# Development of human endogenous retrovirus database, and comprehensive discovery of CRISPR-targeted mobile genetic elements in the human gut metagenome

**Sugimoto, Ryota**

Doctor of Philosophy

Department of Genetics

School of Life Science

The Graduate University for Advanced Studies, SOKENDAI

# Abstract

Viruses are the most numerous biological entity, existing in all environments and infecting all cellular organisms. Compared with cellular life, the evolution and origin of viruses are poorly understood; viruses are enormously diverse, and most lack sequence similarity to cellular genes. To uncover viral sequences without relying on either reference viral sequences from databases or marker genes that characterize specific viral taxa, we developed an analysis pipeline for virus inference based on clustered regularly interspaced short palindromic repeats (CRISPR). CRISPR is a prokaryotic nucleic acid restriction system that stores the memory of previous exposure. Our protocol can infer CRISPR-targeted sequences, including viruses, plasmids, and previously uncharacterized elements, and predict their hosts using unassembled short-read metagenomic sequencing data. By analyzing human gut metagenomic data, we extracted 11,391 terminally redundant CRISPR-targeted sequences, which are likely complete circular genomes. The sequences included 2,154 tailed-phage genomes, together with 257 complete crAssphage genomes, 11 genomes larger than 200 kilobases, 766 genomes of Microviridae species, 56 genomes of Inoviridae species, and 95 previously uncharacterized circular small genomes that have no reliably predicted protein-coding gene. We predicted the host(s) of approximately 70% of the discovered genomes at the taxonomic level of phylum by linking protospacers to taxonomically assigned CRISPR direct repeats. We also investigated CRISPR-targeted RNA sequences. Notably, we found that the Picobirnaviridae species are targeted by CRISPR. The phylogenetic analysis indicated that this viral lineage is evolving rapidly, suggesting that this virus might be escaping from the CRISPR targeting.

# Table of contents

# List of Figures

# List of Tables

# List of Terms

| | |
|---|---|
| CRISPR | Clustered regularly interspaced short palindromic repeats |
| DR | Direct repeats |
| TR | Terminally redundant |
| MCP | Major capsid protein |
| RdRP | RNA dependent RNA polymerase |

# Acknowledgments

I appreciate the members of the National Institute of Genetics (NIG) Human Genetics Lab for providing me with invaluable scientific discussions, friendship, and mentorship. Professor Ituro Inoue taught me scientific thinking, writing, genetics, and the resolution to become a researcher. The co-workers Luca Nishimura and Phuong Thanh Nguyen assisted me with the computational analysis. All other members of the lab provided me with scientific discussions. Professor Hironori Niki, Associate professor Hiroshi Mori, Professor Yasukazu Nakamura, Professor Shigehiro Kuraku, Professor Ken Kurokawa, and Professor emeritus Yasushi Hiromi from NIG provided me with scientific discussions. Doctor Nicholas Parrish from RIKEN provided me with scientific discussions and helped me with writing. Assistant professor Jumpei Ito from Tokyo university provided me with scientific discussions. Doctor Hirofumi Nakaoka from Sasaki Institute provided me with scientific discussions.

# Chapter 1

# Development of herv-tfbs.com; a database for human endogenous retroviruses.

## 1.1 Introduction

My first scientific work was the development of a human endogenous retrovirus database [1]. Under the guidance of former Ph.D. student Ito, we developed a web service that provides information and statistics about the endogenous retroviral elements found in the human genome. This database was named "herv-tfbs.com" and was designed to be interactive and responsive (Figure 1-1).

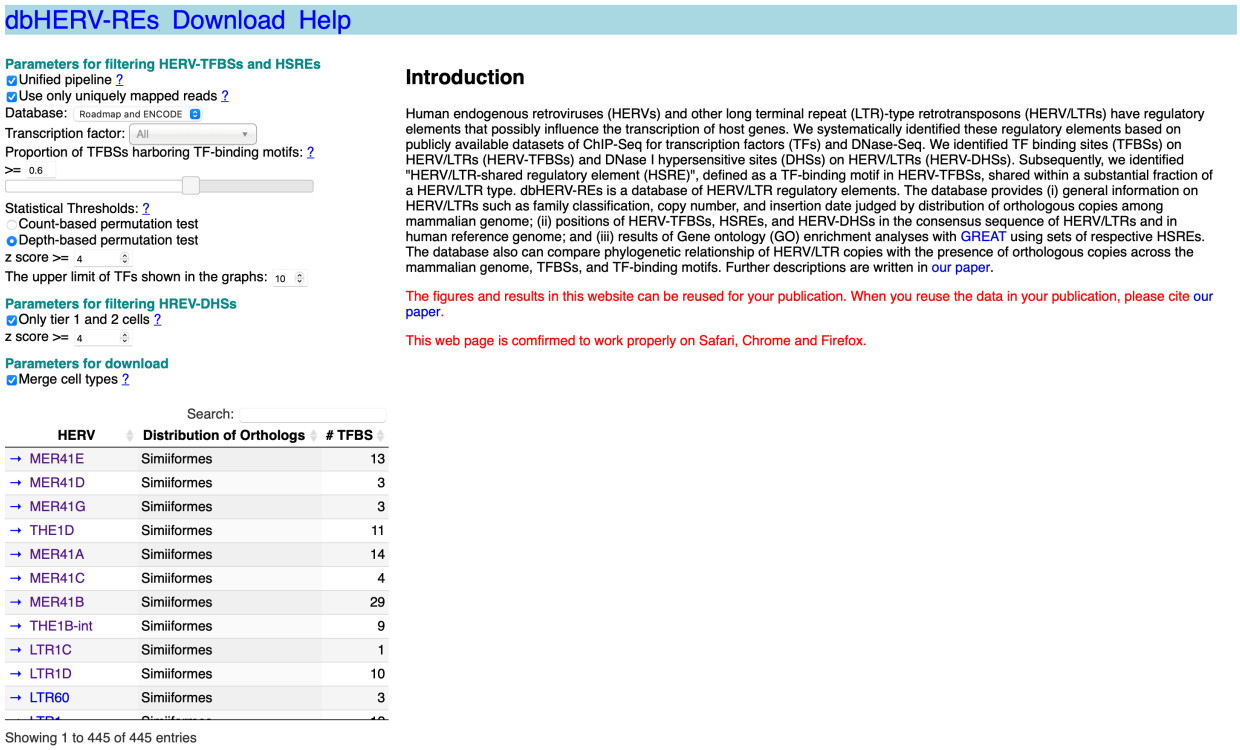



**Figure 1-1. The screenshot of herv-tfbs.com.** The web page can be accessed from regular browsers with the URL herv-tfbs.com.

The server side of this service was implemented with a MySQL relational database server and a Twisted web server running on Amazon Web Service (AWS). The browser side was implemented with jQuery and Plotly written with JavaScript language. The relational tables in the MySQL server were implemented by Jumpei, while web-server scripts and browser scripts were implemented by the author. The service was designed to minimize loading waits by using JavaScript dynamic query and HTML update mechanisms; users are allowed to interact with data and charts without changing or reloading a page. This database has been accessed from numerous countries around the world, suggesting our database implementation was highly successful.

# 1.2 Results

## The access count of herv-tfbs.com is increasing

The database was published in December 2016. Unfortunately, we lost most of the log files before February 2020. After March 2020, herv-tfbs.com has been accessed from 34,825 unique IP addresses. The total access count from unique IP addresses was 77,678. The access count per month is still growing (Figure 1-1). The bulk dataset was downloaded 192 times from 141unique IP addresses.

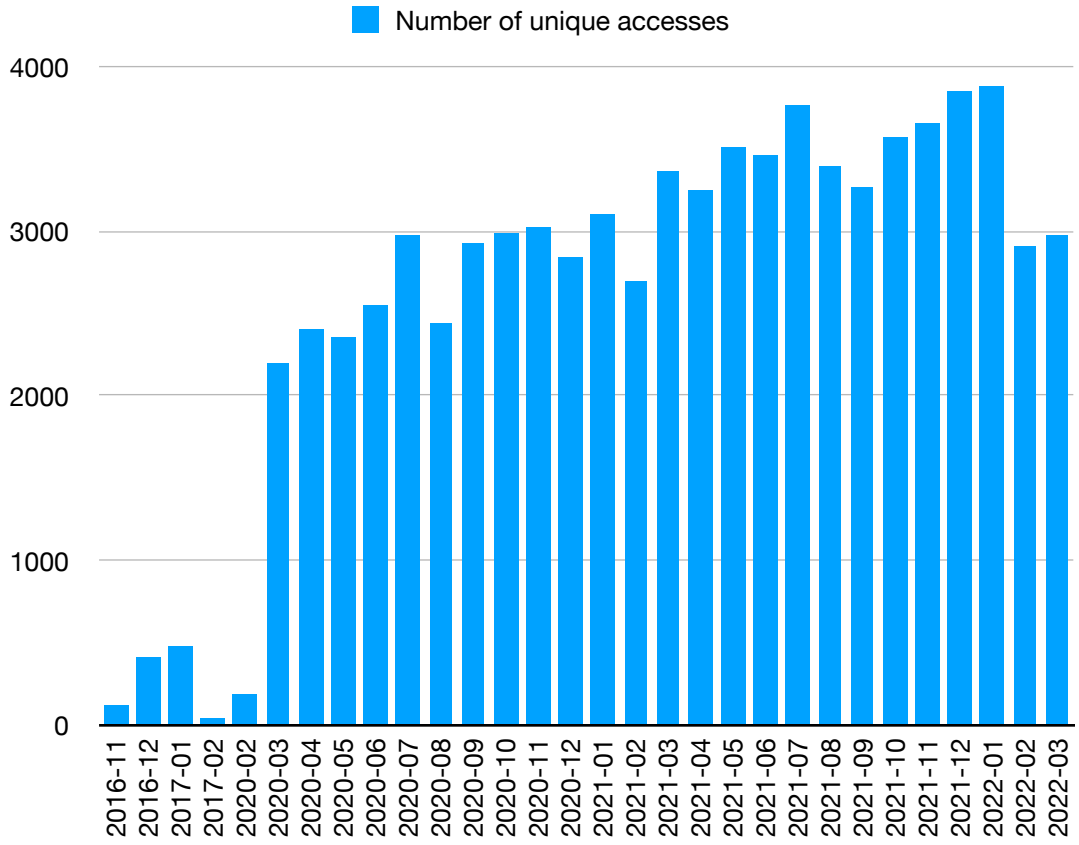



**Figure 1-2. Monthly unique access counts of herv-tfbs.com.** We counted the number of unique accesses based on IP address each day and summed them per month. Access from the AWS

maintenance bot was removed from the count. Unfortunately, the access logs between February 2017 and February 2020 were lost.

The median visit count was one, and the mean visit count was 2.23. 27,166 unique IP addresses visited the database only once. 942 unique IP addresses visited the database more than 10 times, and 481 unique IP addresses visited more than 20 times, suggesting there are several hundred “core” users who frequently visit the database.

## Herv-tfbs.com is globally accessed

The herv-tfbs.com was accessed from 164 countries. The most frequent access was from the United States followed by China and Russia (Figure 1-2).

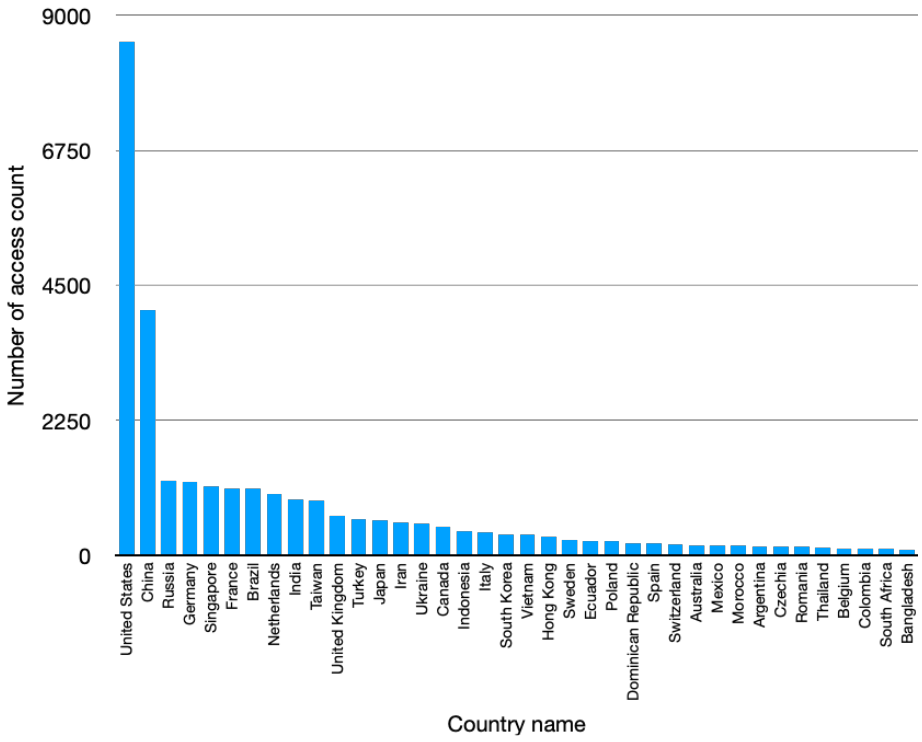


**Figure 1-3. Access counts from each country.** Again, we counted the number of unique accesses based on IP address each day. The countries were derived from IP addresses using the web service ip-api.com.

# 1.2 Conclusion

The herv-tfbs.com has been accessed globally and still increasing its access frequency. While the total access count is still increasing, the majority of users visited the database only once. Based on the visit count and the count of bulk downloads, there seem to be around two to five hundred core users exist. These users likely work in the field of virology and human genetics and found our database could be useful for their research. Overall, we conclude that the database attracted the attention of scientists and provides information valuable to them. We continue listening to the feedback to improve the database in the future.

# Chapter 2

# Comprehensive Discovery of CRISPR-targeted Terminally Redundant DNA Sequences in the Human Gut Metagenomes

## 2.1 Introduction

Viruses are the most abundant and diverse genetic entities on Earth [2]. Uncovering such enormous diversity is the key to understanding the origins and evolution of viruses and is now considered an important milestone in the virology [3, 4]. To collect diverse viral genomes, metagenomics has been utilized in numerous studies. From a given environmental sample, nucleic acids are extracted and sequenced. These sequences presumably include all genetic information of cellular and non-cellular entities existing in the sample. Such collections of genetic information are called metagenomes [5]. From them, one can computationally filter and extract genetic information of viruses that could not be sequenced otherwise, unless there is an established cultivation system which is very difficult or impossible in most cases. Such approaches often referred to as “viral mining”, have been applied to diverse environmental samples such as soil, seawater, and human feces, and led to extract tens of thousands of novel viral genomes. These notable works include

major discoveries such as the identification of crAssphages; a most abundant and prevalent bacteriophage lineage found in human feces [6].

Viral mining is a computational method to extract viral genomic sequences from metagenomes. A commonly used criterion to filter viral genomes from the metagenomic sequences is the presence of the viral signature genes. These genes include various types of capsids, RNA-dependent RNA polymerase, and DNA packaging genes [7–9]. From a given metagenomic contig, predicted amino acid sequences are searched to reference viral protein databases and if a contig is enriched with the viral signatures genes and without cellular genes such as ribosomal RNAs, one could argue that the contig is positively viral. This implementation has proven highly successful by increasing the known viral genome diversity by orders of magnitudes throughout the multiple studies [9, 10]. However, such a protein homology-based method highly relies on the availability of well-annotated viral protein sequences in the database, and arguably reference viral genomes recorded in the database are highly biased. For example, there are currently 14,073 *Caudovirales* species, also known as tailed-phages, genomes are recorded in the NCBI database. In contrast, only 90 *Tubularvirales*, also known as filamentous phages, species genomes are recorded in the NCBI database. This discrepancy is even notable for RNA bacteriophages. Currently, there are only two RNA bacteriophage families; *Cystoviridae* and *Leviviridae*, which comprise 21 species genomes, are recorded in the NCBI database. Several surveys indicated that these recorded viral genomes are only a portion of the entire viral diversity on Earth. Therefore, the viral mining methods that relied on reference genomes might fail to capture entire viral diversity, especially least characterized viral lineages such as filamentous or RNA phages. Furthermore, even for the well-characterized viral clade such as *Caudovirales*, it was proven that the detection of viral signature genes such as capsid can be challenging due to their extreme sequence diversity. For

example, the identification of the capsid gene of crAssphage has taken 4 years from the first publication until its experimental confirmation [11, 12].

Thus, we sought to develop a nonreference-based analytical pipeline to detect viral sequences from metagenomes; however, we were tasked with the question "What information could be used for this purpose?" The underlying biology of the clustered regularly interspaced short palindromic repeats (CRISPR) system, a prokaryotic form of adaptive immunological memory [13], provides a potential resource in this context. After viral infection or horizontal plasmid transfer, some archaeal and bacterial cells incorporate fragments of "nonself" genetic materials in specialized genomic loci between CRISPR direct repeats (DRs). The incorporated sequences, called "spacers," are identical to part of the previously infecting mobile genetic element. Thus, the genetic information encoded in CRISPR spacers can be inferred as likely viral and distinguishable from the genetic material of the organism encoding CRISPR, which is most often cellular, but potentially also viral [8, 14].

CRISPR spacers have been used to detect viral genomes [15–19] and predict viral hosts [20, 21]. They have previously been extracted from assembled bacterial genomes to assess CRISPR "dark matter", revealing that 80%–90% of identifiable material matches known viral genomes [18, 19]. In the current study, we extended this conceptual approach to the enormous amount of unassembled short-read metagenomic data. CRISPR repeats are relatively easily identifiable, particularly compared with unknown viral sequences. This trait allows the search of massive metagenomic datasets for reads comprised in part as CRISPR DR sequences; in turn, unknown sequences inferred as CRISPR spacers can be extracted directly from the raw reads [22, 23].

By analyzing human gut metagenome reads and the contigs assembled from them, we successfully extracted 11,391 terminally redundant (TR) CRISPR-targeted sequences ranging from 894 to 292,414 bases. These sequences are expected to be complete or near-complete circular genomes that can be linked to their CRISPR-targeting hosts. The discovered sequences include 2,154 tailed-phage genomes, together with 257 complete crAssphage genomes [6, 11], 11 genomes larger than 200 kilobases (kb), 766 Microviridae genomes, 56 Inoviridae genomes, 5,658 plasmid-like genomes, and 2,757 uncharacterized genomes. Although the majority of the discovered sequences that were larger than 20 kb were mostly characterized by viral or plasmid genomes, a substantial portion of sequences smaller than 20 kb was not recorded in either plasmid or viral databases. Furthermore, some previously uncharacterized small genomes had notably low coding ratios, which indicates that these elements might have unknown non-coding genetic features. These results demonstrate that our pipeline can discover CRISPR-targeted mobile genetic elements (MGEs) either previously characterized or uncharacterized.

