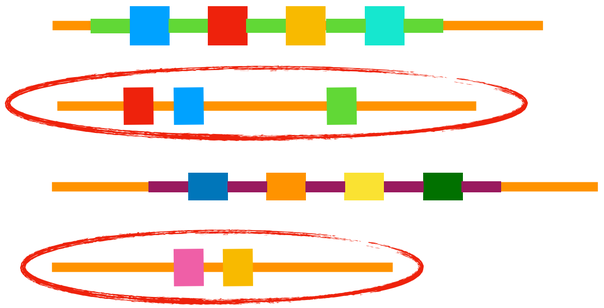
Genome assembly

Contigs



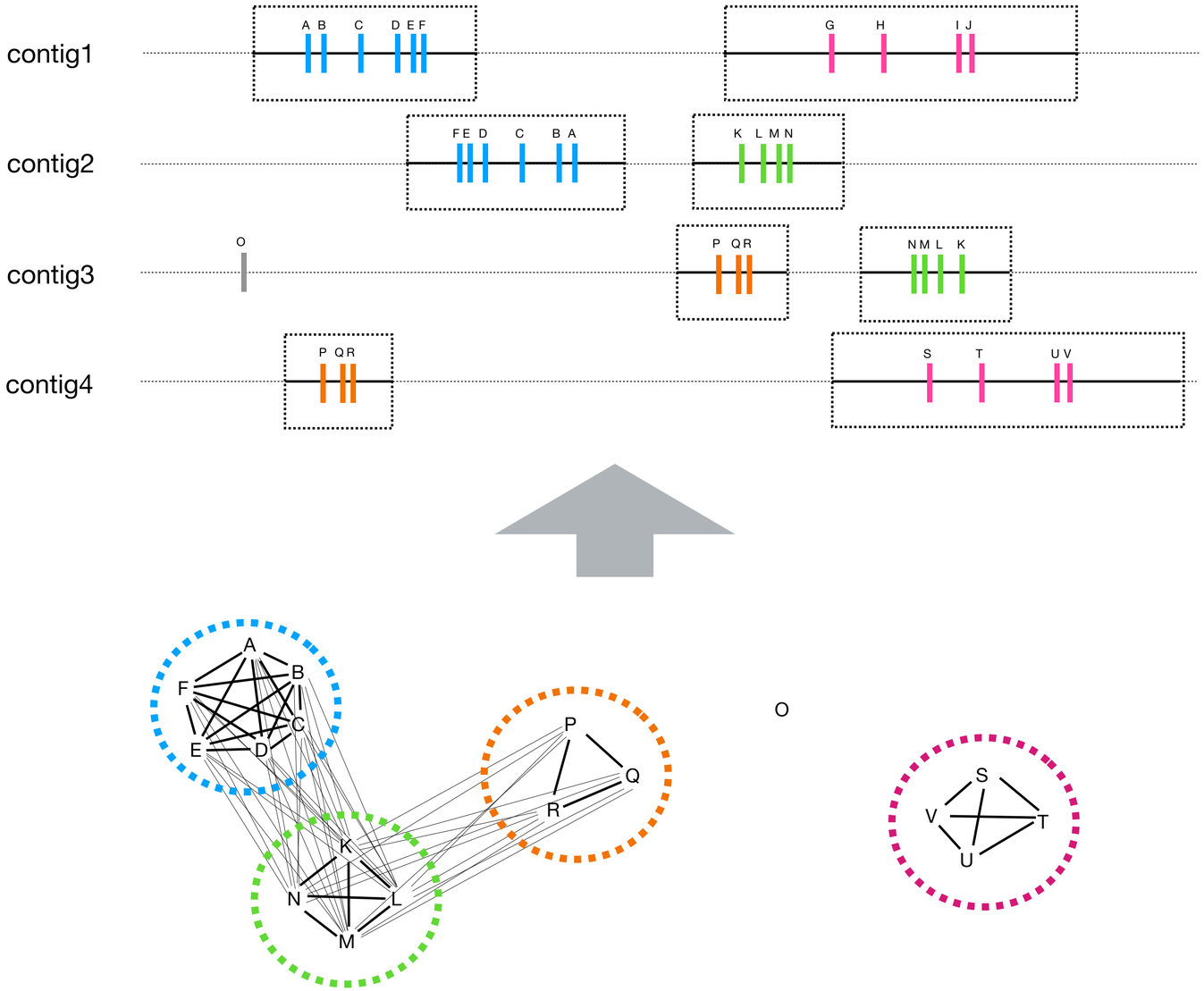
Detect CRISPRs

Collect spacers



Identify contigs targeted by CRISPR

Applying graph-based method to extract CRISPR targeted sequences



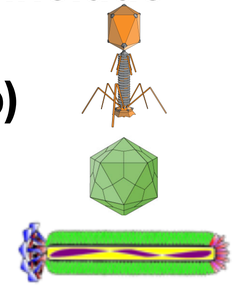
Result

Analyzed 11,817 human gut metagenome dataset (50.7 Tb)
Assembled 180,068,349 (767.7 Gb) contigs
Detected 11,223 CRISPRs
Extracted 1,969,721 CRISPR spacers

