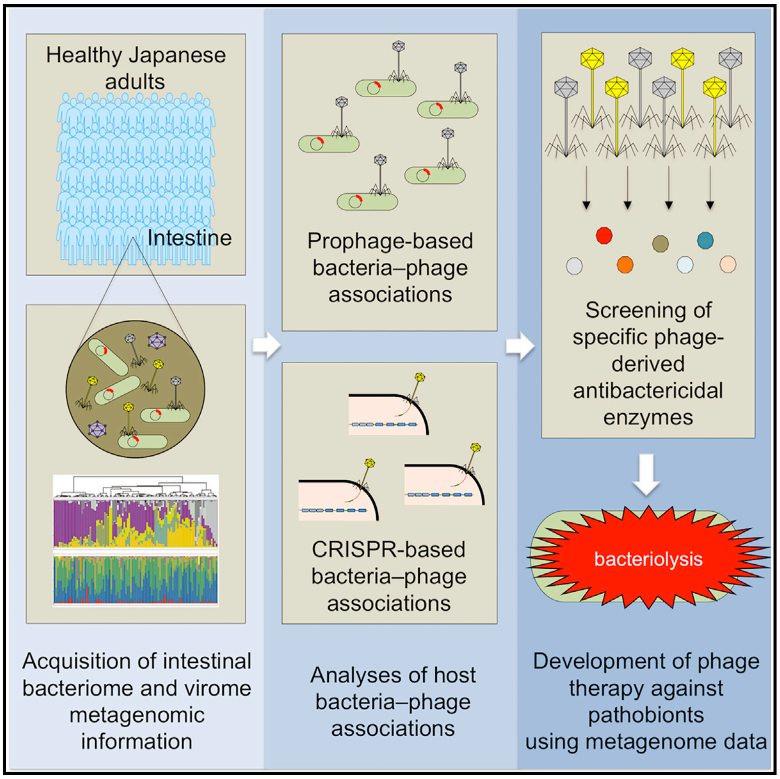
Short Article

Cell Host & Microbe

Metagenome Data on Intestinal Phage-Bacteria Associations Aids the Development of Phage Therapy against Pathobionts

Graphical Abstract


Healthy Japanese adults
Intestine
Acquisition of intestinal bacteriome and virome metagenomic information
Prophage-based bacteria–phage associations
CRISPR-based bacteria–phage associations
Analyses of host bacteria–phage associations
Screening of specific phage-derived antibactericidal enzymes
bacteriolysis
Development of phage therapy against pathobionts using metagenome data

Highlights

- The virome (phages) and bacteriome of 101 healthy Japanese adults were reported
- Host bacteria-phage associations are illustrated in both temperate and virulent phages
- Metagenomic data identify *C. difficile*-specific phage-derived antibacterial enzymes
- Phage-derived endolysins specifically control *C. difficile* infection in mice

Authors

Kosuke Fujimoto, Yasumasa Kimura, Masaki Shimohigoshi, ..., Hiroshi Kiyono, Seiya Imoto, Satoshi Uematsu

Correspondence

imoto@ims.u-tokyo.ac.jp (S.I.), uematsu.satoshi@med.osaka-cu.ac.jp (S.U.)

In Brief

Fujimoto et al. report intestinal bacterial and viral metagenome information from the fecal samples of 101 healthy Japanese individuals. This analysis leverages the determined host bacteria-phage associations to detect phage-derived antibacterial enzymes that specifically control pathobionts. As proof of concept, phage-derived endolysins are shown to control *C. difficile* infection in mice.

Fujimoto et al., 2020, Cell Host & Microbe *28*, 380–389
September 9, 2020 © 2020 Elsevier Inc.
https://doi.org/10.1016/j.chom.2020.06.005

CellPress

Short Article

Metagenome Data on Intestinal Phage-Bacteria Associations Aids the Development of Phage Therapy against Pathobionts

Kosuke Fujimoto,[1,2,3,18] Yasumasa Kimura,[4,18] Masaki Shimohigoshi,[1,18] Takeshi Satoh,[4] Shintaro Sato,[1,5] Georg Tremmel,[6] Miho Uematsu,[1] Yunosuke Kawaguchi,[1] Yuki Usui,[4] Yoshiko Nakano,[1] Tetsuya Hayashi,[1] Koji Kashima,[7] Yoshikazu Yuki,[7] Kiyoshi Yamaguchi,[8] Yoichi Furukawa,[8] Masanori Kakuta,[6] Yutaka Akiyama,[9] Rui Yamaguchi,[6] Sheila E. Crowe,[10] Peter B. Ernst,[11,12,13] Satoru Miyano,[6] Hiroshi Kiyono,[11,12,14,15] Seiya Imoto,[16,17,*] and Satoshi Uematsu[1,2,3,17,19,*]

[1]Department of Immunology and Genomics, Osaka City University Graduate School of Medicine, Osaka 545-8585, Japan
[2]Division of Metagenome Medicine, Human Genome Center, the Institute of Medical Sciences, the University of Tokyo, Tokyo 108-8639, Japan
[3]Division of Innate Immune Regulation, International Research and Development Center for Mucosal Vaccines, the Institute of Medical Sciences, the University of Tokyo, Tokyo 108-8639, Japan
[4]Division of Systems Immunology, the Institute of Medical Sciences, the University of Tokyo, Tokyo 108-8639, Japan
[5]Mucosal Vaccine Project, BIKEN Innovative Vaccine Research Alliance Laboratories, Research Institute for Microbial Diseases, Osaka University, Osaka 565-0871, Japan
[6]Laboratory of DNA Information Analysis, Human Genome Center, the Institute of Medical Sciences, the University of Tokyo, Tokyo 108-8639, Japan
[7]Division of Mucosal Immunology, Department of Microbiology and Immunology, the Institute of Medical Sciences, the University of Tokyo, Tokyo 108-8639, Japan
[8]Division of Clinical Genome Research, the Institute of Medical Sciences, the University of Tokyo, Tokyo 108-8639, Japan
[9]Department of Computer Science, Tokyo Institute of Technology, Tokyo 152-8550, Japan
[10]Department of Medicine, University of California, San Diego, La Jolla, CA 92093, USA
[11]Division of Gastroenterology, Department of Medicine, CU-UCSD Center for Mucosal Immunology, Allergy and Vaccines, University of California, San Diego, La Jolla, CA 92093, USA
[12]Division of Comparative Pathology and Medicine, Department of Pathology, University of California, San Diego, La Jolla, CA 92093, USA
[13]Center for Veterinary Sciences and Comparative Medicine, University of California, San Diego, La Jolla, CA 92093, USA
[14]Department of Mucosal Immunology, IMSUT Distinguished Professor Unit, The Institute of Medical Sciences, The University of Tokyo, Tokyo 108-8639, Japan
[15]International Research and Development Center for Mucosal Vaccines, the Institute of Medical Sciences, the University of Tokyo, Tokyo 108-8639, Japan
[16]Division of Health Medical Intelligence, Human Genome Center, The Institute of Medical Sciences, the University of Tokyo, Tokyo 108-8639, Japan
[17]Collaborative Research Institute for Innovative Microbiology, The University of Tokyo, Tokyo 113-8657, Japan
[18]These authors contributed equally
[19]Lead Contact
*Correspondence: imoto@ims.u-tokyo.ac.jp (S.I.), uematsu.satoshi@med.osaka-cu.ac.jp (S.U.)
https://doi.org/10.1016/j.chom.2020.06.005

SUMMARY

The application of bacteriophages (phages) is proposed as a highly specific therapy for intestinal pathobiont elimination. However, the infectious associations between phages and bacteria in the human intestine, which is essential information for the development of phage therapies, have yet to be fully elucidated. Here, we report the intestinal viral microbiomes (viromes), together with bacterial microbiomes (bacteriomes), in 101 healthy Japanese individuals. Based on the genomic sequences of bacteriomes and viromes from the same fecal samples, the host bacteria-phage associations are illustrated for both temperate and virulent phages. To verify the usefulness of the comprehensive host bacteria-phage information, we screened *Clostridioides difficile*-specific phages and identified antibacterial enzymes whose activity is confirmed both *in vitro* and *in vivo*. These comprehensive metagenome analyses reveal not only host bacteria-phage associations in the human intestine but also provide vital information for the development of phage therapies against intestinal pathobionts.

INTRODUCTION

With advances in intestinal microbial analyses, abnormalities in the human intestinal microflora, known as dysbiosis, have been implicated in various diseases (Belkaid and Hand, 2014; Cho and Blaser, 2012; Gilbert et al., 2016; Lynch and Pedersen, 2016; Rooks and Garrett, 2016). Altered microbial diversity and microbial substitution are frequently observed in dysbiosis, resulting in impairment of the beneficial effects of intestinal microflora on the host and disruption of homeostasis. Under dysbiosis, some symbiotic commensal bacteria acquire virulence traits, proliferate, and become directly involved in the development and progression of disease (Kamada et al., 2012). These bacteria are referred to as "pathobionts," which are distinct from opportunistic pathogens. Recently, characteristic pathobionts, such as adherent invasive *Escherichia coli* in Crohn's disease, *Clostridioides difficile* in pseudomembranous colitis, *Prevotella copri* in rheumatoid arthritis, and *Clostridium ramosum* in obesity and diabetes mellitus, have been identified (Barrios-Villa et al., 2020; Fujimoto et al., 2019; Karlsson et al., 2013; Le Chatelier et al., 2013; Orenstein and Patron, 2019; Scher et al., 2013). Interestingly, *C. ramosum*-specific secretory immunoglobulin A (sIgA) induced by mucosal vaccine prevented obesity and subsequent diabetes in a high-fat-diet-induced obesity model using humanized gnotobiotic mice, suggesting that specific elimination of the pathobiont is a potential treatment for the disease (Fujimoto et al., 2019). If the commensal bacteria were clearly demonstrated to have a causal relationship to the disease development, their elimination would be desirable for disease prevention similar to *Helicobacter pylori* related to stomach ulcers. However, antibiotic usage has the risk of killing beneficial bacteria and promoting dysbiosis. Therefore, the development of methods to specifically manipulate intestinal pathobionts is essential for the treatment of dysbiosis-related diseases. Bacteriophage (phage) therapy is proposed as an alternative treatment for pathobiont elimination because phages possess high specificity to their host bacteria (Mirzaei and Maurice, 2017). There are two modes of replication for phages: the lytic cycle and the lysogenic cycle. In the lysogenic cycle of temperate phages, the phage genome has a greater likelihood of integrating into the bacterial chromosome as a prophage. Even though both types of phage can infect host bacteria, lytic phages are preferable for phage therapy because temperate phages have an inherent ability to mediate the transfer of genes between bacteria. However, it is sometimes challenging to isolate lytic phage specific for anaerobic intestinal bacteria due to difficulties with culturing. In such cases, next-generation phage therapy using phage-derived enzymes, or the generation of artificial phages is considered. Vast numbers of viruses cohabit the human intestine along with bacteria and a major fraction of the intestinal virome is composed of bacteriophages, whose hosts are commensal bacteria (Guerin et al., 2018; Manrique et al., 2016; Minot et al., 2011, 2013; Reyes et al., 2010; Shkoporov et al., 2019; Yutin et al., 2018). Although recently improved sequencing methods have yielded more detailed information on gut viromes (Shkoporov et al., 2019), host bacteria-phage associations in the human intestine are not yet fully understood. To develop future phage therapies specific for viable but non-culturable pathobionts in the intestine, the acquisition of comprehensive metagenomic information about intestinal phages and their hosts is essential.

Here, we obtained shotgun metagenomic sequence data about both the virome and bacteriome of 101 healthy individuals and analyzed the host bacteria-phage associations. Further focusing on *C. difficile* infection (CDI) as a representative dysbiosis-related disease, we identified novel antibacterial enzymes derived from *C. difficile*-specific phages that we used to screen our metagenome data. Our results suggested that the information about host bacteria-phage associations derived from metagenome analysis is useful for the development of phage therapies against intestinal pathobionts.

RESULTS

Acquisition of Metagenomic Information from Healthy Individuals

To clarify the host bacteria-phage associations in the intestine, intestinal viruses and bacteria were metagenomically analyzed in the same feces from 101 healthy Japanese adults. We developed a metagenome analysis pipeline (Figures S1A–S1E) and examined the relative abundance of viruses and bacteria in the feces of healthy Japanese individuals based on metagenomic data obtained by our analysis pipeline (Figure 1). Interestingly, *Microviridae*-abundant, *Caudovirales* including *Siphoviridae*, *Myoviridae*, *Podoviridae*, and uc_Caudovirales-abundant and crAss-like phages including p-crAssphage-abundant samples were found, suggesting that the viral relative abundance varied across individuals. Collectively, we were able to acquire both bacterial and viral metagenomic information from the same feces samples collected from 101 healthy adults.