# 2.2 Results

## Extraction of CRISPR-targeted sequences

We analyzed human gut metagenomes, as they serve as an "ecosystem" with the most abundant metagenomic data available. We downloaded 11,817 human gut metagenome datasets equivalent to 50.7 Tb from the European Nucleotide Archive FTP server. FASTQ files were preprocessed and assembled to 180,068,349 contigs comprising 767.7 Gb of data (Supplementary Table 2-1). We discovered 11,223 unique CRISPR DRs from the assembled contigs that were used to extract CRISPR spacers from raw reads, resulting in 1,969,721 unique CRISPR spacers (Supplementary

Figure 2-1 and Supplementary Data 2-1). These spacers were then used as queries to identify candidate protospacers (i.e., contigs containing the spacer sequence, not within a CRISPR locus). Spacers were mapped to CRISPR masked contigs using a minimum sequence identity threshold of 93%. We chose this identity threshold to capture the escaped mutants, i.e., viruses that escaped CRISPR targeting by introducing mutations to the protospacer loci. To increase specificity, we verified that the 5′- and 3′-adjacent sequences of spacer-mapped positions were not similar to each other or the spacer-associated DR. A total of 164,590,387 candidate protospacer loci, attributed to 1,114,947 unique spacers (56.6% of all unique spacers), were identified (Supplementary Data 2-1). This is a substantially higher discovery rate than that reported previously (~7% [18]) in a study that used National Center for Biotechnology Information (NCBI) nucleotide sequences for protospacer discovery. Although the genuine protospacers from a viral genome are expected to be colocalized in a relatively small region, the false-positive protospacers are expected to be scattered across the metagenome contigs randomly. To further reduce the false-positive hits, spacers were clustered based on protospacer co-occurrence and used to extract contigs targeted by more than 30% of members of a spacer cluster (Supplementary Figure 2-2). This process effectively removes false-positive protospacers that are randomly distributed across the assembled contigs. Finally, 764,883 gapless CRISPR-targeted sequences (15.9 Gb) were extracted. Among them, 11,391 unique sequences were identified as TR [24]; we expected that they were initially complete or near-complete circular MGEs. The size of the CRISPR-targeted TR sequences ranged from 894 to 292,414 bases (Supplementary Table 2-2 and Supplementary Data 2-1).

We then investigated protein-coding genes encoded in CRISPR-targeted sequences and identified 240,369 protein-coding genes among all unique TR sequences. Protein sequences were clustered based on a 30% sequence identity threshold, resulting in 31,204 clusters. Each

representative sequence was used as a query for three jackhmmer iterations, to build Hidden Markov models (HMM), which were then used to search the Protein Data Bank (PDB) [25]. Finally, 10,641 representative sequences, including 110,386 predicted protein sequences, were annotated (HHsearch probability > 80 and E-value < 1e–3) (S1 Data).

## Classification of TR sequences

The evaluation of TR sequence length revealed a multimodal distribution with a distinct trough at 20 kb (Figure 2-1A). For reference, we termed the 8837 TR sequences that were shorter than 20 kb as "small" and the 2554 TR sequences that were longer than 20 kb as "large." This simple classification was previously used to infer capsid morphology [26, 27]. Among the large TR sequences, 2047 (80.1%) encoded HK97 fold capsid proteins, a definitive gene of *Duplodnaviria* [28]. Phage portal proteins were encoded by 2163 large TR sequences (84.7%), indicating that most large TR sequences are from *Caudovirales*, also known as tailed phages (Figure 2-1B). Among the small TR sequences, 766 (8.7%) encoded *Microviridae* major capsid proteins (MCPs) [29], and 56 (0.6%) encoded *Inoviridae* major coat proteins. We propose that this portion of small TR sequences are likely viruses with a non-tailed morphology (Figure 2-1B) [30, 31]. Finally, 107 (1.2%) small TR sequences encoded HK97 fold capsid proteins. We also sought to identify any TR sequences encoding vertical jelly roll fold (vJR) capsids, a definitive gene of *Varidnaviria* [32]; however, we failed to find a significant hit using our search criteria.

We considered that a fraction of the TR sequences that lacked detectable capsid genes included plasmids. Therefore, we examined whether these TR sequences encode proteins that are characteristic of plasmids. We identified 386 large TR sequences (15.1%) and 957 small TR sequences (10.8%) encoding plasmid partitioning proteins A, B, or M. Furthermore, 187 large

(7.3%) and 4554 small (51.5%) TR sequences encoded MoBM relaxase, a protein that is required for initiating conjugation. Thus, 391 large (15.3%) and 5267 small (59.6%) TR sequences are likely plasmids or have life cycles similar to that of plasmids.

To scrutinize other genomic features, such as repeats and noncoding regions, the kmer singleton coverage and coding ratio for each classified and unclassified TR sequence were investigated (Figure 2-1C). Singleton coverage is the number of k-mer singletons from a given contig divided by its length; the value approaches 1 if the sequence does not contain repeats. For the large TR sequences, both classified and unclassified sequences had singleton coverages and coding ratios close to 1, indicating that these sequences are densely occupied by protein-coding genes and have few repeats. Conversely, plasmid-like and unclassified small TR sequences had a wider distribution of singleton coverages and coding ratios, indicating that some of these sequences may have a large proportion of noncoding regions and repeats.

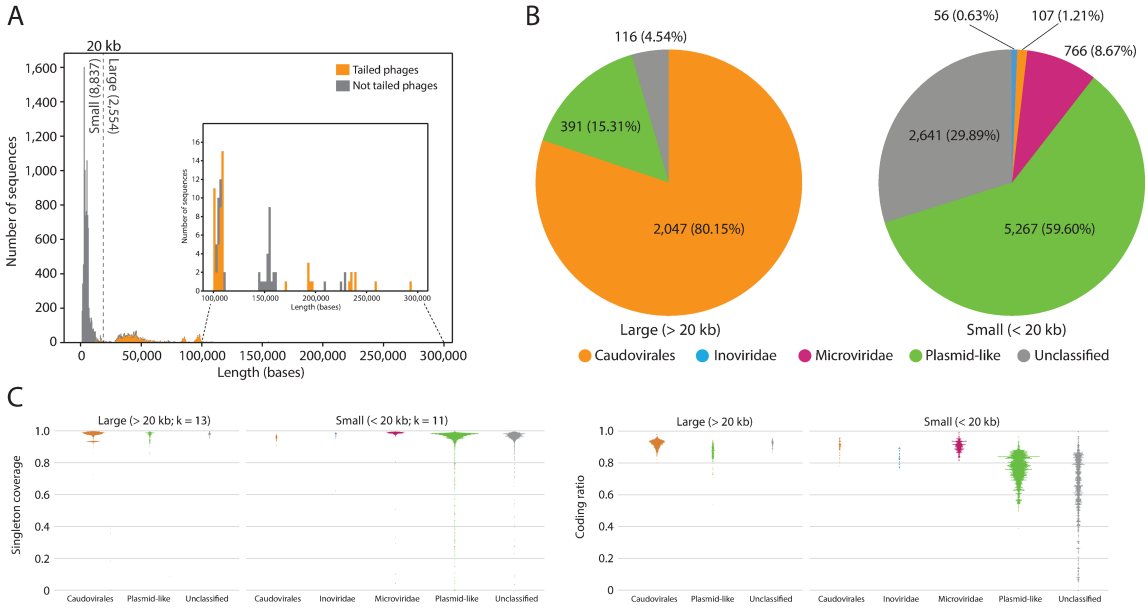



**Figure 2-1. Classification and genetic features of CRISPR-targeted TR sequences. (A) Length** distribution of TR sequences. We used the HK97 capsid and portal proteins as tailed-

phage signature genes. The dotted line at 20 kb represents an arbitrary cut-off between small and large sequences. Sequences longer than 100 kb are shown in the inset. **(B)** Results of the classification of TR sequences. Sequences encoding a detectable capsid gene were classified as a viral taxon according to capsid type, as follows. *Caudovirales*: HK97 fold capsid; *Inoviridae*: *Inoviridae* MCP; and *Microviridae*: *Microviridae* MCP. The capsid-less TR sequences with ParA, ParB, ParM, and/or MoBM were classified as Plasmid-like. The remaining sequences were labeled as "Unclassified." **(C)** Distribution of singleton coverage and coding ratio. Selected k-values were higher in large TR sequences, to avoid doubletons by chance.

## Predicted CRISPR-targeting hosts of TR sequences

As our approach uses CRISPR spacers to extract CRISPR-targeted sequences, we hypothesized that the relationship between a virus and the targeted host could be resolved for the majority of TR sequences. CRISPR DR sequences were searched on RefSeq genomes and taxonomically assigned. CRISPR DRs are shared between distant species through horizontal gene transfer (HGT) [33]. To prevent misassignment of targeting host because of shared DRs, we did not taxonomically assign DRs that were shared between different lineages in a given taxonomic level, and those DRs did not contribute to the targeting host prediction. Based on counts of protospacers associated with taxonomically assigned DRs, 7937 TR sequences (69.7%) were resolved to a targeting host at the phylum level (Figure 2-2A, Supplementary Table 2-2, and Supplementary Figure 2-3), and 6083 TR sequences (53.4%) were resolved to a targeting host at the order level (Supplementary Figure 2-4 and Supplementary Table 2-2). The most frequent host was *Firmicutes*, followed by *Bacteroidetes* and *Actinobacteria*. Notably, these are the most common bacteria in the human intestine [34]. In addition, 1418 TR sequences (12.5%) had putative host ambiguity between multiple phyla. Although some of these TR sequences were associated exclusively with monoderm

or diderm phyla, there was exceptional targeting host ambiguity between *Firmicutes* and *Verrucomicrobia,* which crosses the monoderm–diderm boundary (Supplementary Figure 2-5).

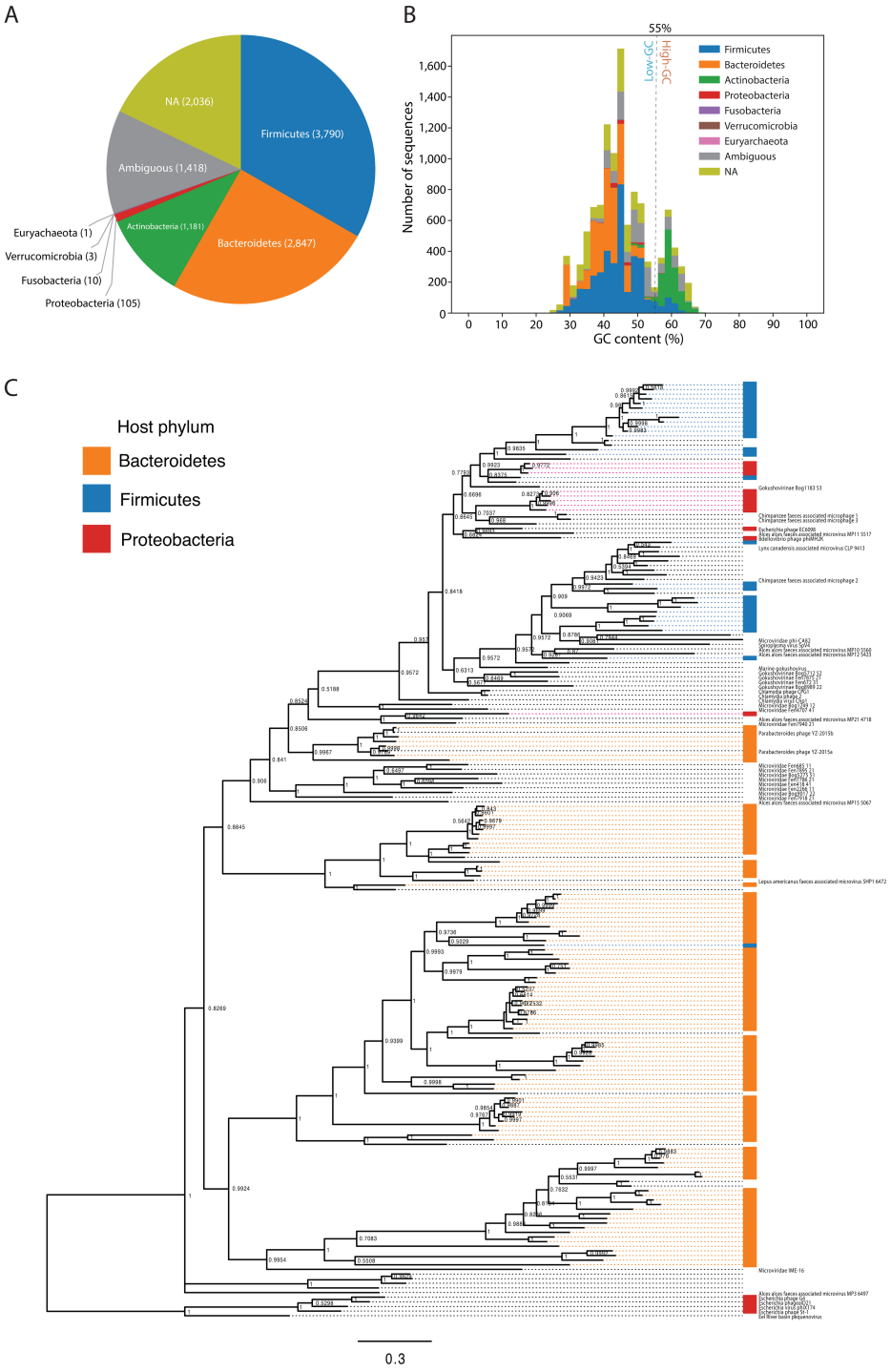


**Figure 2-2. Predicted targeting hosts of CRISPR-targeted TR sequences. (A)** The targeting host composition of TR sequences. Hosts were predicted by mapping CRISPR DR sequences to the RefSeq database. Sequences containing ≥10 protospacer loci but less than 90% associated DR taxa exclusiveness were classified as ambiguous targeting hosts. When ≥10 protospacers could not be assigned to a taxon, the predicted targeting host was denoted as not available (NA). **(B)** Predicted targeting host distribution according to GC content. The dotted line indicates the low- and high-GC content boundary, at 55%. **(C)** Bayesian phylogeny of *Microviridae* major capsid proteins. A total of 159 representative major capsid protein sequences from this study and 43 RefSeq sequences were used for analysis. Taxa without a name denote the *Microviridae* species from this study, and taxa with text denote *Microviridae* species from RefSeq. Taxa were annotated based on predicted targeting hosts. The phi X174 clade was selected as the outgroup.

## Firmicutes–Verrucomicrobia multiple infecting viruses are suspicious

These curious results prompted us to verify whether these sequences are indeed targeted to both *Firmicutes* and *Verrucomicrobia*. We focused on one of these sequences, dubbed amb-1 (note in S2 Table), which is 54,793 bases long and encodes an HK97 fold capsid and portal protein genes, suggesting it is likely a tailed phage. We discovered 676 protospacers in amb-1, 194 of which were associated with the DR sequence "CGTCGCACTCCGCAAGGAGTGCGTGGATTGAAAC" (DR1), whereas the remaining 422 were associated with the DR sequence "GTCGCTCTCCGCAAGGAGAGCGTGGATAGAAATG" (DR2). DR1 and DR2 pairwise alignment yielded only four mismatches (82.9% identity). Our criteria define this level of sequence identity as sufficiently low for separate taxonomic assignments. DR1 aligned perfectly to the *Clostridia* genomes NZ_PSQF01000044 and NZ_NFID01000002.1, and DR2 aligned perfectly to

*Akkermansia muciniphila* genomes. The DR sequence similarity could be explained by the HGT of the CRISPR–Cas system between these species. We found a phage portal and tail gene from the adjacent region of the DR2-aligned positions in CP027011.1, indicating that the CRISPR–Cas system associated with DR2 might have been recently introduced to the *Akkermansia muciniphila* genome by phage integration. The possibility of HGT limits the utility of the DR–host connection for inferring the actual targeting host of amb-1. We searched for signs of HGT between amb-1 and its genuine host, to complement the DR–host connection-based method. We used amb-1 to query the nr database and found that it partially aligned with the *Oscillibacter* genome AP023420.1 (query coverage of 6%, with 67.41% identity), but no significant hit against *Akkermansia muciniphila* was found. Thus, there is no substantial evidence to suggest that amb-1 is being targeted by multiple phyla. Although we focused on one TR sequence in this validation, the signs of phage integration and horizontal transfer of CRISPR–Cas between *Clostridia* and *Akkermansia muciniphila* suggest that the host–DR connection can not be used to infer genuine targeting hosts, particularly in between these species.

## Predicted targeting hosts above the taxonomic level of order are consistent

As we showed on an *ad hoc* basis, the CRISPR–Cas system undergoes frequent HGT between species. Because of unrecorded HGTs that could have occurred very recently, the DR-to-RefSeq alignment method might cause misinterpretations in the prediction of targeting hosts. Therefore, the TR sequences assigned to a single targeting host taxon still require further assessment. To complement DR-to-RefSeq–based host prediction, we used the tRNA genes [35] encoded in TR sequences. We found that 552 TR sequences encoded a total of 1124 tRNA genes. These tRNA gene sequences were searched for RefSeq bacterial genomes (95% minimum query coverage, with

95% minimum sequence identity); 288 tRNA gene sequences were aligned to 82 bacterial species genomes, connecting 97 TR sequences to potential hosts (S1 Data). The comparison of the tRNA-based predicted hosts with the DR-to-RefSeq–based predicted targeting hosts revealed a 93% agreement at the taxonomic level of order and more at higher taxonomic levels (Supplementary Figure 2-6A). The predicted host agreement dropped to 75% at the genus taxonomic level, suggesting that the DR-to-RefSeq–based method is less reliable for taxonomic levels lower than order.

During the review of this manuscript, a study of an enormous gut virome was published [36]. The authors of that study developed the Metagenomic Gut Virome (MGV) catalog, in which the viral genomes were assigned to predicted hosts based on CRISPR spacer alignment. Unlike our method, their CRISPR spacers were extracted from the taxonomically assigned metagenomic contigs recorded in the Unified Human Gastrointestinal Genome database. Thus, we considered that their predicted hosts could be used to assess our results. We found that 2180 TR sequences from our study were also recorded in MGV (S1 Data). For this sequence comparison, we used stricter criteria (85% query coverage, with 95% minimum sequence identity), to avoid host ambiguity between the relatively distant viral species. The predicted hosts were compared for each shared sequence between the two studies, and we found 95% agreement at the taxonomic level of order, and more at higher taxonomic levels (Supplementary Figure 2-6B).

## *Actinobacteria* is the corresponding targeting host of high-GC-content TR sequences

We calculated the GC content [37] of the TR sequences, to determine whether these correspond to the GC content of the targeting host predicted by the DR–host connection-based method (Figure

2-2B). Among the 2050 high-GC (GC% >55%) TR sequences, 1109 (54%) and 249 (12.1%) were predicted to be targeted by *Actinobacteria* and *Firmicutes*, respectively, and the targeting host was undetermined for 222 TR sequences (10.8%). The large fraction of high-GC TR sequences that were predicted to be targeted by *Actinobacteria* likely indicates the genomic adaption of parasitic genetic elements that infect and routinely become targeted by the CRISPR systems of high-GC Gram-positive bacteria (e.g., *Actinobacteria*).

