Detecting viral genomes by using CRISPR adaptive immunological memory

Ryota Sugimoto¹, Nguyen Thanh Phuong¹,², Ruka Nishimura¹,², Ituro Inoue¹
1. National Institute of Genetics, 2. Sokendai University

Introduction
- About $10^{31}$ viruses exist on Earth
- They involved the evolution of cellular organism via horizontal gene transfer

Reference phage genome is limited
- 2292 complete phage genomes are recorded in NCBI

To understand the evolution of viruses, we need more sequences

Problem
How do we discover phage genome sequences from metagenome?

Challenges
1. Phage genomes are highly diverse
2. There is no marker genes
3. Reference sequence is limited

We utilize CRISPR, prokaryotic adaptive immunological memory.

Result
Total 463,248 unique CRISPR spacers
133,358 (28.8%) protospacers

Discovered 298 novel viral genomes

Features
- Genome length: 349,676
- Number of ORFs: 364
- Discovered from 3 independent samples
- 8~12 protospacers
- The genome is positively circular
- Predicted host is Clostridiales order, Clostridium sp. AM09-51 (ANI=98.78%)

Notable genes
- DNA polymerase III alpha and beta subunits
- N6 adenine methyltransferase
- tRNA and tmRNA
- Capsids
- TypeI Cas genes and CRISPR

Method
- Analyzed 511 human gut metagenome (9.8Tbp)
- Collected spacers
- Pool of spacers

Distribution of the genome length

A case-study: CRISPR-Cas encoding large phage

Next
- Detection of non-terminally redundant viral genomes
- Detection of RNA phages
- Construction of a database for viral genomes and protein families