Using them, we detected 11,391 CRISPR targeted sequences

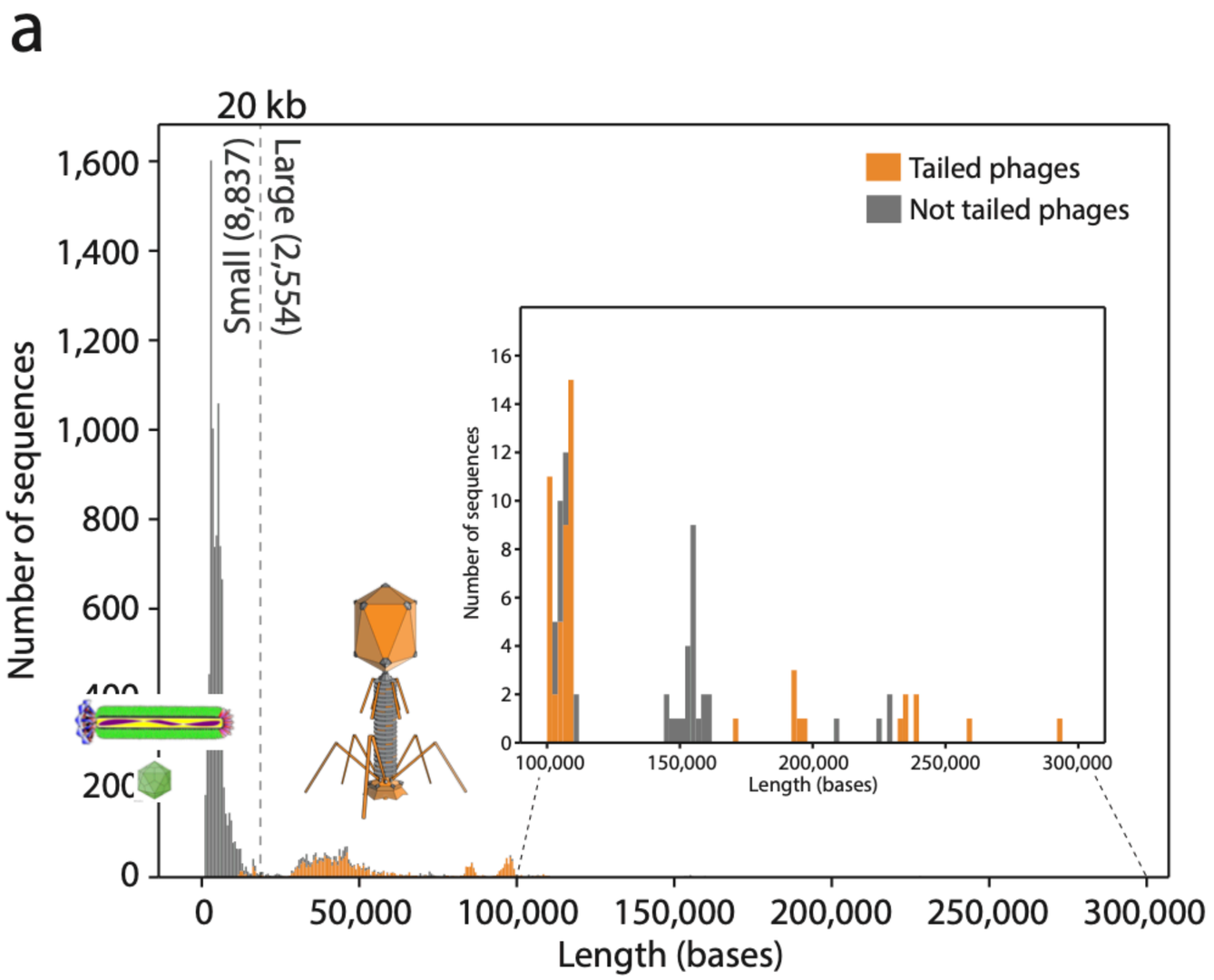
The extracted sequences include

- **257 crAssphage**
- **5 huge phages (> 200 kb)**
- **766 *Microviridae***
- **56 *Inoviridae***