## *Microviridae* species encountered a cross-phylum host-switching event

Host range within a viral lineage was further assessed by phylogenetic analysis of the TR sequences that were determined to represent *Microviridae* species. The predicted targeting hosts of putative *Microviridae* species from our study were *Bacteroides*, *Firmicutes*, and *Proteobacteria* (S2 Table), indicating a broad host range for *Microviridae* species. The molecular phylogeny of the *Microviridae* major capsid protein segregated sequences based on their targeting host (Figure 2-2C). Interestingly, most of the known *Escherichia coli*-infecting viral species (such as phiX174) and other *Proteobacteria*-infecting species (such as phi MH2K [38]) were used as a reference in this phylogeny were split into two clades, and the clade containing the latter was within a clade of TR sequences targeted by *Firmicutes*. The nested phylogenetic structure of capsids encoded by TR sequences that were predicted to represent *Microviridae* species may indicate that host switching, a critically important topic in viral evolution [39, 40], occurred within the viral lineage.

## CRISPR-targeted noncoding elements

Intrigued by the lower coding ratio detected in some unclassified small TR sequences, we selected two representative sequences with a notably low coding ratio (<0.3) for further manual inspection.

These two sequences had distinguishable sequence similarity and GC content. For simplicity, we dubbed them circ-1 and circ-2 (note in S2 Table). Circ-1 represented 84 small TR sequences (S1 Data) with a GC content of 37.4%, a length of 1356 bases, and 23 unique protospacers (Supplementary Figure 2-7A). Circ-2 represented 11 small TR sequences (S1 Data) with a GC content of 62.6%, a length of 1872 bases, and nine unique protospacers (Supplementary Figure 2-8). The genome comparisons indicated that circ-1 and circ-2 were nearly complete sequences of circular or tandem genomes (Supplementary Figure 2-9). Next, we searched for a conserved gene among similar genomes. From circ-1 and its similar sequences, no consistently shared gene was predicted. Circ-2 and its similar sequences shared one coding gene that was 114 codons in length; however, the protein sequence showed no significant hit in the PDB database (an ORFan). Thus, circ-1 and circ-2 seem to be CRISPR-targeted DNA elements without a reliably detected or annotated coding gene.

The predicted targeting host of circ-1 was *Veillonella*, a common gut *Firmicutes*, based on the connection between the CRISPR DR and the associated protospacers. The DR sequence was aligned to the *Veillonella* genomes LR778174.1 and AP022321.1 (Supplementary Figure 2-7C). The DR-aligned loci encode Cas9, Cas1, and Cas2. Therefore, the spacers aligned to circ-1 are likely derived from genuine Class 2 CRISPR–Cas systems [33, 41] encoded in the *Veillonella* genomes. The protospacer adjacent motif (PAM) was TTTN (Supplementary Figure 2-7B), as calculated from the adjacent sequences of protospacers on circ-1. Twelve protospacers on circ-1 were adjacent to this motif (Supplementary Figure 2-7A), indicating that the CRISPR–Cas system restricts this DNA element. The GC content of LR778174.1 was 38.8%, which was close to the circ-1 GC content. Moreover, circ-1 was not aligned to any bacterial genomes, including

*Veillonella*, indicating that this element is not encoded in a cellular genome. We concluded that circ-1 is likely an extrachromosomal element restricted by the *Veillonella* CRISPR–Cas system.

We could not identify the circ-2 targeting host because the DR sequence yielded no significant hits, and circ-2 itself also had no significant hits. However, the high-GC content of circ-2 indicates that its possible host is *Actinobacteria*.

## Comparison of CRISPR-targeted TR sequences with available viral and plasmid sequences

The TR sequences identified in this study were compared with sequences included in virus and plasmid genome databases, including RefSeq virus [42], RefSeq plasmid, IMG/VR [43, 44], and GVD [9] (Figure 2-3). IMG/VR is the largest database of uncultivated viral genomes. GVD is a database of viral genomes that were discovered from human gut metagenome datasets using VirSorter and VirFinder, both of which rely on protein homology, which complements our approach to the discovery of viral genomes. Therefore, we consider that these databases are appropriate to validate our results.

Among the 2554 large TR sequences identified here, we found that 1726 TR sequences (67.6%) were represented in RefSeq virus, IMG/VR, or GVD (Figure 2-3A) using a threshold of 85% sequence identity with at least a 75% aligned fraction of the query sequence to a unique subject sequence. These sequences included 257 crAssphage genomes ranging in size from 92,182 to 100,327 bases (average, 96,984.8 bases). Members of *Bacteroidales* were predicted as targeting hosts of crAssphages in this study, which is consistent with those hosts reported to propagate crAssphage in previous studies [6, 11]. We also found 154 large TR sequences (6%) listed in RefSeq plasmid, seven of which were also listed in the IMG/VR database. Thus, we determined

that there is an adequate classification between plasmids and viruses for large genomes. Notably, we discovered 11 TR sequences larger than 200 kb, two of which corresponded to recently reported "huge phage" genomes [8] (note in S2 Table); moreover, five of these very large TR sequences encoded HK97 fold capsid proteins. Finally, 681 large TR sequences (26.7%) did not yield database hits using our criteria. We concluded that many of the large TR sequences are already represented in virus or plasmid databases, except for those greater than 200 kb, which were only recently reported.

In contrast with the large TR sequences, most of the small TR sequences were not represented in the databases (Figure 2-3B). Among the 8837 small TR sequences identified here, 491 (5.6%) were listed in IMG/VR, whereas 2573 (29.1%) were listed in GVD. Only 256 small TR sequences (2.9%) were listed in both IMG/VR and GVD. Finally, 5518 small TR sequences (62.4%) yielded no database hits, indicating that the majority of small elements targeted by CRISPR remain unexplored, even in the intensively studied human gut metagenome. However, IMG/VR filters out genomes shorter than 5 kb, to minimize the rate of false-positive predictions [44], which likely explains the substantially lower representation of small TR sequences in IMG/VR.

Among the 942 RefSeq plasmid-listed sequences, 444 small TR sequences were listed in RefSeq virus, IMG/VR, and/or GVD, suggesting a possible misclassification of viruses and plasmids within these databases. We investigated the database representation of the CRISPR-targeted TR sequences predicted to represent *Microviridae* and *Inoviridae*. Among the 766 putative *Microviridae* genomes from this study, 639 genomes (83.4%) were represented in at least one viral database, and none were listed among RefSeq plasmids. In contrast, among the 56 putative *Inoviridae* genomes, none were listed in either viral or plasmid databases. To further

assess our putative *Inoviridae* genomes, we compared these genomes with recently reported *Inoviridae* genomes discovered using a machine-learning approach [45]. We found that 21 genomes from our study were highly similar to the genomes from the previous study, supporting our prediction that these sequences were indeed *Inoviridae* genomes. Finally, we compared the TR sequences against the Integrative and Conjugative Element (ICEberg) database and obtained nine hits. None of these sequences encoded a detectable capsid protein.

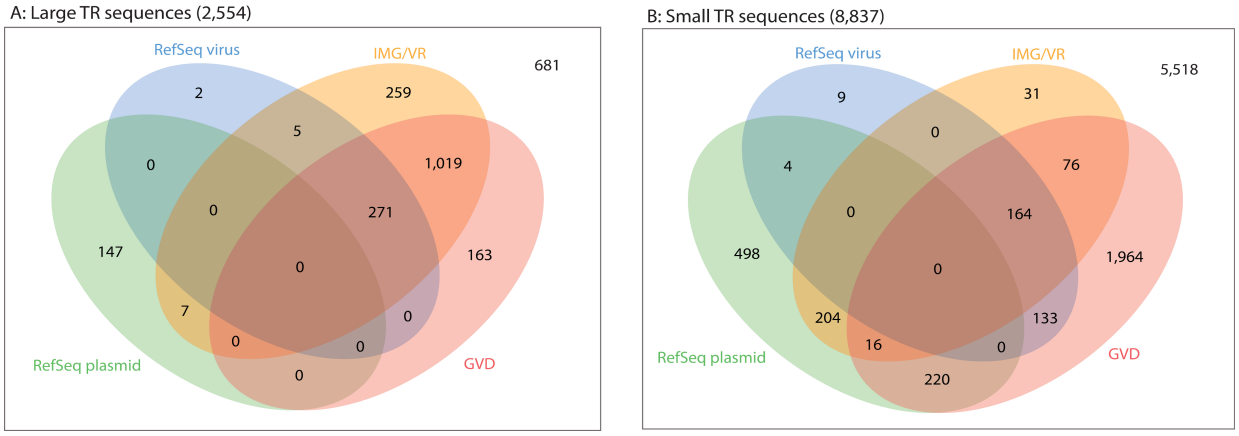



**Figure 2-3. Venn diagrams of database comparisons for (A) large and (B) small TR sequences.** Each TR sequence was compared to RefSeq virus, RefSeq plasmid, IMG/VR, and GVD using BLASTN. The database hit minimum criteria were set to 85% sequence identity with 75% aligned fraction of the query sequence to a unique subject sequence.

## Comparison with the prediction results of VirSorter

We compared our findings with the prediction results of the VirSorter program. The VirSorter program adopts a homology-based strategy to detect viral genomes; therefore, we considered that this program complemented our strategy. All unique TR sequences were fed into the VirSorter program using default parameters. The program predicted 730 *Microviridae* major capsid-protein-coding TR sequences (95.3%), 916 HK97 fold capsid-protein-coding TR sequences (42.5%), no

*Inoviridae* major coat-protein-coding TR sequences, and 109 TR sequences without a detectable capsid predicted to be positively viral (category ≤6). The VirSorter program had a good agreement with the predicted *Microviridae* species identified in our analysis. Conversely, the program predicted about half of HK97 capsid-protein-coding TR sequences and no *Inoviridae* MCP-coding sequences as being positively viral. Notably, among the 257 TR sequences with hits to the crAssphage reference genome, only 10 sequences were predicted to be positively viral (category 3 and 6).

## Classification of Inoviridae major coat-protein-encoding TR sequences

Among the discovered capsid/coat-protein-encoding TR sequences, the sequences encoding *Inoviridae* major coat proteins were notably unlisted in the databases; thus, we investigated these sequences regarding whether they contain a novel clade of the viral lineage. Among the 56 *Inoviridae* major coat-protein-encoding TR sequences, 54 encoded the Zonula occludens toxin (Zot). From the Zot-encoding sequences, we selected eight representative genomes, dubbed Ino-01 to Ino-08 (S1 Data and note in S1 Table), by clustering Zot amino acid sequences using a 50% sequence similarity threshold; the genus demarcation criteria for the *Inoviridae* family were as proposed by the International Committee on Taxonomy of Viruses (ICTV). The phylogenetic tree of the Zot domains formed a distinct clade, dubbed Clade-1, which contained six and only discovered genomes (Ino-01 to Ino-06) (Figure 2-4A). Two other representatives, Ino-07 and Ino-08, were placed in the clades with RefSeq-recorded genomes. The consensus-predicted targeting host class of Clade-1, except Ino-03 and Ino-06, was *Clostridiales*, and the consensus-predicted targeting host phylum of Clade-1, except Ino-03, was *Firmicutes*. Outside Clade-1, the predicted

targeting host phylum of Ino-07 was *Proteobacteria*, and the predicted targeting host class of Ino-08 was *Lactobacillales*.

The phylogeny of Zot domains suggests that Clade-1 is the most diversified Zot protein subfamily in the human gut metagenome, and the encoding genomes showed notable diversity as well. The genome lengths of Clade-1 ranged from 5,635 to 9,374 bases, and the GC content ranged from 27% to 41%. Some of the Clade-1 genomes showed nonconventional gene organizations (Figure 2-4B). The *Zot* genes of Ino-03 and Ino-04 were encoded between the major coat and the probable minor coat gene, which typically is the longest gene and is encoded right after the major coat gene. Ino-06 encoded the tyrosine recombinase after the replication initiator gene, suggesting that this virus uses site-specific recombination for viral genome integration into the host chromosome. Accordingly, Ino-06 was partially aligned (43% query coverage with 69.58% sequence identity) to a tRNA locus of *Lachnospira eligens* strain 2789STDY5834875 (accession: NZ_CZBU01000012.1). The aligned region was located at the 3′ end of the *tRNA-lys* gene; this observation demonstrated a previously reported integration mechanism of *Inoviridae* species [46]. Ino-01 encoded an endonuclease in the opposite strand to the essential genes. This endonuclease was partially homologous to homing-endonuclease; thus, we speculated that this virus might use an intron-like mechanism to integrate its genomes into the host chromosome. However, we failed to find a similar sequence to Ino-01 among the RefSeq bacterial genomes. Finally, Ino-04 encoded a HUH-endonuclease domain replication initiator that was non-homologous to the other Rep genes and uncommon in typical *Inoviridae* species [47].

Another notable feature of this study was that all representatives targeted by *Firmicutes* encoded a stand-alone Zot domain, whereas the Ino-07 *Zot* gene was fused to an

unknown domain (Figure 2-4B); a similar structure could be found in other *Proteobacteria*-infecting filamentous phages, such as *M13* and *If1*.

Recently, ICTV updated the taxonomy and definition of the filamentous phage clades. A taxonomy level previously called *Inoviridae* in *ssDNA virus* is now a family of the newly defined Tubulavirales order. This order currently includes three families: *Inoviridae*, *Paulinoviridae*, and *Plectroviridae*. These families were defined based on the analysis of the gene-sharing network of filamentous phage genomes, to represent the enormous HGTs between the closely related species within this order [45, 48]. To assign the discovered genomes to the known families based on the gene-sharing method, we used a classification program provided by ICTV [49]. Unexpectedly, the Clade-1 genomes were assigned to two families (Figure 2-4A); four *Paulinoviridae*, one *Inoviridae*, and one unclassified family. Outside Clade-1, Ino-07 was assigned to the *Inoviridae* family, which was consistent with the other species in the same clade; lastly, Ino-08 was unclassified.

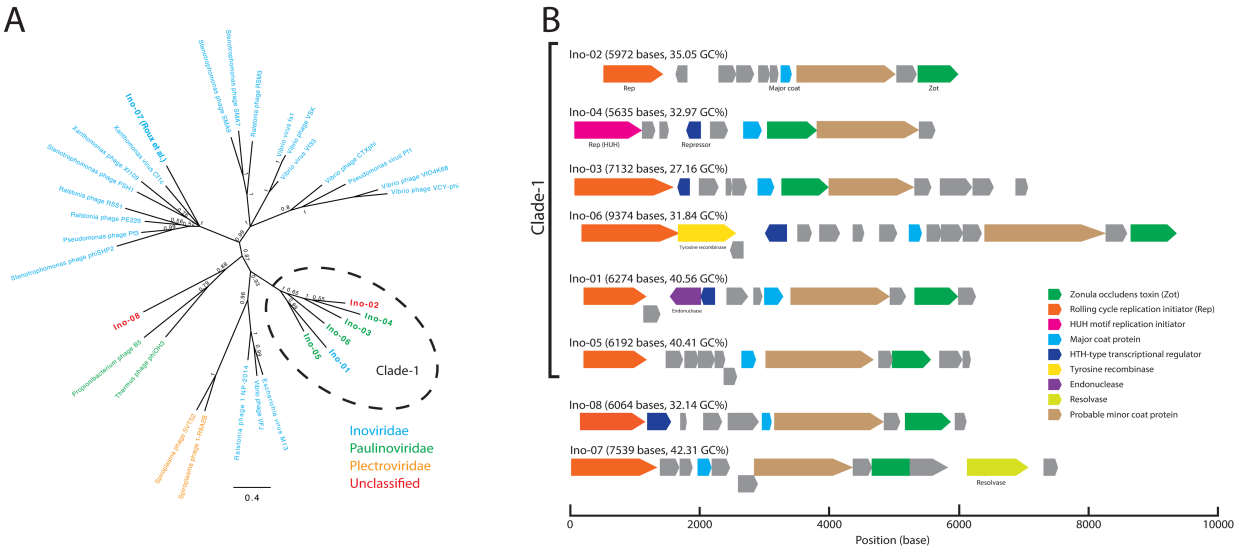



**Figure 2-4. Classifications and genome organizations of the discovered *Inoviridae* species. (A)** Bayesian phylogeny of Zot domains. Representatives were selected from RefSeq and the Inoviridae major coat-protein-encoding TR sequences by clustering Zot amino acid sequences

using a 50% identity threshold. Each taxon was colored according to its corresponding family. The families of the discovered genomes (Ino-01 to Ino-08) were predicted using the ICTV-provided taxonomic classification program. A sequence reported in a previous study (Ino-07) is denoted in parenthesis. **(B)** Genome organizations of the discovered Inoviridae species. All sequences were phased to align so that the Rep genes appear first. The predicted ORFs are colored according to the annotation results.

## Gene-content-based hierarchical clustering of CRISPR-targeted TR sequences

We hierarchically clustered TR sequences based on gene content to scrutinize CRISPR-targeted TR sequences in a genomic context. For this analysis, we selected the top 1000 genes that were recurrently observed among large and small TR sequences. The clustering results for large TR sequences (Figure 2-5A) showed that the majority of genomes, with a variety of gene contents, had already been listed in viral databases. This finding further supports the assertion that these databases contain at least representatives that are similar, at broad taxonomic levels, to the large viruses present in the human gut. Specific gene contents were observed for crAssphages, which Bacteroidetes target exclusively. In addition, RefSeq plasmid hit sequences formed an exclusive cluster with conjugation-related genes. As the conjugation proteins included pili formation and intercellular DNA transfer, these sequences are likely F plasmids. TR sequences were predicted to be targeted by monoderm and diderm hosts clustered separately, indicating little gene flow between them. Except for the likely F plasmid sequences, most of these sequences encoded HK97 fold capsids. Combined with the fact that most of the large TR sequences encoded portal proteins, this result further supports the conclusion that the majority of the large TR sequences are *Caudovirales*.

In contrast to the large TR sequences, we found that the small TR sequences remained largely enigmatic when clustered according to gene content (Figure 2-5B). Few clusters were represented in IMG/VR, and although GVD covers a relatively broader range, representatives were still missing or sparse for some clusters. Several clusters were also listed in both plasmid and viral databases. As many clusters did not encode detectable capsid genes, both clusters representing *Microviridae* evinced capsid genes. Another pattern shown by this analysis was that high-GC-content TR sequences were frequently observed with *Actinobacteria* as the predicted targeting host.

