De novo virus inference and host prediction from metagenome using CRISPR spacers

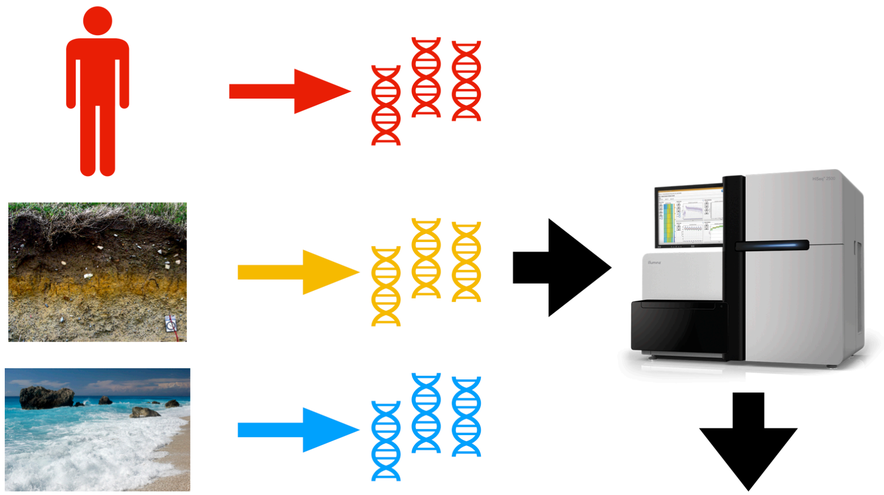
○Ryota Sugimoto[1], Luca Nishimura[1,2], Jumpei Ito[3], Nicholas F. Parrish[4],
Hiroshi Mori[1], Ken Kurokawa[1], Hirofumi Nakaoka[5] and Ituro Inoue[1]

1) Human genetics laboratory, National Institute of Genetics, Research Organization of Information and Systems
2) The Graduate University for Advanced Studies, SOKENDAI
3) Division of Systems Virology, Department of Infectious Disease Control, International Research Center for Infectious Diseases, Institute of Medical Science, The University of Tokyo
4) Genome Immunobiology RIKEN Hakubi Research Team, Center for Integrative Medical Sciences, RIKEN
5) Department of Cancer Genome Research, Sasaki Institute

Introduction

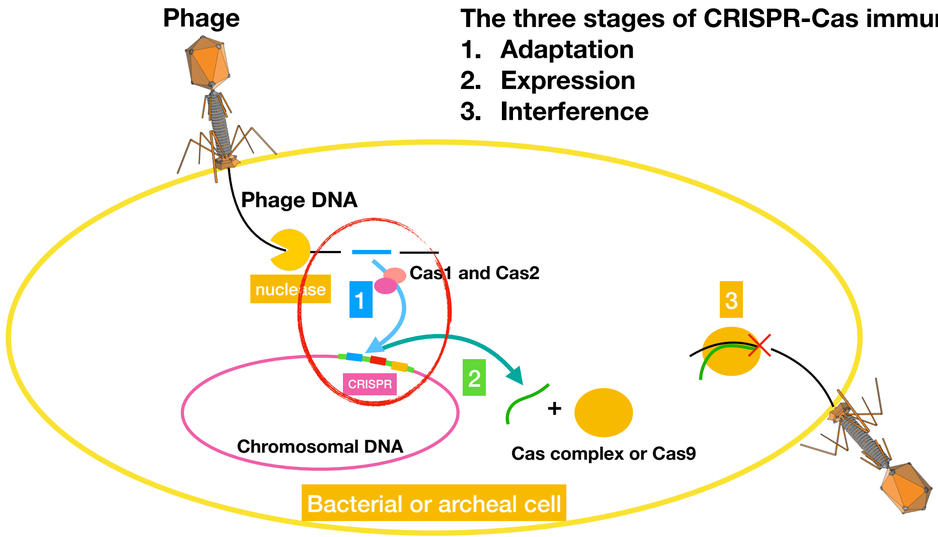
About 10^31 viruses exist on Earth
The evolution and origin(s) of viruses are poorly understood
We need a lot of genomic sequences from variety of viral lineages

Metagenome is mixture of bacterial, archeal, eukaryotic and viral genome sequences