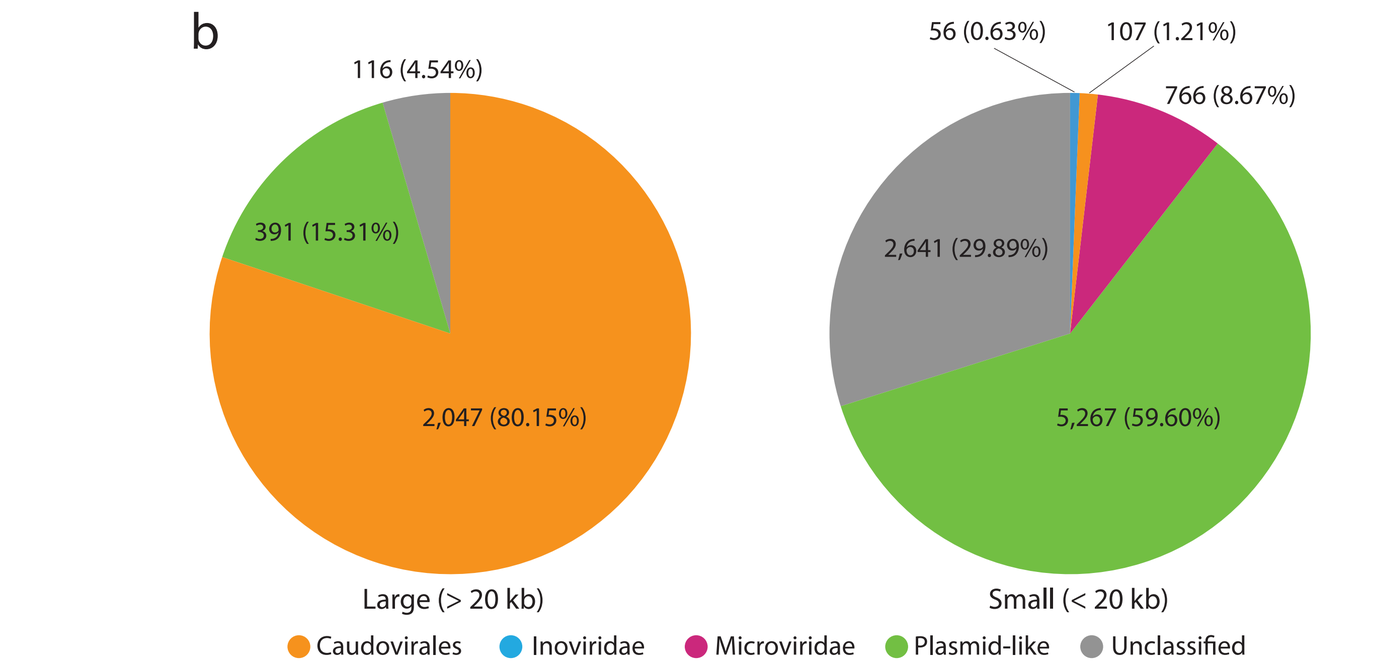


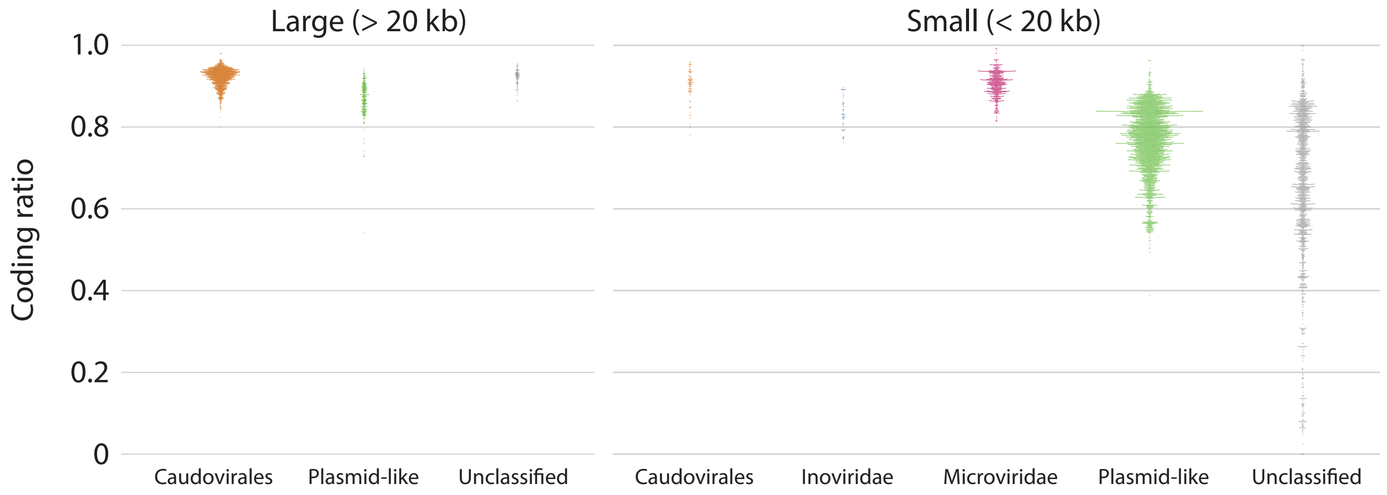
We also identified the hosts of 7,937 (69.7%) CRISPR targeted sequences