Prophage-Based Host Bacteria-Phage Associations

We next analyzed the host bacteria-phage associations in the gut. We comprehensively searched for prophage sequences using either bacterial sequence or viral sequence data (Figure 2A). Among the four representative intestinal bacterial phyla, including Firmicutes, Bacteroidetes, Proteobacteria, and Actinobacteria, the number of identified prophages was highest in Firmicutes (Figure 2B). We especially focused on host bacteria-phage associations detected by the circular viral contigs since such phages may initiate the active lytic cycle releasing viral particles from host bacteria. In total, 114 prophages corresponding to circular viral contigs were successfully identified. For example, the whole sequence of a circular viral contig classified as uc_Caudovirales (57,169 nt in length) was found as a prophage sequence (99.997% identity) in a bacterial contig classified as *Bacteroides uniformis* (615,322 nt in length) (Figure 2C). We also confirmed the presence of a phage integrase gene located at the end of the prophage sequence (Figures 2C, S2A, and S2B). Among the 114 prophages associated with circular viral contigs, 61, 46, 5, and 2 prophages were found in Firmicutes, Bacteroidetes, Proteobacteria, and Actinobacteria, respectively (Figure 2D). Similar to previous studies (Krupovic and Forterre, 2011; Reyes et al., 2012), the phages classified as *Caudovirales* showed major interactions with bacteria. Our analysis also revealed that *Microviridae* sequences were integrated in the genomes of Firmicutes, Bacteroidetes, and Proteobacteria (Figure 2D). Interestingly, one prophage related to crAss-like phage was detected in Firmicutes (Figures

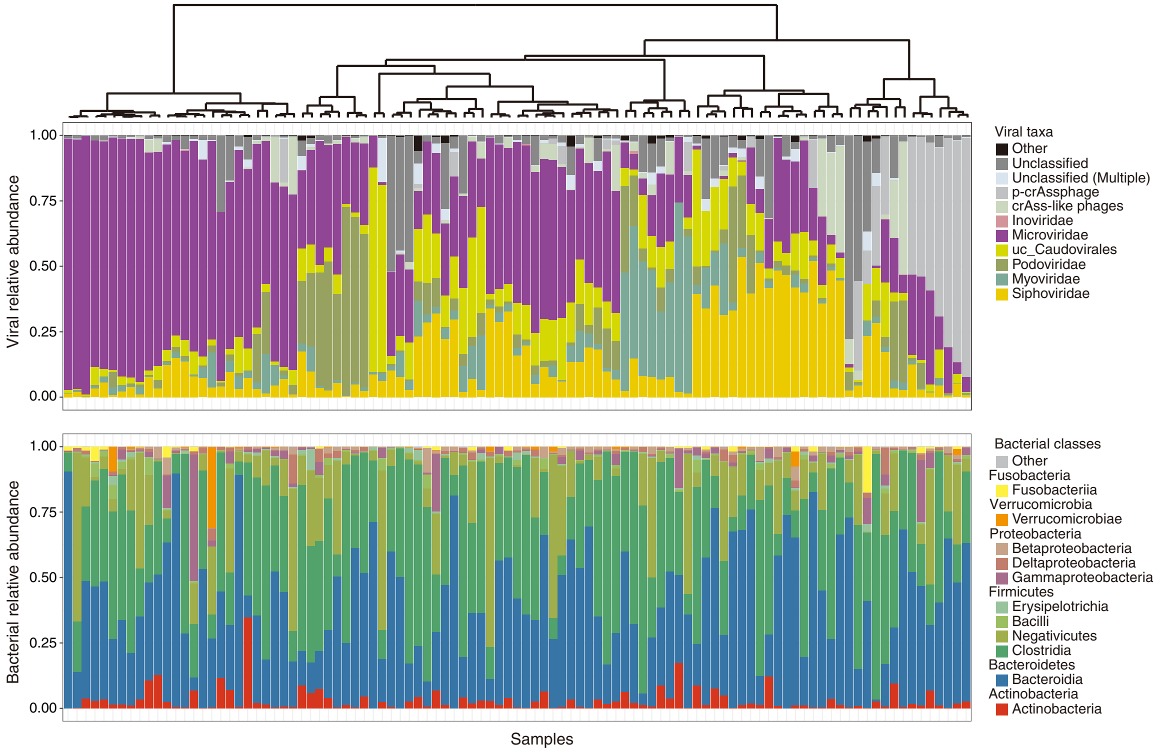


Figure 1. Clustering Analysis of the Virome and Bacteriome Data

Relative abundances of viral taxa (top) and bacterial classes (bottom). Hierarchical clustering (Euclidean distance and Ward linkage) with virome compositions on the top was used.

2D and S2C). The crAss-like phage contig was 159 kb in length (much longer than the p-crAssphage genome of 97 kb) but possessed a DNA polymerase crAss-like phage marker (Figure S2C). By contrast, p-crAssphage was not detected as a prophage (Figure 2D) despite the fact that there were p-crAssphage-rich populations among the healthy individuals (Figure 1). To date, using only the bacteriome, it has proven difficult to identify prophage-based host bacteria-phage associations due to insufficient viral databases. Here, by analyzing the bacteriome and virome from the same fecal samples, we could identify prophage-based host bacteria-phage associations effectively.

CRISPR Spacer-Based Host Bacteria-Phage Associations

Since the CRISPR loci together with CRISPR-associated (*cas*) genes comprise a bacterial adaptive defense system against phages (Barrangou et al., 2007; Makarova et al., 2011), CRISPR spacers are considered as records of past infections and can be used to investigate infectious associations between gut bacteria and temperate phages as well as virulent phages. We inferred CRISPR repeats and spacers on the bacterial contigs using CRISPRDetect (Biswas et al., 2016) (Figures 3A and S3). About 20% of the spacers (16,145) were aligned to the viral contigs and 36%, 9%, 5%, 35%, 9%, and 0.3% of these were associated with *Siphoviridae*, *Myoviridae*, *Podoviridae*, uc_Caudovirales, *Microviridae*, and crAss-like phage contigs, respectively (Figure 3B, left). There were 19 spacers associated with multiple viral groups, which corresponded to the sequence fragments common in different viral groups: *Siphoviridae* and crAss-like phages (5%), *Myoviridae* and crAss-like phages (11%), uc_Caudovirales and crAss-like phages (42%), *Siphoviridae* and *Microviridae* (26%), uc_Caudovirales and *Microviridae* (5%), and crAss-like phages and p-crAssphage (11%) (Figure 3B, right). We next elucidated which bacterial phyla were linked to the spacer-target phages. Whereas the spacers against *Caudovirales* including *Siphoviridae*, *Myoviridae*, *Podoviridae*, and uc_Caudovirales, were predominantly from Firmicutes (68%, 73%, 71%, and 65%, respectively), the spacers against *Microviridae* were mainly from Bacteroidetes (62%) (Figure 3C). In the case of the spacers against crAss-like phages, 75% and 24% of them were from Bacteroidetes and Firmicutes, respectively. Interestingly, all of the CRISPR spacers against p-crAssphage were from Bacteroidetes.

We next aligned the spacers to the sequences of ORFs on the viral contigs to identify and then annotate the target ORFs (Figure 3A). In the case of *Caudovirales* including *Siphoviridae*, *Myoviridae*, *Podoviridae*, and uc_Caudovirales, phage structural protein genes, such as the phage tail, phage capsid, and phage portal, and enzymatic genes were the most frequent target gene categories of CRISPR spacers (Figure 3D). By contrast, the spacers against *Microviridae* targeted the genes of replication protein VP4, major capsid protein VP1, and DNA pilot protein VP5, all of which are characteristic genes of *Microviridae* (Figure 3D). Although most of the target ORFs were not annotated in crAss-like phages, four enzymatic genes including the

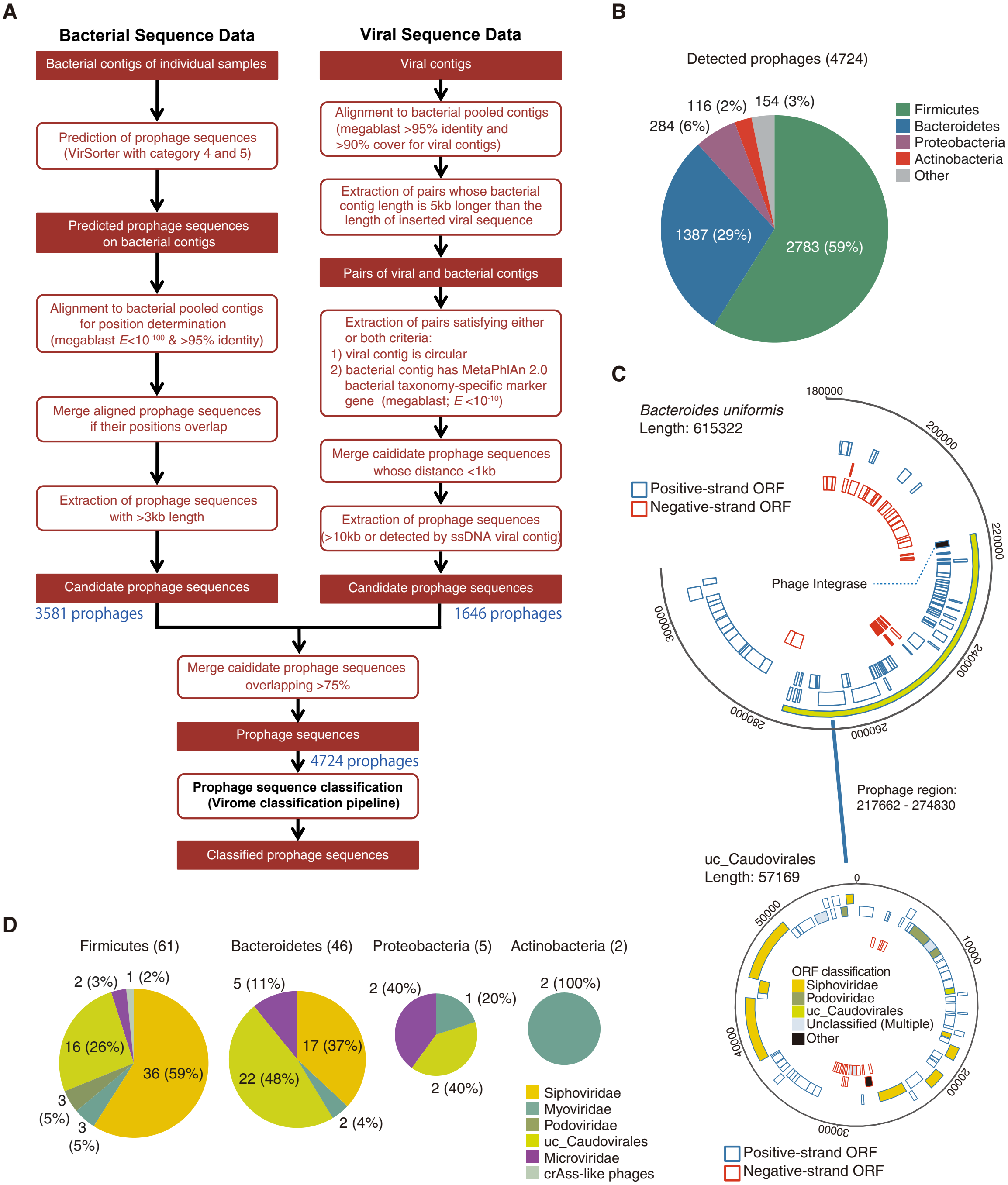

A
Bacterial Sequence Data
Bacterial contigs of individual samples
Prediction of prophage sequences (VirSorter with category 4 and 5)
Predicted prophage sequences on bacterial contigs
Alignment to bacterial pooled contigs for position determination (megablast E<10^-100 & >95% identity)
Merge aligned prophage sequences if their positions overlap
Extraction of prophage sequences with >3kb length
Candidate prophage sequences
3581 prophages
Viral Sequence Data
Viral contigs
Alignment to bacterial pooled contigs (megablast >95% identity and >90% cover for viral contigs)
Extraction of pairs whose bacterial contig length is 5kb longer than the length of inserted viral sequence
Pairs of viral and bacterial contigs
Extraction of pairs satisfying either or both criteria:
1) viral contig is circular
2) bacterial contig has MetaPhlAn 2.0 bacterial taxonomy-specific marker gene (megablast; E<10^-10)
Merge caididate prophage sequences whose distance <1kb
Extraction of prophage sequences (>10kb or detected by ssDNA viral contig)
Candidate prophage sequences
1646 prophages
Merge caididate prophage sequences overlapping >75%
Prophage sequences
4724 prophages
Prophage sequence classification (Virome classification pipeline)
Classified prophage sequences
B
Detected prophages (4724)
116 (2%)
154 (3%)
284 (6%)
1387 (29%)
2783 (59%)
Firmicutes
Bacteroidetes
Proteobacteria
Actinobacteria
Other
C
Bacteroides uniformis
Length: 615322
Positive-strand ORF
Negative-strand ORF
180000
200000
220000
240000
260000
280000
300000
Phage Integrase
Prophage region:
217662 - 274830
uc_Caudovirales
Length: 57169
0
10000
20000
30000
40000
50000
ORF classification
Siphoviridae
Podoviridae
uc_Caudovirales
Unclassified (Multiple)
Other
Positive-strand ORF
Negative-strand ORF
D
Firmicutes (61)
2 (3%)
1 (2%)
16 (26%)
36 (59%)
3 (5%)
3 (5%)
Bacteroidetes (46)
5 (11%)
17 (37%)
22 (48%)
2 (4%)
Proteobacteria (5)
2 (40%)
1 (20%)
2 (40%)
Actinobacteria (2)
2 (100%)
Siphoviridae
Myoviridae
Podoviridae
uc_Caudovirales
Microviridae
crAss-like phages

Figure 2. Analysis of Prophages

(A) Prophage detection pipeline.