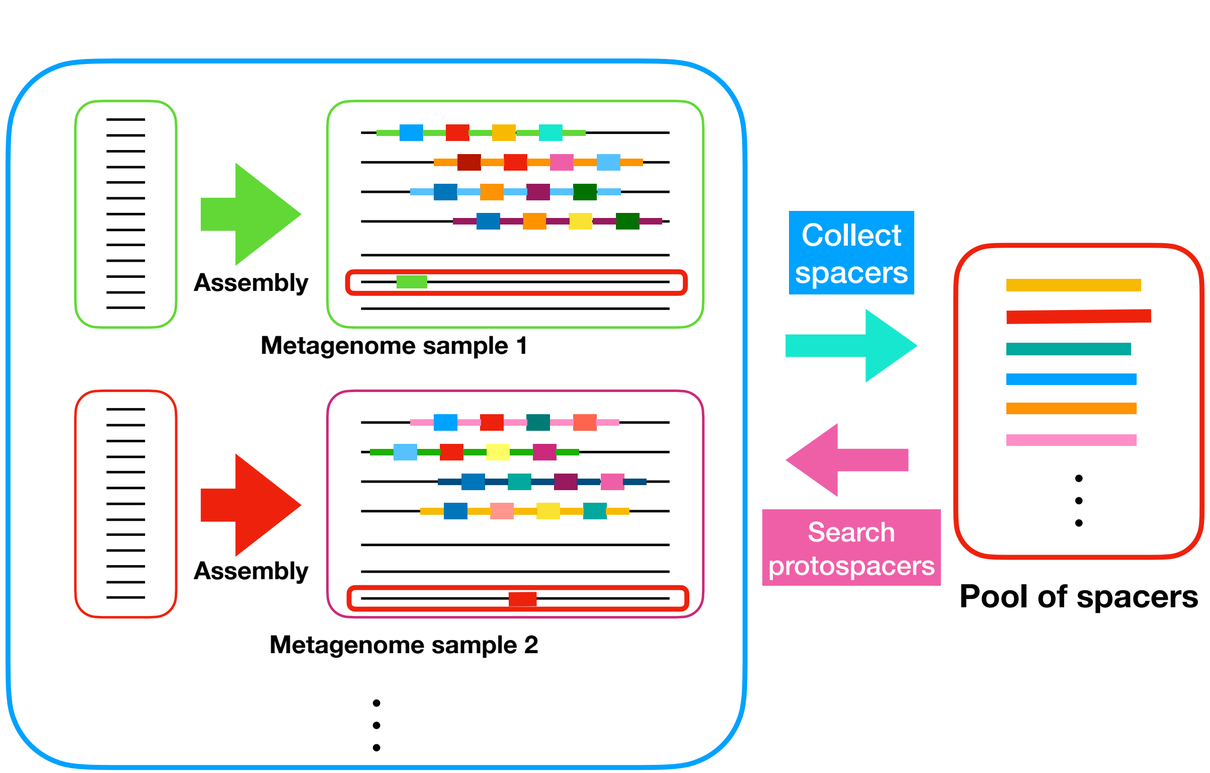
How do we detect viral sequences from metagenome?

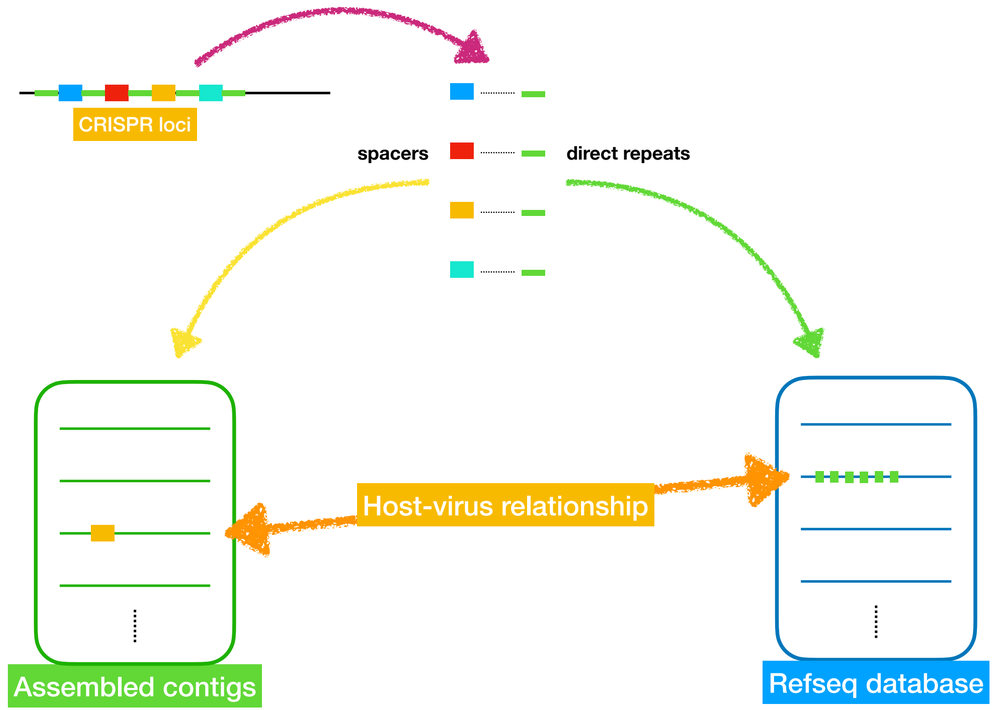
We use CRISPR, prokaryotic adaptive immunological memory!