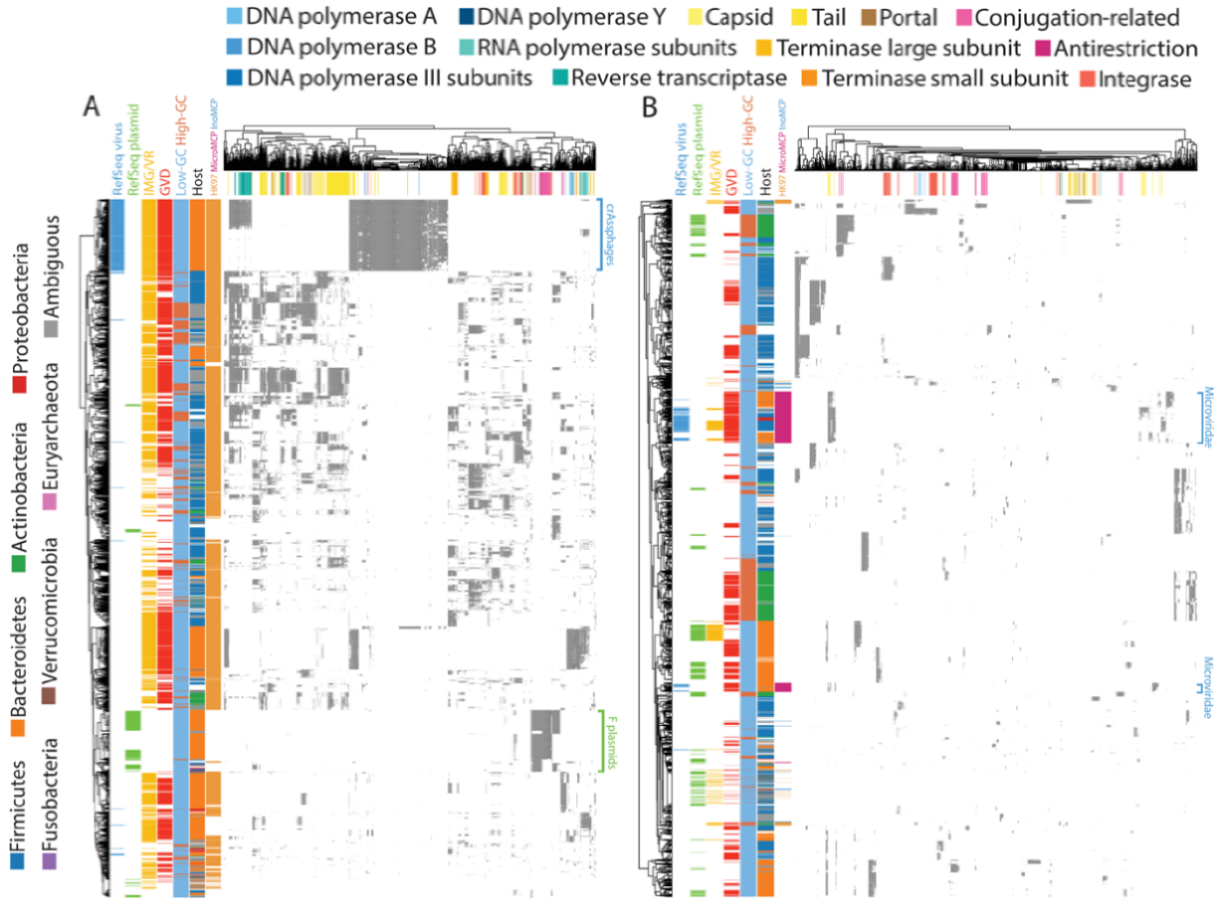



**Figure 2-5. Hierarchical clustering of (A) large and (B) small TR sequences based on gene content.** Heatmaps represent the gene content of TR sequences, in which each row is a TR sequence and each column is a gene cluster. The gray areas in the heatmap indicate sequences

encoding a gene that is homologous to the gene cluster. Note that one gene can be homologous to multiple gene clusters. Sequences are annotated by database containing similar sequences, GC content, host, and capsid genes. Capsid genes are colored differently according to their types, as indicated in the figure; HK97, Microviridae major capsid protein (MicroMCP), and Inoviridae major coat protein (InoMCP). Gene clusters were annotated by searching corresponding HMMs in the UniRef50 database. Several notable RefSeq-listed clusters are denoted on the right side of the heatmaps.

## Remnant CRISPR spacers and contribution of CRISPR-targeted sequences to the identified spacers

Viruses and other MGEs can escape CRISPR targeting by acquiring mutations in protospacer loci [50, 51]. Although the corresponding spacers are no longer effective, they can remain in the host genome. To investigate these potential "remnant" CRISPR spacers, we mapped all unique CRISPR spacers to TR sequences and scrambled sequences using various sequence identity thresholds (Figure 2-6). The scrambled sequences were used to monitor false-positive matches arising by chance (see Materials and Methods). Based on the observation of incremental false-positive matches of spacers to scrambled sequences, we considered that a sequence identity threshold of 84% is adequate for the mapping of some putative remnant spacers, with very few false-positive matches. At an 84% sequence identity threshold, 269,808 and 126,616 spacers were mapped to large and small TR sequences, respectively. Altogether, 20.1% of all unique spacers (396,424 spacers) were mapped to TR sequences. Compared with the identity threshold that was applied initially (93%), 91.9% more spacers were mapped to small TR sequences using the relaxed threshold, whereas only 42.7% more spacers were mapped to large TR sequences. These results suggest that a substantial fraction of CRISPR spacers imperfectly match protospacers within

circular MGEs in the human gut, potentially reflecting an “escape mutation” phenomenon. According to percentage, small TR sequences can be explained in this manner better than large TR sequences.

Although we focused on TR sequences because of the high confidence of the genomic completeness, spacers could be derived from incompletely reconstructed or noncircular genomes. To estimate the contribution of the discovered CRISPR-targeted sequences to the identified CRISPR spacers, all unique spacers were mapped to all representative CRISPR-targeted sequences using an 84% sequence identity threshold. Under these conditions, 971,224 spacers (49.3% of all unique spacers) were mapped.

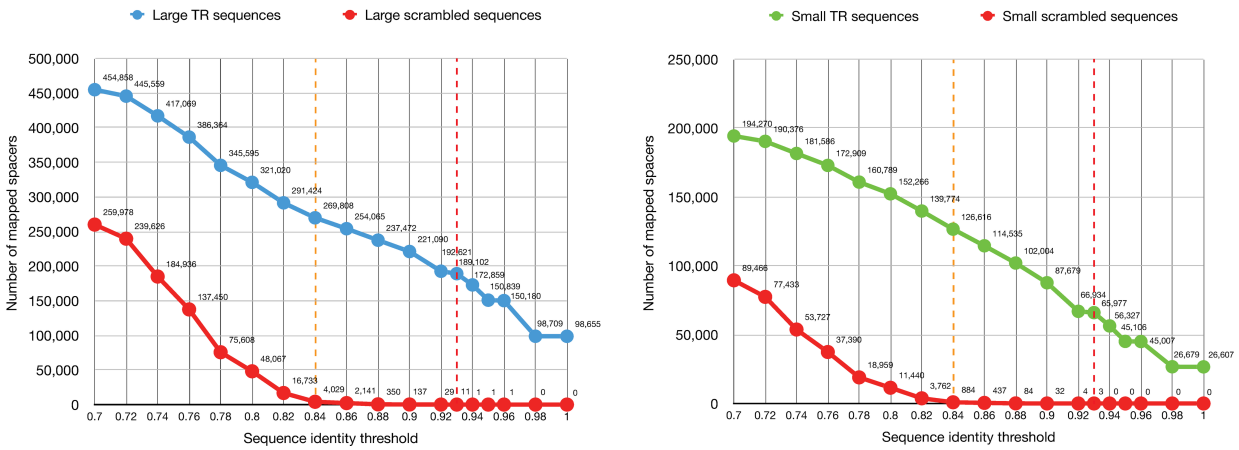



**Figure 2-6. Number of mapped spacers according to sequence identity threshold.** All unique CRISPR spacers were mapped to large TR sequences, small TR sequences, and scrambled sequences. The relaxed sequence identity thresholds applied initially are denoted as red- and orange-colored dashed lines. The spacer mapping process was identical to the protospacer discovery process (see Materials and Methods).

# 2.3 Discussion

By analyzing the vast size of metagenome sequences, we extracted a large amount of CRISPR-targeted presumably complete sequences of circular genomes. This analysis intended to discover a viral genome that could not be discovered using the conventional homology-based method. Although most of the discovered genomes with detectable capsid genes were previously recognized viral lineages, substantial portions of particularly small TR sequences remained unclassified (Figure 2-1B). The coding ratio of these unclassified sequences exhibited a broad distribution, and some were exceptionally low; thus, we speculated that these sequences might have unknown genetic features that differ from the conventional protein-coding genes. We selected two small genomes, circ-1 and circ-2, and inferred that one of them was likely extrachromosomal and targeted by the *Veillonella* species CRISPR–Cas system (Supplementary Figure 2-7). These presumable noncoding extrachromosomal DNA elements resemble satellites, which are DNA/RNA elements that replicate with the assistance of the host and/or other MGEs. The one that comes to mind first is the viroid, which is a plant pathogenic circular RNA element that lacks coding genes. However, we cannot infer or relate the discovered genome to known viroids in any biological or evolutionary means based on our current scarce knowledge. Recently reported "satellite plasmids" [52], a plasmid state that lacks autonomous replication genes, also share similarities with these genomes. However, satellite plasmids are an evolutionarily transient state that eventually are lost from the cell population. The genome length and sequence similarity of circ-1 and circ-2 were maintained across various samples, implying that these entire genomic sequences might have unknown functions. Another similar element is represented by circular noncoding RNAs (circRNAs) [53, 54]. However, the samples we analyzed in this study did not include RNA sequences, and circRNAs are expressed from cellular genomes, a finding that

conflicts with the fact that circ-1 was not aligned to any bacterial genomes, including the presumed host. The function, mobility, and potential pathogenicity of circ-1 and circ-2 remain entirely unknown at this point. Further experimental research and discovery of reassembling DNA/RNA elements are required.

A substantial number of the discovered sequences encoded the HK97 fold capsid, *Microviridae* major capsid, and *Inoviridae* major coat proteins, allowing us to validate that these portions of the discovered genomes were indeed viral. However, in this analysis, vJR capsids were suspiciously absent, even though vJR capsid-coding viruses, or *Varidnaviria*, are ubiquitous across many environments [55–57]. Currently, we cannot explain why they do not propagate in human gut common bacteria/archaea populations. To date, only two families and nine species belonging to *Varidnaviria* are known to infect bacteria. It is plausible that the lack of a reliable reference genome hampers the detection of vJR capsid genes in the current state. Considering that nearly half of the detected genes were not annotated with our pipeline, and about 30% of small TR sequences remain unclassified, these remaining sequences may encode vJR capsid genes that cannot be detected based on the currently limited known sequence diversity. A recent application applied a machine-learning approach to this problem and achieved a notable result [58]. The folding-based method could soon complement the sequence-similarity-based method to discover extraordinarily distant homologs.

The targeting host prediction results suggest that approximately 70% of the discovered sequences are targeted by specific host phyla (Figure 2-2A). Targeting host phyla was ambiguous for 12.5% of the TR sequences; however, most of the targeting host ambiguity observed between *Firmicutes* and *Verrucomicrobia* was suspicious because of the likely horizontal transfer of the CRISPR–Cas system between these species. There is still considerable ambiguity within

monoderm phyla. We are uncertain whether these elements infect multiple hosts at present or recently host switched, or whether some genomes became CRISPR-targeted because of abortive infections [59]. TR sequences assigned to a single host taxon were further assessed using a tRNA-based method and cross-study comparison, which yielded good agreement levels above the taxonomic level of order. The human gut metagenome has been intensively sequenced over the past decade, and the database likely captures most of the CRISPR–Cas loci known at present, thus allowing us to confidently predict the targeting hosts of the discovered genomes at higher taxonomic levels (above order). However, in a different environment, i.e., poorly sequenced, the DR-to-RefSeq–based method could lead to significant misinterpretations because of unrecorded HGTs of CRISPR–Cas systems. Therefore, multiple methods should be applied to infer the host of the discovered genomes from such environments. In addition, we state that the spacer-based host prediction method does not directly connect the spacer-aligned sequence to its currently infecting host. A spacer aligned to a sequence is a record of the infection history that occurred in the past, and viruses and MGEs possibly undergo HGTs and switch hosts. Expanding this approach to more diverse samples and observing the evolution of CRISPR loci, in particular, might allow the inference of the evolutionary history and genetic factors involved in viral/MGE host-switching events.

The disagreement of VirSorter prediction results for HK97-coding sequences could be explained by the high diversity of their gene contents. We found that, among the 257 TR sequences that were similar to the crAssphage reference genome, only 10 sequences were predicted to be positively viral. crAssphages are known to have a variety of gene sets in addition to the essential core gene set, even in a closely related lineage. This might complicate the prediction mechanism of the program, leading to the output of less-confident prediction results. This hypothesis also

explains the prediction results of *Microviridae* species. These viruses have small genomes that are densely occupied by a few essential genes that are conserved across distant lineages, and such less-diverse gene sets could yield a good prediction agreement. Finally, none of the *Inoviridae* major coat-protein-coding TR sequences, including previously characterized genomes, were predicted to be positively viral. As demonstrated here, none of these *Inoviridae* MCP-coding TR sequences were listed in either viral or plasmid databases. Therefore, this prediction result could be explained by the lack of reference sequences, which precludes the building of a sufficiently sensitive viral gene database for internal use by the program.

The Zot domains of the discovered *Inoviridae* species formed a distinct clade in the phylogenetic tree. The genomes in this clade had notable diversity regarding length and gene component. The predicted families of these genomes were inconsistent, suggesting that further investigation to classify these genomes is required. Filamentous phages, or *Tubulavirales*, are known to undergo intensive HGTs [45, 47]. To uncover the whole picture of the complex network of filamentous phage evolution, one might need to build a complete catalog of the molecular evolution of phage genes, which requires diverse sets of genes collected from various samples. Furthermore, some genes of filamentous phages, including *Rep*, seem to be acquired from non-capsid-protein-coding MGEs, such as plasmids [47]. Applying our method to diverse samples would expand the diversity of virus-MGE shared genes, which could be used to resolve the evolutionary networks of viral genomes.

The results of spacer mapping using a looser criterion suggested that at least one-fifth of the discovered CRISPR spacers originated from TR sequences or their recognizable evolutionary predecessors, whereas about half of the CRISPR spacers originated from our discovered CRISPR-targeted sequences, including both TR and non-TR. The source of nearly half

of the CRISPR spacers encoded by residents of the human gut remains unknown, suggesting that additional protospacer reservoirs, whether extinct or simply unsampled, remain uncharacterized.

# 2.4 Conclusion

We demonstrated that CRISPR spacers can be used to detect viral genomes and other MGEs from metagenome sequences. Using spacers to infer with confidence the sequences that are targeted by CRISPR, we substantially expanded the diversity of MGEs identifiable from the human gut metagenome, which has been a topic of intense investigation for virus discovery. Comparing the sequences predicted by this approach against viral databases showed that our protocol effectively detected viral genomes without requiring similarity to any known viral sequence. Although the majority of large (>20 kb) genomes were predicted as *Caudovirales* with high confidence based on sequence homology, we found that the majority of small (<20 kb) genomes remained unclassified because of a lack of similar genomes in annotated databases. Applying this conceptual advancement to additional metagenomic datasets will increase the breadth of the lens through which we can study the diversity of Earth's virome.

# Appendix A

# Chapter 2 Supplementary Information

## A.1 Materials and Methods

### Materials

Sequencing data were selected based on NCBI metadata. The filtering parameters used for the query were as follows: layout = PAIRED, platform = ILLUMINA, selection = RANDOM, strategy = WGS, source = METAGENOME, NCBI Taxonomy = 408170 (human gut metagenome), and minimum library size = 1 Gb. If a sample contained multiple runs, we selected the run with the most bases, to simplify the analytical pipeline and avoid possible bias to protospacer counts from nearly identical metagenomes.

### Database versions and download dates

RefSeq Release 98 was downloaded on January 10, 2020, and IMG/VR Release Jan. 2018 was downloaded on October 21, 2019. GVD was downloaded on March 11, 2020, and UniRef50 was downloaded on December 16, 2019. The PDB database preprocessed by HHsuite was downloaded on September 16, 2020. Metaclust [60] was downloaded on November 10, 2020. The VirSorter database was downloaded on June 7, 2020.

## Metagenome assembly

All downloaded paired FASTQ files were preprocessed based on the guidance provided in BBTools [61] (version 38.73). Adapters, phi X, and human sequences were removed using BBDuk and BBMap. Sequencing errors were corrected using Tadpole. Each preprocessed pair of FASTQ files was assembled using SPAdes [62] (version 3.12) with the -meta option. Contigs smaller than 1 kb were discarded.

## Detection of CRISPR and spacer extraction

Assembled contigs were scanned with CRISPRDetect [63] (version 2.2) to extract CRISPR DRs, which were deduplicated using CD-HIT-EST [64] (version 4.7) and used to mask the raw reads using BBDuk. We extracted CRISPR spacers from the raw reads to maximize spacer capture from the library. Sequences located between the masked regions within the raw reads were considered CRISPR spacers and were extracted by a simple Python program (available in our source code repository), and then deduplicated.

## Detection of protospacer loci

All DRs were mapped to contigs using BBMap with a 93% minimum sequence identity. The DR mapped positions and their flanking 60 bases were masked as CRISPR loci. Next, the identified spacers were mapped to all CRISPR masked contigs with a 93% minimum sequence identity. Rather than excluding all contigs with CRISPR loci, we exclusively masked CRISPR loci to identify viruses that encode CRISPR systems that are themselves targeted by other CRISPR systems. To increase specificity, we aligned the 5′ and 3′ adjacent regions of spacer-mapped positions. These adjacent sequences were also aligned to the DR sequence associated with the mapped spacer. We discarded loci in which any alignment score divided by the length was higher

than 0.5, using the following alignment parameters: match = 1, mismatch = −1, gap = −1, and gap extension = −1. The remaining positions were considered authentic protospacer loci.

## Co-occurrence-based spacer clustering

We clustered spacers in two steps (Supplementary Figure 2-2). First, we clustered protospacer loci located within 50 kb of another protospacer locus. We then clustered spacers based on the co-occurrence of protospacers represented as a graph. In this graph, protospacers are nodes, and the edges represent the co-occurrence of connected protospacers. The weights of edges were the observed counts of co-occurrence of the connected protospacers, as defined in the previous clustering. Graph communities were detected using a Markov clustering algorithm [65] (options: -I 4 -pi 0.4; version 14-137). Clusters with a size smaller than 10 and a global clustering coefficient lower than 0.5 were discarded. Finally, 12,749 clusters comprising 591,189 spacers were derived.

## Extraction of CRISPR-targeted sequences

Contiguous regions of contigs targeted by more than 30% of the members of a spacer cluster were marked as a bed file using BEDTools [66]. To join the fragmented clusters, adjacent regions within 1 kb were concatenated. Marked regions were extracted, and sequences containing assembly gaps were discarded. Finally, both ends of each extracted sequence were compared, to identify TR sequences using a Python program utilizing the Biopython [67] package (available in our source code repository).

## Deduplication of CRISPR-targeted sequences

TR sequences were clustered using PSI-CD-HIT (options: -c 0.95 -aS 0.95 -aL 0.95 -G 1 -g 1 -prog blastn -circle 1). The remaining CRISPR-targeted sequences were clustered twice using

linclust [68] (options: --cluster-mode 2 --cov-mode 1 -c 0.9 --min-seq-id 0.95), then clustered again using PSI-CD-HIT (options: -c 0.9 -aS 0.95 -G 1 -g 1 -prog blastn -circle 1).