(B) Viral taxonomic distribution of prophages detected from the bacterial contigs.

(C) Bacterial contig, in which the whole sequence of a circular viral contig was found as a prophage.

(D) Viral taxonomic distribution of activated prophages (whole sequence of the circular viral contigs) for four bacterial phyla.

peptidase, methylase, helicase, and DNA ligase were identified as the targets (Figure 3D). Thus, analyzing the CRISPR spacers on bacterial contigs contributes toward our understanding of the complex phage-bacteria associations. All these data suggested that the gut microbiota is formed as a result of multiple, complex interactions between the host bacteria and phages.

Screening of Phage-Derived Antibacterial Enzymes Based on Metagenome Data

We further investigated whether the information about host bacteria-phage associations obtained by metagenome analysis could be used for the detection of pathobiont-specific phages or antibacterial agents that may be useful for eliminating

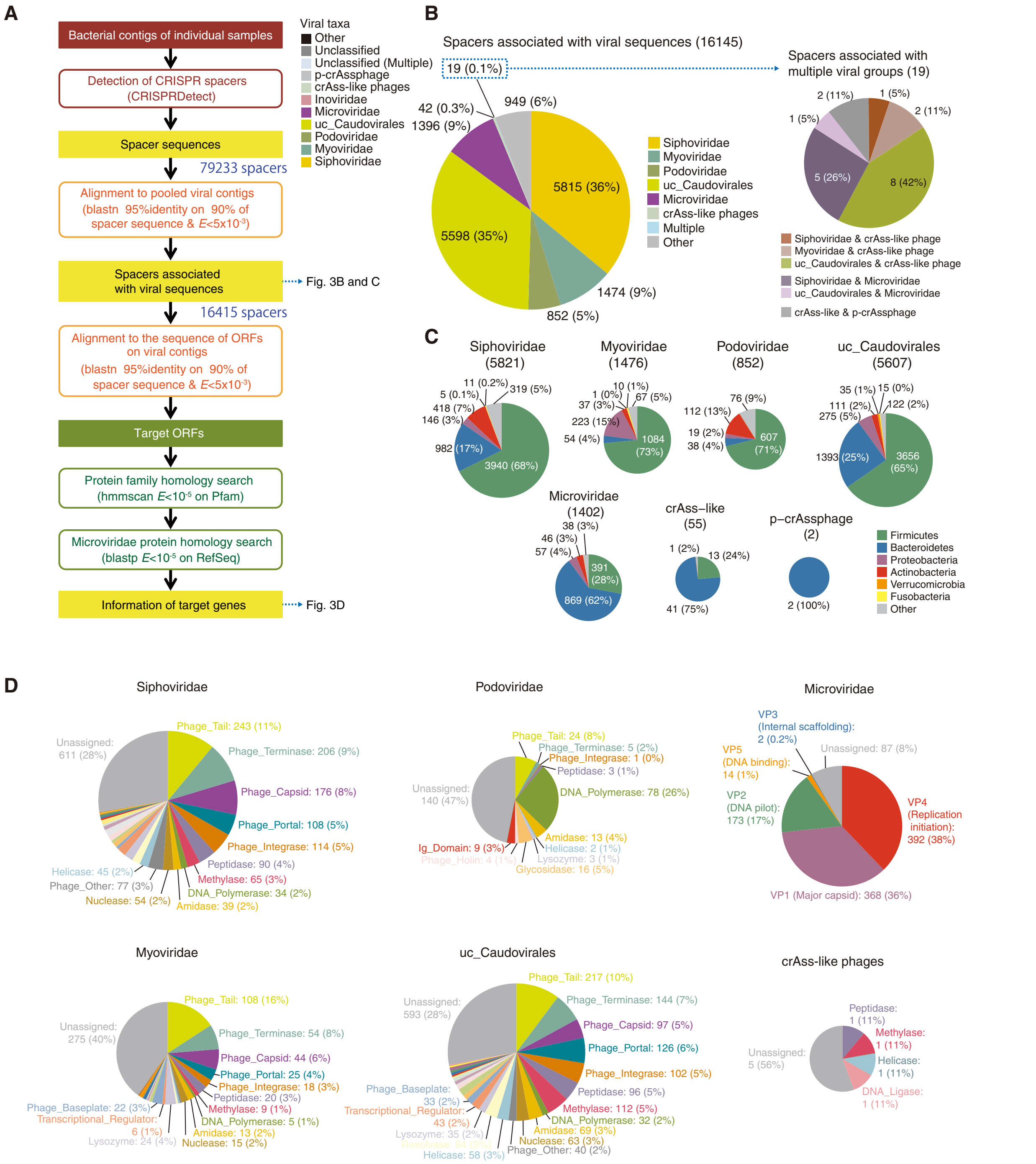


Figure 3. Analysis of the Infectious History Based on the CRISPR Spacers

(A) CRISPR spacer detection pipeline.

(B) Distribution of CRISPR spacers associated with viral sequences (left). Distribution of CRISPR spacers associated with multiple viral groups (right).

(C) Distribution of bacterial phyla with CRISPR spacers related to *Siphoviridae*, *Myoviridae*, *Podoviridae*, uc_Caudovirales, *Microviridae*, crAss-like phage, and crAssphage.

(D) Distribution of viral gene categories targeted by CRISPR spacers in *Siphoviridae*, *Myoviridae*, *Podoviridae*, uc_Caudovirales, *Microviridae*, and crass-like phage.

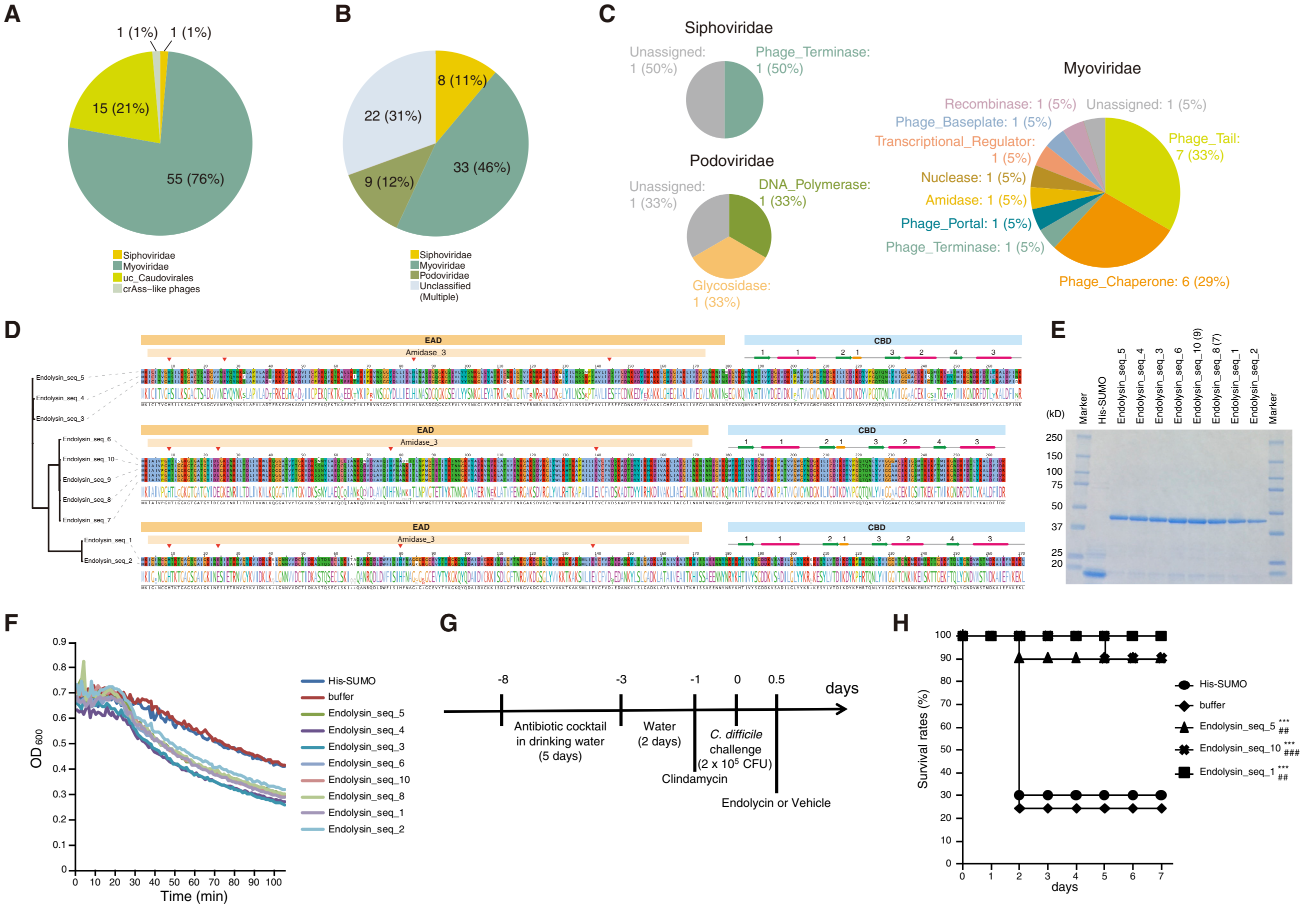


Figure 4. Analysis of *C. difficile* Phage-Derived Endolysins

(A) Viral taxonomic distribution of prophages detected from the *C. difficile* contigs in 101 healthy individuals and 17 cultured *C. difficile* strains.

(B) Distribution of the *C. difficile* CRISPR spacers associated with viral sequences in 101 healthy individuals and 17 cultured *C. difficile* strains.

(C) Distribution of viral gene categories targeted by CRISPR spacers in *Siphoviridae*, *Myoviridae*, and *Podoviridae*.

(D) Multiple alignment of the detected *C. difficile* endolysin sequences containing Pfam annotation Amidase_3 and the CBD. The aligned amino acid sequences (top) and their conservation scores (middle) and consensus sequences (bottom) are shown.

(E) Expression and purification of endolysin. A representative image from two independent experiments is shown.

(F) Bacteriolytic capacity of endolysins against *C. difficile* strain ATCC 43255. A representative experiment from two independent experiments is shown.

(G) Experimental design schematics for the acute *C. difficile* murine infection model.