Using CRISPR spacers, we can infer viral genomes from metagenome!

Method





Result 1

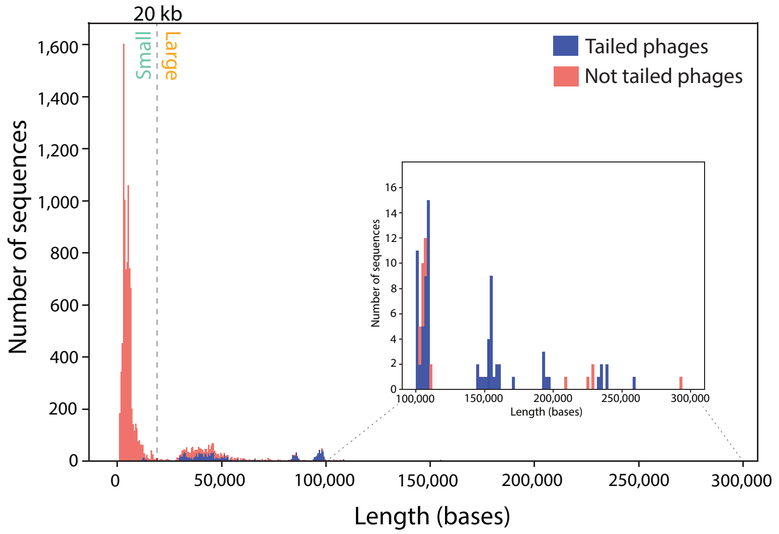
11,817 human gut metagenome datasets (50.7 Tb) were analyzed
180,068,349 assembled contigs (767.7 Gb)
11,223 unique CRISPR direct repeats
1,969,721 unique CRISPR spacers

11,391 unique nearly complete CRISPR targeted sequences

Including, 257 crAssphages, 11 genomes larger than 200 kb, 766 Microviridae,
114 Inoviridae and many entirely novel genomes

Result 2

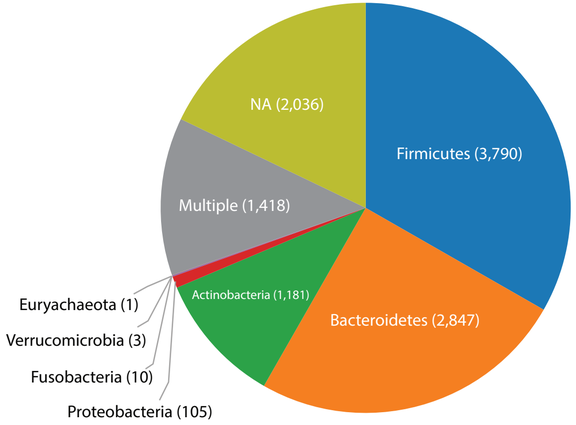
Length distribution of CRISPR targeted genomes