## Gene prediction and annotation of CRISPR-targeted sequences

Protein-coding genes were predicted from TR sequences using Prodigal (version 2.6.3) with the -p meta option. Each TR sequence was concatenated in silico, and unique predicted genes were selected to recover truncated genes. Predicted protein sequences with partial frags were discarded. The remaining protein sequences were clustered based on a 30% sequence identity threshold using mmseqs [69] (version e1a1c1226ef22ac3d0da8e8f71adb8fd2388a249). HMMs were constructed from each representative sequence using three iterations of jackhmmer [70] (version 3.2.1) against the Metaclust database. The constructed HMMs were then used as queries to search PDB (probability, >80; E-value, <1e–3) using HHsearch (version 3.1.0).

## Assessment of capsid-protein-detection sensitivity and specificity

The sensitivity and specificity of capsid protein detection were assessed using TR sequences similar to the RefSeq-recorded viral and plasmid genomes. Among the 588 TR sequences recorded in the RefSeq virus, we detected capsid genes from 577 TR sequences (271 HK97 capsid genes, 306 *Microviridae* MCP genes, and 0 *Inoviridae* MCP genes) (98.13%). Conversely, among the 1096 TR sequences recorded in RefSeq plasmid, we detected capsid genes from 6 TR sequences (6 HK97 capsid genes) (0.55%). Our pipeline successfully detected capsids from reference recorded viral genomes and did not detect them from nonviral genomes with agreeable measures. Accordingly, we conclude that our pipeline has acceptable sensitivity and specificity.

## Targeting host prediction

DRs were mapped to RefSeq bacterial and archaeal genomes using BBMap. A locus with more than three consecutive DR hits within 100 bases was considered an authentic CRISPR locus associated with the mapped DR. DRs mapped to multiple taxa at a given taxonomic level were not taxonomically assigned to that level. The DRs assigned to taxa were used to predict the targeting host. We counted the protospacers linked to taxonomically assigned DRs within a TR sequence. If the count of a given taxon was ≥10 and exhibited higher than 90% exclusiveness, we considered that the corresponding taxon was a targeting host of a given contig. Host predictions were performed for each taxonomic level: species, genus, family, order, class, phylum, and domain.

## tRNA prediction from TR sequences

TR sequences were fed into the ARAGORN program [71] using the -gcbact option, which corresponds to the Bacterial/Plant chloroplast genetic code.

## Gene-content-based hierarchical clustering of TR sequences

TR sequences were scanned by HMMs derived from the clustering results of the predicted protein sequences using both TR and non-TR sequences. The scanned result was represented by a binary matrix (score > 60). We selected the top 1000 genes that were recurrently observed within TR sequences. The matrix was hierarchically clustered using ComplexHeatmap [72] (version 2.5.3) and then annotated according to database hits, host, GC content, capsid types, and predicted gene functions. Clustering was performed separately for large and small TR sequences.

## Phylogenetic analysis of *Microviridae* MCP

Representative MicroMCP sequences were selected by clustering all capsid proteins based on an 85% sequence identity threshold throughout the entire length of the protein. Representative and reference protein sequences were aligned using MAFFT [73] (version 7.310) and then trimmed using trimAl [74] (version 1.4). Aligned sequences were used for Bayesian phylogenetic analysis using MrBayes [75] (version 3.2.7). A mixed substitution model with a uniform prior that converged to Blosum62 (posterior probability = 1.000) was selected. All other priors were set to the default state. Two Markov chain Monte Carlo chains with identical priors were run over ten million generations and sampled every 500 generations. The standard deviation of split frequencies approached zero (0.007837) over the run. The phylogenetic tree was visualized using FigTree [76].

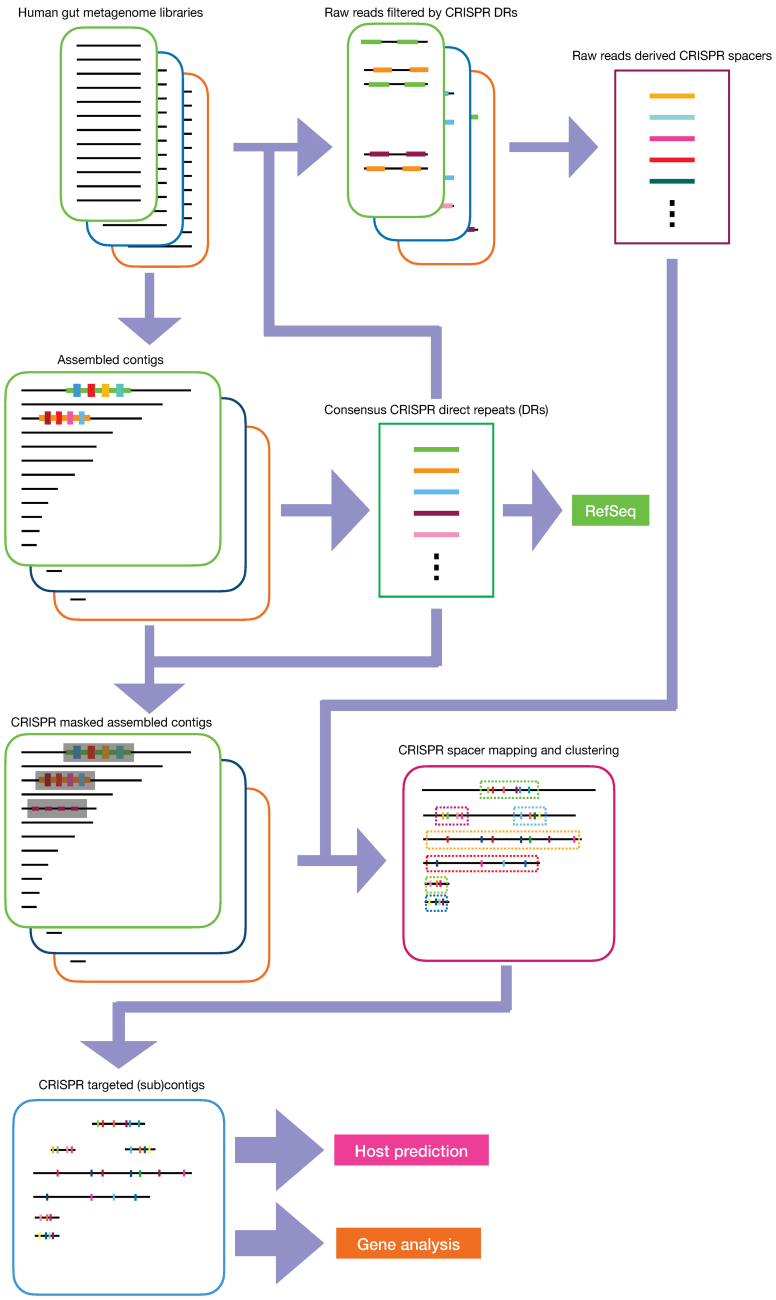
## Phylogenetic analysis of the Zot domain

Representative Zot protein sequences were selected via clustering based on a 50% sequence identity threshold throughout the entire length of the protein. Representative sequences were aligned using MAFFT with the –localpair option. Aligned domains were manually inspected and extracted, then trimmed using trimAl. Aligned domain sequences were used for Bayesian phylogenetic analysis using MrBayes. A mixed substitution model with a uniform prior that converged to Blosum62 (posterior probability = 1.000) was selected. All other priors were set to the default state. Two Markov chain Monte Carlo chains with identical priors were run over twelve million generations and sampled every 500 generations. The standard deviation of split frequencies approached zero (0.002185) over the run. The phylogenetic tree was visualized using FigTree.

## Generation of scrambled sequences

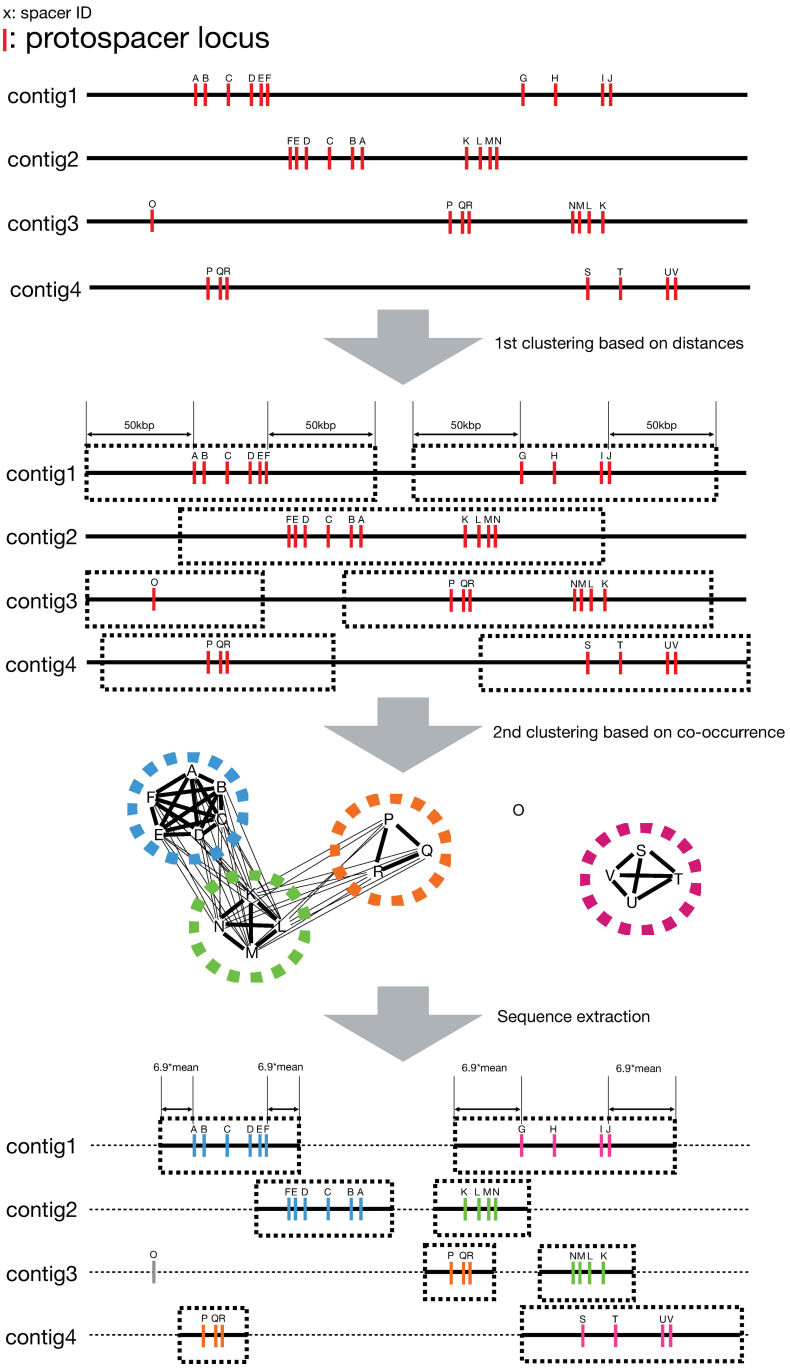
Scrambled sequences are random sequences that were identical to the TR sequences in length. The sequences were generated based on the sampled nucleotide frequencies from the TR sequences using Biopython (available in our source code repository).

# A.2 Supplementally Figures



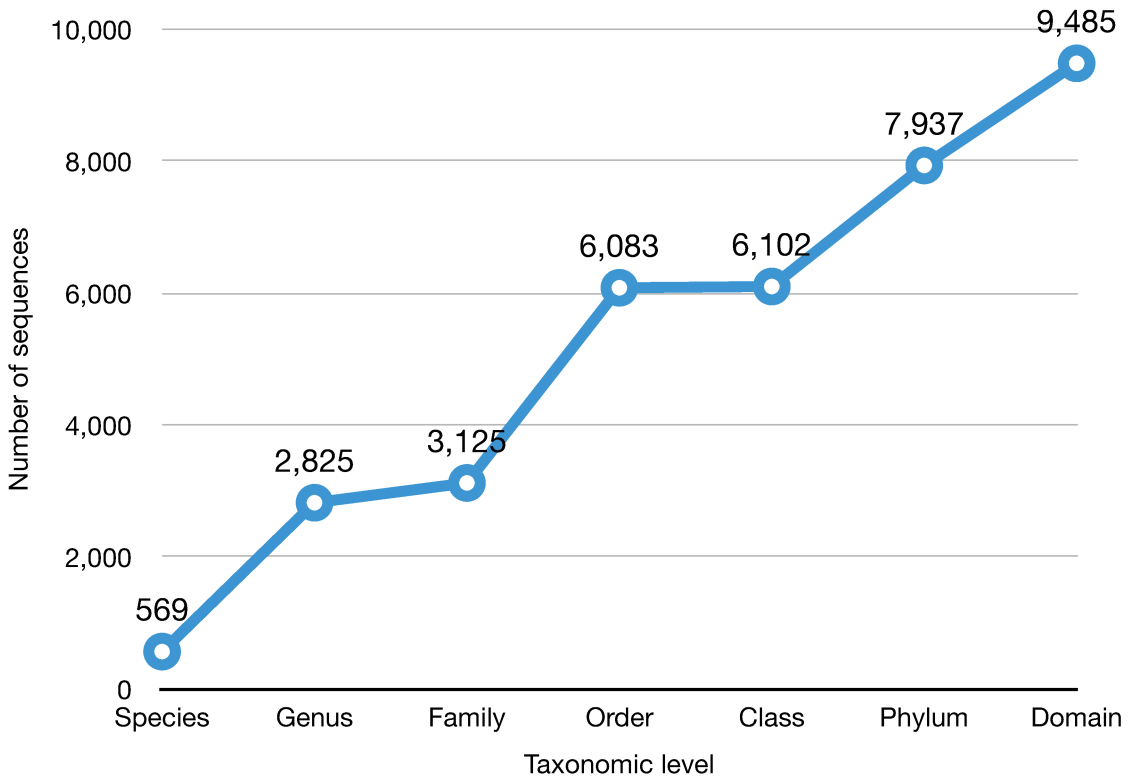

**Supplementary Figure 2-1. Basic workflow used for viral genome detection.** Human gut metagenome libraries were preprocessed to remove adapters, phi X, and human sequences. After correcting sequencing errors, libraries were assembled. Clustered regularly interspaced short palindromic repeats (CRISPR) loci were discovered from the assembled contigs. Consensus direct

repeats (DRs) from the discovered CRISPR loci were used to extract spacers, mask the CRISPR loci, and predict the host. All unique CRISPR spacers were mapped to contigs to discover the protospacer loci. Spacers were clustered based on the co-occurrence of the associated protospacers. Sequences targeted by more than 30% of the members of a spacer cluster were extracted and used for further analysis.

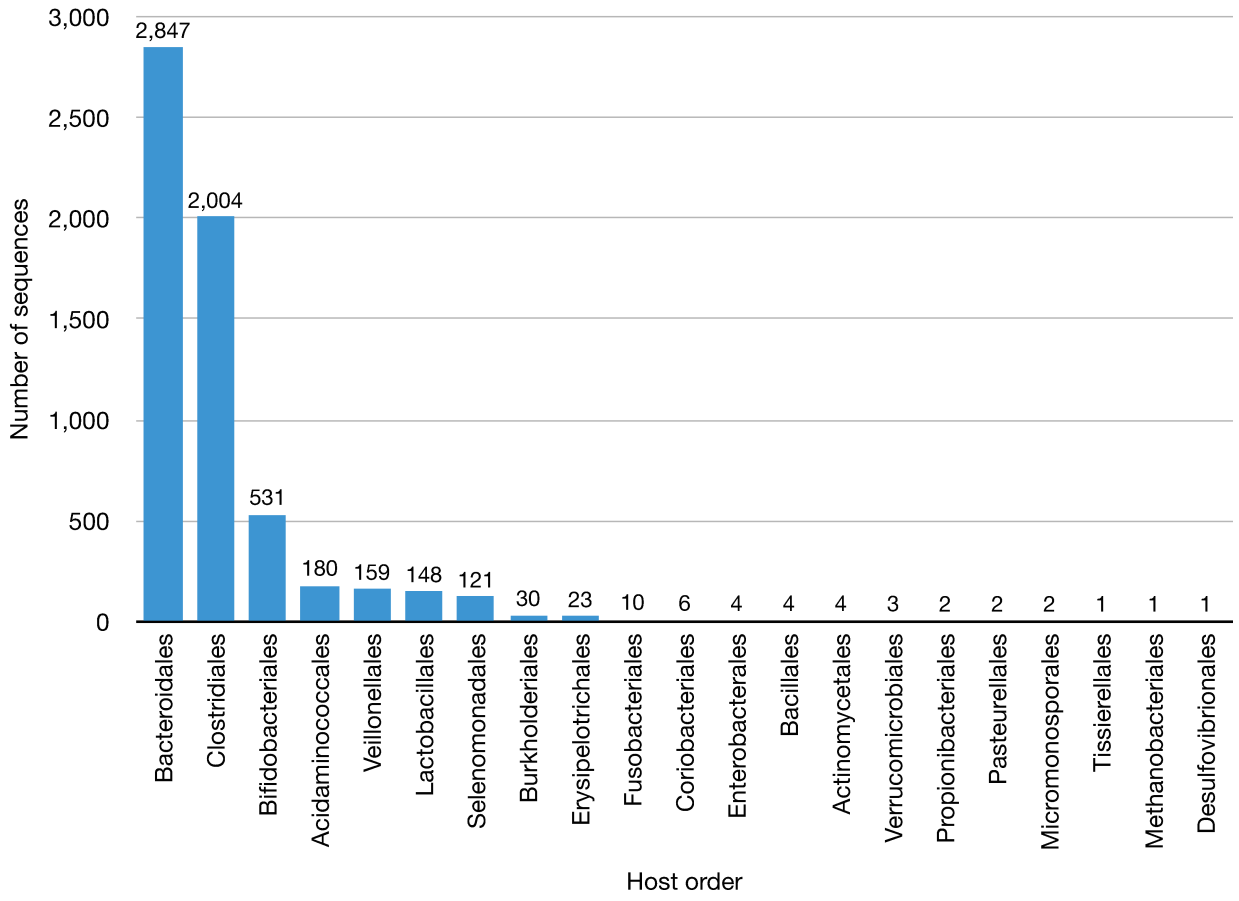



**Supplementary Figure 2-2. Spacer clustering based on the co-occurrence of protospacers.** Initially, protospacer loci were clustered based on the distance between them. Within initial clusters, co-occurrences of protospacers were counted and used to construct an undirected graph. The nodes (spacers) in the undirected graph were further clustered using the Markov clustering