(H) Kaplan–Meier survival plots of antibiotic-treated mice exposed to *C. difficile* (2×10^5 CFU on day 0) and treated with the indicated endolysin or vehicle including His-SUMO or buffer (50 mM Na_3PO_4, 0.15 M NaCl, pH 7.0) (day 0.5). His-SUMO, Endolysin_seq_5, Endolysin_seq_10, and Endolysin_seq_1: n = 10 mice/group, buffer: n = 20 mice/group. The p values are derived from log-rank tests; $^{***}p < 0.001$ (versus buffer), $^{\#\#}p < 0.01$ and $^{\#\#\#}p < 0.001$ (versus His-SUMO).

pathobionts. *C. difficile* is a Gram-positive, spore-forming anaerobic bacterium that is the representative cause of nosocomial diarrhea following antibiotic treatment. Acute CDI is treated by antibiotics, such as metronidazole, vancomycin, and fidaxomicin. However, antibiotic treatment carries the risk of dysbiosis exacerbation due to the simultaneous killing of beneficial bacteria. In addition, it provides an opportunity for the emergence of antibiotic-resistant *C. difficile*, which is closely related with the high recurrence rates (20%–30%) of CDI (Lessa et al., 2015). Therefore, the development of alternative therapies for CDI is a crucial task. We therefore analyzed *C. difficile* contigs from 101 metagenomes of healthy Japanese individuals and the sequence data from 17 clinical isolates of *C. difficile* and extracted 72 prophage sequences. Among them, 55, 15, 1, and 1 prophages were classified as *Myoviridae*, uc_Caudovirales, *Siphoviridae*, and crAss-like phage, respectively (Figure 4A). We next examined the phage taxa targeted by the CRISPR spacers detected on the *C. difficile* contigs. Consistent with the prophage analysis, *Caudovirales*, such as *Siphoviridae*, *Myoviridae*, and *Podoviridae*, were their main targets (Figure 4B). We further showed the viral gene categories of *Myoviridae*, *Podoviridae*, and *Siphoviridae* targeted by the spacers (Figure 4C). Interestingly, *C. difficile* seems to acquire resistance especially against *Myoviridae* by incorporating various gene sequences as spacers (Figure 4C).

A number of temperate phages that infect *C. difficile* have been identified and most belonged to either the *Myoviridae* or *Siphoviridae* families (Ackermann and Prangishvili, 2012). Interestingly, purely lytic phages suitable for phage therapy have not been identified to date (Monteiro et al., 2019). However, usage of the bactericidal machinery from *C. difficile*-specific phages is considered to be a promising strategy to control the infection. Endolysins, which consist of two domains, an enzymatically active domain (EAD) and a cell wall-binding domain (CBD) that recognizes specific cell surface features, can specifically hydrolyze peptidoglycan (Nelson et al., 2012; Pohane and Jain, 2015). Endolysins are also reported to be effective bactericidal enzymes (Love et al., 2018; Roach and Donovan, 2015). We then identified 10 open reading frames ORFs homologous to known endolysins from the detected prophage regions in the *C. difficile* contigs (Figure 4D). All of the identified endolysins had Amidase_3 family (PF01520) in their EADs (Figure 4D) and all four known active sites were conserved (Figure 4D, red triangles). The endolysin CD27L is a well-characterized endolysin of *Clostridium* virus phiCD27. It possesses a $\beta\alpha\beta\beta\alpha\beta\alpha$ repeat structure in its CBD and lyses *C. difficile* strain ATCC 43255 (Chen et al., 2008; Dunne et al., 2014). Although no Pfam family was annotated on the CBDs of the identified endolysins, multiple sequence alignments along with CD27L revealed high conservation in the CBD region (Dunne et al., 2014) (Figure S4). Examining the conservation of the identified endolysins confirmed high similarity among the endolysin sequences, suggesting the existence of sequence patterns of functional endolysins against *C. difficile* (Figure 4D). We generated His-SUMO-tagged endolysins (Figure 4E), all of which had effective lytic activities *in vitro* (Figure 4F). Furthermore, His-tagged-Endolysin_seq_5, Endolysin_seq_10, and Endolysin_seq_1, which were representative of three lengths of amino acid sequences in Figure 4D, were effective against fatal CDI infection (Chen et al., 2008) (Figures 4G and 4H). Taken together, the endolysins detected in the *C. difficile* prophage sequences from the metagenome data have potential clinical applications for CDI.

DISCUSSION

Here, we obtained intestinal bacterial and viral metagenome information from the fecal samples of 101 healthy individuals through the development of a virome analysis pipeline. By examining the bacterial and viral genomic sequences from the same fecal samples, prophage-based host bacteria-phage associations were detected effectively. In addition, the analysis of CRISPR spacers on bacterial contigs elucidated the potential bacterial hosts of the intestinal phages, including virulent phages. Based on this information about host bacteria-phage associations, we screened *C. difficile*-specific phages and identified novel antibacterial enzymes whose activity was confirmed both *in vitro* and *in vivo*. Thus, our series of analyses revealed host bacteria-phage associations in the human intestine and provided crucial information for the development of phage therapies against intestinal pathobionts.

Multidrug-resistant virulent *C. difficile* strains have been reported (Carman et al., 2018; Li et al., 2019; Spigaglia et al., 2018); therefore, the development of alternative therapies, such as a phage therapy, for CDI is needed. Since *C. difficile* is a Gram-positive bacterium, endolysins were expected to be effective for CDI. In this study, we focused on lysogenic phages, the hosts of which were identified by the analysis of prophage-based host bacteria-phage associations. Such phages, which can be detected not only in bacterial genomes as prophages but also in the viral fraction as viral particles, were considered to possess lytic modes within their life cycles. Therefore, antibacterial enzymes identified from their genomes are potentially useful for lysing their specific host bacteria. In fact, we identified 10 new applicable endolysins for CDI from *C. difficile*-specific phages detected among our metagenome data (Figure 4D). The identified endolysins could effectively lyse high-level toxin-producing *C. difficile* strains *in vitro* (Figure 4F) and protect against fatal CDI infection (Figure 4H). This method can be applied to other intestinal pathobionts, which have prophage sequences in their genomes and for which the corresponding phages exist as particles in the feces.

In summary, metagenome analysis-based information about host bacteria-phage associations is useful for the detection of phage-derived antibacterial enzymes that specifically control pathobionts. Such information can also prove valuable when engineering phages against pathobionts, such as searching for stronger enzymes or modifying tail fibers. The accumulation of more metagenomic information on intestinal phages and bacteria will open up the possibility of developing treatments for a variety of dysbiosis-related diseases.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead Contact
  - ○ Materials Availability
  - ○ Data and Code Availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Human Subjects
  - ○ Mice
- METHOD DETAILS
  - ○ Human Fecal Sample Fractionation
  - ○ Treatment of the Viral Fraction
  - ○ Treatment of the Bacterial Fraction
  - ○ DNA Extraction
  - ○ Bacterial DNA Sequencing
  - ○ Viral DNA Sequencing
  - ○ Sequenced Data Processing
  - ○ Comparison of Assemblers
  - ○ Metagenome Assembly
  - ○ Extraction of Viral Contigs
  - ○ Viral Nucleotide and Protein Database
  - ○ Viral Classification by Nucleotide Alignment
  - ○ Gene Prediction
  - ○ crAss-like Phage Detection
  - ○ Tentative Viral Contig Classification
  - ○ Viral Classification with Pfam Structural Proteins
  - ○ Bacterial Taxonomic Assignment
  - ○ Calculation of the Read Coverage of Contigs
  - ○ Clustering of Virome Samples
  - ○ Analysis of Prophages
  - ○ Validation of the Assembled Contigs
  - ○ Analysis of CRISPR Spacers
  - ○ Analysis of *C. difficile* Phage-Derived Endolysins

- ○ SDS-PAGE
- ○ Lytic Activity Analysis
- ○ Endolysin Therapy in the Mouse Acute CDI Model
- ● QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.chom.2020.06.005.

ACKNOWLEDGMENTS

We thank S. Yin, B. Baigalmaa, H. Ohmiya, and S. Hatakeyama for technical assistance and K. Ogawa, Y. Kumagai, M. Maeda, and K. Suetsugu for secretarial assistance. We also thank M. Mejima, S. Kurokawa, Y. Suzuki, S. Minakawa, N. Takeyama, T. Azegami, and Y. Fukuyama for sample preparation, and Colin Hill (APC Microbiome Ireland & School of Microbiology, University College Cork, Cork T12 YT20, Ireland) and his colleagues provided scientific insights. The super-computing resource was provided by Human Genome Center, Institute of Medical Sciences, The University of Tokyo (http://sc.hgc.jp/shirokane.html). This study was supported by the Ministry of Education, Culture, Sports, Science and Technology of Japan (Grant-in-Aid for Challenging exploratory Research [to S.U.]), the Japan Agency for Medical Research and Development (AMED) (to S.U.; JP18ak0101069), the Takeda Science Foundation (to S.U.), the Canon Foundation (to S.U.), a grant from Meiji Co., Ltd. (to S.U.), Grant-in-Aid for Early-Career Scientists (to K.F.) from the JSPS, Grants-in-Aid for Young Scientists (B) (to Y. Kimura) from the JSPS, Grants-in-Aid for Young Scientists (to Y. Kimura) from the JSPS, Grants-in-Aid for Young Scientists (B) (to T.S.) from the JSPS, Priority Issue on Post-K computer (hp150265, hp160219, and hp170227), strategic programs for innovative research field 1 (hp150232) from MEXT (to S.M.), and the Center of Innovation Program from Japan Science and Technology Agency (JST) (to S.M.).

AUTHOR CONTRIBUTIONS

K.F., Y. Kimura, S.I., and S.U. conceived and designed the study; K.F. and M.S. mainly performed the experiments; Y. Kimura mainly performed data analysis; Y. Kimura and S.I. developed the data analysis pipeline; T.S., Y. Kawaguchi, Y.U., Y.N., T.H., K.K., and Y.Y. performed sample preparation; G.T., M.U., K.Y., M.K., Y.A., and R.Y. helped with the data analyses; S.S., Y.F., S.C., P.E., S.M., and H.K. provided scientific insights and critical review of the manuscript; K.F., Y. Kimura, S.I., and S.U. wrote and edited the manuscript; S.U. also directed the research.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 21, 2020
Revised: May 12, 2020
Accepted: June 10, 2020
Published: July 10, 2020

REFERENCES

Ackermann, H.W., and Prangishvili, D. (2012). Prokaryote viruses studied by electron microscopy. Arch. Virol. *157*, 1843–1849.

Atarashi, K., Suda, W., Luo, C., Kawaguchi, T., Motoo, I., Narushima, S., Kiguchi, Y., Yasuma, K., Watanabe, E., Tanoue, T., et al. (2017). Ectopic colonization of oral bacteria in the intestine drives TH1 cell induction and inflammation. Science *358*, 359–365.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. *19*, 455–477.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., and Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. Science *315*, 1709–1712.

Barrios-Villa, E., Martínez de la Peña, C.F., Lozano-Zaraín, P., Cevallos, M.A., Torres, C., Torres, A.G., and Rocha-Gracia, R.D.C. (2020). Comparative genomics of a subset of adherent/invasive Escherichia coli strains isolated from individuals without inflammatory bowel disease. Genomics *112*, 1813–1820.

Belkaid, Y., and Hand, T.W. (2014). Role of the microbiota in immunity and inflammation. Cell *157*, 121–141.

Biswas, A., Staals, R.H., Morales, S.E., Fineran, P.C., and Brown, C.M. (2016). CRISPRDetect: a flexible algorithm to define CRISPR arrays. BMC Genomics *17*, 356.

Brum, J.R., Ignacio-Espinoza, J.C., Roux, S., Doulcier, G., Acinas, S.G., Alberti, A., Chaffron, S., Cruaud, C., de Vargas, C., Gasol, J.M., et al. (2015). Ocean plankton. Patterns and ecological drivers of ocean viral communities. Science *348*, 1261498.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). Blast+: architecture and applications. BMC Bioinformatics *10*, 421.

Carman, R.J., Daskalovitz, H.M., Lyerly, M.W., Davis, M.Y., Goodykoontz, M.V., and Boone, J.H. (2018). Multidrug resistant Clostridium difficile ribotype 027 in southwestern Virginia, 2007 to 2013. Anaerobe *52*, 16–21.

Chen, X., Katchar, K., Goldsmith, J.D., Nanthakumar, N., Cheknis, A., Gerding, D.N., and Kelly, C.P. (2008). A mouse model of Clostridium difficile-associated disease. Gastroenterology *135*, 1984–1992.

Cho, I., and Blaser, M.J. (2012). The human microbiome: at the interface of health and disease. Nat. Rev. Genet. *13*, 260–270.

Deng, L., Ignacio-Espinoza, J.C., Gregory, A.C., Poulos, B.T., Weitz, J.S., Hugenholtz, P., and Sullivan, M.B. (2014). Viral tagging reveals discrete populations in Synechococcus viral genome sequence space. Nature *513*, 242–245.

Dunne, M., Mertens, H.D., Garefalaki, V., Jeffries, C.M., Thompson, A., Lemke, E.A., Svergun, D.I., Mayer, M.J., Narbad, A., and Meijers, R. (2014). The CD27L and CTP1L endolysins targeting Clostridia contain a built-in trigger and release factor. PLoS Pathog. *10*, e1004228.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. *39*, W29–W37.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., et al. (2016). The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. *44*, D279–D285.