Length distribution of discovered genomes



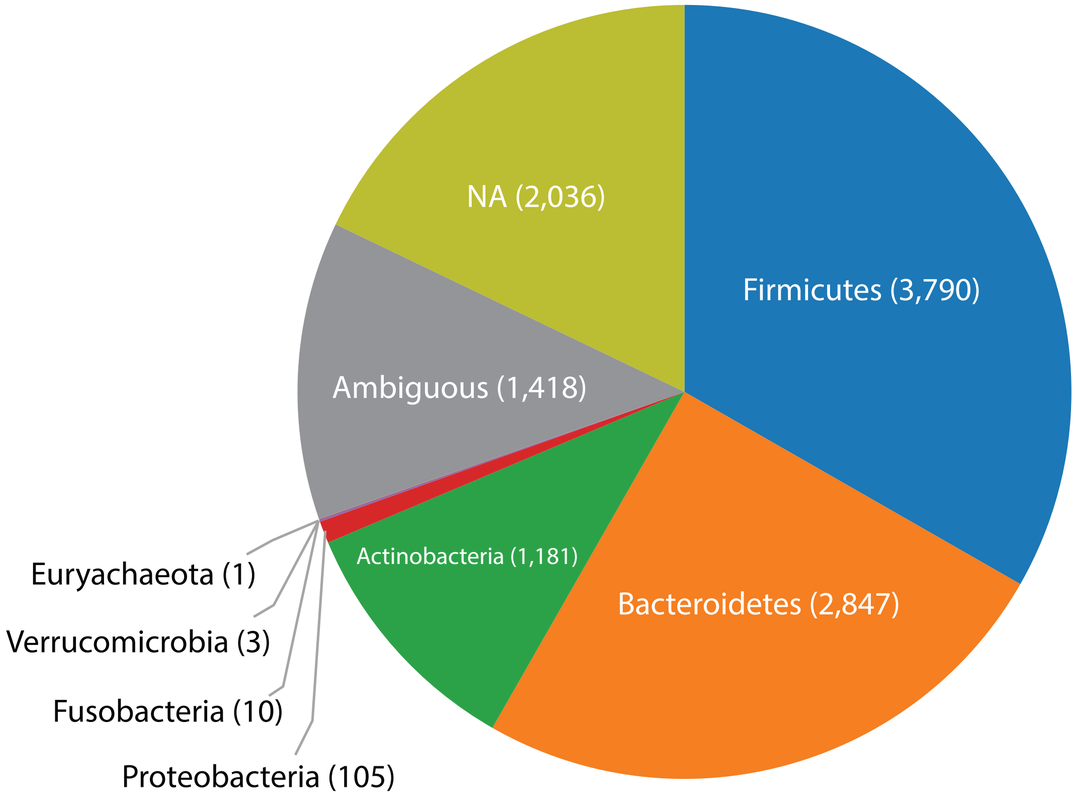
Classification of discovered CRISPR targeted sequences