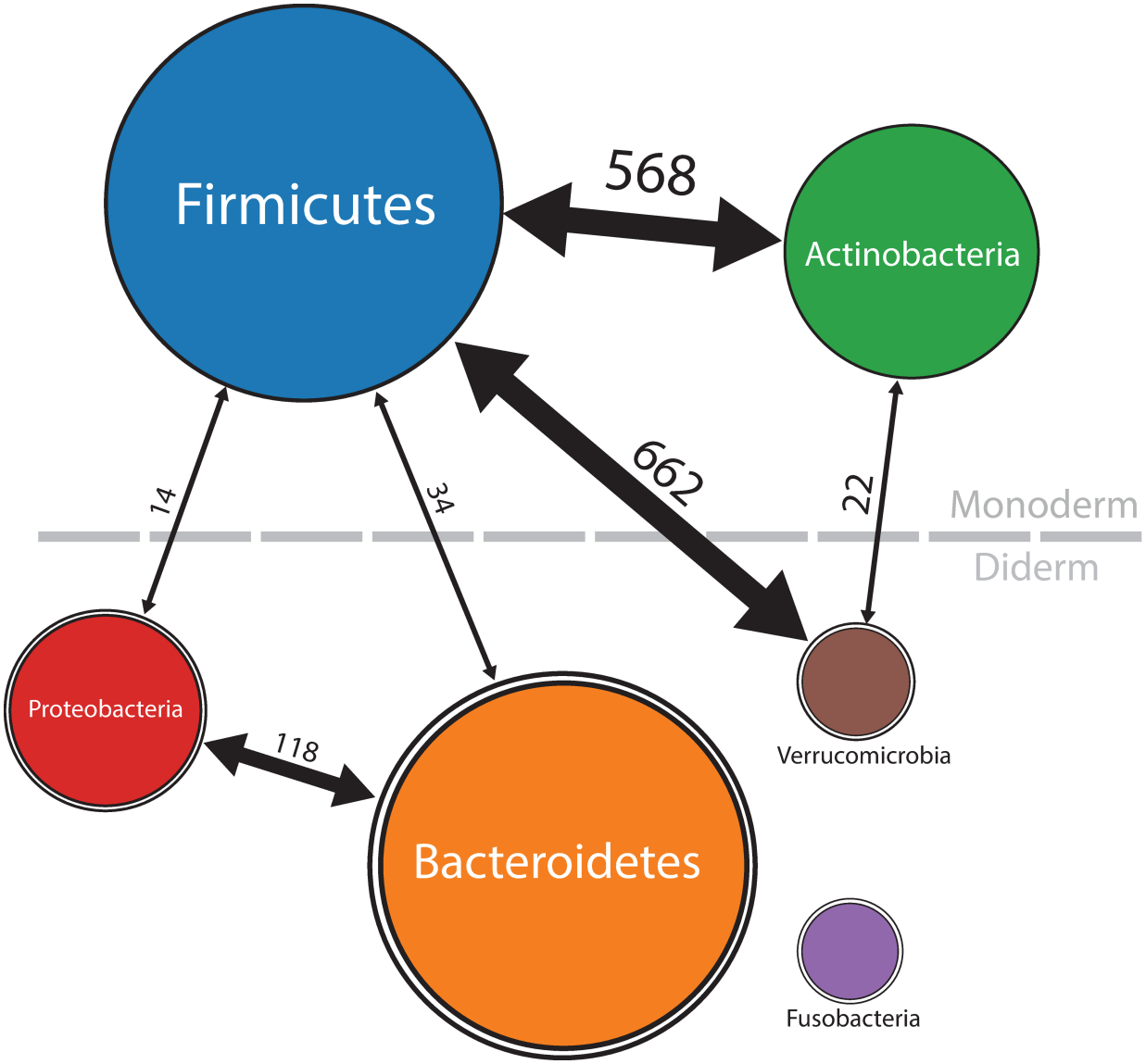
algorithm. The mean distances between adjacent protospacer loci within clusters were calculated and used to extract CRISPR-targeted sequences. The length and number of protospacers shown here are conceptual and not based on observed data.



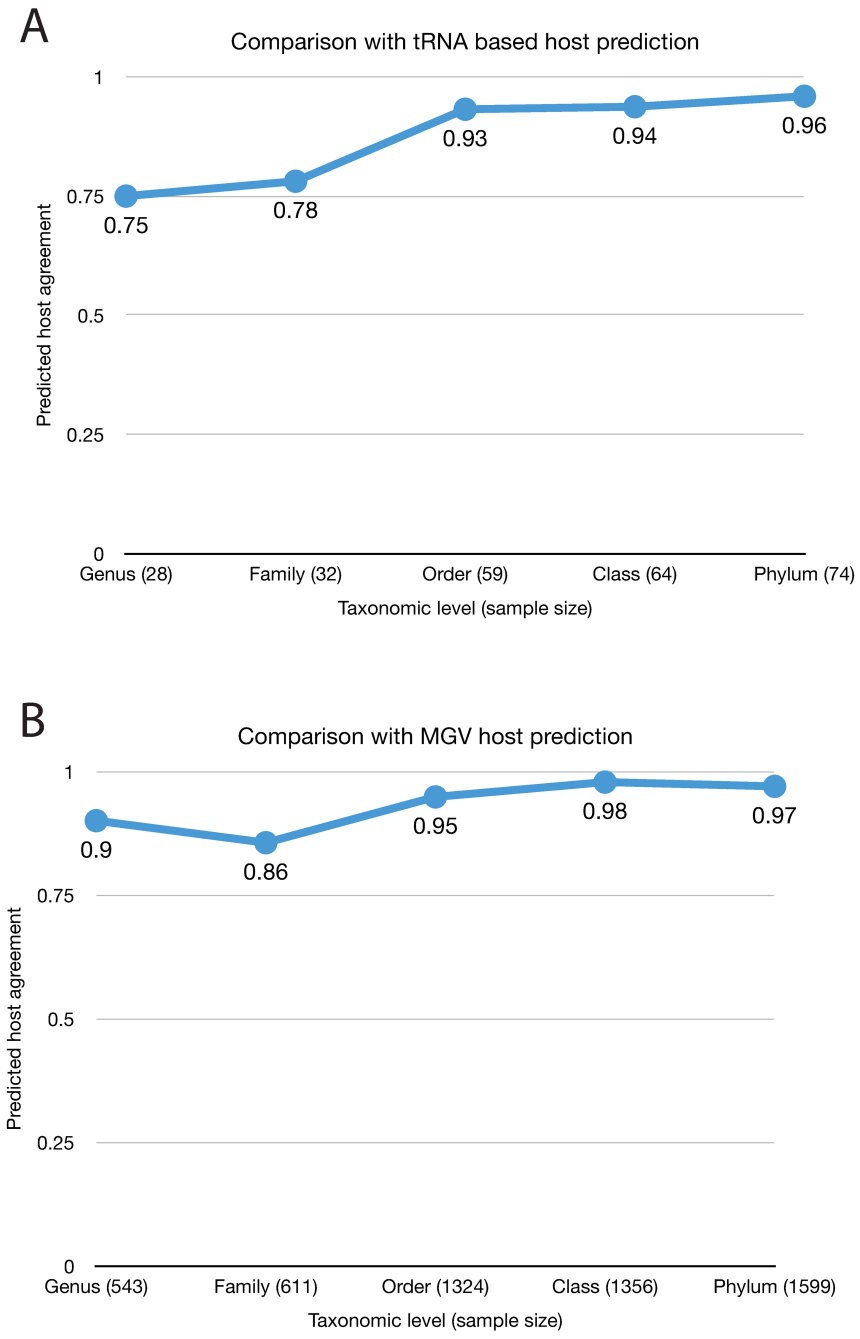

**Supplementary Figure 2-3. Number of sequences with a predicted targeting host according to each taxonomic level.**



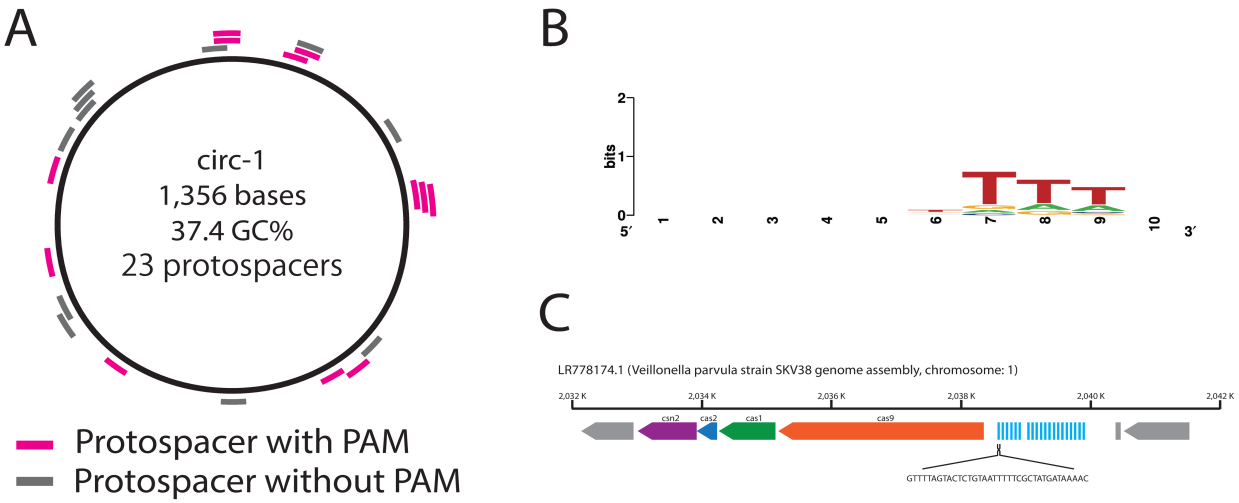

**Supplementary Figure 2-4. Number of sequences with a predicted CRISPR-targeting host at the taxonomic level of order.**



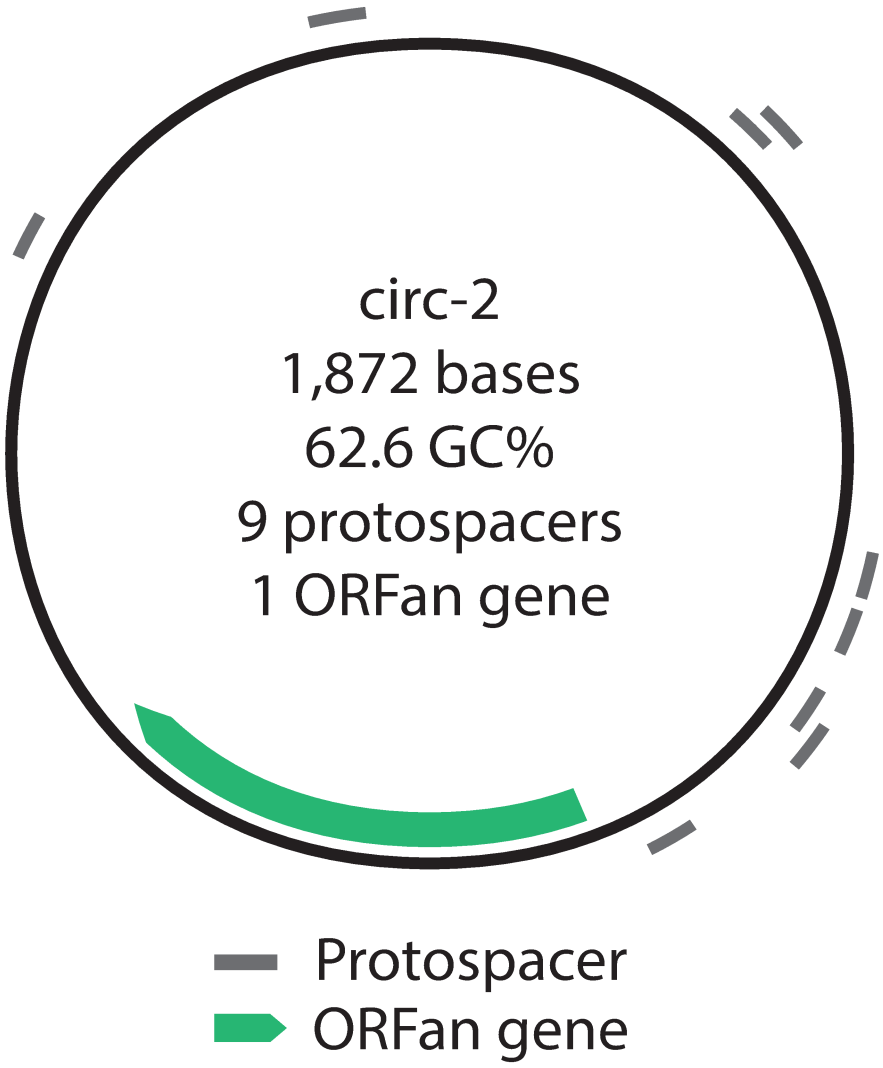

**Supplementary Figure 2-5. Heterogeneous distribution of TR sequences targeting host ambiguity.** Circle size approximately represents the popularity of the respective host. The bidirectional arrows connect the top two phyla according to host-assigned protospacer counts (i.e., protospacers most often associated with CRISPR DRs are assigned to these two phyla). The numbers on the arrows are counts of the number of TR sequences associated with the connected phyla.



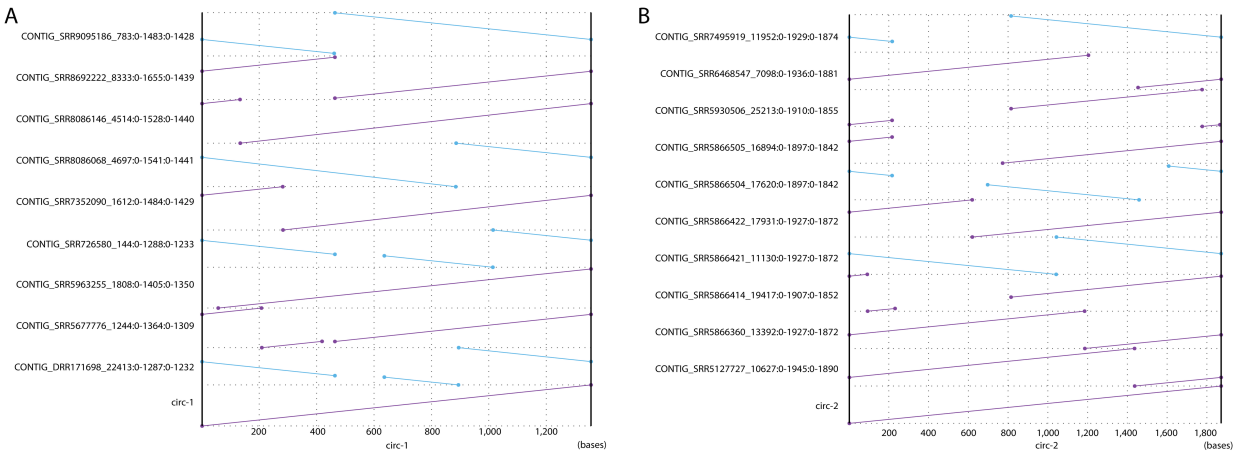

**Supplementary Figure 2-6. (A) Host prediction comparison between DR-based and tRNA-based methods. (B) Host prediction comparison between MGV and this study.**




**Supplementary Figure 2-7. circ-1 protospacers, associated PAM, and Cas genes. (A)** Genomic map of circ-1. The circle represents the circular genome of circ-1. The positions of protospacers are indicated outside the circle. The protospacers with and without PAM were colored magenta and dark gray, respectively. **(B)** PAM of circ-1 protospacers. Both adjacent sequences of protospacer positions up to 10 bases were collected and then aligned, to generate a logo using WebLogo [77]. **(C)** DR-aligned locus of LR778174.1. The cyan bars are the DR-aligned positions. The genes related to the Class 2 Cas system were annotated using colors.




**Supplementary Figure 2-8. circ-2 protospacers and *ORFan* gene.** Genomic map of circ-2. The positions of protospacers and an ORFan gene are depicted outside and inside the circle, respectively.




**Supplementary Figure 2-9. (A) circ-1 and (B) circ-2 dot plot representations of genome comparisons.** The representative and similar genomes were aligned using nucmer [78], then plotted using mummerplot. For circ-1, the 10 most-similar genomes were selected.

# A.3 Supplementary Tables

**Supplementary Table 2-1. Samples and assembly summary.** This table is available in the Zenodo repository: https://doi.org/10.5281/zenodo.6354110

**Supplementary Table 2-2. CRISPR-targeted TR sequence summary.** This table is available in the Zenodo repository: https://doi.org/10.5281/zenodo.6354110

# A.4 Supplementary Scripts

All scripts used in this study are available in the Zenodo repository: https://doi.org/10.5281/zenodo.6621424

# A.5 Supplementary Data

**Supplementary Data 2-1. Dataset including the discovered CRISPR spacers, direct repeats, protospacers, co-occurrence-based spacer clustering results, predicted protein sequences, built HMMs, database comparison results, phylogenetic analysis results, predicted targeting hosts, and CRISPR-targeted TR sequences.** The dataset is available in the Zenodo repository: https://doi.org/10.5281/zenodo.6503687

# Chapter 3

# CRISPR-targeted RNA-dependent RNA polymerase coding RNA sequences in the human gut metatranscriptomes

## 3.1 Introduction

RNA bacteriophages are one of the most poorly studied genetic entities. Currently, only two families; *Cystoviridae* and *Fiersviridae,* comprising 27 species genomes are recorded in the NCBI RefSeq database. Several surveys predicted that the current knowledge of the RNA viral genomes is only a portion of the entire diversity [7, 79]. To expand the knowledge of RNA bacteriophage diversity, we investigated RNA sequences in the human gut metatranscriptome. We extracted non-transcribed RNA sequences by comparing them to the DNA sequences from the same sampling points. Then we searched for CRISPR targeted sequences from RNA-dependent RNA polymerase (RdRP) coding non-transcribed sequences. Interestingly, we found that *Picobirnaviridae* species, which host is currently controversial [80, 81], are being targeted by CRISPR. The substitution rate of this virus lineage was nearly $10^{-2}$ substitutions per site per year, which might indicate that this viral lineage is escaping the CRISPR-targeting by the incredible evolution speed.

# 3.2 Results

## Extraction of non-transcribed RNA sequences

Firstly, we extracted RNA sequences that were not transcribed from DNA sequences in the human gut metagenome. To remove such transcripts from metatranscriptome, we used omics data of the human gut microbiome published by the Integrative Human Microbiome Project (IHMP) [82]. This dataset includes both human gut metagenome and metatranscriptome sequences periodically sampled from the subject individuals, allowing us to extract non-transcribed RNA sequences by comparing the DNA and RNA sequences.

From the IHMP web page, we downloaded paired FASTQ files comprising 1,312 metagenome sequences and 762 metatranscriptome sequences, sampled from 104 individuals over 1,289 (average 12.4 per individual) temporal sampling points. For each sampling point of the individual, transcriptome sequences were preprocessed and then assembled. If a sampling point contains more than one paired FASTQ file, these files were pooled together before assembly. The assembled contigs are clustered with 100% sequences similarity thresholds, resulting in 8,299,028 contigs comprising about 4.3 billion bases. We speculated that most of these assembled transcriptome contigs are transcripts of cellular/viral DNA. To remove such transcribed sequences, we compared the transcriptome contigs with kmers from the metagenomic sequences. From all preprocessed metagenome sequences, 1.38 trillion non-singleton kmers ($k = 43$) were collected. The transcriptome contigs sharing one or more of these kmers were removed, resulting in 140,272 contigs comprising about 50.6 million bases (98.8% bases are removed). We consider that these remaining contigs are sequences not transcribed in the human gut.