Fujimoto, K., Kawaguchi, Y., Shimohigoshi, M., Gotoh, Y., Nakano, Y., Usui, Y., Hayashi, T., Kimura, Y., Uematsu, M., Yamamoto, T., et al. (2019). Antigen-specific mucosal immunity regulates development of intestinal bacteria-mediated diseases. Gastroenterology *157*, 1530–1543.e4.

Gilbert, J.A., Quinn, R.A., Debelius, J., Xu, Z.Z., Morton, J., Garg, N., Jansson, J.K., Dorrestein, P.C., and Knight, R. (2016). Microbiome-wide association studies link dynamic microbial consortia to disease. Nature *535*, 94–103.

Gregor, I., Dröge, J., Schirmer, M., Quince, C., and McHardy, A.C. (2016). PhyloPythiaS+: a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. PeerJ *4*, e1603.

Guerin, E., Shkoporov, A., Stockdale, S.R., Clooney, A.G., Ryan, F.J., Sutton, T.D.S., Draper, L.A., Gonzalez-Tortuero, E., Ross, R.P., and Hill, C. (2018). Biology and taxonomy of crAss-like bacteriophages, the most abundant virus in the human gut. Cell Host Microbe *24*, 653–664.e6.

Guo, H., Zhu, J., Tan, Y., Li, C., Chen, Z., Sun, S., and Liu, G. (2016). Self-assembly of virus-like particles of rabbit hemorrhagic disease virus capsid protein expressed in Escherichia coli and their immunogenicity in rabbits. Antiviral Res. *131*, 85–91.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics *29*, 1072–1075.

Hyatt, D., LoCascio, P.F., Hauser, L.J., and Uberbacher, E.C. (2012). Gene and translation initiation site prediction in metagenomic sequences. Bioinformatics *28*, 2223–2230.

Kakuta, M., Suzuki, S., Izawa, K., Ishida, T., and Akiyama, Y. (2017). A massively parallel sequence similarity search for metagenomic sequencing data. Int. J. Mol. Sci. *18*, 2124.

Kamada, N., Kim, Y.G., Sham, H.P., Vallance, B.A., Puente, J.L., Martens, E.C., and Núñez, G. (2012). Regulated virulence controls the ability of a pathogen to compete with the gut microbiota. Science *336*, 1325–1329.

Kang, D.W., Adams, J.B., Gregory, A.C., Borody, T., Chittick, L., Fasano, A., Khoruts, A., Geis, E., Maldonado, J., McDonough-Means, S., et al. (2017). Microbiota transfer therapy alters gut ecosystem and improves gastrointestinal and autism symptoms: an open-label study. Microbiome *5*, 10.

Karlsson, F.H., Tremaroli, V., Nookaew, I., Bergström, G., Behre, C.J., Fagerberg, B., Nielsen, J., and Bäckhed, F. (2013). Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature *498*, 99–103.

Kimura, Y., Soma, T., Kasahara, N., Delobel, D., Hanami, T., Tanaka, Y., de Hoon, M.J., Hayashizaki, Y., Usui, K., and Harbers, M. (2016). Edesign: primer and enhanced internal probe design tool for quantitative PCR experiments and genotyping assays. PLoS One *11*, e0146950.

Krupovic, M., and Forterre, P. (2011). Microviridae goes temperate: Microvirus-related proviruses reside in the genomes of Bacteroidetes. PLoS One *6*, e19893.

Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., Almeida, M., Arumugam, M., Batto, J.M., Kennedy, S., et al. (2013). Richness of human gut microbiome correlates with metabolic markers. Nature *500*, 541–546.

Lessa, F.C., Mu, Y., Bamberg, W.M., Beldavs, Z.G., Dumyati, G.K., Dunn, J.R., Farley, M.M., Holzbauer, S.M., Meek, J.I., Phipps, E.C., et al. (2015). Burden of Clostridium difficile infection in the United States. N. Engl. J. Med. *372*, 825–834.

Li, D., Liu, C.M., Luo, R., Sadakane, K., and Lam, T.W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics *31*, 1674–1676.

Li, H., Li, W.G., Zhang, W.Z., Yu, S.B., Liu, Z.J., Zhang, X., Wu, Y., and Lu, J.X. (2019). Antibiotic resistance of clinical isolates of Clostridioides difficile in China and its association with geographical regions and patient age. Anaerobe *60*, 102094.

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

Love, M.J., Bhandari, D., Dobson, R.C.J., and Billington, C. (2018). Potential for bacteriophage endolysins to supplement or replace antibiotics in food production and clinical care. Antibiotics (Basel) *7*, 17.

Lynch, S.V., and Pedersen, O. (2016). The human intestinal microbiome in health and disease. N. Engl. J. Med. *375*, 2369–2379.

Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J., Wolf, Y.I., Yakunin, A.F., et al. (2011). Evolution and classification of the CRISPR-Cas systems. Nat. Rev. Microbiol. *9*, 467–477.

Manrique, P., Bolduc, B., Walk, S.T., van der Oost, J., de Vos, W.M., and Young, M.J. (2016). Healthy human gut phageome. Proc. Natl. Acad. Sci. USA *113*, 10400–10405.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. *17*, 10–12.

Mikheenko, A., Saveliev, V., and Gurevich, A. (2016). MetaQUAST: evaluation of metagenome assemblies. Bioinformatics *32*, 1088–1090.

Minot, S., Bryson, A., Chehoud, C., Wu, G.D., Lewis, J.D., and Bushman, F.D. (2013). Rapid evolution of the human gut virome. Proc. Natl. Acad. Sci. USA *110*, 12450–12455.

Minot, S., Sinha, R., Chen, J., Li, H., Keilbaugh, S.A., Wu, G.D., Lewis, J.D., and Bushman, F.D. (2011). The human gut virome: inter-individual variation and dynamic response to diet. Genome Res. *21*, 1616–1625.

Mirzaei, M.K., and Maurice, C.F. (2017). Menage a trois in the human gut: interactions between host, bacteria and phages. Nat. Rev. Microbiol. *15*, 397–408.

Monteiro, R., Pires, D.P., Costa, A.R., and Azeredo, J. (2019). Phage therapy: going temperate? Trends Microbiol. *27*, 368–378.

Morita, H., Kuwahara, T., Ohshima, K., Sasamoto, H., Itoh, K., Hattori, M., Hayashi, T., and Takami, H. (2007). An improved DNA isolation method for metagenomic analysis of the microbial flora of the human intestine. Microb. Environ. *22*, 214–222.

Nakamoto, N., Sasaki, N., Aoki, R., Miyamoto, K., Suda, W., Teratani, T., Suzuki, T., Koda, Y., Chu, P.S., Taniki, N., et al. (2019). Gut pathobionts underlie intestinal barrier dysfunction and liver T helper 17 cell immune response in primary sclerosing cholangitis. Nat. Microbiol. *4*, 492–503.

Nelson, D.C., Schmelcher, M., Rodriguez-Rubio, L., Klumpp, J., Pritchard, D.G., Dong, S., and Donovan, D.M. (2012). Endolysins as antimicrobials. Adv. Virus Res. *83*, 299–365.

Nikolenko, S.I., Korobeynikov, A.I., and Alekseyev, M.A. (2013). BayesHammer: Bayesian clustering for error correction in single-cell sequencing. BMC Genomics *14*, S7.

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. (2017). metaSPAdes: a new versatile metagenomic assembler. Genome Res. *27*, 824–834.

Orenstein, R., and Patron, R.L. (2019). Clostridioides difficile therapeutics: guidelines and beyond. Ther. Adv. Infect. Dis. *6*, 2049936119868548.

Peng, Y., Leung, H.C., Yiu, S.M., and Chin, F.Y. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics *28*, 1420–1428.

Pohane, A.A., and Jain, V. (2015). Insights into the regulation of bacteriophage endolysin: multiple means to the same end. Microbiology *161*, 2269–2276.

Ren, J., Ahlgren, N.A., Lu, Y.Y., Fuhrman, J.A., and Sun, F. (2017). VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome *5*, 69.

Reyes, A., Haynes, M., Hanson, N., Angly, F.E., Heath, A.C., Rohwer, F., and Gordon, J.I. (2010). Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature *466*, 334–338.

Reyes, A., Semenkovich, N.P., Whiteson, K., Rohwer, F., and Gordon, J.I. (2012). Going viral: next-generation sequencing applied to phage populations in the human gut. Nat. Rev. Microbiol. *10*, 607–617.

Roach, D.R., and Donovan, D.M. (2015). Antimicrobial bacteriophage-derived proteins and therapeutic applications. Bacteriophage *5*, e1062590.

Rooks, M.G., and Garrett, W.S. (2016). Gut microbiota, metabolites and host immunity. Nat. Rev. Immunol. *16*, 341–352.

Roux, S., Emerson, J.B., Eloe-Fadrosh, E.A., and Sullivan, M.B. (2017). Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. PeerJ *5*, e3817.

Roux, S., Hallam, S.J., Woyke, T., and Sullivan, M.B. (2015). Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. eLife *4*, e08490.

Roux, S., Solonenko, N.E., Dang, V.T., Poulos, B.T., Schwenck, S.M., Goldsmith, D.B., Coleman, M.L., Breitbart, M., and Sullivan, M.B. (2016). Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. PeerJ *4*, e2777.

Scher, J.U., Sczesnak, A., Longman, R.S., Segata, N., Ubeda, C., Bielski, C., Rostron, T., Cerundolo, V., Pamer, E.G., Abramson, S.B., et al. (2013). Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. eLife *2*, e01202.

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics *27*, 863–864.

Shkoporov, A.N., Clooney, A.G., Sutton, T.D.S., Ryan, F.J., Daly, K.M., Nolan, J.A., McDonnell, S.A., Khokhlova, E.V., Draper, L.A., Forde, A., et al. (2019). The human gut virome is highly diverse, stable, and individual specific. Cell Host Microbe *26*, 527–541.e5.

Spigaglia, P., Mastrantonio, P., and Barbanti, F. (2018). Antibiotic resistances of Clostridium difficile. Adv. Exp. Med. Biol. *1050*, 137–159.

Tanoue, T., Morita, S., Plichta, D.R., Skelly, A.N., Suda, W., Sugiura, Y., Narushima, S., Vlamakis, H., Motoo, I., Sugita, K., et al. (2019). A defined commensal consortium elicits CD8 T cells and anti-cancer immunity. Nature *565*, 600–605.

Truong, D.T., Franzosa, E.A., Tickle, T.L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., and Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat. Methods *12*, 902–903.

Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., and Barton, G.J. (2009). Jalview Version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189–1191.

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., and Madden, T.L. (2012). Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics *13*, 134.

Yutin, N., Makarova, K.S., Gussow, A.B., Krupovic, M., Segall, A., Edwards, R.A., and Koonin, E.V. (2018). Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. Nat. Microbiol. *3*, 38–46.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and Virus Strains		
Enterobacteria phage f1	Biological Resource Center, NITE	NBRC 20010
Escherichia virus phiX174	Biological Resource Center, NITE	NBRC 103405
Enterobacteria phage T3	Biological Resource Center, NITE	NBRC 20003
Escherichia virus Lambda	Biological Resource Center, NITE	NBRC 20016
Clostridioides difficile strain	the American Type Culture Collection in Manassas	ATCC 43255
Clostridioides difficile strain	the Japan Collection of Microorganisms, RIKEN BRC	JCM 1296, JCM 5243, JCM 5244, JCM 5245, JCM 5246, JCM 5247, JCM 5248, JCM 5249, JCM 5250, JCM 5251, JCM 5252, JCM 5253, JCM 5254, JCM 5256, JCM 5257, and JCM 5258
BL21	Takara	Cat#9126
Chemicals, Peptides, and Recombinant Proteins		
NaCl	Nacalai Tesque	Cat#09649-15
Tris-HCl	Nacalai Tesque	Cat#35436-01
CaCl•2H$_2$O	Nacalai Tesque	Cat#09653-45
Gelatin	Nacalai Tesque	Cat#16631-05
Na_3PO_4	Nacalai Tesque	Cat#31804-75
imidazole	Nacalai Tesque	Cat#19004-35
DNase I	Roche	Cat#4716728001
Benzonase	Novagen	Cat#71205-3
Baseline-Zero DNase	Epicentre	Cat#DB0715K
EDTA	Nacalai Tesque	Cat#06894-85
Recombinant human lysozyme	Sigma Aldrich	Cat#L1667
Achromopeptidase	Wako	Cat#014-09661
Proteinase K	Nacalai Tesque	Cat#15679-06
SDS	Nacalai Tesque	Cat#31606-75
Phenol/chloroform/isoamyl alcohol	Nacalai Tesque	Cat#25970-56
Chloroform	Nacalai Tesque	Cat#08401-65
Sodium acetate	Nacalai Tesque	Cat#31119-65
Isopropanol	Nacalai Tesque	Cat#29113-95
Glycogen	Roche	Cat#10901393001
Ethanol	Nacalai Tesque	Cat#09666-85
Brain Heart Infusion	BD	Cat#BD237500
Gifu Anaerobic Broth	Nissui	Cat#05422
*Bam*HI	TOYOBO	Cat#BAH-111
*Sal*I	TOYOBO	Cat#SAL-111
LB Broth	Nacalai Tesque	Cat#20068-75
Isopropyl β-D-thiogalactoside	Nacalai Tesque	Cat#19742-94
xTractor™ buffer	Takara	Cat#635625
kanamycin	Nacalai Tesque	Cat#19860-44
gentamicin	Nacalai Tesque	Cat#16637-74
colistin	Sigma-Aldrich	Cat#C4461
metronidazole	Nacalai Tesque	Cat#23254-22
vancomycin	DUCHEFA Biochemie	Cat#V0155.0005