Majority of genomes longer than 20 kb are likely tailed phages

Result 3

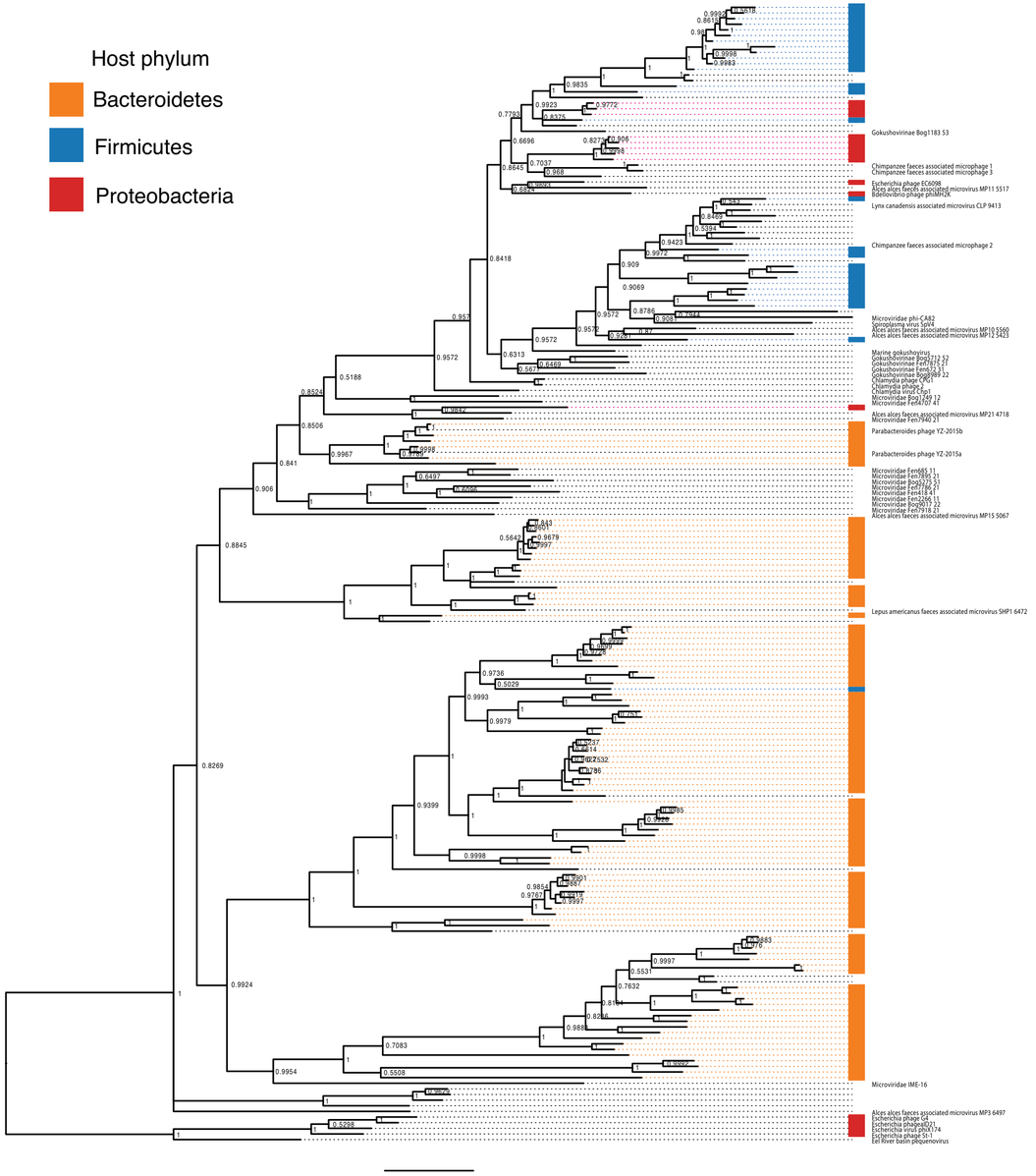
Predicted targeting hosts composition



The host composition is analogous to microbe composition in human gut

Result 4

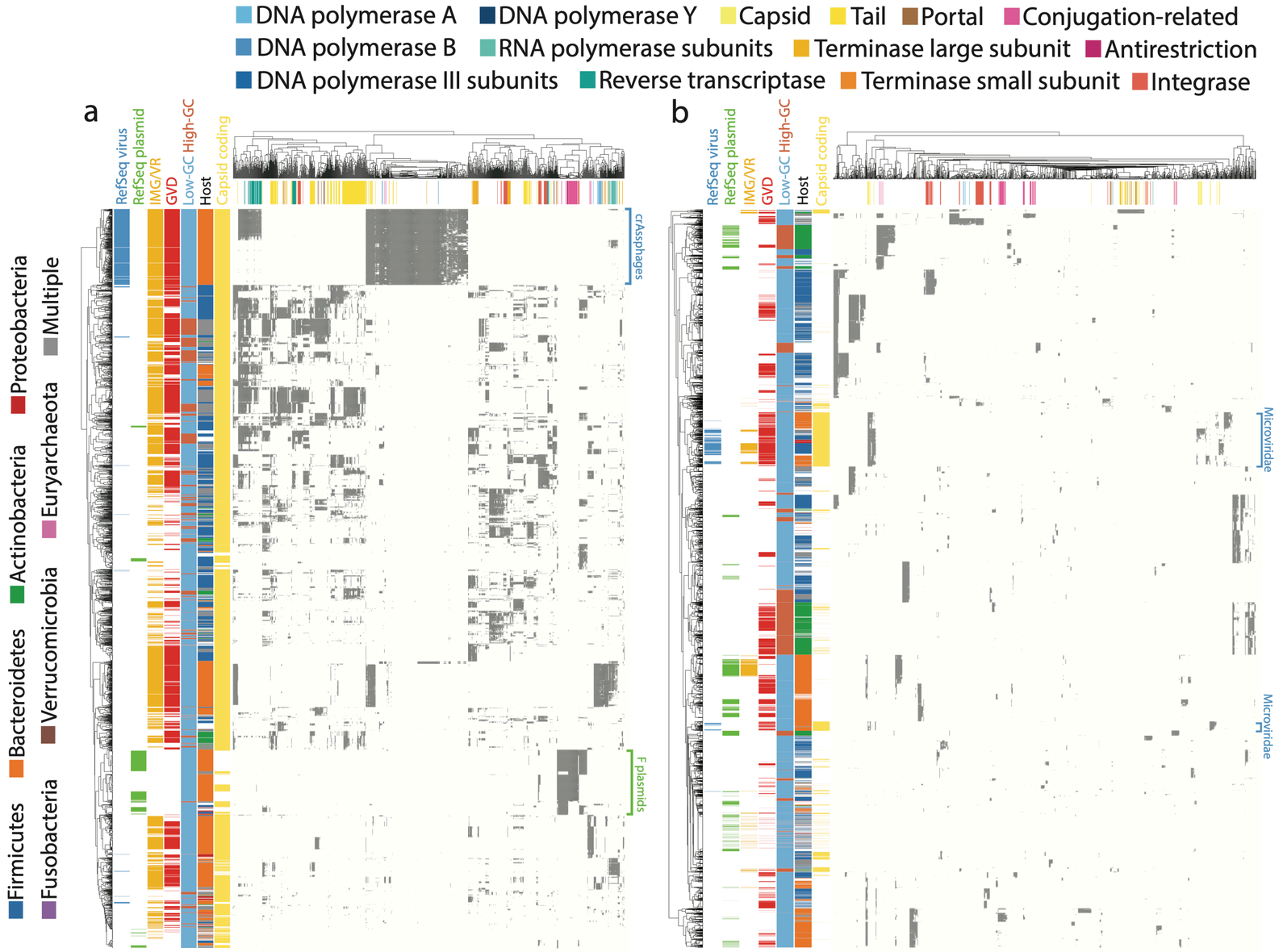
Molecular phylogeny of *Microviridae* major capsid protein