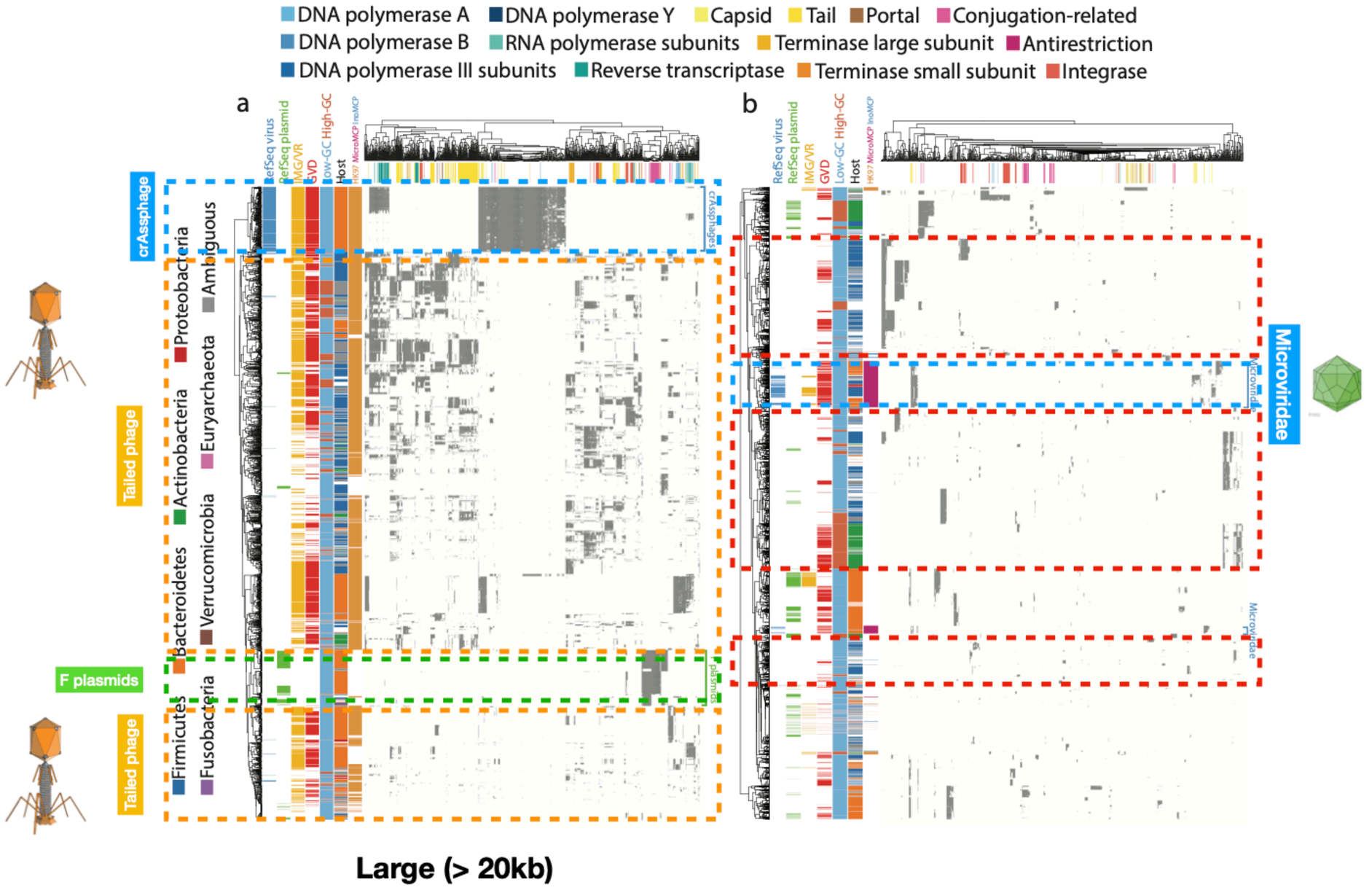


Predicted hosts

a



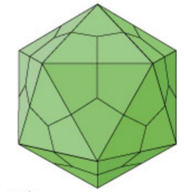
Hierarchical clustering based on gene components