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
ampicillin	Nacalai Tesque	Cat#07239-32
clindamycin	Wako	Cat#037-24531
Critical Commercial Assays		
QuantiFluor dsDNA system	Promega	Cat#E2670
KAPA HyperPlus Kit	KAPA Biosystems	Cat#KK8512
Accel-NGS 1S Plus DNA Library Kit	Swift Biosciences	Cat#10096
KAPA Illumina Library Quantification kit	KAPA Biosystems	Cat#KK4824
DNA-12000 Kit	SHIMADZU	Cat#S292-36600-91
MiSeq reagent kit v3	Illumina	Cat#MS-102-3003
HiSeq Rapid SBS Kit v2-HS	Illumina	Cat#FC-402-4023
HiSeq PE Rapid Cluster Kit v2-HS	Illumina	Cat no. PE-402-4002
PowerSoil DNA Isolation Kit	QIAGEN	Cat#12888-100
Capturem™ His-Tagged Purification Maxiprep columns	Takara	Cat#635719
HiTrap^{TM} desalting columns	Amersham Biosciences	Cat#17140801
Amicon® Ultra-15 10K	Merck Millipore	Cat#UFC901096
Protein Assay CBB Solution	Nacalai Tesque	Cat#11617
Q-Stain	NIPPON Genetics	Cat#NE-FG-QS1
Experimental Models: Organisms/Strains		
Mouse: C57BL/6 (SPF)	SLC Japan	N/A
Oligonucleotides		
NEBNext multiplex Oligos for Illumina	New England BioLabs	Cat#E6609S
1S Plus Dual Index Kit	Swift Biosciences	Cat#18096
Software and Algorithms		
cutadapt	Martin, 2011	http://cutadapt.readthedocs.io/en/stable/index.html
PRINSEQ	Schmieder and Edwards, 2011	http://prinseq.sourceforge.net/
BayesHammer	Nikolenko et al., 2013	http://bioinf.spbau.ru/spades/bayeshammer
SPAdes	Bankevich et al., 2012	http://cab.spbu.ru/software/spades/
MetaSPAdes	Nurk et al., 2017	http://cab.spbu.ru/software/spades/
IDBA-UD	Peng et al., 2012	https://i.cs.hku.hk/~alse/hkubrg/projects/idba_ud/
MEGAHIT	Li et al., 2015	https://github.com/voutcn/megahit
MetaQUAST	Mikheenko et al., 2016	http://quast.sourceforge.net/metaquast
CD-HIT-EST	Li and Godzik, 2006	http://weizhongli-lab.org/cd-hit/
BLAST+	Camacho et al., 2009	https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download
VirSorter	Roux et al., 2015	https://github.com/simroux/VirSorter
VirFinder	Ren et al., 2017	https://github.com/jessieren/VirFinder
MetaProdigal	Hyatt et al., 2012	https://github.com/hyattpd/Prodigal
blast2lca	https://github.com/emepyc/Blast2lca	https://github.com/emepyc/Blast2lca
GHOST-MP	Kakuta et al., 2017	http://www.bi.cs.titech.ac.jp/ghostmp/index.html
HMMER3	Finn et al., 2011	http://hmmer.org/
PhyloPythiaS+	Gregor et al., 2016	https://github.com/algbioi/ppsplus
MetaPhlAn 2.0	Truong et al., 2015	http://huttenhower.sph.harvard.edu/metaphlan2
CRISPRDetect	Biswas et al., 2016	http://brownlabtools.otago.ac.nz/CRISPRDetect/predict_crispr_array.html
Primer-BLAST	Ye et al., 2012	https://www.ncbi.nlm.nih.gov/tools/primer-blast/
Edesign	Kimura et al., 2016	https://github.com/yasumasak/edesign
BBtools	Joint Genome Institute	https://jgi.doe.gov/data-and-tools/bbtools/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
R software	The R Foundation for Statistical Computing	https://www.r-project.org/
R package *vegan*	https://cran.r-project.org/web/packages/vegan/	https://cran.r-project.org/web/packages/vegan/
R package *qvalue*	https://github.com/StoreyLab/qvalue	https://github.com/StoreyLab/qvalue
NCBI RefSeq sequences	https://www.ncbi.nlm.nih.gov/genomes/	https://www.ncbi.nlm.nih.gov/genomes/
NCBI taxonomy	ftp://ftp.ncbi.nih.gov/pub/taxonomy/	ftp://ftp.ncbi.nih.gov/pub/taxonomy/
Pfam	Finn et al., 2016	http://pfam.xfam.org/
MUSCLE	Edgar, 2004	https://www.drive5.com/muscle/
Jalview	Waterhouse et al., 2009	https://www.jalview.org
Other		
microTUBE-15 AFA Beads Screw-Cap	Covaris	Cat#520145
pCold-SUMO	Creative Biogene	Cat#VET1156

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Satoshi Uematsu (uematsu.satoshi@med.osaka-cu.ac.jp).

Materials Availability

The materials that support the findings of this study are available from the corresponding authors upon reasonable request. Please contact the Lead Contact for additional information.

Data and Code Availability

All sequencing data that support the findings of this study have not been deposited in a public repository because of the ethical concerns, but are available from the corresponding authors upon reasonable request, with the approval of our ethics committee. Please contact the Lead Contact for additional information.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human Subjects

In total, 101 Japanese male (n=96) and female (n=5) healthy volunteers (aged over 20 years) provided fecal samples. The study protocol was approved by the Ethics Committee of the Institute of Medical Science, The University of Tokyo, and signed informed consent was obtained from each participant (27-41-0917 and 29-44-B1010).

Mice

Six-week-old female specific-pathogen-free C57BL/6 mice were purchased from SLC, Shizuoka, Japan. Mice were housed in a temperature-controlled (23°C ± 2°C) room with a dark period from 20:00 to 08:00 h. They were allowed free access to sterile water and standard laboratory mouse chow. All of the animal experiments were performed with the approval of the Animal Care and Use Committees of The University of Tokyo and Osaka City University.

C. difficile *Strains*

C. difficile strains JCM 1296, JCM 5243, JCM 5244, JCM 5245, JCM 5246, JCM 5247, JCM 5248, JCM 5249, JCM 5250, JCM 5251, JCM 5252, JCM 5253, JCM 5254, JCM 5256, JCM 5257, and JCM 5258 were provided by the Japan Collection of Microorganisms, RIKEN BRC, which is participating in the National BioResource Project of the MEXT, Japan. Strain ATCC 43255, a high-level toxin-producing strain, was obtained from the American Type Culture Collection in Manassas, Virginia, USA. Each *C. difficile* colony was isolated from brain–heart infusion (BHI) plates or Gifu anaerobic medium plates and cultured in a Ruskinn Bugbox Plus anaerobic chamber (Baker Co., Sanford, ME, USA) supplied with 10% H_2, 10% CO_2, and 80% N_2 at 37°C overnight in BHI broth or Gifu anaerobic medium. Bacterial DNA from each strain was extracted using the PowerSoil DNA Isolation Kit (QIAGEN, Germantown, MD, USA) in accordance with the manufacturer's protocol.

METHOD DETAILS

Human Fecal Sample Fractionation

The human fecal samples were immediately stored at 4°C under anaerobic conditions after collection. One gram of the samples was stored at −80°C until use. Stocks of human fecal samples were homogenized in 3 ml of SM-plus buffer (100 mM NaCl, 50 mM Tris-HCl (pH 7.4), 8 mM $MgSO_4 \cdot 7H_2O$, 5 mM $CaCl \cdot 2H_2O$, and 0.01% (w/v) gelatine in distilled water) with vortex mixing, and were passed through a 100-μm cell strainer. SM-plus buffer was passed through a 0.22-μm syringe filter before usage at each step. The debris on the cell strainer was washed twice with 3 ml of SM-plus buffer, and the filtrates were centrifuged at 6,000 × *g* for 5 min. The supernatants were transferred to new tubes and the pellets were suspended in 1 ml of SM-plus buffer. After centrifugation at 6,000 × *g* for 5 min, the supernatants were combined with those obtained at the previous step. Combined supernatants were centrifuged at 6,000 × *g* for 15 min again and were carefully recovered as the "viral fraction". The remaining pellets were suspended in 1 ml of SM-plus buffer, and 100 μl of the suspensions were used as the "bacterial fraction".

Treatment of the Viral Fraction

The supernatant fractions from human feces or 1 g of mouse feces along with equal amounts of the four phages (*Enterobacteria* phage f1 (NBRC 20010), *Escherichia* virus phiX174 (NBRC 103405), *Enterobacteria* phage T3 (NBRC 20003), and *Escherichia* virus Lambda (NBRC 20016)) were passed through a 0.45-μm syringe filter to remove contaminated debris. Cell-free DNA in the samples was degraded by incubation with DNase mix (1 U/ml DNase I, (Roche, Basel, Switzerland), 10 U/ml Benzonase (Merck, Dermstadt, Germany), and 1 U/ml Baseline-Zero DNase (Epicentre, Madison, WI, USA)) at 37°C for 1 h. After stopping the DNase reaction by the addition of EDTA (final conc. 20 mM), the samples were subjected to DNA extraction. To establish the absence to negligible bacterial DNA contamination using our method for viral DNA purification from fecal samples, we performed quality-control assays using 16S rDNA PCR (Reyes et al., 2010).

Treatment of the Bacterial Fraction

To extract DNAs from the bacterial fraction, we used the protocol reported previously (Morita et al., 2007) with some modifications. This method is widely used for metagenome analyses (Atarashi et al., 2017; Nakamoto et al., 2019; Tanoue et al., 2019). The samples were incubated with 1 ml of SM-plus buffer containing 20 mM EDTA, 100 μg/ml recombinant human lysozyme (Sigma Aldrich, St Louis, MO, USA), and 0.5 U/ml achromopeptidase at 37°C for 1 h and were subjected to DNA extraction.

DNA Extraction

Each supernatant sample from the viral and bacterial fractions was incubated with a 1400 volume of 20 mg/ml proteinase K (Nacalai Tesque, Kyoto, Japan) and a 1/20 volume of 10% SDS at 55°C for 1 h. Thereafter, the samples were added to an equal volume of phenol/chloroform/isoamyl alcohol and mixed vigorously. After centrifugation at 16,000 × *g* for 5 min, the aqueous phase of the samples was transferred to new tubes, followed by chloroform extraction. Again, the aqueous phase of the samples was transferred to new tubes and mixed with a 1/10 volume of 3 M sodium acetate and an equal volume of isopropanol. For the viral fractions, 40 μg of glycogen (Roche) was added as a co-precipitate. The samples were centrifuged at 16,000 × *g* for 15 min. After discarding the supernatants, the pellets were washed with 70% ethanol and centrifuged at 16,000 × *g* for 5 min. The supernatants were removed completely, and the pellets were air-dried for 5 min. The DNAs were resuspended in 10 mM Tris-HCl buffer (pH 8.0), and the concentrations were quantified with the Quantus dsDNA kit (Promega, Madison, WI, USA).

Bacterial DNA Sequencing

DNA libraries were prepared with the KAPA HyperPlus Kit (KAPA Biosystems, Indianapolis, IN, USA) following the manufacturer's instructions, except that NEBNext multiplex Oligos were used for Illumina (New England BioLabs, Ipswich, MA, USA) at the adapter ligation and barcoding steps. The prepared target library size was 450–550 nucleotides. The concentration of the library was quantified with the KAPA Illumina Library Quantification kit (KAPA Biosystems) and adjusted to the mean library size measured by MCE-202 MultiNA (Shimadzu, Kyoto, Japan). The libraries were pooled and sequenced on a HiSeq2500 sequencer (2 × 250 paired-end reads, HiSeq Rapid SBS Kit v2 (Illumina, San Diego, CA, USA)). For each run, 10 libraries for the bacterial fractions were pooled at equimolar concentrations.

