Thousands of previously unknown phages discovered in whole-community human gut metagenomes

Sean Benler
National Institutes of Health
Natalya Yutin
National Institutes of Health
Dmitry Antipov
St Petersburg State University
Mikhail Raykov
Saint-Petersburg State University: Sankt-peterburgskij gosudarstvennyj universitet
Sergey Shmakov
National Institutes of Health
Ayal B Gussow
National Institutes of Health
Pavel A Pevzner
University of California San Diego
Eugene Koonin (koonin@ncbi.nlm.nih.gov)
https://orcid.org/0000-0003-3943-8299

Research

Keywords: dsDNA , microbiomes, metagenome mining, hallmark genes

Posted Date: February 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-89426/v2

Version of Record: A version of this preprint was published at Microbiome on March 29th, 2021. See the published version at https://doi.org/10.1186/s40168-021-01017-w.

Abstract

Background: Double-stranded DNA bacteriophages (dsDNA phages) play pivotal roles in structuring human gut microbiomes; yet, the gut virome is far from being fully characterized, and additional groups of phages, including highly abundant ones, continue to be discovered by metagenome mining. A multilevel framework for taxonomic classification of viruses was recently adopted, facilitating the classification of phages into evolutionary informative taxonomic units based on hallmark genes. Together with advanced approaches for sequence assembly and powerful methods of sequence analysis, this revised framework offers the opportunity to discover and classify unknown phage taxa in the human gut.

Results: A search of human gut metagenomes for circular contigs encoding phage hallmark genes resulted in the identification of 3,738 apparently complete phage genomes that represent 451 putative genera. Several of these phage genera are only distantly related to previously identified phages and are likely to found new families. Two of the candidate families, “Flandersviridae” and “Quimbyviridae”, include some of the most common and abundant members of the human gut virome that infect *Bacteroides, Parabacteroides* and *Prevotella*. The third proposed family, “Gratiaviridae”, consists of less abundant phages that are distantly related to the families *Autographiviridae*, *Drexlerviridae* and *Chaseviridae*. Analysis of CRISPR spacers indicates that phages of all three putative families infect bacteria of the phylum Bacteroidetes. Comparative genomic analysis of the three candidate phage families revealed features without precedent in phage genomes. Some “Quimbyviridae” phages possess Diversity-Generating Retroelements (DGRs) that generate hypervariable target genes nested within defense-related genes, whereas the previously known targets of phage-encoded DGRs are structural genes. Several “Flandersviridae” phages encode enzymes of the isoprenoid pathway, a lipid biosynthesis pathway that so far has not been known to be manipulated by phages. The “Gratiaviridae” phages encode a HipA-family protein kinase and glycosyltransferase, suggesting these phages modify the host cell wall, preventing superinfection by other phages. Hundreds of phages in these three and other families are shown to encode catalases and iron-sequestering enzymes that can be predicted to enhance cellular tolerance to reactive oxygen species.

Conclusions: Analysis of phage genomes identified in whole-community human gut metagenomes resulted in the delineation of at least three new candidate families of *Caudovirales* and revealed diverse putative mechanisms underlying phage-host interactions in the human gut. Addition of these phylogenetically classified, diverse and distinct phages to public databases will facilitate taxonomic decomposition and functional characterization of human gut viromes.

Background

The bulk of the human-associated virome resides in the distal gastrointestinal tract and is composed of tailed double-stranded (ds) DNA bacteriophages (dsDNA phages) [1-3] that, in the recent virus megataxonomy, are classified as the class *Caudoviricetes* under the phylum *Uroviricota* [4]. The ternary

interactions between phages, bacteria and their human hosts are being elucidated at an increasing pace through experiments on model systems and sequencing of the uncultured community of viruses (virome) [5-9]. Comparisons of the human gut virome within and between individuals unveils remarkable longitudinal stability and high diversity of resident phages [2, 10, 11]. Although the human gut offers a rich source of phage genomic diversity, the virome so far has been explored to a much lesser extent than the whole community (metagenome), composed of viruses, Bacteria and Archaea. The rapid growth of the public whole-community metagenomic data offers the opportunity to identify numerous novel phage genomes lurking in metagenomes.

Tailed dsDNA phages encapsidate their genome as a linear molecule, but depending on the terminal genomic arrangement, many complete phage genomes assemble into a 'circular' contig (i.e., a contig with direct terminal repeats) [12]. Thus, circularity can be used as one feature to identify putative complete phage genomes in viromes and metagenomes. However, the comparatively small size of dsDNA phage genomes (~50 kb, on average) [13] and the estimated low virus-to-microbe ratio in the gut (1 : 10) [1] jointly translate into a relatively small amount of phage DNA present in whole community metagenomic libraries [14]. Moreover, similar-sized plasmids also assemble into circular contigs [15]. A recently developed computational method aims to address this problem by focusing specifically on the assembly of circular phage genomes and their automatic discrimination from plasmids based on gene content [16]. The genetic repertoire shared between plasmids and phages, for example, the *parABS* partitioning system encoded by both *Escherichia coli* phage P1 and plasmids [17], can obfuscate their automatic annotation-based discrimination and necessitate manual curation. Despite these challenges, there is a pressing need to reduce the amount of viral "dark matter" in the human gut by identifying and classifying phages for reference-based analyses [18, 19].

The global organization of the virosphere was recently captured in a comprehensive, unified framework using protein domains encoded by viral hallmark genes to infer evolutionary connections between major groups of viruses [4] and subsequently approved by the International Committee on the Taxonomy of Viruses (ICTV) as the comprehensive, multi-rank taxonomy of viruses. In particular, dsDNA viruses possess either the HK97 fold or the double jelly-roll fold in their major capsid proteins, along with distinct ATPases involved in capsid maturation, and thus appear to have independent origins, justifying their separation into two realms (the highest virus taxon rank) [4]. Tailed dsDNA phages, with their HK97 major capsid proteins, comprise the order *Caudovirales* within the class *Caudoviricetes*, under the phylum *Uroviricota* (that also include the distantly related herpesviruses of animals) and are further classified into 9 families. With the now formally recognized ability to classify viruses from sequence data alone [20], phylogenomic analysis of uncultured phage genomes can delineate novel taxa.

Here, we describe 3,738 completely assembled phage genomes discovered by analysis of 5,742 whole-community human gut metagenomes. Using abundance, taxonomy and genomic composition as criteria to select genomes for further scrutiny, three groups of phages, all infecting bacteria of the phylum Bacteroidetes comprising potential new families, were analyzed in detail. All these candidate families, named "Quimbyviridae", "Flandersviridae", and "Gratiaviridae" consist of phages infecting bacteria of the

phylum Bacteroidetes, and the first two are widely distributed and abundant in human gut viromes. The phages in these families and others yet to be classified encode enzymes that are involved in the response of cells to oxidative stress, implicating phages in the tolerance of anaerobes to oxygen. Furthermore, comparative genomic analysis exposed genetic cassettes that are unique to some genera in each family and thus appear to be relatively recent acquisitions involved in phage-host interactions. Addition of all the phage genomes identified here to public databases will substantially expand the known phage diversity and augment taxonomic classification of the human gut virome.

Methods

Identification of phage genomes in human gut metagenomes

5,742 whole-community metagenome assemblies generated from human fecal samples were downloaded from the NCBI Assembly database (accessed 8/2019). To limit the search space to likely complete genomes, 95,663 'circular' contigs (50-200 bp direct overlap at contig ends) were extracted from these assemblies. Next, 304 phage-specific protein alignments from the CDD database [21] and 117 custom alignments (Yutin et al., *in press*) were converted to Hidden Markov Models (HMMs) using hmmpress (v. 3.2.1). Proteins in the 95,663 contigs were predicted by Prodigal (v. 2.6.3) [22] in the metagenomic mode and searched against the set of 304 phage-specific HMMs using hmmsearch, with the relaxed e-value cutoff of < 0.05. Contigs with at least one hit (n = 4,907) were selected for a second round of searches after correcting for re-assigned codons, as follows. All contigs were searched for the presence of tRNAs using tRNA-scan-SE (v. 2.0) [23]. In 212 contigs, an amber stop codon-suppressor tRNA was identified. ORFs were re-predicted for these contigs with the amber stop codon re-assigned to glutamine, given that this reassignment is most commonly observed in human gut phages [24, 25]. The re-translated contigs were added back to the database and all contigs were subjected to a second profile search with a stricter e-value cutoff (< 0.01). Contigs were classified as phages when exceeding 3 kbp in length and possessing at least one ORF that matched a capsid, portal or large terminase subunit protein profile below the e-value threshold. The phage classifications were cross-checked with Seeker [26] and ViralVerify [16]. In cases where both tools classified a contig as non-phage, the protein annotations were examined manually, revealing four contigs of ambiguous identity that were discarded.

Collection of phage genomes in GenBank

Taxonomic accession codes corresponding to all prokaryotic viruses were collected from the NCBI Taxonomy database and used to extract sequences longer than 3 kbp from the non-redundant nucleotide database (accessed 09/2019). The protein predictions for each genome sequence were retrieved using the 'efetch' functionality in the entrez direct command line tools [27]. Genomic sequences lacking protein predictions were discarded.