From the non-transcribed contigs, protein-coding genes were predicted and clustered with a 30% sequence similarity threshold, resulting in 60,430 representative protein sequences. The representative sequences were used for building HMMs which are then used as queries for searching the PDB database. Finally, 4,391 protein sequences were annotated (E-value < 1e-5).

## Taxonomy of RNA-dependent RNA polymerase in the human gut

RdRP is considered a hallmark gene for RNA viruses [3, 7, 79]. 117 representative protein sequences, predicted from 7,321 unique contigs, were homologous to RdRP. These RdRP protein sequences were taxonomically assigned by searching to RefSeq viral protein database (E-value < $10^{-10}$) (Figure 3-1). The majority of the sequences were assigned to plant viruses such as *Virgaviridae,* likely derived from viruses infecting vegetable foods. Several sequences were assigned to well-known human infecting viruses such as *Picornaviridae* (*Enterovirus*). About a quarter of the RdRP sequences were unclassified.

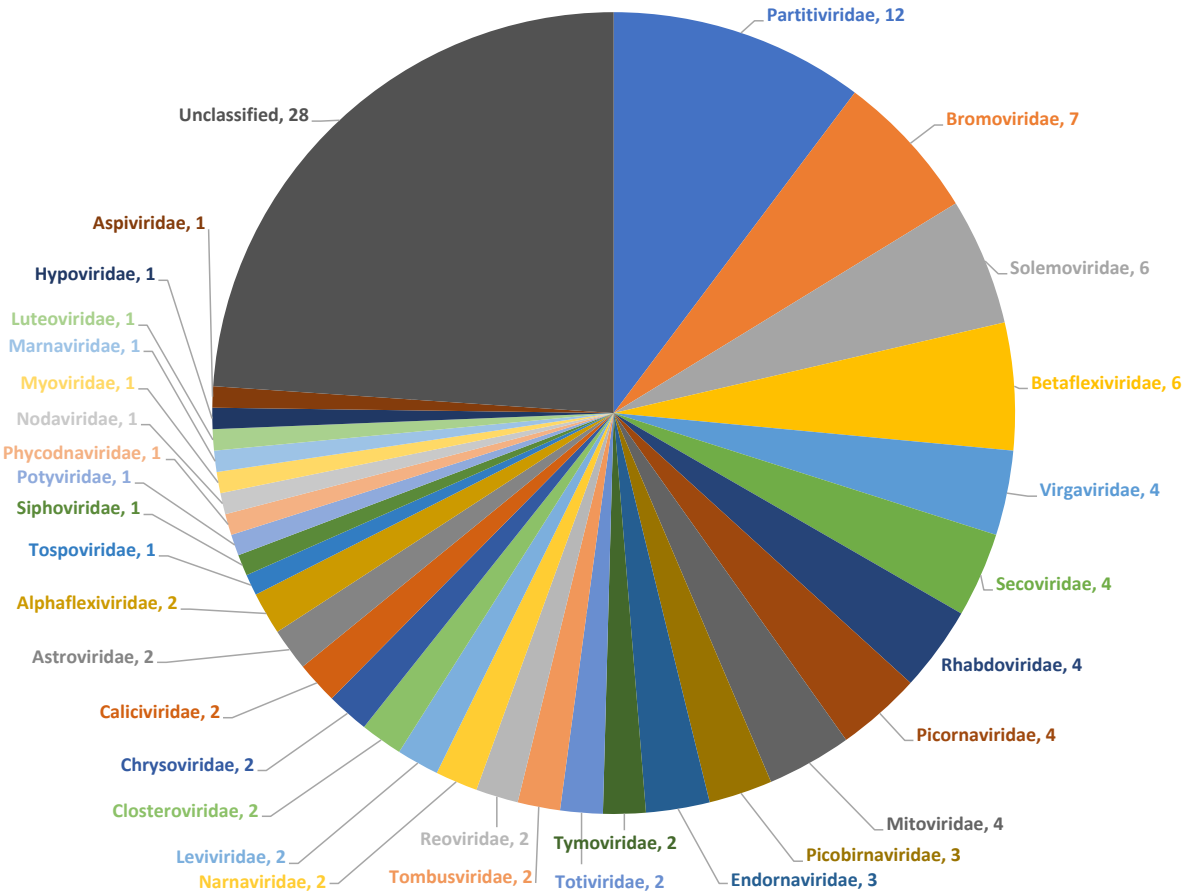



**Figure 3-1. Taxonomic assignments of RdRP in the human gut.** Pie charts represent the numbers of the representative RdRP protein sequences taxonomically assigned to the corresponding viral families. Sequences without a hit within the threshold (E-value < $10^{-10}$) or aligned to multiple viral families were assigned to Unclassified.

## *Picobirnaviruses* targeted by the CRISPR-Cas system

Next, we investigated the RdRP coding sequences targeted by CRISPR. From the IHMP metagenome preprocessed reads, we extracted about 27 million reads containing CRISPR direct repeats (DRs). From them, we extracted 253,563 unique spacers. These spacers were aligned to RdRP coding non-transcribed contigs. Interestingly, we found that nine RdRP encoding contigs

taxonomically assigned to *Picobirnaviridae* were containing a protospacer, suggesting that these *Picobirnaviridae* lineages might be infecting prokaryotic cells (Figure 3-2).


**Figure 3-2. A protospacer within the *Picobirnaviridae* genome.** The sequence in the middle is a raw read containing the CRISPR DR. The DR is highlighted green and shows the pair-wise alignment to the reference DR depicted above. The bottom sequence is a part of a *Picobirnaviridae* RdRP coding contig. The spacer in the read and the protospacer in the contig are highlighted in red along with the pair-wise alignment.

To predict the CRISPR targeting host, the DR associated with the protospacer found from the *Picobirnaviridae* contigs was searched in the assembled contigs and RefSeq genomes. No identical hits to the assembled contigs longer than 100 bases were found. On the other hand, hits with one or more mismatched were found from RefSeq recorded *Lachnospiraceae* family genomes (NZ_NFLQ01000008.1).

## CRISPR-targeted RdRP coding sequence is a genuine *Picobirnaviridae* species

To further confirm that the CRISPR-targeted sequences are a lineage of *Picobirnaviridae* species, we constructed the Bayesian phylogeny from the RdRP protein sequences taxonomically assigned to the *Picobirnaviridae* family and the RefSeq recorded *Picobirnaviridae* species RdRP protein sequences (Figure 3-3).

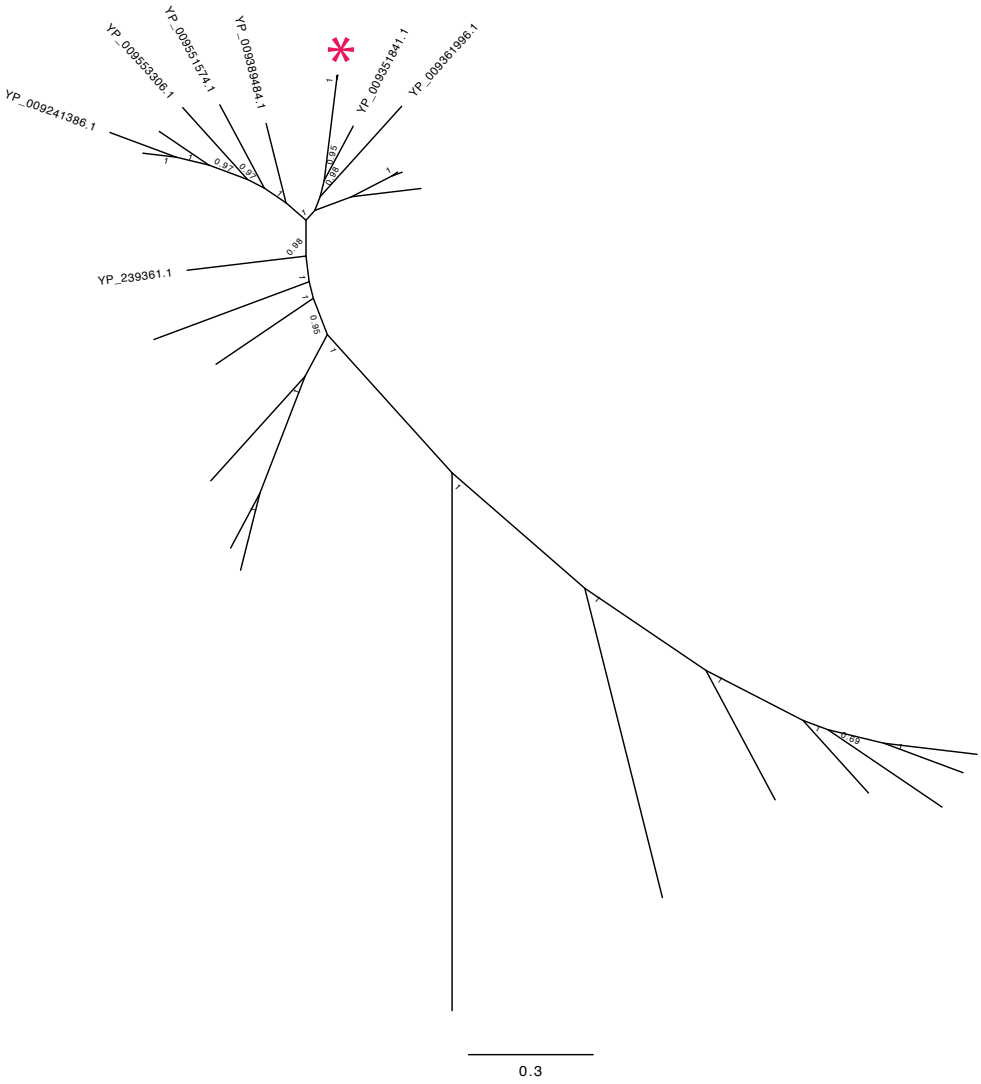



**Figure 3-3. Bayesian phylogeny of *Picobirnaviridae* RdRP.** The tips without labels are sequences discovered in this study, and the tips labeled with the NCBI accessions are the *Picobirnaviridae* sequences derived from the RefSeq database. The RdRP contigs containing protospacers were labeled with a red-colored asterisk.

In the phylogeny, the CRISPR-targeted sequences were placed within the branches of RefSeq recorded *Picobirnaviridae* sequences suggesting that this sequence is a genuine lineage of novel *Picobirnaviridae* species.

## Discovered *Picobirnaviridae* genomes have prokaryotic ribosome binding motifs

Next, we investigated the presence of a ribosome binding site upstream of the *Picobirnaviridae* RdRP genes. We extracted upstream (20 bases) of RdRP genes and start codon from the discovered *Picobirnaviridae* species genomes. 34 unique sequences were aligned and used to build an HMM, which was used to plot a logo. We found that a strong signal was present upstream of the RdRP genes (Figure 3-4).

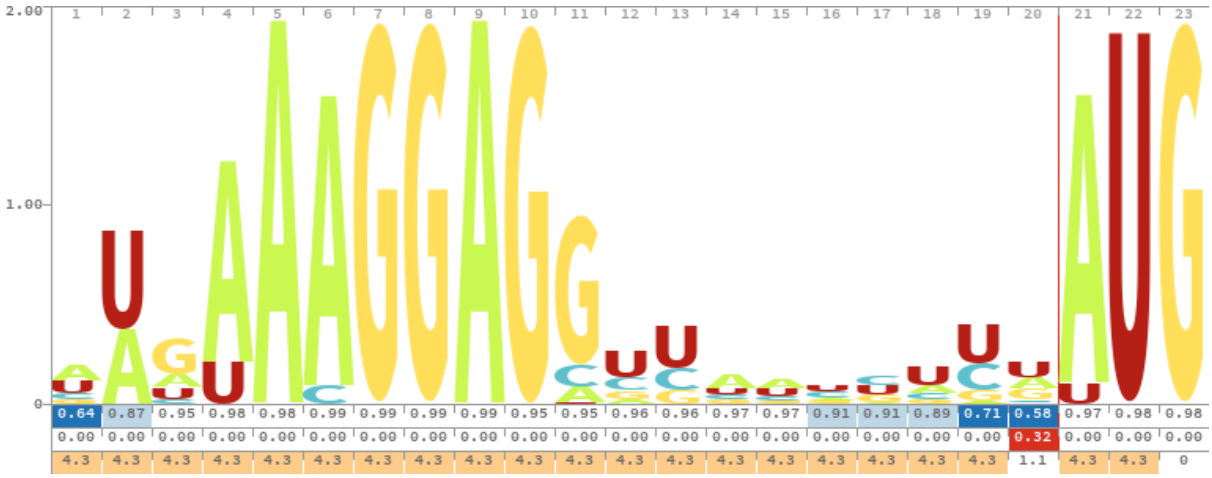



**Figure 3-4. Motif upstream of the discovered *Picobirnaviridae* RdRP genes.** The logo was generated from the 20 bases upstream and the start codon of the RdRP genes. Each column is the position within the HMM, and the height of each character represents the information content of that position.

The motif includes an AGGAGG sequence, a prokaryotic ribosomal binding motif known as the Shine-Dalgarno (SD) sequence, suggesting that the *Picobirnaviridae* RdRP genes might be translated by prokaryotic ribosomes. This result is consistent with the previous study

using the NCBI recorded genomes [80]. We also manually confirmed that the CRISPR-targeted genome contained the perfect SD sequence (AGGAGG) nine bases upstream of the RdRP gene.

## *Picobirnaviridae* genomes are rapidly evolving in the human gut

While we found nine protospacers by aligning 253,563 spacers extracted in this study to the RdRP coding contigs, none of the spacers extracted from our previous study which compose more than two million sequences were aligned to the discovered RdRP coding contigs. From these observations, we hypothesized that the RNA phages are escaping CRISPR-targeting with an incredibly fast evolution. To investigate the evolutionary speed of the discovered *Picobirnaviridae* genomes, we calculated a dated phylogenetic tree using the cDNA sequences of RdRP genes and the sampling date provided by the IHMP metadata (Figure 3-5).

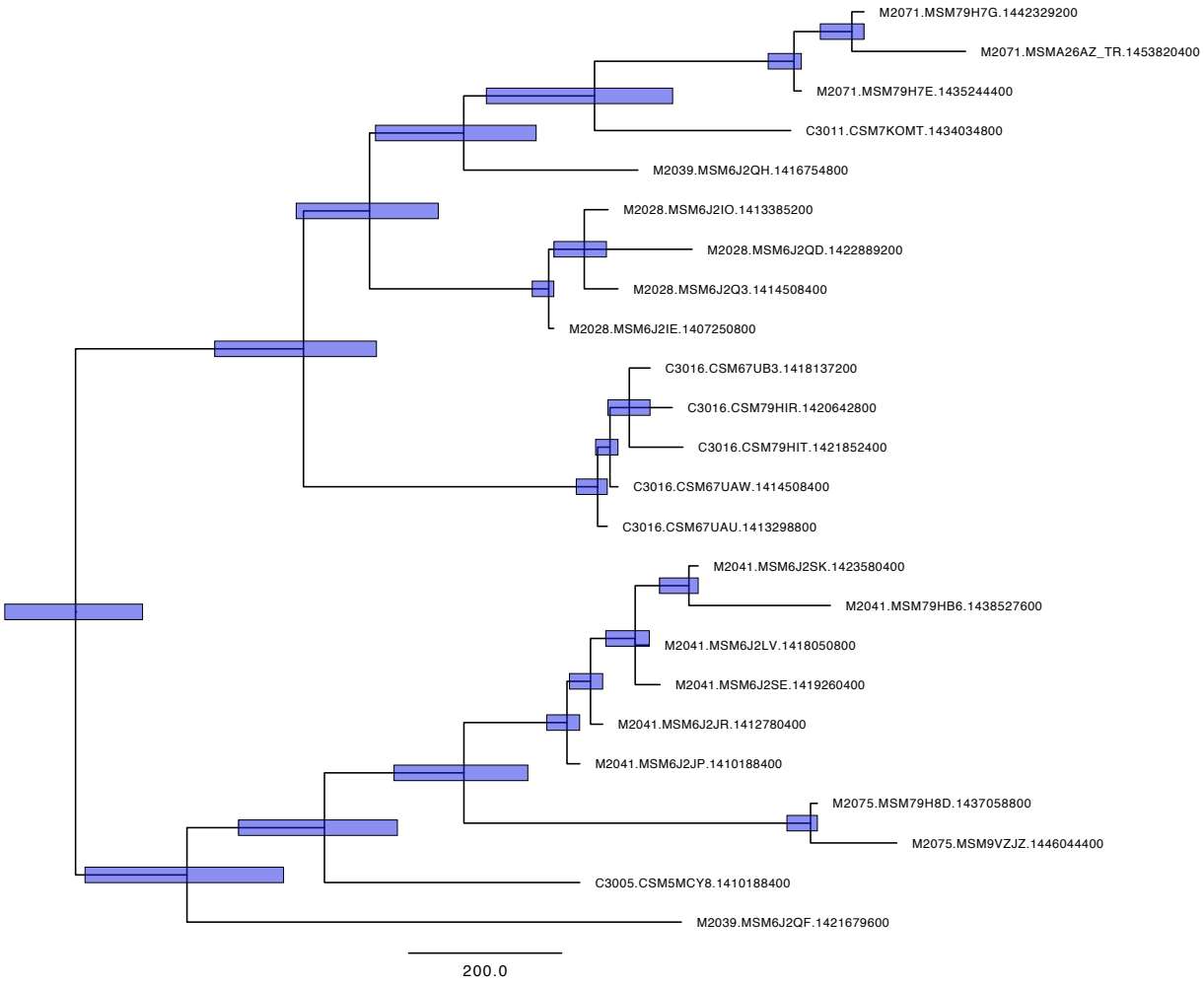



**Figure 3-5. Bayesian phylogeny of *Picobirnaviridae* RdRP cDNA sequences.** The unit of the x-axis is a day. The tip ages were fixed to the sampling date and labeled with human subject identifiers, library names, and sampling time points. The blue rectangles indicate the node heights' 95% highest posterior density (HPD). The phylogeny was calculated under a strict molecular clock model (a uniform clock rate applied all over the tree).

In the phylogenetic tree, the *Picobirnaviridae* strains sampled from the same human individuals were clustered together, suggesting that those lineages are evolving within the individual's gut environments. The mean of the clock rate posterior distribution was 0.000026

substitutions per site per day (0.000021 is the lower bound and 0.000032 is the upper bound of the 95% HPD interval), which is equivalent to 0.00949 substitutions per site per year.