Viral DNA Sequencing

Extracted viral DNAs were fragmented identically using the Covaris system (M220) (Covaris, Woburn, MA, USA). The prepared fragment size was about 350 nucleotides. After fragmentation, DNA libraries were prepared with the Swift 1S Plus DNA Library Kit for Illumina. The concentration of the library was quantified with the Illumina Library Quantification kit (KAPA Biosystems) and adjusted to the mean library size measured by MCE-202 MultiNA (Shimadzu). The libraries were pooled and sequenced on an Illumina MiSeq sequencer (2 × 300 paired-end reads, MiSeq reagent kit v3) for the spike-in viromes (Figures S1B–S1E) or a HiSeq2500 sequencer (2 × 250 paired-end reads, HiSeq Rapid SBS Kit v2) for human viromes. Four libraries from the viral fraction were pooled at equimolar concentrations per run on the MiSeq. For each run on the HiSeq2500 sequencer, 40 libraries for the viral fractions were pooled at equimolar concentrations.

Sequenced Data Processing

Sequencing yielded a total of 256 Gb and 1,530 Gb of sequence data from the viral and bacterial fractions, respectively (5,060,445 ± 2,577,253 and 30,878,304 ± 14,204,793 of paired-end reads per sample from the viral and bacterial fractions, respectively). Sequencing reads were demultiplexed using the Illumina CASAVA software and were processed using the following three steps: 1) adaptor sequence trimming, 2) nucleotide trimming and removal of duplicates, and 3) error base correction. First, adaptor sequences were removed by cutadapt software (http://cutadapt.readthedocs.io/en/stable/index.html) (v.1.2.1). In the second step, the first and last 10 nucleotides of each read were removed, then within 20 nt from both ends, low quality nucleotides with a Phred quality score <20 were trimmed, and polynucleotides at the end of the sequence were also trimmed. After trimming, sequences shorter than 75 nt, low complexity sequences (DUST score >7), exact duplicates, sequences containing N, and singletons were filtered out. This step was performed by PRINSEQ software (http://prinseq.sourceforge.net/) (lite v.0.20.4) (-trim_right 10 -trim_left 10 -trim_qual_right 20 -trim_qual_left 20 -trim_qual_window 20 -trim_ns_right 1 -min_len 75 -lc_method dust -lc_threshold 7 -ns_max_n 0 -derep 1). Third, correction of sequencing errors based on the Hamming graph and Bayesian subclustering were performed using BayesHammer software (Nikolenko et al., 2013) (SPAdes v.3.11.0) (spades.py –only-error-correction).

Comparison of Assemblers

There is no assembler dedicated to intestinal virome analyses, therefore we compared four assemblers, SPAdes (Bankevich et al., 2012), MetaSPAdes (Nurk et al., 2017), IDBA-UD (Peng et al., 2012), and MEGAHIT (Li et al., 2015), in terms of accuracy and length in contig generation and adopted MetaSPAdes as the most suitable assembler for our pipeline (Figures S1B–S1E).

In detail, the quality-filtered and error-corrected reads from the spike-in metagenome samples were assembled with the four different assembly methods: SPAdes (Bankevich et al., 2012), MetaSPAdes (Nurk et al., 2017), IDBA-UD (Peng et al., 2012), and MEGAHIT (Li et al., 2015), for comparison. SPAdes v3.11.0 was used with MismatchCorrector (options: –careful –only-assembler). MetaSPAdes v3.11.0 (options: –meta –only-assembler) was used with default settings. IDBA-UD v.1.1.1 was used with pre-read error correction (option: –pre-correction) using fasta reads (converted from fastq reads with the tool "fq2fa"). MEGAHIT v1.1.2 was used with default settings. Each of the contigs constructed by the four assemblers were evaluated with MetaQUAST (Mikheenko et al., 2016) v4.5 (options: –min-alignment 500 -m 1500). The sequences of the four spike viral genomes (accession numbers J02448.1, NC_001422.1, NC_003298.1, and NC_001416.1 for *Enterobacteria* phage f1, *Escherichia* virus phiX174, *Enterobacteria* phage T3, and *Escherichia* virus Lambda, respectively) were downloaded from NCBI (https://www.ncbi.nlm.nih.gov/nuccore/) and were used as reference genomes for computing MetaQUAST statistics. As evaluation metrics, we focused on the NGA50 static, the genome fraction, the number of local misassemblies, and the total length of the contigs. To compute NGA50 (Gurevich et al., 2013), the contigs were first broken into smaller segments at the identified misassembly breakpoints. The NGA50 for the given reference genomes is the maximal size of the broken aligned segments where all aligned segments of equal or larger size cover 50% of the total reference genome length. Broken aligned segments were obtained by breaking contigs at misassembly events and removing all unaligned bases. The genome fraction measures the percentage of a reference genome covered by aligned contigs. The number of local misassemblies is the error metrics reflective of the assembly quality. The total length of contigs denotes the total number of bases in the assembly.

Metagenome Assembly

The quality-filtered and error-corrected reads in each sample were assembled with MetaSPAdes (Nurk et al., 2017) v3.11.0 with default k-mer lengths (options: –meta –only-assembler) (Figure S1A). To compare the abundance of contigs across the samples, the assembled contigs (with lengths $\geq$1 kb for the viral fraction, and $\geq$5 kb for the bacterial fraction) from individual samples were pooled. CD-HIT-EST (Li and Godzik, 2006) (v.4.6) was used to cluster pooled contigs at 95% global average nucleotide identity (-c 0.95 -G 1 -n 10 -mask NX) (Figure S1A). The contigs from the viral fraction and the bacterial fraction were treated separately. From the non-redundant pooled contigs, circular contigs were identified by detecting overlaps in the 5′ and 3′ end sequences (more than 50 nucleotides overlap at 100% identity) of the contigs using megablast (BLAST+ v.2.7.1) (Camacho et al., 2009). Each of the detected circular contigs was trimmed to remove redundant parts. Circular contigs longer than 1.5 kb and linear contigs longer than 5 kb were used for the analyses (Figure S1A).

Extraction of Viral Contigs

The contigs constructed from the viral fraction were screened with a gene enrichment based method VirSorter (Roux et al., 2015) (v1.0.3) and a k-mer frequency based method VirFinder (Ren et al., 2017) (v1.1) to identify and remove bacteria-like contigs. VirSorter was performed using both RefSeqABVir (–db 1) and Viromes (–db 2) databases using the "virome decontamination" mode (–virome 1) to extract the viral contigs (category 1, 2 or 3). VirFinder was performed using a default prediction model and $p < 0.05$ as a threshold. The viral contigs detected by either or both methods were used for further analyses (Figure S1A). After this process, the remaining contigs were defined as viral contigs and used for the following analyses.

Viral Nucleotide and Protein Database

To classify the viral contigs, we prepared viral genome and protein databases. The viral RefSeq sequences (v84, containing 9,497 genomes and 231,157 proteins) were downloaded from NCBI (https://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=10239). Taxonomic lineage information was assigned to each viral sequence using the NCBI taxonomy data (ftp://ftp.ncbi.

nih.gov/pub/taxonomy/, downloaded on November 30, 2017). Since only 57 viral contigs were classified (Figure S1A), the ORFs for the rest of the viral contigs were predicted for classification as below.

Viral Classification by Nucleotide Alignment

We first classified viral contigs using viral RefSeq genomes (first part of Figure 1A shown in red). The viral contig sequences were searched against the viral RefSeq genomes (v84) using megablast (BLAST+ v.2.7.1) with an *E* value $<10^{-10}$. Since some viral contig sequences contained multiple fragments that were mapped to different viral genomes, significant alignments up to the top five were considered to determine viral taxonomy. If the top five alignments covered more than 50% of a contig sequence, the lowest common ancestor (LCA) of the top five hits was determined using blast2lca (https://github.com/emepyc/Blast2lca) (modified to use accession.version identifiers) with the NCBI taxonomy (downloaded on November 30, 2017) and was assigned to the contig. Since the classification of p-crAssphage is conducted together with the detection of crAss-like marker genes in a later step of the pipeline, contigs classified as p-crAssphage at this step were set as unclassified.

Gene Prediction

The contigs that were not classified in the previous step (megablast with viral RefSeq genomes) were analyzed according to their open reading frames (ORFs). The ORFs on the contigs were predicted using MetaProdigal (Hyatt et al., 2012) (v2.6.3) with the metagenomics procedure (-p meta). To predict genes spanning the 3′ to 5′ ends of a circular contig, a temporary version of the circular contig was used in the ORF prediction, where the first 1500 nucleotides were duplicated and added at the end of the contig.

crAss-like Phage Detection

Using the annotation of ORFs on contigs, we first detected crAss-like phage contigs by following the method for crAss-like phage detection reported in a previous study (Guerin et al., 2018). The amino acid sequences of the prototypical crAssphage (p-crAssphage, NC_024711.1) genetic signatures, the polymerase (UGP_018) and the terminase (UGP_092), were queried using blastp (BLAST+ v.2.7.1) against the ORFs of the viral contigs with an *E* value $<10^{-5}$ and an alignment length $\geq$350. The viral contigs containing a blastp hit of either the p-crAssphage polymerase or terminase with a minimum contig length of 70 kb were classified as crAss-like phages. The similarities between the crAss-like phage contigs and the p-crAssphage genome (NC_024711.1) were assessed using megablast (BLAST+ v.2.7.1). The crAss-like phage contigs that showed $\geq$95% identity over 80% of the contig sequence (criteria previously used for genome-based species separation (Brum et al., 2015; Deng et al., 2014) with the p-crAssphage genome) were classified as p-crAssphage (shown by the gray background in Figure 1A).

Tentative Viral Contig Classification

To annotate the predicted ORFs, the amino acid sequences of the ORFs were queried by GHOST-MP (Kakuta et al., 2017) (v.1.3.4) against the viral RefSeq protein (v84) with an *E* value $<10^{-5}$ and a bitscore >50. The predicted ORFs were also queried by hmmscan in HMMER3 (Finn et al., 2011) (v3.1b2) against the PfamA (Finn et al., 2016) (v31.0) database with an *E* value $<10^{-5}$.

The viral Refseq proteins with the top three closest homologies (*E* value $<10^{-5}$ and bitscore >50) were considered for each ORF. The taxonomic lineages of the three viral proteins were compared from the species level to the order level. For each level, if two or more of the hits shared the same taxon, the ORF was assigned with that taxon. To analyses the taxonomy of the entire contig, taxonomic lineages of the classified ORFs within the contig were compared. From the species level to the order level, a taxon that was common in more than 50% of the classified ORFs was assigned to the contig, analogous to a previously reported method (Kang et al., 2017).

Viral Classification with Pfam Structural Proteins

To improve the tentative classification of viral contigs, we additionally used the phage structural proteins in the PfamA annotation (*E* value $<10^{-5}$). The contigs were classified as *Caudovirales*, *Myoviridae*, or *Microviridae*, if the contigs possessed a phage tail protein, a phage tail sheath protein, or a *Microviridae* capsid protein gene, respectively. The Pfam entries of the phage structural proteins used for the classification are listed in Table S1.

If a taxon at a certain rank was undetermined but instead determined at a higher rank, the upper-level taxon name with the prefix "uc_" (unclassified_) was used for its classification at the lower rank (e.g., a contig with no family assignment but classified as *Caudovirales* at the order level was labeled as uc_Caudovirales at the family level). A contig that could not be classified was labeled as "Unclassified". A contig that contained homologous proteins from different orders of viruses was labeled as "Unclassified (Multiple)". Through this pipeline, 9,889 viral contigs (145 with crAss-like phage markers, 291 with the *Myoviridae* sheath, 1,706 with the *Caudovirales* tail, 568 with the *Microviridae* capsid, and 7,179 with other viral proteins) could be classified using their ORF annotations (Figure S1A).

Bacterial Taxonomic Assignment

We assigned bacterial taxonomy to the contigs using PhyloPythiaS+ (Gregor et al., 2016). The whole PhyloPythiaS+ pipeline was run (options: -n -g -o s16 mg -t -p c -r -s) using their reference database 'NCBI201502' with the configuration parameters: maxLeaf-

Clades=500 and minPercentInLeaf=0.05 (the others parameters were set as default). For obtaining the taxonomic profile in each sample, the quality-filtered and error-corrected reads were mapped to microbial taxonomy-specific marker genes with MetaPhlAn 2.0 (Truong et al., 2015).