Dereplication and annotation of phage genomes

The collections of GenBank and human gut phage genomes were each dereplicated at 95% average nucleotide identity across 80% of the genome length using dRep (v. 2.6.2) [28] and its associated dependencies, Mash [29] and FastANI [30], with all other settings left as default. The proteins from these contigs were collected and clustered at 95% amino acid identity across 50% of the protein length using mmclust [31]. The representative protein sequences were combined into a single BLAST database and compared against the multiple sequence alignments (MSAs) in the CDD database [21] with PSI-BLAST [32] at an evalue cutoff of 0.01. If the representative protein sequence produced a significant result, the representative and all constituent members of the protein cluster were annotated using the best hit.

Phylogenetic reconstruction

Alignments of the large terminase subunit (TerL), capsid or portal protein were constructed as previously described [33]. Marker proteins from the metagenomic phages were combined with markers from GenBank phages into a single database and initially clustered to 50% amino acid identity using mmclust [31]. The clusters were aligned using MUSCLE [34]; cluster alignments were then compared to each other using HHsearch (v. 3.0) [35]. The cluster-cluster similarity scores were converted to distances as $-\ln(S_{A,B}/\min(S_{A,A},S_{B,B}))$, where $S_{A,B}$ is similarity between the profiles A and B, then, an unweighted pair group method with arithmetic mean (UPGMA) dendrogram was constructed using the estimated cluster distances. Tips of the tree (depth <1.5) were used to guide the pairwise alignment of the clusters at the tree leaves with HHalign, creating larger protein clusters. The resulting alignments were filtered to remove sites with more than 50% gaps and a homogeneity lower than 0.1 [36]. The filtered alignment was used to construct an approximate maximum-likelihood tree using FastTree [37], with the Whelan-Goldman models of amino acid evolution and gamma-distributed site rates. Examination of the trees identified 353 nearly identical PhiX-174 sequences that were removed from subsequent analyses as a contamination from a sequencing reagent.

Phage genome analysis

A gene-sharing network of phage genomes was constructed using Vcontact2 (v. 0.9.19) [38], with default search settings against the database of dereplicated GenBank phage genomes. The results were imported into Cytoscape (v. 3.8) [39] for visualization.

The ORFs for selected groups of phages (see the main text) were additionally annotated through HHblits searches against the Uniprot database clustered to 30% identity and the PDB database clustered to 70% identity (available at http://wwwuser.gwdg.de/~compbiol/data/hhsuite/databases/hhsuite_dbs/, accessed 02/2020) [40]. Genomes encoding a predicted reverse transcriptase (RT) were examined for the presence of repeats corresponding to a diversity-generating retroelement using DGRScan with default settings [41]. To identify repeats outside of the 10 kb RT-centered window (the default window of DGRScan), the template repeats were used as BLASTn queries against the encoding genome with the following parameters: -dust no -perc_identity 75 -qcov_hsp_perc 50 -ungapped -word_size 4.

Fractional abundance of phage genomes in metagenomes

Dereplicated phage genomes from the NCBI Genbank database were combined with the dereplicated gut phages into a single database and indexed for read recruitment using Bowtie2 [42]. A collection of 1,241 human gut viromes were downloaded from the NCBI SRA using the SRA-toolkit (v. 2.10) and quality filtered with fastp (v. 0.20.1) [43]. The quality-filtered virome reads were mapped to a database containing the reference human genome (GCF_000001405), phiX-174 (NC_001422.1) and cloning vectors (available from ftp://ftp.ncbi.nlm.nih.gov/pub/UniVec/) using Bowtie2 with default parameters. Unaligned, "decontaminated" reads were then recruited to the phage database using Bowtie2 with default parameters, except for the following additions "–no-unal –maxins 1000000". The length-normalized fractional abundance of each phage genome in each virome was calculated as described previously [1].

Host prediction from CRISPR-spacer matches

A database of CRISPR spacers was compiled from previous surveys of CRISPR-Cas systems [44, 45]. Each spacer was used as a BLASTN [46] query against the phage genomes, using word size of 8 and low complexity filtering disabled. A phage-host prediction was inferred if the spacer was 95% identical over 95% of its length to a phage sequence.

Prediction of anti-CRISPR proteins

Identification of anti-CRISPR proteins (Acrs) was carried as out as previously described [47]. Briefly, each protein was assigned a score by the Acr prediction model ranging between 0 and 1, where a higher score corresponds to a higher likelihood of the protein being an Acr. The proteins were then clustered at 50% amino acid identity and considered a candidate Acr if they satisfied the following criteria: 1) received a mean score of 0.9 or above, 2) are present in a directon of 5 or fewer genes, 3) at least one of the directons encodes an HTH domain-containing protein and 4) the cluster does not produce a hit with an HHpred probability greater than 0.9 to any PDB or CDD database sequence [21].

Results

Identification of novel phage genomes from whole-community human gut metagenomes

The collection of 5,742 whole-community assembled metagenomes was searched for the presence of complete phage genomes. To limit the search space to likely closed genomes, 'circular' contigs were extracted from these assemblies that contained direct repeats at their termini, within the typical *k*-mer size used to assemble short reads (50 – 200 bp, n = 95,663). Each contig was searched for open reading frames (ORFs) matching a known phage marker profile (i.e, the terminase large subunit, major capsid protein, or portal protein). In total, 3,738 contigs encode at least one ORF that passed the e-value and length cutoff criteria (Methods) (**Additional file 1**). Dereplication at approximately 95% mean nucleotide identity reduced the number of phage marker-matching contigs to 1,886 (**Additional file 2**). A subset of 664 contigs encoded all three markers, 531 encoded two of the three and the remaining 691 possessed a single detectable marker (**Additional file 3**). The putative phage contigs had a median length of 44.9 kb, which is consistent with the recent estimates of the median genome size of dsDNA phages [48]. To

exclude any contaminating contigs (e.g., a plasmid harboring an integrated phage), each was assessed with ViralVerify [16] and Seeker [26], two bioinformatic tools trained to discriminate phage genomes from other sequences. These tools classified all but 36 of the selected contigs as phages with varying levels of confidence (**Additional file 2)**. Upon manual examination for typical phage genes other than the markers, four of the 36 unassigned contigs were discarded and the remainder were found to represent false negative classifications by these tools as judged by the presence of signature phage genes. Although we cannot rule out the possibility that some non-phage contigs were retained erroneously, the results collectively suggest that the set of circular marker-matching contigs predominantly consists of complete phage genomes.

To determine the host ranges of the phages, a database of CRISPR spacers from prokaryotic genomes was used to query the metagenomic phages for potential matches. In total, 553 (29%) of the dereplicated phage genomes were found to be targeted by at least one CRISPR-Cas system allowing host prediction (**Additional file 4**). The most common predicted hosts were *Firmicutes* (323 phages), followed by *Bacteroidetes* (143), *Actinobacteria* (43), *Proteobacteria* (41) and *Verrucomicrobia* (4). Among the identified phages, 111 were predicted to infect at least two different bacterial genera, consistent with other studies demonstrating that related bacteria possess CRISPR spacers targeting the same phage [44, 49]. Notably, 359 of the dereplicated phages harbor at least one protospacer identical to another gut phage (**Additional file 4)**, indicating that a single CRISPR spacer often can confer immunity against multiple phages.

Many phages have been found to encode Anti-CRISPR proteins (Acrs) to parry CRISPR-Cas defenses [50-53]. Given their function in counter-defense, Acrs evolve rapidly and show limited sequence similarity to experimentally characterized Acrs, making inference challenging [50]. However, a machine-learning based method has been recently developed that utilizes genomic context to identify candidate Acrs [47]. Application of this method showed that 41 phages, 16 of which were found to be targeted by a CRISPR-Cas system of their inferred host, encoded at least one candidate Acr, (**Additional file 5**). The highest-scoring Acrs belong to four phages that are targeted by *Bifidobacterium* CRISPR-Cas systems. All four phages are $\geq$ 97% identical over $\geq$ 90% of their length at the nucleotide level to uncharacterized prophages in cultured *Bifidobacterium* isolates **(Additional file 5)**, confirming their host-tropism assignment via CRISPR spacer-protospacer matches. In these phages, two candidate Acr-encoding genes lie between the large terminase subunit and integrase (**Additional file 6**). The localization of the Acr-encoding genes suggests they are expressed not only upon initial entry into the host cell and during lysogeny [54], but also upon transition to the lytic program to prevent cleavage of progeny phage genomes by CRISPR-Cas, as demonstrated experimentally in *Listeria*-infecting phages [55]. Transcription of Acrs is typically regulated by HTH domain-containing proteins termed Acr-associated proteins (Acas) [56]. Indeed, in the *Bifidobacterium* phages identified here, a short HTH domain-encoding ORF is located immediately downstream of the Acrs and can be predicted to regulate the expression of these two genes throughout the phage lifecycle. While these uncharacterized *Bifidobacterium* phages possess the characteristic features of Acr loci, the great majority of the phages identified in this work did not harbor any detectable Acrs yet were targeted by CRISPR-Cas (**Additional file 4)**. Some of these phages might encode distinct

Acrs undetectable by the method we used that was trained on a collection of previously characterized Acrs, whereas others might employ alternative anti-CRISPR strategies.

Taxonomic decomposition of the gut phages identifies previously unknown putative families

Phylogenetic trees were constructed for the large terminase subunit (TerL), major capsid protein (MCP) and portal protein encoded in each phage genome using an iterative approach to construct the underlying alignments [33]. The trees were constructed alongside reference proteins derived from phage genomes extracted from the NCBI GenBank database. Reflecting the set of protein profiles employed to identify the phage contigs, 1,480 (78%) genomes were assigned to the phylum *Uroviricota*, 360 (19%) to the *Phixviricota* and 46 (2%) to the *Loebvirae* (**Additional file 3**). The phylum *Phixviricota* includes *Escherichia coli* phage phiX174 that is used as a sequencing reagent; however, the 360 *Phixviricota* phages detected in this analysis do not include any sequences closely related to phiX174 (see Methods). The remaining analyses focus on the taxonomic decomposition of the phages that belong to *Uroviricota*, given that these contigs represent by far the largest fraction of the recovered genomes and that the *Loebvirae* and *Phixviricota* phyla are the subject of recent taxonomic analyses [57, 58].