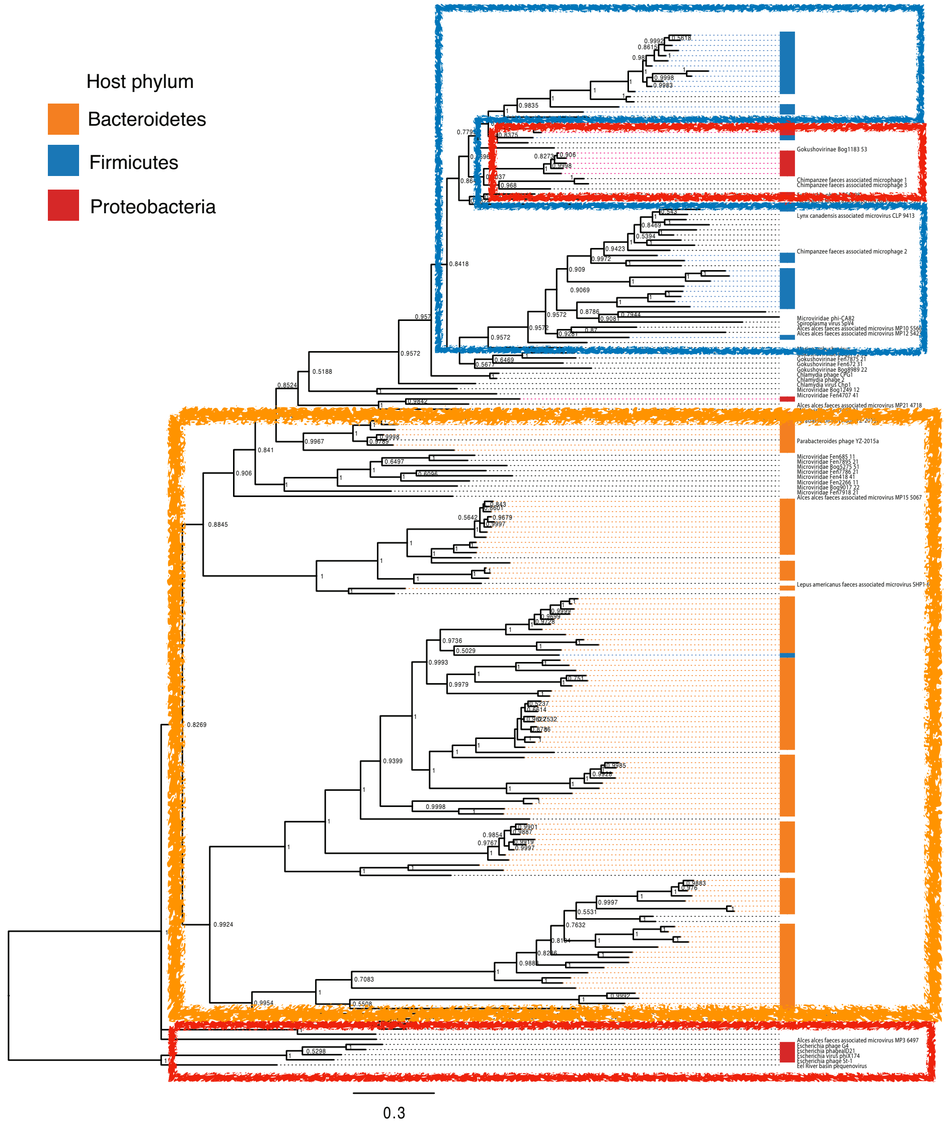


Large (> 20kb)

Discovered numerous uncharacterized genomes

Phylogeny of *Microviridae* species





Summary

- Discovered about 10 thousands CRISPR targeted genomes
- Substantial portion of them encode capsids
- 70% of them were host resolved at phylum level
- Discovered many novel small genomes

Next steps

- Further inspection of novel genomes
- Discover RNA viruses and archaeal viruses
- Construct database for viral genomes and proteins