# 3.3 Discussion

We attempted to discover RNA phages from human gut metatranscriptome sequences using the CRISPR targeting and the RdRP coding non-transcribed transcriptome contigs. Interestingly, a lineage of *Picobirnaviridae* species contained protospacers. Furthermore, we showed that these viruses have prokaryotic ribosomal binding sites upstream of the RdRP genes. These results support that the *Picobirnaviridae* species infects Prokaryotic cells. We questioned why so few protospacers were identified from the discovered RdRP coding non-transcribed sequences and hypothesized that RNA phages are escaping from CRISPR-targeting by incredibly rapid evolution, leaving the previously acquired spacers not complementary and unable to initiate the restriction of the mutated RNA sequences. The estimated clock rate of the discovered *Picobirnaviridae* species was nearly $10^{-2}$ substitutions per site per year. However, this estimation might be biased due to the time-dependent rate phenomenon [83–86]. This phenomenon could occur by sampling bias and/or ignoring the effect of selection over a longer period. To estimate the long-term evolutionary speed, samples from mummified tissues, coprolite, and calculus could be used to sequence the ancient phage DNA and RNA [87]. To the author's knowledge, there is no study about the substitution rate of RNA phages comparable to this result. A human infecting ssRNA virus SARS-CoV-2 was estimated to be around $1\sim1.5 \times 10^{-3}$ substitutions per site per year [88–91], suggesting that the *Picobirnaviridae* species are evolving nearly a magnitude faster than the known human infecting RNA virus. If the RNA phages are evolving at such a rapid rate and cells are acquiring CRISPR spacers directly from the RNA phage genomes, we should be able to observe the rapid evolution

of the CRISPR loci on the host cell population as well. However, we failed to find the original CRISPR locus of the Picobirnaviridae-associated protospacer due to the fragmentation of the assembled metagenome contigs. This issue might be caused by both insufficient depths to capture the low-populated cells and the high diversity of the CRISPR locus in the population. With ultra-deep DNA and RNA sequencing technology, we might be able to directly observe the evolutionary arms race between the RNA phage genome and the CRISPR locus. If the CRISPR locus is highly diversified in the cell population, the long-read technology might be helpful to capture spacer acquisition patterns. With the information on the RNA phage-associated CRISPR locus, we should be able to narrow the host range of the subject RNA phage which could be used to construct an RNA phage culture system. Finally, nearly a quarter of the discovered RdRP protein sequences were taxonomically unassigned indicating that there are still uncovered RNA virus sequences presenting in the human gut environment.

# Appendix B

# Chapter 3 Supplementary Information

## B.1 Materials and Methods

### Materials

Both metagenomic and metatranscriptomic sequences were downloaded from the IHMP web page. These data were selected based on the availability of both RNA and DNA sequences. For example, individuals lacking RNA sequences were removed from the analysis.

### Sequence preprocessing

All downloaded paired FASTQ files were preprocessed based on the guidance provided in BBTools [61] (version 38.73). Adapters, phi X, and human sequences were removed using BBDuk and BBMap. Sequencing errors were corrected using Tadpole.

### Kmer extraction from metagenome sequences

From all preprocessed IHMP metagenomic sequences, non-singleton kmers (k = 43) were collected using Jellyfish [92]. We started collecting canonical kmer (specified by the -C option) from each library, then all built kmer files were merged into a single file which is used for filtering the non-transcribed RNA sequences.

## Metatranscriptome assemblies and extraction of non-transcribed contigs

Each preprocessed pair of metatranscriptome FASTQ files was assembled using SPAdes [62] (version 3.15) with the –rnaviral option. From the assembled contigs, we collected kmers (k = 43) from both strands and searched these kmers for the metagenomic kmers previously described. Contigs containing one or more shared kmers were removed. A custom-made python program was used for this kmer comparison.

## Gene annotation of non-transcribed RNA sequences

Protein-coding genes were predicted from non-transcribed RNA sequences using Prodigal (version 2.6.3) with the -p meta option. The predicted protein sequences were clustered based on a 30% sequence identity threshold using mmseqs [69] (version 96d452cb432fc4674991a48952deaf24d1787e77). HMMs were constructed from each representative sequence using three iterations of jackhmmer [70] (version 3.3.1) search to the Metaclust database. The constructed HMMs were then used as queries to search the PDB database using HHsearch (version 3.1.0).

## Taxonomic assignment of RdRP protein sequences

The predicted protein sequences annotated as RdRP were used as queries to search the RefSeq viral protein database using the BLASTN program. The sequences were assigned to the taxonomic family with hits less than $10^{-10}$ E-values.

## Phylogenetic analysis of RdRP protein sequences

The RdRP protein sequences from the discovered and the RefSeq database recorded Picobirnaviridae genomes were aligned together using MUSCLE [93]. Aligned sequences were used for Bayesian phylogenetic analysis using MrBayes [75] (version 3.2.7). A mixed substitution model with a uniform prior that converged to the WAG model (posterior probability = 1.000) was selected. All other priors were set to the default state. Two Markov chain Monte Carlo chains with identical priors were run over ten million generations and sampled every 500 generations. The standard deviation of split frequencies approached zero (0.000520) over the run. The phylogenetic tree was visualized using FigTree [76].

## Dated phylogenetic analysis of RdRP coding sequences

From the multiple sequence alignment result of the RdRP protein sequence, we generated codon alignment using PAL2NAL [94]. Aligned codon sequences were used for Bayesian phylogenetic analysis using MrBayes. The ages of each sequence were fixed to the corresponding sampling date derived from the IHMP metadata. The GTR + I + Γ model was selected for the substitution model. Branch lengths were set to conform uniform clock model, and the clock prior distribution was set to the Normal distribution with the mean at 0.001 and the standard deviation of 0.1. The prior tree height was set to the Gamma distribution with a mean of 1000 and a standard deviation of 50. Four Markov chain Monte Carlo chains with identical priors were run over a million generations and sampled every 50 generations. The standard deviation of split frequencies approached zero (0.005153) over the run. The phylogenetic tree was visualized using FigTree.

# References


1. Ito J, Sugimoto R, Nakaoka H, Yamada S, Kimura T, Hayano T et al. Systematic identification and characterization of regulatory elements derived from human endogenous retroviruses. PLoS Genet. 2017;13:e1006883.

2. Wommack KE, Colwell RR. Virioplankton: viruses in aquatic ecosystems. Microbiol Mol Biol Rev. 2000;64:69-114.

3. Koonin EV, Senkevich TG, Dolja VV. The ancient Virus World and evolution of cells. Biol Direct. 2006;1:29.

4. Krupovic M, Dolja VV, Koonin EV. Origin of viruses: primordial replicators recruiting capsids from hosts. Nature Reviews Microbiology. 2019;17:449-458.

5. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol. 1998;5:R245-9.

6. Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GG, Boling L et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nat Commun. 2014;5:4498.

7. Callanan J, Stockdale SR, Shkoporov A, Draper LA, Ross RP, Hill C. Expansion of known ssRNA phage genomes: From tens to over a thousand. Sci Adv. 2020;6:eaay5981.

8. Al-Shayeb B, Sachdeva R, Chen L-X, Ward F, Munk P, Devoto A et al. Clades of huge phages from across Earth's ecosystems. Nature. 2020;578:425-431.

9. Gregory AC, Zablocki O, Zayed AA, Howell A, Bolduc B, Sullivan MB. The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut. Cell Host & Microbe. 2020

10. Roux S, Enault F, Hurwitz BL, Sullivan MB. VirSorter: mining viral signal from microbial genomic data. PeerJ. 2015;3:e985.

11. Yutin N, Makarova KS, Gussow AB, Krupovic M, Segall A, Edwards RA et al. Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. Nat Microbiol. 2018;3:38-46.

12. Shkoporov AN, Khokhlova EV, Fitzgerald CB, Stockdale SR, Draper LA, Ross RP et al. ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. Nat Commun. 2018;9:4781.

13. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S et al. CRISPR provides acquired resistance against viruses in prokaryotes. Science. 2007;315:1709-1712.

14. Seed KD, Lazinski DW, Calderwood SB, Camilli A. A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. Nature. 2013;494:489-491.

15. Andersson AF, Banfield JF. Virus population dynamics and acquired virus resistance in natural microbial communities. Science. 2008;320:1047-1050.

16. Snyder JC, Bateson MM, Lavin M, Young MJ. Use of Cellular CRISPR (Clusters of Regularly Interspaced Short Palindromic Repeats) Spacer-Based Microarrays for Detection of Viruses in Environmental Samples. Applied and Environmental Microbiology. 2010;76:7251-7258.

17. Zhang Q, Rho M, Tang H, Doak TG, Ye Y. CRISPR-Cas systems target a diverse collection of invasive mobile genetic elements in human microbiomes. Genome Biol. 2013;14:R40.

18. Shmakov SA, Sitnik V, Makarova KS, Wolf YI, Severinov KV, Koonin EV. The CRISPR spacer space is dominated by sequences from species-specific mobilomes. MBio. 2017;8:e01397-17.

19. Shmakov SA, Wolf YI, Savitskaya E, Severinov KV, Koonin EV. Mapping CRISPR spaceromes reveals vast host-specific viromes of prokaryotes. Commun Biol. 2020;3:321.

20. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N et al. Uncovering Earth's virome. Nature. 2016;536:425-430.

21. Stern A, Mick E, Tirosh I, Sagy O, Sorek R. CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. Genome Res. 2012;22:1985-1994.

22. Skennerton CT, Imelfort M, Tyson GW. Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. Nucleic Acids Res. 2013;41:e105.

23. Moller AG, Liang C. MetaCRAST: reference-guided extraction of CRISPR spacers from unassembled metagenomes. PeerJ. 2017;5:e3788.

24. STREISINGER G, EDGAR RS, DENHARDT GH. CHROMOSOME STRUCTURE IN PHAGE T4. I. CIRCULARITY OF THE LINKAGE MAP. Proc Natl Acad Sci U S A. 1964;51:775-779.

25. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H et al. The protein data bank. Nucleic acids research. 2000;28:235-242.

26. Cui J, Schlub TE, Holmes EC. An allometric relationship between the genome length and virion volume of viruses. J Virol. 2014;88:6403-6410.

27. Hua J, Huet A, Lopez CA, Toropova K, Pope WH, Duda RL et al. Capsids and Genomes of Jumbo-Sized Bacteriophages Reveal the Evolutionary Reach of the HK97 Fold. mBio. 2017;8

28. Koonin EV DVV, Krupovic M, Varsani A, Wolf YI, Yutin N, Zerbini M, Kuhn JH. Create a megataxonomic framework, filling all principal/primary taxonomic ranks, for dsDNA viruses encoding HK97-type major capsid proteins. 2019

29. Roux S, Krupovic M, Poulet A, Debroas D, Enault F. Evolution and diversity of the Microviridae viral family through a collection of 81 new complete genomes assembled from virome reads. PloS one. 2012;7:e40418.

30. Ackermann HW. Tailed bacteriophages: the order caudovirales. Adv Virus Res. 1998;51:135-201.

31. Ackermann HW. Phage classification and characterization. Methods Mol Biol. 2009;501:127-140.

32. Koonin EV DVV, Krupovic M, Varsani A, Wolf YI, Yutin N, Zerbini M, Kuhn JH. Create a megataxonomic framework, filling all principal taxonomic ranks, for DNA viruses encoding vertical jelly roll-type major capsid proteins. 2019

33. Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ et al. An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol. 2015;13:722-736.

34. Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R. Diversity, stability and resilience of the human gut microbiota. Nature. 2012;489:220-230.

35. Bailly-Bechet M, Vergassola M, Rocha E. Causes for the intriguing presence of tRNAs in phages. Genome Res. 2007;17:1486-1495.

36. Nayfach S, Páez-Espino D, Call L, Low SJ, Sberro H, Ivanova NN et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. Nature Microbiology. 2021;6:960-970.

37. Almpanis A, Swain M, Gatherer D, McEwan N. Correlation between bacterial G+C content, genome size and the G+C content of associated plasmids and bacteriophages. Microb Genom. 2018;4

38. Brentlinger KL, Hafenstein S, Novak CR, Fane BA, Borgon R, McKenna R et al. Microviridae, a family divided: isolation, characterization, and genome sequence of φMH2K, a bacteriophage of the obligate intracellular parasitic bacterium Bdellovibrio bacteriovorus. Journal of bacteriology. 2002;184:1089-1094.

39. Sharp PM, Hahn BH. Origins of HIV and the AIDS pandemic. Cold Spring Harb Perspect Med. 2011;1:a006841.

40. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. nature. 2020;579:270-273.

41. Shmakov S, Smargon A, Scott D, Cox D, Pyzocha N, Yan W et al. Diversity and evolution of class 2 CRISPR-Cas systems. Nat Rev Microbiol. 2017;15:169-182.

42. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2007;35:D61-5.

43. Paez-Espino D, Chen IA, Palaniappan K, Ratner A, Chu K, Szeto E et al. IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses. Nucleic Acids Res. 2017;45:D457-D465.

44. Paez-Espino D, Roux S, Chen IA, Palaniappan K, Ratner A, Chu K et al. IMG/VR v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes. Nucleic Acids Res. 2019;47:D678-D686.

45. Roux S, Krupovic M, Daly RA, Borges AL, Nayfach S, Schulz F et al. Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. Nat Microbiol. 2019;4:1895-1906.

46. Askora A, Abdel-Haliem MEF, Yamada T. Site-specific recombination systems in filamentous phages. Molecular Genetics and Genomics. 2012;287:525-530.

47. Kazlauskas D, Varsani A, Koonin EV, Krupovic M. Multiple origins of prokaryotic and eukaryotic single-stranded DNA viruses from bacterial and archaeal plasmids. Nat Commun. 2019;10:3425.

48. Roux S, Krupovic M. Create one new family (Paulinoviridae) including two genera moved from the family Inoviridae (Tubulavirales).

49. Roux S. Inovirus_classifier. https://githubcom/simroux/Inovirus/tree/master/Inovirus_classifier.

50. Deveau H, Barrangou R, Garneau JE, Labonté J, Fremaux C, Boyaval P et al. Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. Journal of bacteriology. 2008;190:1390-1400.

51. Fineran PC, Gerritzen MJ, Suárez-Diez M, Künne T, Boekhorst J, van Hijum SA et al. Degenerate target sites mediate rapid primed CRISPR adaptation. Proc Natl Acad Sci U S A. 2014;111:E1629-38.

52. Zhang X, Deatherage DE, Zheng H, Georgoulis SJ, Barrick JE. Evolution of satellite plasmids can prolong the maintenance of newly acquired accessory genes in bacteria. Nat Commun. 2019;10:5809.

53. Capel B, Swain A, Nicolis S, Hacker A, Walter M, Koopman P et al. Circular transcripts of the testis-determining gene Sry in adult mouse testis. Cell. 1993;73:1019-1030.

54. Ashwal-Fluss R, Meyer M, Pamudurti NR, Ivanov A, Bartok O, Hanan M et al. circRNA biogenesis competes with pre-mRNA splicing. Mol Cell. 2014;56:55-66.

55. Krupovic M, Bamford DH, Koonin EV. Conservation of major and minor jelly-roll capsid proteins in Polinton (Maverick) transposons suggests that they are bona fide viruses. Biology Direct. 2014;9:6.

56. San Martín C, van Raaij MJ. The so far farthest reaches of the double jelly roll capsid protein fold. Virol J. 2018;15:181.

57. Yutin N, Bäckström D, Ettema TJG, Krupovic M, Koonin EV. Vast diversity of prokaryotic virus genomes encoding double jelly-roll major capsid proteins uncovered by genomic and metagenomic sequence analysis. Virol J. 2018;15:67.

58. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T et al. Improved protein structure prediction using potentials from deep learning. Nature. 2020;577:706-710.

59. Watson BNJ, Vercoe RB, Salmond GPC, Westra ER, Staals RHJ, Fineran PC. Type IF CRISPR-Cas resistance against virulent phages results in abortive infection and provides population-level immunity. Nature communications. 2019;10:1-8.

60. Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. Nat Commun. 2018;9:2542.

61. Bushnell B. BBTools software package. URL http://sourceforge net/projects/bbmap. 2014

62. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology. 2012;19:455-477.

63. Biswas A, Staals RHJ, Morales SE, Fineran PC, Brown CM. CRISPRDetect: a flexible algorithm to define CRISPR arrays. BMC genomics. 2016;17:356.

64. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22:1658-1659.

65. Van Dongen SM. Graph clustering by flow simulation [dissertation]. 2000.

66. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841-842.

67. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25:1422-1423.

68. Steinegger M, Söding J. Linclust: clustering billions of protein sequences per day on a single server. bioRxiv. 2017104034.

69. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature biotechnology. 2017;35:1026-1028.

70. Eddy SR, team HMMERD. HMMER User's Guide. 2019

71. Laslett D, Canback B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. 2004;32:11-16.

72. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016;32:2847-2849.

73. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic acids research. 2002;30:3059-3066.

74. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972-1973.

75. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 2001;17:754-755.

76. Andrew R. FigTree. https://githubcom/rambaut/figtree/.

77. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14:1188-1190.

78. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. PLoS computational biology. 2018;14:e1005944.

79. Edgar RC, Taylor J, Lin V, Altman T, Barbera P, Meleshko D et al. Petabase-scale sequence alignment catalyses viral discovery. Nature. 2022;602:142-147.

80. Krishnamurthy SR, Wang D. Extensive conservation of prokaryotic ribosomal binding sites in known and novel picobirnaviruses. Virology. 2018;516:108-114.

81. Ghosh S, Malik YS. The True Host/s of Picobirnaviruses. Front Vet Sci. 2020;7:615293.

82. Integrative HMPRNC. The Integrative Human Microbiome Project. Nature. 2019;569:641-648.

83. Aiewsakun P, Katzourakis A. Time-Dependent Rate Phenomenon in Viruses. J Virol. 2016;90:7184-7195.

84. Duchêne S, Holmes EC, Ho SY. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. Proc Biol Sci. 2014;281:20140732.

85. Ho SY, Lanfear R, Bromham L, Phillips MJ, Soubrier J, Rodrigo AG et al. Time-dependent rates of molecular evolution. Mol Ecol. 2011;20:3087-3101.

86. Ghafari M, Simmonds P, Pybus OG, Katzourakis A. A mechanistic evolutionary model explains the time-dependent pattern of substitution rates in viruses. Curr Biol. 2021;31:4689-4696.e5.

87. Nishimura L, Sugimoto R, Inoue J, Nakaoka H, Kanzawa-Kiriyama H, Shinoda KI et al. Identification of ancient viruses from metagenomic data of the Jomon people. J Hum Genet. 2021;66:287-296.

88. Singh D, Yi SV. On the origin and evolution of SARS-CoV-2. Exp Mol Med. 2021;53:537-547.

89. Chaw SM, Tai JH, Chen SL, Hsieh CH, Chang SY, Yeh SH et al. The origin and underlying driving forces of the SARS-CoV-2 outbreak. J Biomed Sci. 2020;27:73.

90. Li X, Zai J, Zhao Q, Nie Q, Li Y, Foley BT et al. Evolutionary history, potential intermediate animal host, and cross-species analyses of SARS-CoV-2. J Med Virol. 2020;92:602-611.

91. Duchene S, Featherstone L, Haritopoulou-Sinanidou M, Rambaut A, Lemey P, Baele G. Temporal signal and the phylodynamic threshold of SARS-CoV-2. Virus Evol. 2020;6:veaa061.

92. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27:764-770.

93. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004;5:113.

94. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 2006;34:W609-12.