The phylum *Uroviricota* is organized into a single class (*Caudoviricetes*) and order (*Caudovirales*), but a new order encompassing the crAss-like phages, a common and apparently most abundant group of phages in the human gut virome, is being proposed [59]. Our profiles recovered 141 phage genomes (dereplicated from 601 total genomes) that displayed phylogenetic relationship with the crAss-like phages and are the subject of a separate study (Yutin *et al.*, in press). The *Caudovirales* are presently organized into 9 families, but 3 of these (*Myoviridae, Podoviridae, Siphoviridae*) are expansive and demonstrably polyphyletic [60-62] and were thus not used for family-level taxonomic assignment although the remaining 6 families represent only a small fraction of the phages available in GenBank. The phylogenetic tree of TerL, a hallmark protein that is frequently used for phylogenetic reconstruction of *Caudovirales* phages and appears to be the best phylogenetic marker thanks to its ubiquity among phages and high level of sequence conservation [62, 63], reveals only 34 gut phages that belong to one of these 6 ICTV-accepted families (**Figure 1** and **Additional file 3**). The remainder of these unclassified phages are likely to found new families presently composed entirely of uncultured phages or belong to families with a cultured representative that have yet to be defined under the new multi-rank taxonomy of viruses.

Selection of candidate families for comparative genomics

The taxonomic analysis based on phage hallmark proteins demonstrates that few phages in the human gut belong to a currently accepted ICTV-family. To prioritize candidate families for in-depth analysis, we next complemented the hallmark gene-based taxonomic analysis with whole-genome comparisons and abundance calculations of each phage relative to GenBank phages.

A gene-sharing network was constructed with the phages recovered from metagenomes and those deposited in GenBank. Edges are drawn between two viral genomes, represented as nodes, based on the

number of ORFs that share significant sequence similarity [38]. Most of the metagenome-recovered phages bore multiple connections within the network to GenBank phages, in agreement with the manual curation of these contigs as genuine phage genomes (**Figure 1B**). However, two large groups of phages (tentatively labelled “Flandersviridae” and “Gratiaviridae”) were weakly connected to the larger network, reflecting disparate genome content. The divergence of the gene content of these phages from those of previously known phages and their distinct position in phylogenetic trees (**Figure 1A** and see below) indicate that they represent novel genera, and likely, new families.

To quantify the fractional abundance [1] of each phage in the human gut viral community, reads from a collection of 1,241 human gut viromes were competitively mapped against a database containing the metagenome-recovered and GenBank phages. The majority of the genomes do not recruit any reads (“detection”) from more than 2% of the viromes (Q1-Q3, 0-2% of viromes) (**Figure 1C-D**), consistent with the previously reported individuality of the human gut viromes [2, 7, 11]. A notable exception are the crAss-like phages [64] that recruit at least one read from about one-third of the viromes (Q1 - Q3, 9 - 28%), in agreement with previous reports of their cosmopolitan distribution [65, 66]. One uncharacterized *Caudovirales* genome was frequently observed in the collection of human gut viromes (54%, **Figure 1C**), suggesting that this phage is also cosmopolitan. To rule out the possibility that the observed frequency stemmed from non-specific read mapping to one or a few loci, rather than the complete genome, the coverage of sequencing reads across the genome (accession OMAC01000147.1) was examined. The broad coverage of this genome in the viromes confirms that its frequent detection is not an artifact, although several loci present in the reference sequence were absent in the viromes (**Additional file 7**). The exceptional detection of this uncharacterized phage (hereafter referred to as Quimbyvirus, after the character Mayor Quimby from the *Simpsons*) in the human gut viral community warrants its detailed examination.

Thus, three groups of phages were selected for in-depth analysis based on their distinct positions in the phylogenetic trees of the marker genes (all three groups), combined with divergent gene contents (“Flandersviridae” and “Gratiaviridae”) and high abundance in the human gut viral community (“Flandersviridae” and “Quimbyviridae”). A comparative genomic analysis of each candidate family is presented below, case-by-case.

“Quimbyviridae” phages are abundant, hypervariable phages infecting *Bacteroides*

In the TerL phylogenetic tree, Quimbyvirus belongs to a group of phages whose closest characterized relatives include the *Vequintavirinae* and *Ounavirinae* subfamilies, under the now defunct *Myoviridae* family. To elucidate the taxonomic affiliation of Quimbyvirus, genomes from adjacent branches were examined (**Figure 2**). The median genome length of Quimby-like phages is 75.2 kb, close to the genome size of a branch basal to the Quimby-like branch, “group 4986” (72 kb), but smaller than the genomes of other phages in adjacent branches, *Ounavirinae* (88 kb) and *Vequintavirinae* (145 kb). Despite the similarity in genome size, phylogenetic reconstruction of the portal protein and MCP separate the Quimby-like phages from group 4986 (**Additional file 8**). Moreover, most Quimby-like phages encode a

DnaG-family primase and DnaB-family helicase that are both absent in group 4986. However, in one branch of Quimby-like phages, the primase was lost from the replication module. The genomes of this branch encode a protein adjacent to the DnaB-family helicase with significant structural similarity to the winged helix-turn-helix domain of RepA (HHpred probability, 96.5) (**Figure 2**). RepA-family proteins mediate replication of plasmids by interacting with host DnaG primases [67], suggesting that the RepA-like protein coopts the host primase during replication, triggering the loss of the phage-encoded *dnaG* in this lineage. Consistent with a RepA-mediated episomal replication strategy, no integrase is identifiable in the genomes on this branch yet the phages encode numerous antirepressors, proteins involved in the lysis-lysogeny decision of temperate phages [68, 69]. The rest of the Quimby-like phages harbor a full-length, three-domain tyrosine integrase, indicating that these phages integrate into their host cell genome (**Figure 3**). Based on the topologies of the TerL, portal, MCP and DnaG trees, we propose that Quimby-like phages represent a novel taxonomic group at the family rank (henceforth, the "Quimbyviridae"). The potential differences in replication strategies (episomal vs. integrated) combined with the topologies of the phylogenetic trees of marker proteins suggest that "Quimbyviridae" splits into two distinct subfamilies.

The Quimbyvirus genome aligns with a cryptic prophage of the bacterium *Bacteroides dorei* (CP011531.1), with 95% nucleotide sequence identity across 92% of its length, indicating that *B. dorei*, a common constituent of human gut microbiomes [70], carries a prophage closely related to Quimbyvirus. Inspection of the alignment shows that Quimbyvirus site-specifically integrates into the tRNA-Asp gene of *B. dorei*, a typical site of prophage integration [71]. The hosts of the other "Quimbyviridae" phages, determined through CRISPR-spacer analysis, include the *Prevotella*, *Bacteroides* and *Parabacteroides* genera within the phylum *Bacteroidetes* and the *Lachnospiraceae* within the phylum *Firmicutes*. In contrast, the hosts of group 4986 do not include any *Bacteroidetes*. The differences in the inferred host ranges support separating group 4986 from 'Quimbyviridae" phages and suggests that group 4986 might represent a novel family, but these genomes were not investigated further.

Some of the "Quimbyviridae" phages harbor diversity-generating retroelements (DGRs), a cassette of genes that selectively mutate a short locus, known as the variable repeat, that is part of a C-type lectin or an immunoglobulin-like domain [72, 73]. Targeted mutation of these domains yields proteins with altered binding affinities and specificities [74]. The DGR cassette in *Bordetella* phage BPP-1 of the genus *Rauchvirus* is the only experimentally studied DGR system in a phage, where diversification of the C-type lectin domain-containing tail fiber gene enables adsorption to different host cell receptors [75]. In Quimbyvirus, the RT component of the DGR is encoded by overlapping ORFs in all three frames (ORFs 52-54), suggesting that the active RT is produced by two programmed frameshifts. Although overlapping ORFs and programmed frameshifts have been identified in many compact tailed phage genomes [76-79], DGR RTs have thus far only been predicted to be encoded by a single ORF. To discern if the frameshifts render the RT inactive, the variable repeats were examined for adenine-specific substitutions, a hallmark of DGR-mediated variation [73]. The two variable repeats reside in ORF 47 and 80 of the Quimbyvirus genome, which both encode proteins containing C-type lectin domains, the canonical target of DGRs [72] *(***Figure 2***)*. Alignment of the variable repeats with their cognate template repeats from nearly identical

Quimbyvirus genomes (≥ 95 % average nucleotide identity) allowed the detection of 22 adenine sites in the variable repeat exhibiting substitutions whereas all other bases were nearly perfectly conserved (**Additional file 9**). Collectively, these results suggest that the frameshifted RT possesses the selective infidelity that characterizes DGR-mediated hypervariation.