***Microviridae* species might have encountered cross-phyla host-switching**

Result 5

Gene contents based clustering of CRISPR targeted genomes

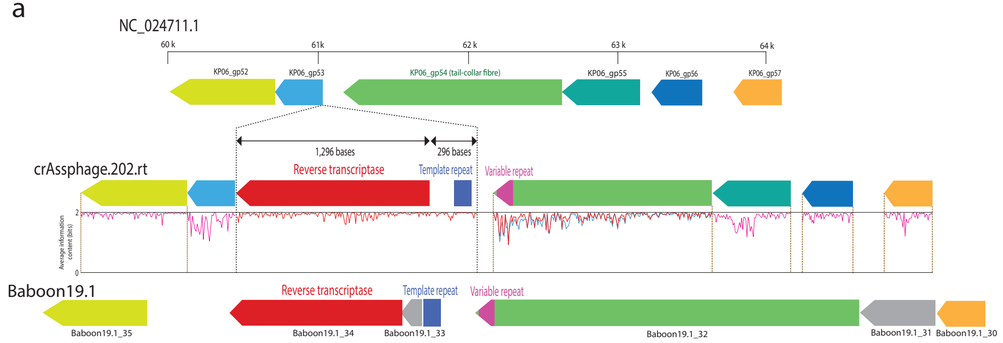


Most of large genomes were already in virus or plasmid databases
In contrast, majority of small genomes were novel