Calculation of the Read Coverage of Contigs

The quality-filtered and error-corrected reads were mapped to the non-redundant pooled contigs using the bbmap tool from BBtools with $\geq$95% identity and the ambiguous mapping option (ambiguous=random). A contig was considered as "detected" in a sample if more than 75% of the contig length was covered by mapped reads, as recommended in a previous study (Roux et al., 2017). The abundance of a contig was calculated as the average contig coverage (number of nucleotides mapped to the contig divided by the contig length), where abundance of a non "detected" contig was set to 0, and normalized by the total number of nucleotides of the mapped reads in a sample, to have a total number of nucleotides equal to 10^9. In viral library preparation, dsDNA genomes provide twice as many templates as ssDNA genomes per single virus through the dsDNA denaturation step. Therefore, in a viral fraction sample, the read coverage of contigs that were not classified as ssDNA viruses was divided by two, as reported in a previous study (Roux et al., 2016).

Clustering of Virome Samples

The relative abundances of the viral contigs in each sample were computed to have a total abundance equal to 1 and were summed at the viral family level. Although the other analyses treated crAss-like phages and p-crAssphage separately, here p-crAssphage was treated the same as the crAss-like phage family. Hierarchical clustering of the virome samples was performed based on the family-level relative abundances using the Euclidean distance and Ward linkage. The relative abundances of the unclassified contigs were not included in the distance calculation.

Analysis of Prophages

Prophage sequences in the bacterial contigs were detected by a combination of the following two approaches. In the first approach, prophage sequences were predicted according to known viral signatures with VirSorter (Roux et al., 2015) (v1.0.3). The bacterial contigs ($\geq$5 kb) from individual samples were analyzed by VirSorter using both RefSeqABVir (–db 1) and Viromes (–db 2). To remove virus-derived sequences from the bacterial contigs, predicted prophage sequences of VirSorter categories 4 or 5 (presence of viral hallmark genes or enrichment of viral-like genes in a prophage region) were extracted. The positions of the predicted prophage sequences on the non-redundant pooled bacterial contigs were obtained through megablast (BLAST+ v.2.7.1) searches (*E* value $<10^{-100}$ and $\geq$95% identity), and prophage sequences were merged if their positions overlapped. Last, prophage sequences longer than 3 kb were extracted and listed as candidate prophage sequences. In the second approach, prophages were searched using the viral contigs generated in this study. The viral contigs were queried against the bacterial contigs using megablast (BLAST+ v.2.7.1) with 90% of the viral contig length aligned at a minimum identity of 95%. We considered the aligned sequence to be a prophage if it satisfied the following criteria: i) the bacterial contig sequence was 5 kb longer than the aligned viral contig sequence; and ii) the aligned viral contig was detected as circular or the bacterial contig contained a MetaPhlAn 2.0 (Truong et al., 2015) bacterial taxonomy-specific marker gene (megablast; *E* value $<10^{-10}$). Since the detected prophage sequences could include partial sequences from a single prophage region, the detected prophage sequences located within 1 kb on a contig were merged. Last, the detected prophage sequences longer than 10 kb or the prophage sequences detected by ssDNA phage contigs (known ssDNA phage genomes are shorter than 10 kb) were extracted and listed as candidate prophage sequences.

The candidate prophage sequences from the first and second approaches were compared to define the final prophage sequences. We first considered the prophage sequences detected by the circular viral contigs using the second approach to be highly reliable. Then the prophage sequences from the first approach that overlapped with the circular viral contig-derived prophages (>75% overlap in either sequence) were removed from the candidate list. We next examined the rest of the candidate prophage sequences. If the prophage sequences from the two approaches overlapped (>75% overlap in either sequence), the overlapping prophage sequences were merged. The merged and non-overlapping candidate prophage sequences, as well as the circular viral contig-derived prophage sequences, were defined as the final prophage sequences.

To assess the "activated" prophages that were induced and released from host bacteria, the detected prophage sequences were compared with the viral contig sequences. A prophage was considered as "activated" if the prophage sequence was aligned (megablast; $\geq$95% identity) with a viral contig sequence (>75% overlap in either sequence) (all of the prophages detected by the second approach may be "activated" prophages).

Validation of the Assembled Contigs

To validate the assembled contig sequences, PCR primer pairs for the prophage boundary regions in the bacterial contig shown in Figure S2A were designed using Primer-BLAST (Ye et al., 2012) and Edesign (v2.0.1) (Kimura et al., 2016). PCR was performed with the following primers: 2D-1-Fw: 5'- GGCAGTCTTTCACCATCTTCG-3'; 2D-1-Rv: 5'- GCGGTCACAAATGGAACGG-3'; 2D-2-Fw: 5'-TGGAAAAGCATTGTGCCGAC-3'; 2D-2-Rv: 5'-TCAACTCAACACAGAACGATTGA-3' (Figure S2B). All of the PCR amplicons were analyzed using MCE-202 MultiNA with DNA-12000 and sequenced by the Sanger method.

Analysis of CRISPR Spacers

Since CRISPR regions are diverse across individuals, the bacterial contigs from individual samples were analyzed to determine the CRISPR regions. CRISPR repeats and spacers on bacterial contigs ($\geq$5 kb) from individual samples were predicted using the CRISPR array identification program CRISPRDetect (Biswas et al., 2016) (-array_quality_score_cutoff 3). The identified CRISPR spacers were linked to the representative (pooled) bacterial contigs using clustering information from CD-HIT-EST performed during contig pooling and the redundancy removal process.

To identify the target phages of the CRISPR spacers, we queried the predicted spacers using blastn (BLAST+ v.2.7.1) (Camacho et al., 2009) against the viral contigs and extracted the aligned spacers with >90% of their length aligned with a minimum identity level of 95% and a maximum *E* value of 5×10^{-3}. The former two criteria were the most important; however, the *E* value cutoff aids the removal of short spacers that may be false positives. The viral contigs constructed by our virome pipeline were used for the target search. For analysis of the spacers detected in the 17 cultured *C. difficile* strains, we also used the viral contigs stored in the public viral database IMG/VR (version IMG_VR_2018-01-01_3), with viral classification performed by our pipeline.

Viral genes targeted by CRISPR spacers were investigated by additionally querying the spacers using blastn (BLAST+ v.2.7.1) against the nucleotide sequences of the ORFs. When a spacer was aligned to an ORF with 90% or more of its length showing a minimum identity level of 95% and a maximum *E* value of 5×10^{-3}, the spacer was considered to be targeting an ORF. Otherwise it was considered to be targeting a non-ORF region. ORFs were annotated with the Pfam protein families (*E* value $<10^{-5}$) and were assigned related gene categories using a customized Pfam–category annotation table shown as Table S2. Since the Pfam annotations to *Microviridae* ORFs were found almost exclusively for the major capsid protein (PF02305) but rarely for other proteins, the *Microviridae* ORFs were annotated by searching against the RefSeq proteins (v84) of *Microviridae* using blastp (BLAST+ v.2.7.1) with an *E* value $<10^{-5}$. The annotations were categorized using the nomenclature used for *Chlamydia* phage genomes (i.e., VP1: major capsid protein, VP2: DNA pilot protein, VP3: internal scaffolding protein, VP4: replication initiation protein, and VP5: DNA binding protein).

Analysis of *C. difficile* Phage-Derived Endolysins

The ORFs that putatively encode endolysin protein were extracted from the *C. difficile* contigs from the 101 healthy individuals and the prophage regions found in the 17 cultured *C. difficile* strains using their ORF annotations. The viral RefSeq protein annotations containing "Endolysin"/"endolysin" or the PfamA annotations, Amidase_2, Amidase_3, or Amidase_5, were extracted. To avoid false positives, the ORFs were selected using the following criteria: (i) the ORF was annotated with a significant *E*-value $<10^{-15}$; (ii) the ORF length was <330 amino acids; and (iii) the ORF contained >50 amino acids other than the enzymatically active domain (EAD, annotated as Amidase) for the cell wall-binding domain (CBD). To investigate their conserved sequences, the ORFs were grouped according to their Pfam family (Amidase_2 or Amidase_3, no significant alignment was found for Amidase_5) and multiple alignments of the protein sequences in each group were performed with MUSCLE (Edgar, 2004) with default settings. Jalview (Waterhouse et al., 2009) was used to visualize the multiple alignment results.

Gene prediction was performed as described above and each gene is shown in Figure 4D. His-SUMO-tagged endolysins were obtained by PCR using genomic DNA from human feces or *C. difficile* as a template and were ligated into the *Bam*HI and *Sal*I sites of the pCold-SUMO expression vector (Guo et al., 2016) (Creative Biogene, Shirley, NY). The primer sequences are shown in Table S3. The plasmids were then transformed into BL21(DE3) cells, which were grown in LB medium containing 100 μg/ml ampicillin at 37°C until the OD_{600} reached 0.4–0.6. Isopropyl β-D-thiogalactoside (IPTG) (Nacalai Tesque) was added until the final concentration reached 1 mM, and the culture media was incubated at 16°C for 18 h at 120 rpm. After centrifugation at 3,000 × *g* for 15 min, the pellet was washed with sterile deionized water. After centrifugation at 3,000 × *g* for 15 min, the precipitate was resuspended in xTractor™ buffer (Takara, Shiga, Japan). The lysate was incubated with 10 U/ml of DNase I (Roche) and 100 μg/ml of hLysozyme (Sigma-Aldrich) for 30 min and disrupted by sonication (20 s pulse, 80 s rest, over 10 min). After centrifugation at 12,000 × *g* for 30 min, the supernatant was sterile-filtered through 0.45 and 0.2 μm filters. The target proteins were purified through Capturem™ His-Tagged Purification Maxiprep columns (Takara). The lysate was loaded onto the column, which was equilibrated with xTractor™ buffer, and then centrifuged at 2,000 × *g* for 3 min at room temperature. The column was washed with wash buffer (20 mM Na_3PO_4, 150 mM NaCl, pH 7.6), and the target protein was eluted with elution buffer (20 mM Na_3PO_4, 500 mM NaCl, 500 mM imidazole, pH 7.6).

All samples were desalted by HiTrap™ desalting columns (Amersham Biosciences) with HiTrap buffer (50 mM Na_3PO_4, 0.15 M NaCl, pH 7.0). The protein solution was loaded into the concentration tube (Amicon® Ultra-15 10K) (Millipore) and centrifuged at 5,000 × *g* for 20 min at room temperature. The concentration of the target protein was measured using Protein Assay CBB Solution (Nacalai Tesque).

SDS-PAGE

The samples collected after purification and concentration, as described above, were subjected to 5%–20% SDS-PAGE and incubated with Q-Stain (NIPPON Genetics Co., Ltd., Tokyo, Japan).

Lytic Activity Analysis

C. difficile strain ATCC 43255 was grown under anaerobic conditions and was harvested by centrifugation at 3,000 × *g* for 15 min. Cell pellets were washed and resuspended in HiTrap buffer (50 mM Na_3PO_4, 0.15 M NaCl, pH 7.0). The lytic activity of endolysin was calculated based on reduction at an optical density of 600 nm (OD_{600}) as measured in a TVS062CA BioPhoto recorder (ADVANTEC, Tokyo, Japan), with the OD_{600} measured every 1 min. For each sample, 400 μg of endolysin (20 μg/ml final concentration) was added to the cell resuspension.

Endolysin Therapy in the Mouse Acute CDI Model

The protocol used to established the *in vivo* murine acute CDI model was modified from a previously described method (Chen et al., 2008). An antibiotic mixture of kanamycin (0.4 mg/ml) (Nacalai Tesque), gentamicin (0.035 mg/ml) (Nacalai Tesque), colistin (850 U/ml) (Sigma-Aldrich), metronidazole (0.215 mg/ml) (Nacalai Tesque), and vancomycin (0.045 mg/ml) (DUCHEFA Biochemie, Haarlem, The Netherlands) was prepared in drinking water. Mice were allowed to drink the antibiotic cocktail for 5 days, followed by autoclaved water for 2 days. Mice received a single dose of clindamycin (20 mg/kg) intraperitoneally 1 day before *C. difficile* (strain ATCC 43255) inoculation. At day zero, mice were challenged with 200 μl of 2×10^5 CFU *C. difficile* cells via oral gavage. Twelve hours after *C. difficile* challenge, 400 μg of endolysin (JCM 5246_03, JCM 5252_02, and JCM 5252_07) in 200 μl HiTrap buffer, 400 μg of His-SUMO protein in 200 μl HiTrap buffer, or 200 μl of HiTrap buffer was administered intrarectally to mice under anesthesia. All mice were followed for 7 days, and the mouse survival data were analyzed by Kaplan–Meier curves. Statistical analyses were performed with R software.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical details related to sequence data and experiments can be found in the figure legends and methods. P values of < 0.05 were considered significant.