The first variable repeat resides in the C-terminus of ORF 51 that is located downstream of the tail fiber genes, suggesting that this gene codes for a structural component of the virion, similar to the hypervariable tail fiber of phage BPP-1 [75, 80]. The second DGR target locus is in ORF 84 that is distal to the phage structural gene module and is expressed from the opposite DNA strand, suggestive of a non-structural protein. The genomic neighborhood of ORF 80 includes genes coding for a nuclease, four methyltransferases and a tRNA ligase within 7 kb. The nuclease shows significant sequence and structural similarity to *E. coli mutY* (HHpred probability, 97.3, **Additional file 10**), a DNA glycosylase involved in base excision repair. The methyltransferases are most similar to adenine- and cytosine-modifying enzymes (HHpred probability 100 and 99.9, respectively, **Additional file 10**) that likely prevent cleavage by host restriction endonucleases. Similarly, the tRNA ligase might repair tRNAs cleaved by host anticodon endonucleases [81]. Overall, the adjacency of ORF 84 with defense- and counterdefense-related genes implies that this hypervariable phage protein plays a role in the phage-host conflicts; however, the exact functions of the DGR and hypervariable target proteins during the life cycle of “Quimbyviridae” phages remain to be investigated.

“Flandersviridae” phages are common and abundant in whole-community metagenomes

Analysis of the phylogenetic trees of TerL identified a deep branch of 29 gut phages (dereplicated from 196 total genomes) that joins the family *Ackermannviridae* (**Figure 3A**). Annotation of the ORFs encoded by the 29 representative contigs demonstrated that the genomes are colinear, confirming that they belong to a cohesive group (**Figure 3B**). The cohesiveness of this group was confirmed by the gene-sharing network, where these genomes form a coherent cluster that has few connections to the larger network (**Figure 1B**), reflecting distant (if any) similarity between most of the proteins encoded by these phages and proteins of phages in GenBank. The median genome size of the phages in this group is 85.2 kb, compared to 157.7 kb among the *Ackermannviridae* phages. There is a conserved module of structural genes that encode the MCP, portal, sheath and baseplate proteins, TerL and the virion maturation proteinase. The presence of a contractile tail sheath indicates that these viruses possess contractile tails similar to those in the family *Ackermannviridae*, in agreement with the TerL phylogeny. Several of the genes within the structural block contain immunoglobulin-like or C-type lectin domains (e.g., BACON and GH5, respectively), which are predicted to play a role in adhesion of the virion to bacterial cells or host-associated mucosal glycans [82-85]. Downstream of the structural block is a module of genes involved in DNA replication that includes a DnaB-family helicase, DnaG-family primase and DNA polymerase I (PolA). The *polA* gene is widely distributed among dsDNA phages and therefore serves as a useful marker for delineating the diversity of phage replication modules [86]. Phylogenetic reconstruction of both *polA* and *dnaG* encoded by these phages confirmed their monophyly (**Additional file 11**). Following the replication module is an approximately 20 kb long locus containing ORFs that showed no detectable similarity to

functionally characterized proteins. Two of the phages harbor matches to CRISPR spacers encoded by *Bacteroides* and *Parabacteroides* spp., indicating these bacteria serve as hosts. Based on the large terminase and *polA* phylogeny, colinearity of their genomes and differences from known phages in both genome size and content, we propose that these *Bacteroides*-infecting phages represent a novel taxonomic group, with a family rank hereafter “Flandersviridae” (after the region where some of the metagenomes were sampled).

Although all members of the “Flandersviridae” are syntenic, some contain an insertion of two adjacent genes encoding nucleotidyltransferase superfamily enzymes within the DNA replication module. One enzyme belongs to the *ispD* family that is involved in the biosynthesis of isoprenoids [87, 88], and the other is a *licD* family enzyme that is responsible for the addition of phosphorylcholine to teichoic acids present in bacterial cell walls [89] (**Figure 3**). To our knowledge, neither of these enzymes has been reported in phages previously. Given that only some members of the “Flandersviridae” possess these genes, they are unlikely to perform essential functions in phage reproduction, and instead could be implicated in phage-host interactions. The *licD* family enzyme might modify teichoic acids to prevent superinfection by other phages, given that these polysaccharides serve as receptors for some phages to adsorb to the host cells [90]. The role of *ispD* is less clear because *ispD* family enzymes catalyze one step in the biosynthesis of isopentenyl pyrophosphate, a building block for a large variety of diverse isoprenoids [91]. Phages manipulate host metabolic networks including central carbon metabolism, nucleotide metabolism and translation [92]; the discovery of *ispD* present in the “Flandersviridae” phage genomes might add to this list the isoprenoid biosynthetic pathway.

Complete “Flandersviridae” phage genomes were recovered from 249 whole-community human gut metagenomic assemblies. Their frequent assembly into closed contigs suggests that these phages might persist in their host cells as extrachromosomal circular DNA molecules, similar to phage P1 [93]. However, neither genes involved in DNA partitioning nor lysis-lysogeny switches are readily identifiable in the “Flandersviridae” genomes. Thus, this group of phages might be obligately lytic although discerning the lifestyle of a phage from the genome sequence alone is challenging [94]. Regardless of their lifestyle, the frequent recovery of these phages from whole-community metagenomes implies that they are common members of the human gut virome. Indeed, the “Flandersviridae” phages reach similar detection frequency as the crAss-like phages (**Figure 1D**) although there are fewer Flanders-like phages in the database. Like the “Quimbyviridae”, the even coverage of sequencing reads across one “Flandersviridae” genome (accession OLOC01000071.1) confirms its detection is not artifactual (**Additional file 12**). The high fractional abundance and detection of Flanders-like phages in viromes generally agrees with their frequent assembly from whole community metagenomes although they were not the most abundant (see Discussion). Overall, Flanders-like phages represent a previously undetected phage group that is widely distributed in human gut viromes.

“Gratiaviridae”, a putative novel family of phages infecting *Bacteroides*

A deeply branching cluster of 18 genomes (dereplicated from 45 total) is basal to the families Autographiviridae, Drexlerviridae and Chaseviridae on the TerL phylogenetic tree (**Figure 4A)**. Although not commonly present in gut viromes (**Figure 1D**), the deep relationship between these contigs and established phage families prompts in depth genome analysis of these putative phages. All 18 genomes encode a DnaG-family primase and a DnaE-family polymerase, and phylogenetic reconstruction for these genes demonstrates monophyly of these phages; the sole exception is the *dnaE* gene of bacteriophage phiST, a marine *Cellulophaga*-infecting phage that belongs to the polyphyletic, currently defunct Siphoviridae family [95] (**Additional file 13**). The *dnaG* and *dnaE* genes are nested within a module of other replication-associated genes that include superfamily I and II helicases, SbcCD exonucleases and a RecA family ATPase (**Figure 4B**). The structural module is composed of genes that encode an MCP, capsid maturation protease, portal protein, baseplate proteins and a contractile tail sheath protein. Although these genomes are not strictly colinear as observed for the “Flandersviridae” phages, the overall similarity of the proteins encoded by these phages is apparent in the gene-sharing network where they form a coherent cluster that shares some edges with the crAss-like phages (**Figure 1B**). Similar to crAss-like phages, the predicted hosts suggested by CRISPR-spacer matches are the *Bacteroides* and *Parabacteroides* genera (**Additional file 4)**. Taken together, the phylogenetic and genomic organization of these phages indicate that they represent a new family, provisionally named “Gratiaviridae” (after the pioneering phage biologist Dr. Andre Gratia).

In addition to structural and replication proteins, “Gratiaviridae” phages encode several enzymes of the ferritin-like diiron-carboxylate superfamily. The ferritin-like enzymes encoded by these phages belong to two families, namely, DNA protecting proteins (DPS) and manganese-catalases. Manganese-catalases have not been documented in phage genomes and DPS-like enzymes have only been observed in seven *Lactobacillus*-infecting phages [96]. Both enzymes are involved in the tolerance of anaerobes to oxidative stress. Catalases detoxify hydrogen peroxide to oxygen and water, enhancing survival of anaerobic *Bacteroides* in the presence of oxygen [97]. DPS enzymes catalyze a reaction between oxygen and free iron to yield insoluble iron oxide, lowering the concentration of both intracellular oxygen and free iron levels that would otherwise react with hydrogen peroxide and produce a hydroxyl radical, the most toxic reactive oxygen species [98, 99]. “Gratiaviridae” phages might deploy catalase- and DPS-like enzymes during infection to enhance the tolerance of their strictly anaerobic *Bacteroides* hosts to oxidative damage. Notably, these enzymes were not restricted to the “Gratiaviridae” but could be identified in 196 (manganese catalase) and 36 (DPS) other phage genomes, including the “Flandersviridae”. The frequent identification of these enzymes in gut phage genomes underscores the importance of intracellular iron and reactive oxygen species concentration for productive infections in an anaerobic environment.

Five of the “Gratiaviridae” phages encode a protein containing a serine/threonine protein kinase domain with distant but significant sequence similarity to HipA family kinases (HHpred probability 99, **Additional file 10**). Whereas HipA family kinases are present in numerous, phylogenetically distinct bacterial genomes as the toxin component of a distinct variety of type II toxin-antitoxin systems [100, 101], there are only two characterized examples of protein kinases encoded by phages. The protein kinase of T7-like phages phosphorylates RNA polymerase and RNAse III early during infection as part of the takeover of

the host cell transcriptional and translational machinery [102-104]. In contrast, the protein kinase of *E. coli* phage 933W is expressed during lysogeny and mediates abortive infection upon superinfection of the host cell by phage HK97 [105]. The HipA-like kinase is unlikely to function early during infection like the kinase of T7-like phages because, in all five “Gratiaviridae” phages, the kinase is encoded between the portal protein and MCP genes, which are expressed late during infection in numerous cultured phages [106, 107]. Instead, the kinase might confer immunity to heterotypic phage infection, analogous to the kinase encoded by 933W [105]. In support of an immunity-related role, an AAA-family ATPase and a glycosyltransferase are encoded immediately upstream of the kinase in all five phage genomes (**Figure 4**). Glycosyltransferases are encoded within capsular polysaccharide biosynthetic loci [108] and phase variation of the capsular polysaccharides confers immunity from phages that rely on these molecules for adsorption [109]. The specific roles of the HipA-family kinase, ATPase and glycosyltransferase are unknown but, collectively, these enzymes might modify host cell capsules, granting temporary immunity to heterotypic phage infection while the morphogenesis of “Gratiaviridae” progeny virions completes.