Result 6

DGR locus and phylogeny of discovered crAssphage genomes

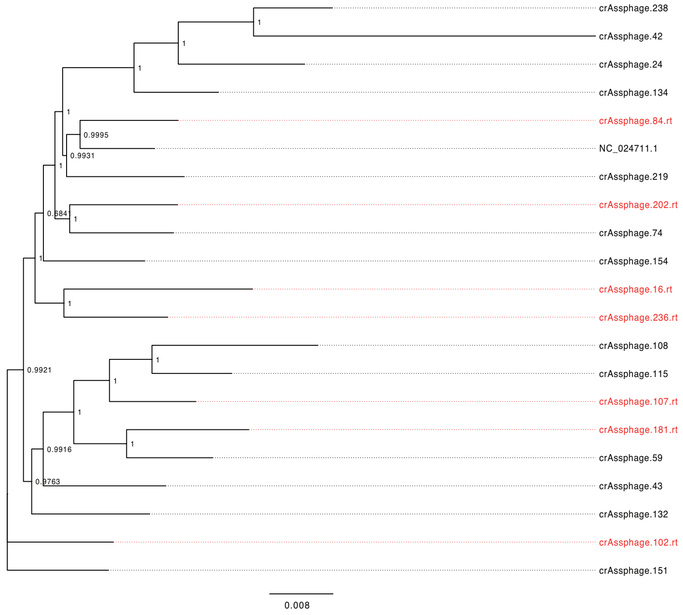
a



b

```
Baboon19.1:25837-25939 TAATAGTAGTCAGACTGACAATGGGTCTCGGGTTAGTCTTCTTTATCTTAATGCTAATAATGCTGTTAGTAACGCTAAGTCTAATAATGGGTTTACTAAAATG
                       |..|.||.|||..|||||.|.|||||||||||||||.|||||||||.||||..||||..|..|||||||.|||.||||..||||.||..|||||||.||||.||
Baboon19.1:25486-25588 TGTTGGTGGTCGTACTGTCGATGGGTCTCGGGTTGGTCTTCTTTTTCTTGCTGCTTCTGGTGCTGTTGGTAGCGCTGGGTCTGATGGTGGGTTTGCTAAGGTG
```

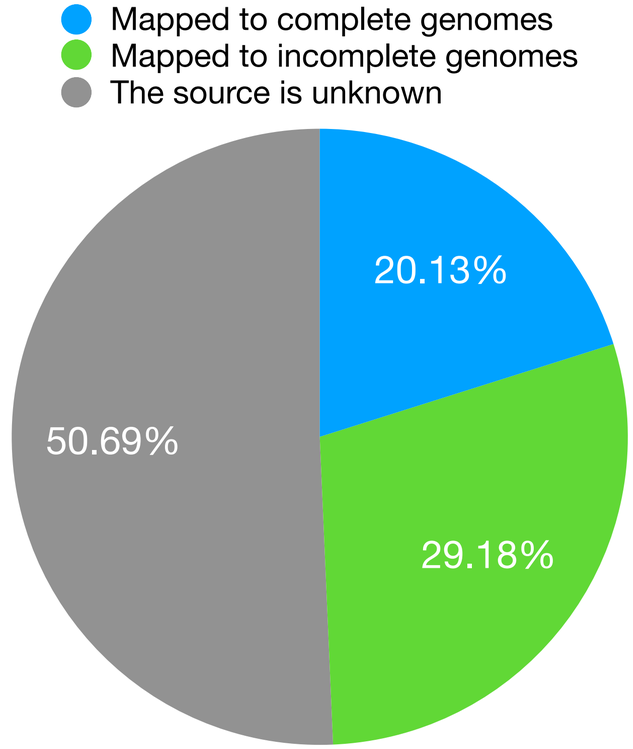
c



crAssphage DGR is orthologous, however recently lost multiple times

Result 7

The contribution of discovered genomes to CRISPR spacers



Majority of the source of CRISPR spacers still not investigated