Discussion

A search of human gut metagenomes identified 3,738 putative complete phage genomes. In an attempt to recover complete phage genomes, this analysis restricts the search space to metagenomic contigs with direct terminal repeats which are present at the termini of some phage genomes that consequently form circular assemblies [12]. Circular assemblies can also arise upon sequencing a concatemer of DNA present during phage DNA replication and packaging [12], in which case the direct repeats are a technical artifact and are the same length as the k-mer size used to assemble the contigs. Phages with different replication and DNA packaging strategies, such as members of the phyla *Preplasmiviricota, Dividoviricota*, or *Escherichia* phage Mu [12], that lack direct repeats do not yield circular assemblies and thus were not detected here. As a result, the set of phage genomes recovered by this strategy is both biased and an underestimate. The results are also skewed towards smaller genomes that are more likely to assemble into a single contig although, in one metagenome, a 294 kb phage genome was identified (**Additional file 1)**. Despite these limitations, phylogenetic and comparative genomic analyses suggest that this set of contigs includes many previously unnoticed lineages of phages, some, most likely, at the family rank.

The family-rank phage lineages proposed here were defined using a combination of approaches. Principally, phylogenetic analysis of the large terminase subunit (TerL), a hallmark gene of the *Uroviricota* phylum, revealed branches of genomes distinct from any of those reported in the GenBank database (Fig. 1A). Supporting the phylogenetic results, the genomes of phages on adjacent branches were of similar length and largely syntenic, whereas distant branches possess entirely different architectures (**Figs. 2-4**).

The phages in two of the three proposed families (“Flandersviridae” and “Gratiaviridae”), the genomes are largely disconnected from other phages in the gene-sharing network. The phages in the family “Quimbyviridae” share genes with numerous phylogenetically distinct phages (**Fig. 1C**), although not enough to warrant automated assignment into the same “viral cluster” (**Additional File 2)** [38]. Phages

that infect phylogenetically related hosts share genes more frequently with one another than they do with phages of distantly related hosts [110, 111]. The proximity of *Quimbyviridae* with other phages in the gene sharing network, most likely, reflects a similar preference for *Bacteroides* hosts (**Additional file 4)**. Phylogenetic reconstruction of hallmark genes helped to delineate the *Quimbyviridae* as a distinct group which was otherwise obscured by the numerous connections in the gene sharing network. Overall, a combination of phylogenetic analysis of hallmark genes and gene sharing analysis will facilitate the taxonomic classification of the gut viral community into higher levels of organization.

Two groups of phages were selected for in-depth analysis based on their frequent recovery in metagenomes and viromes. Complete genomes of “Flandersviridae” and “Quimbyviridae” phages were assembled in 249 and 20 whole-community metagenomes, respectively. Yet, Quimbyvirus was more frequently detected in the viromes than any “Flandersviridae” phage (**Figure 1D**). The discrepancy can be attributed to several factors, including sampling bias, the greater number of “Flandersviridae” genomes in the reference database “diluting” the number of mapped reads per genome, or the presence of variable loci (e.g., the variable repeats of DGRs) that break contig assemblies [112]. Regardless, both groups encompass abundant members of the human gut virome. Predictably, the hosts of these phages include *Bacteroides spp.*, which are some of the most dominant bacterial taxa of the human gut [113] and serve as hosts for other common human gut phages [66, 114]. Much of the uncharacterized “dark matter” in these phage genomes is likely to be dedicated to preventing superinfection of the *Bacteroides* host cells by such phages and to counter the host defenses. Although in general defense systems in *Bacteroidetes* remain poorly characterized, most of the bacteria possess active CRISPR-Cas systems, and numerous CRISPR spacers targeting the phages analyzed here were detected using stringent thresholds with a low estimated false discovery rate (0.06) [44]. This implies that many if not most of the phages infecting *Bacteroidetes* would encode Acrs. However, the currently available prediction method that was trained on the sequences of previously identified Acrs detected putative Acrs only in a small minority of these phages. The remaining phages of *Bacteroidetes* might encode distinct Acrs or employ alternative anti-CRISPR strategies.

Several phage genera possess DGRs, including Quimbyvirus. Metagenomic surveys have shown that DGRs are enriched in the viruses that inhabit gastrointestinal environments [112, 115]. Combined with the induction of DGR-carrying phages from human gut bacteria [116, 117], these observations reflect a prominent role of hypervariability underlying phage-host interactions in the gastrointestinal environment. Notably, Quimbyvirus and another DGR-carrying phage (Hankyphage, BK010646.1), lysogenize the same *Bacteroides* species and both phages are frequently detected in human gut viromes [117]. The commonalities aside, the Quimbyvirus DGR RT is encoded by three overlapping reading frames and targets two proteins, one in the structural module and one in a defense-related island. DGRs have been associated with putative defense and signaling systems in cyanobacterial and gammaproteobacterial genomes [118, 119], but beyond the presence of the C-type lectin fold, the hypervariable proteins possess few other recognizable domains that obfuscate their precise roles.

The third group analyzed in this study, the “Gratiaviridae”, is not abundant but occupies a deep position on the TerL tree relative to the *Autographiviridae*, *Chaseviridae* and *Drexlerviridae* families. Analysis of the “Gratiaviridae” genomes will facilitate the future organization of these families into higher taxonomic ranks, potentially, at the order level. Furthermore, analysis of the “Gratiaviridae” genomes demonstrated the presence of catalase- and DPS-family enzymes that arbitrate cellular responses to oxidative stress [120]. Oxygen concentrations vary along the length of the gastrointestinal tract, where the concentration is lower in the distal vs. proximal gut [121]. Oxygen also diffuses from tissues radially into the lumen [122] and, in combination with other factors, these gradients affect the structure and composition of the gastrointestinal microbiota [123]. The acquisition of oxygen detoxifying-enzymes by the “Gratiaviridae” and other gut phages signals a need to supplement their host cell’s tolerance to oxidative damage during infection, which might be especially important for cells that reside near the tissue surface where oxygen exposure is higher.

A unique feature of some “Gratiaviridae” phages is a HipA-family protein kinase. The T7-like phages (within the *Autographiviridae* family) and *Escherichia* phage 933W (currently unclassified at the family level) encode PKC-family protein kinases that function during host cell takeover and abortive infection, respectively [102, 105]. A third, CotH-family protein kinase domain is occasionally observed in phage genomes where it is fused to a hypervariable C-type lectin domain [72, 115], but these proteins are currently unstudied. The “Gratiaviridae” phages recruited a fourth family of protein kinases that, together with the phage encoded glysosyltransferase, might modify the host cell envelope, contributing to the prevention of superinfection.

Conclusions

In summary, comparative genomic analysis of the phages described here, along with the complementary analysis of crAss-like phages (Yutin *et. al,* in press), substantially increases the characterized diversity of phages, primarily, those infecting Bacteroidetes bacteria, which are major components of the human gut microbiome. These findings also expand the repertoire of phage gene functions, notably, by adding the isoprenoid metabolic pathway, catalase-like enzymes, HipA family protein kinases and hypervariable genes implicated in defense. All of these open multiple directions for experimental study.

Declarations

Ethics approval and consent to participate – Not applicable

Consent for publication – Not applicable

Availability of data and materials – All phage genomes are available in the NCBI GenBank database using the accession numbers listed in **Additional file 1**. The genomes, protein predictions and annotations are also provided at ftp://ftp.ncbi.nih.gov/pub/yutinn/benler_2020/gut_phages/. Underlying alignments

of marker protein sequences used to generate phylogenetic trees and the source data used to generate Figs. 1C-D are also available on the FTP site.

Competing interests - The authors declare that they have no competing interests.

Funding – Funding for this project was provided by the Intramural Research Program of the National Institutes of Health (National Library of Medicine). MR & DA were supported by St. Petersburg State University, Russia (grant ID PURE 51555639). The funding body had no role in the collection, analysis, and interpretation of data or writing of the manuscript.

Authors contributions –NY, MR and DA analyzed the gut metagenomes; SS and AG executed the CRISPR analyses; SB analyzed the phage genomes and wrote the manuscript; PP and EVK conceived the project and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements – We are grateful for the assistance from Yuri Wolf with phylogenetic reconstruction of the phage hallmark genes and to Koonin group members for useful discussions. This work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov).

References

1. Cobián Güemes AG, Youle M, Cantú VA, Felts B, Nulton J, Rohwer F: **Viruses as Winners in the Game of Life**. *Annual Review of Virology* 2016, **3**(1):197-214.
2. Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, Gordon JI: **Viruses in the faecal microbiota of monozygotic twins and their mothers**. *Nature* 2010, **466**(7304):334-338.
3. Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, Salamon P, Rohwer F: **Metagenomic analyses of an uncultured viral community from human feces**. *Journal of bacteriology* 2003, **185**(20):6220-6223.
4. Koonin EV, Dolja VV, Krupovic M, Varsani A, Wolf YI, Yutin N, Zerbini FM, Kuhn JH: **Global Organization and Proposed Megataxonomy of the Virus World**. *Microbiology and Molecular Biology Reviews* 2020, **84**(2):e00061-00019.
5. Barr JJ, Auro R, Furlan M, Whiteson KL, Erb ML, Pogliano J, Stotland A, Wolkowicz R, Cutting AS, Doran KS *et al*: **Bacteriophage adhering to mucus provide a non–host-derived immunity**. *Proceedings of the National Academy of Sciences* 2013, **110**(26):10771-10776.
6. Roach DR, Leung CY, Henry M, Morello E, Singh D, Di Santo JP, Weitz JS, Debarbieux L: **Synergy between the Host Immune System and Bacteriophage Is Essential for Successful Phage Therapy against an Acute Respiratory Pathogen**. *Cell Host & Microbe* 2017, **22**(1):38-47.e34.
7. Moreno-Gallego JL, Chou S-P, Di Rienzi SC, Goodrich JK, Spector TD, Bell JT, Youngblut ND, Hewson I, Reyes A, Ley RE: **Virome Diversity Correlates with Intestinal Microbiome Diversity in Adult Monozygotic Twins**. *Cell Host & Microbe* 2019, **25**(2):261-272.e265.

8. Gogokhia L, Buhrke K, Bell R, Hoffman B, Brown DG, Hanke-Gogokhia C, Ajami NJ, Wong MC, Ghazaryan A, Valentine JF *et al*: **Expansion of Bacteriophages Is Linked to Aggravated Intestinal Inflammation and Colitis**. *Cell Host & Microbe* 2019, **25**(2):285-299.e288.
9. Waller AS, Yamada T, Kristensen DM, Kultima JR, Sunagawa S, Koonin EV, Bork P: **Classification and quantification of bacteriophage taxa in human gut metagenomes**. *The ISME Journal* 2014, **8**(7):1391-1402.
10. Minot S, Bryson A, Chehoud C, Wu GD, Lewis JD, Bushman FD: **Rapid evolution of the human gut virome**. *Proceedings of the National Academy of Sciences* 2013, **110**(30):12450-12455.
11. Shkoporov AN, Clooney AG, Sutton TDS, Ryan FJ, Daly KM, Nolan JA, McDonnell SA, Khokhlova EV, Draper LA, Forde A *et al*: **The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific**. *Cell Host & Microbe* 2019, **26**(4):527-541.e525.
12. Casjens SR, Gilcrease EB: **Determining DNA Packaging Strategy by Analysis of the Termini of the Chromosomes in Tailed-Bacteriophage Virions**. In: *Bacteriophages Methods in Molecular Biology.* Edited by Clokie M.R., A.M. K, vol. 502: Humana Press; 2009.
13. Al-Shayeb B, Sachdeva R, Chen L-X, Ward F, Munk P, Devoto A, Castelle CJ, Olm MR, Bouma-Gregson K, Amano Y *et al*: **Clades of huge phages from across Earth's ecosystems**. *Nature* 2020, **578**(7795):425-431.
14. Shkoporov AN, Ryan FJ, Draper LA, Forde A, Stockdale SR, Daly KM, McDonnell SA, Nolan JA, Sutton TDS, Dalmasso M *et al*: **Reproducible protocols for metagenomic analysis of human faecal phageomes**. *Microbiome* 2018, **6**(1):68.
15. Antipov D, Raiko M, Lapidus A, Pevzner PA: **Plasmid detection and assembly in genomic and metagenomic data sets**. *Genome Research* 2019, **29**(6):961-968.
16. Antipov D, Raiko M, Lapidus A, Pevzner PA: **metaviralSPAdes: assembly of viruses from metagenomic data**. *Bioinformatics* 2020.
17. Gideon SG, Wright A: **DNA Segregation in Bacteria**. *Annual Review of Microbiology* 2000, **54**(1):681-708.
18. De Sordi L, Lourenço M, Debarbieux L: **The Battle Within: Interactions of Bacteriophages and Bacteria in the Gastrointestinal Tract**. *Cell Host & Microbe* 2019, **25**(2):210-218.
19. Beller L, Matthijnssens J: **What is (not) known about the dynamics of the human gut virome in health and disease**. *Current Opinion in Virology* 2019, **37**:52-57.
20. Simmonds P, Adams MJ, Benkő M, Breitbart M, Brister JR, Carstens EB, Davison AJ, Delwart E, Gorbalenya AE, Harrach B *et al*: **Virus taxonomy in the age of metagenomics**. *Nature Reviews Microbiology* 2017, **15**(3):161-168.
21. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS *et al*: **CDD/SPARCLE: the conserved domain database in 2020**. *Nucleic Acids Research* 2019, **48**(D1):D265-D268.
22. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ: **Prodigal: prokaryotic gene recognition and translation initiation site identification**. *BMC Bioinformatics* 2010, **11**(1):119.

23. Lowe TM, Eddy SR: **tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence**. *Nucleic Acids Research* 1997, **25**(5):955-964.
24. Devoto AE, Santini JM, Olm MR, Anantharaman K, Munk P, Tung J, Archie EA, Turnbaugh PJ, Seed KD, Blekhman R *et al*: **Megaphages infect Prevotella and variants are widespread in gut microbiomes**. *Nature Microbiology* 2019, **4**(4):693-700.
25. Ivanova NN, Schwientek P, Tripp HJ, Rinke C, Pati A, Huntemann M, Visel A, Woyke T, Kyrpides NC, Rubin EM: **Stop codon reassignments in the wild**. *Science* 2014, **344**(6186):909-913.
26. Auslander N, Gussow AB, Benler S, Wolf YI, Koonin EV: **Seeker: Alignment-free identification of bacteriophage genomes by deep learning**. *bioRxiv* 2020:2020.2004.2004.025783.
27. Kans J: **Entrez Direct: E-utilities on the Unix Command Line**. Bethesda (MD): National Center for Biotechnology Information; 2010.
28. Olm MR, Brown CT, Brooks B, Banfield JF: **dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication**. *The ISME Journal* 2017, **11**(12):2864-2868.
29. Ondov BD, Starrett GJ, Sappington A, Kostic A, Koren S, Buck CB, Phillippy AM: **Mash Screen: high-throughput sequence containment estimation for genome discovery**. *Genome Biology* 2019, **20**(1):232.
30. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S: **High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries**. *Nature Communications* 2018, **9**(1):5114.
31. Steinegger M, Söding J: **Clustering huge protein sequence sets in linear time**. *Nature Communications* 2018, **9**(1):2542.
32. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Research* 1997, **25**(17):3389-3402.
33. Wolf YI, Kazlauskas D, Iranzo J, Lucía-Sanz A, Kuhn JH, Krupovic M, Dolja VV, Koonin EV: **Origins and Evolution of the Global RNA Virome**. *mBio* 2018, **9**(6):e02329-02318.
34. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Research* 2004, **32**(5):1792-1797.
35. Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J: **HH-suite3 for fast remote homology detection and deep protein annotation**. *BMC Bioinformatics* 2019, **20**(1):473.
36. Yutin N, Makarova KS, Mekhedov SL, Wolf YI, Koonin EV: **The Deep Archaeal Roots of Eukaryotes**. *Molecular Biology and Evolution* 2008, **25**(8):1619-1630.
37. Price MN, Dehal PS, Arkin AP: **FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments**. *PLOS ONE* 2010, **5**(3):e9490.
38. Bin Jang H, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, Brister JR, Kropinski AM, Krupovic M, Lavigne R *et al*: **Taxonomic assignment of uncultivated prokaryotic virus genomes is**

enabled by gene-sharing networks. *Nature Biotechnology* 2019, **37**(6):632-639.

39. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks**. *Genome Research* 2003, **13**(11):2498-2504.
40. Mirdita M, von den Driesch L, Galiez C, Martin MJ, Söding J, Steinegger M: **Uniclust databases of clustered and deeply annotated protein sequences and alignments**. *Nucleic Acids Research* 2016, **45**(D1):D170-D176.
41. Ye Y: **Identification of Diversity-Generating Retroelements in Human Microbiomes**. *International Journal of Molecular Sciences* 2014, **15**(8).
42. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nature Methods* 2012, **9**(4):357-359.
43. Chen S, Zhou Y, Chen Y, Gu J: **fastp: an ultra-fast all-in-one FASTQ preprocessor**. *Bioinformatics* 2018, **34**(17):i884-i890.
44. 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**(5):e01397-01317.
45. Shmakov SA, Wolf YI, Savitskaya E, Severinov KV, Koonin EV: **Mapping CRISPR spaceromes reveals vast host-specific viromes of prokaryotes**. *Communications Biology* 2020, **3**(1):321.
46. Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schäffer AA: **Database indexing for production MegaBLAST searches**. *Bioinformatics* 2008, **24**(16):1757-1764.
47. Gussow AB, Park AE, Borges AL, Shmakov SA, Makarova KS, Wolf YI, Bondy-Denomy J, Koonin EV: **Machine-learning approach expands the repertoire of anti-CRISPR protein families**. *Nature Communications* 2020, **11**(1):3784.
48. Luque A, Benler S, Lee DY, Brown C, White S: **The Missing Tailed Phages: Prediction of Small Capsid Candidates**. *Microorganisms* 2020, **8**(12):1944.
49. Soto-Perez P, Bisanz JE, Berry JD, Lam KN, Bondy-Denomy J, Turnbaugh PJ: **CRISPR-Cas System of a Prevalent Human Gut Bacterium Reveals Hyper-targeting against Phages in a Human Virome Catalog**. *Cell Host & Microbe* 2019, **26**(3):325-335.e325.
50. Hynes AP, Rousseau GM, Agudelo D, Goulet A, Amigues B, Loehr J, Romero DA, Fremaux C, Horvath P, Doyon Y *et al*: **Widespread anti-CRISPR proteins in virulent bacteriophages inhibit a range of Cas9 proteins**. *Nature Communications* 2018, **9**(1):2919.
51. Marino ND, Zhang JY, Borges AL, Sousa AA, Leon LM, Rauch BJ, Walton RT, Berry JD, Joung JK, Kleinstiver BP *et al*: **Discovery of widespread type I and type V CRISPR-Cas inhibitors**. *Science* 2018, **362**(6411):240-242.
52. Pawluk A, Davidson AR, Maxwell KL: **Anti-CRISPR: discovery, mechanism and function**. *Nature Reviews Microbiology* 2018, **16**(1):12-17.
53. Hwang S, Maxwell KL: **Meet the Anti-CRISPRs: Widespread Protein Inhibitors of CRISPR-Cas Systems**. *The CRISPR Journal* 2019, **2**(1):23-30.

54. Osuna BA, Karambelkar S, Mahendra C, Christie KA, Garcia B, Davidson AR, Kleinstiver BP, Kilcher S, Bondy-Denomy J: **Listeria Phages Induce Cas9 Degradation to Protect Lysogenic Genomes**. *Cell Host & Microbe* 2020.

55. Osuna BA, Karambelkar S, Mahendra C, Sarbach A, Johnson MC, Kilcher S, Bondy-Denomy J: **Critical Anti-CRISPR Locus Repression by a Bi-functional Cas9 Inhibitor**. *Cell Host & Microbe* 2020.

56. Stanley SY, Borges AL, Chen K-H, Swaney DL, Krogan NJ, Bondy-Denomy J, Davidson AR: **Anti-CRISPR-associated proteins are crucial repressors of anti-CRISPR transcription**. *Cell* 2019, **178**(6):1452-1464. e1413.

57. Roux S, Krupovic M, Daly RA, Borges AL, Nayfach S, Schulz F, Sharrar A, Matheus Carnevali PB, Cheng J-F, Ivanova NN *et al*: **Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes**. *Nature Microbiology* 2019, **4**(11):1895-1906.

58. Wang H, Ling Y, Shan T, Yang S, Xu H, Deng X, Delwart E, Zhang W: **Gut virome of mammals and birds reveals high genetic diversity of the family Microviridae**. *Virus Evolution* 2019, **5**(1).

59. Koonin EV, Yutin N: **The crAss-like Phage Group: How Metagenomics Reshaped the Human Virome**. *Trends in Microbiology* 2020, **28**(5):349-359.

60. Iranzo J, Krupovic M, Koonin EV: **The Double-Stranded DNA Virosphere as a Modular Hierarchical Network of Gene Sharing**. *mBio* 2016, **7**(4):e00978-00916.

61. Aiewsakun P, Adriaenssens EM, Lavigne R, Kropinski AM, Simmonds P: **Evaluation of the genomic diversity of viruses infecting bacteria, archaea and eukaryotes using a common bioinformatic platform: steps towards a unified taxonomy**. *Journal of General Virology* 2018, **99**(9):1331-1343.

62. Barylski J, Enault F, Dutilh BE, Schuller MB, Edwards RA, Gillis A, Klumpp J, Knezevic P, Krupovic M, Kuhn JH *et al*: **Analysis of Spounaviruses as a Case Study for the Overdue Reclassification of Tailed Phages**. *Systematic Biology* 2019, **69**(1):110-123.

63. Low SJ, Džunková M, Chaumeil P-A, Parks DH, Hugenholtz P: **Evaluation of a concatenated protein phylogeny for classification of tailed double-stranded DNA viruses belonging to the order Caudovirales**. *Nature Microbiology* 2019, **4**(8):1306-1315.

64. Yutin N, Makarova KS, Gussow AB, Krupovic M, Segall A, Edwards RA, Koonin EV: **Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut**. *Nature Microbiology* 2018, **3**(1):38-46.

65. Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GGZ, Boling L, Barr JJ, Speth DR, Seguritan V, Aziz RK *et al*: **A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes**. *Nature Communications* 2014, **5**(1):4498.

66. Edwards RA, Vega AA, Norman HM, Ohaeri M, Levi K, Dinsdale EA, Cinek O, Aziz RK, McNair K, Barr JJ *et al*: **Global phylogeography and ancient evolution of the widespread human gut virus crAssphage**. *Nature Microbiology* 2019, **4**(10):1727-1736.

67. Schumacher MA, Tonthat NK, Kwong SM, Chinnam Nb, Liu MA, Skurray RA, Firth N: **Mechanism of staphylococcal multiresistance plasmid replication origin assembly by the RepA protein**. *Proceedings of the National Academy of Sciences* 2014, **111**(25):9121-9126.

68. Botstein D, Lew KK, Jarvik V, Swanson CA: **Role of antirepressor in the bipartite control of repression and immunity by bacteriophage P22**. *Journal of Molecular Biology* 1975, **91**(4):439-462.
69. Argov T, Sapir SR, Pasechnek A, Azulay G, Stadnyuk O, Rabinovich L, Sigal N, Borovok I, Herskovits AA: **Coordination of cohabiting phage elements supports bacteria–phage cooperation**. *Nature Communications* 2019, **10**(1).
70. Davis-Richardson AG, Ardissone AN, Dias R, Simell V, Leonard MT, Kemppainen KM, Drew JC, Schatz D, Atkinson MA, Kolaczkowski B *et al*: **Bacteroides dorei dominates gut microbiome prior to autoimmunity in Finnish children at high risk for type 1 diabetes**. *Frontiers in Microbiology* 2014, **5**(678).
71. Campbell A: **Comparative Molecular Biology of Lambdoid Phages**. 1994, **48**(1):193-222.
72. Wu L, Gingery M, Abebe M, Arambula D, Czornyj E, Handa S, Khan H, Liu M, Pohlschroder M, Shaw KL *et al*: **Diversity-generating retroelements: natural variation, classification and evolution inferred from a large-scale genomic survey**. *Nucleic Acids Research* 2017, **46**(1):11-24.
73. Handa S, Jiang Y, Tao S, Foreman R, Schinazi RF, Miller JF, Ghosh P: **Template-assisted synthesis of adenine-mutagenized cDNA by a retroelement protein complex**. *Nucleic Acids Research* 2018, **46**(18):9711-9725.
74. Miller JL, Coq JL, Hodes A, Barbalat R, Miller JF, Ghosh P: **Selective Ligand Recognition by a Diversity-Generating Retroelement Variable Protein**. *PLOS Biology* 2008, **6**(6):e131.
75. Liu M, Deora R, Doulatov SR, Gingery M, Eiserling FA, Preston A, Maskell DJ, Simons RW, Cotter PA, Parkhill J *et al*: **Reverse Transcriptase-Mediated Tropism Switching in Bordetella Bacteriophage**. *Science* 2002, **295**(5562):2091-2094.
76. Xu J, Hendrix RW, Duda RL: **Conserved Translational Frameshift in dsDNA Bacteriophage Tail Assembly Genes**. *Molecular Cell* 2004, **16**(1):11-21.
77. Levin ME, Hendrix RW, Casjens SR: **A Programmed Translational Frameshift is Required for the Synthesis of a Bacteriophage λ Tail Assembly Protein**. *Journal of Molecular Biology* 1993, **234**(1):124-139.
78. Baranov PV, Fayet O, Hendrix RW, Atkins JF: **Recoding in bacteriophages and bacterial IS elements**. *Trends in Genetics* 2006, **22**(3):174-181.
79. McNair K, Zhou C, Dinsdale EA, Souza B, Edwards RA: **PHANOTATE: a novel approach to gene identification in phage genomes**. *Bioinformatics* 2019, **35**(22):4537-4542.
80. Doulatov S, Hodes A, Dai L, Mandhana N, Liu M, Deora R, Simons RW, Zimmerly S, Miller JF: **Tropism switching in Bordetella bacteriophage defines a family of diversity-generating retroelements**. *Nature* 2004, **431**(7007):476-481.
81. Meineke B, Shuman S: **Determinants of the cytotoxicity of PrrC anticodon nuclease and its amelioration by tRNA repair**. *RNA* 2012, **18**(1):145-154.
82. de Jonge PA, von Meijenfeldt FAB, van Rooijen LE, Brouns SJJ, Dutilh BE: **Evolution of BACON Domain Tandem Repeats in crAssphage and Novel Gut Bacteriophage Lineages**. *Viruses* 2019, **11**(12):1085.

83. Pell LG, Gasmi-Seabrook GMC, Morais M, Neudecker P, Kanelis V, Bona D, Donaldson LW, Edwards AM, Howell PL, Davidson AR *et al*: **The Solution Structure of the C-Terminal Ig-like Domain of the Bacteriophage λ Tail Tube Protein**. *Journal of Molecular Biology* 2010, **403**(3):468-479.
84. Fraser JS, Yu Z, Maxwell KL, Davidson AR: **Ig-Like Domains on Bacteriophages: A Tale of Promiscuity and Deceit**. *Journal of Molecular Biology* 2006, **359**(2):496-507.
85. Barr JJ, Auro R, Sam-Soon N, Kassegne S, Peters G, Bonilla N, Hatay M, Mourtada S, Bailey B, Youle M *et al*: **Subdiffusive motion of bacteriophage in mucosal surfaces increases the frequency of bacterial encounters**. *Proceedings of the National Academy of Sciences* 2015, **112**(44):13675-13680.
86. Nasko DJ, Chopyk J, Sakowski EG, Ferrell BD, Polson SW, Wommack KE: **Family A DNA Polymerase Phylogeny Uncovers Diversity and Replication Gene Organization in the Virioplankton**. *Frontiers in Microbiology* 2018, **9**(3053).
87. Campbell TL, Brown ED: **Characterization of the Depletion of 2-C-Methyl-d-Erythritol-2,4-Cyclodiphosphate Synthase in Escherichia coli and Bacillus subtilis**. *Journal of Bacteriology* 2002, **184**(20):5609-5618.
88. Heuston S, Begley M, Gahan CGM, Hill C: **Isoprenoid biosynthesis in bacterial pathogens**. *Microbiology* 2012, **158**(6):1389-1401.
89. Zhang J-R, Idanpaan-Heikkila I, Fischer W, Tuomanen EI: **Pneumococcal licD2 gene is involved in phosphorylcholine metabolism**. *Molecular Microbiology* 1999, **31**(5):1477-1488.
90. Lopez R, Garcia E, Garcia P, Ronda C, Tomasz A: **Choline-containing bacteriophage receptors in Streptococcus pneumoniae**. *Journal of Bacteriology* 1982, **151**(3):1581-1590.
91. Odom AR: **Five Questions about Non-Mevalonate Isoprenoid Biosynthesis**. *PLOS Pathogens* 2011, **7**(12):e1002323.
92. Breitbart M, Bonnain C, Malki K, Sawaya NA: **Phage puppet masters of the marine microbial realm**. *Nature Microbiology* 2018, **3**(7):754-766.
93. Łobocka MB, Rose DJ, Plunkett G, Rusin M, Samojedny A, Lehnherr H, Yarmolinsky MB, Blattner FR: **Genome of Bacteriophage P1**. *Journal of Bacteriology* 2004, **186**(21):7032-7068.
94. McNair K, Bailey BA, Edwards RA: **PHACTS, a computational approach to classifying the lifestyle of phages**. *Bioinformatics* 2012, **28**(5):614-618.
95. Holmfeldt K, Solonenko N, Shah M, Corrier K, Riemann L, VerBerkmoes NC, Sullivan MB: **Twelve previously unknown phage genera are ubiquitous in global oceans**. *Proceedings of the National Academy of Sciences* 2013, **110**(31):12798-12803.
96. Kyrkou I, Byth Carstens A, Ellegaard-Jensen L, Kot W, Zervas A, Djurhuus AM, Neve H, Hansen M, Hestbjerg Hansen L: **Expanding the Diversity of Myoviridae Phages Infecting Lactobacillus plantarum—A Novel Lineage of Lactobacillus Phages Comprising Five New Members**. *Viruses* 2019, **11**(7):611.
97. Rocha ER, Selby T, Coleman JP, Smith CJ: **Oxidative stress response in an anaerobe, Bacteroides fragilis: a role for catalase in protection against hydrogen peroxide**. *Journal of Bacteriology* 1996, **178**(23):6895-6903.

98. Betteken MI, Rocha ER, Smith CJ: **Dps and DpsL Mediate Survival In Vitro and In Vivo during the Prolonged Oxidative Stress Response in Bacteroides fragilis**. *Journal of Bacteriology* 2015, **197**(20):3329-3338.
99. Gauss GH, Reott MA, Rocha ER, Young MJ, Douglas T, Smith CJ, Lawrence CM: **Characterization of the Bacteroides fragilis Gene Product Identifies a Bacterial DPS-Like Protein and Suggests Evolutionary Links in the Ferritin Superfamily**. *Journal of Bacteriology* 2012, **194**(1):15-27.
100. Germain E, Castro-Roa D, Zenkin N, Gerdes K: **Molecular Mechanism of Bacterial Persistence by HipA**. *Molecular Cell* 2013, **52**(2):248-254.
101. Huang CY, Gonzalez-Lopez C, Henry C, Mijakovic I, Ryan KR: **hipBA toxin-antitoxin systems mediate persistence in Caulobacter crescentus**. *Scientific Reports* 2020, **10**(1):2865.
102. Rahmsdorf HJ, Pai SH, Ponta P, Herrlich P, Roskoski R, Schweiger M, Studier FW: **Protein Kinase Induction in Escherichia coli by Bacteriophage T7**. *Proceedings of the National Academy of Sciences* 1974, **71**(2):586-589.
103. Gone S, Alfonso-Prieto M, Paudyal S, Nicholson AW: **Mechanism of Ribonuclease III Catalytic Regulation by Serine Phosphorylation**. *Scientific Reports* 2016, **6**(1):25448.
104. Severinova E, Severinov K: **Localization of the Escherichia coli RNA Polymerase β' Subunit Residue Phosphorylated by Bacteriophage T7 Kinase Gp0.7**. *Journal of Bacteriology* 2006, **188**(10):3470-3476.
105. Friedman DI, Mozola CC, Beeri K, Ko C-C, Reynolds JL: **Activation of a prophage-encoded tyrosine kinase by a heterologous infecting phage results in a self-inflicted abortive infection**. *Molecular Microbiology* 2011, **82**(3):567-577.
106. Miller ES, Kutter E, Mosig G, Arisaka F, Kunisawa T, Rüger W: **Bacteriophage T4 Genome**. *Microbiology and Molecular Biology Reviews* 2003, **67**(1):86-156.
107. Casjens S: **Prophages and bacterial genomics: what have we learned so far?** *Molecular Microbiology* 2003, **49**(2):277-300.
108. Whitfield C, Wear SS, Sande C: **Assembly of Bacterial Capsular Polysaccharides and Exopolysaccharides**. *Annual Review of Microbiology* 2020, **74**(1):521-543.
109. Porter NT, Hryckowian AJ, Merrill BD, Fuentes JJ, Gardner JO, Glowacki RWP, Singh S, Crawford RD, Snitkin ES, Sonnenburg JL *et al*: **Phase-variable capsular polysaccharides and lipoproteins modify bacteriophage susceptibility in Bacteroides thetaiotaomicron**. *Nature Microbiology* 2020, **5**(9):1170-1181.
110. Hatfull GF: **Actinobacteriophages: Genomics, Dynamics, and Applications**. *Annual Review of Virology* 2020, **7**(1):37-61.
111. Shapiro JW, Putonti C: **Gene Co-occurrence Networks Reflect Bacteriophage Ecology and Evolution**. *mBio* 2018, **9**(2):e01870-01817.
112. Minot S, Grunberg S, Wu GD, Lewis JD, Bushman FD: **Hypervariable loci in the human gut virome**. *Proceedings of the National Academy of Sciences* 2012, **109**(10):3962-3966.

113. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, Creasy HH, Earl AM, FitzGerald MG, Fulton RS *et al*: **Structure, function and diversity of the healthy human microbiome**. *Nature* 2012, **486**(7402):207-214.

114. Hryckowian AJ, Merrill BD, Porter NT, Van Treuren W, Nelson EJ, Garlena RA, Russell DA, Martens EC, Sonnenburg JL: **Bacteroides thetaiotaomicron-Infecting Bacteriophage Isolates Inform Sequence-Based Host Range Predictions**. *Cell Host & Microbe* 2020, **28**(3):371-379.e375.

115. Roux S, Paul BG, Bagby SC, Allen MA, Attwood G, Cavicchioli R, Chistoserdova L, Hallam SJ, Hernandez ME, Hess M *et al*: **Ecology and molecular targets of hypermutation in the global microbiome**. *bioRxiv* 2020:2020.2004.2001.020958.

116. Cornuault JK, Petit M-A, Mariadassou M, Benevides L, Moncaut E, Langella P, Sokol H, De Paepe M: **Phages infecting Faecalibacterium prausnitzii belong to novel viral genera that help to decipher intestinal viromes**. *Microbiome* 2018, **6**(1):65.

117. Benler S, Cobián-Güemes AG, McNair K, Hung S-H, Levi K, Edwards R, Rohwer F: **A diversity-generating retroelement encoded by a globally ubiquitous Bacteroides phage**. *Microbiome* 2018, **6**(1):191.

118. Kaur G, Burroughs AM, Iyer LM, Aravind L: **Highly regulated, diversifying NTP-dependent biological conflict systems with implications for the emergence of multicellularity**. *Elife* 2020, **9**.

119. Vallota-Eastman A, Arrington EC, Meeken S, Roux S, Dasari K, Rosen S, Miller JF, Valentine DL, Paul BG: **Role of Diversity-Generating Retroelements for Regulatory Pathway Tuning in Cyanobacteria**. *bioRxiv* 2020:2020.2005.2026.117283.

120. Valguarnera E, Wardenburg JB: **Good Gone Bad: One Toxin Away From Disease for Bacteroides fragilis**. *Journal of Molecular Biology* 2020, **432**(4):765-785.

121. Friedman ES, Bittinger K, Esipova TV, Hou L, Chau L, Jiang J, Mesaros C, Lund PJ, Liang X, FitzGerald GA *et al*: **Microbes vs. chemistry in the origin of the anaerobic gut lumen**. *Proceedings of the National Academy of Sciences* 2018, **115**(16):4170-4175.

122. Albenberg L, Esipova TV, Judge CP, Bittinger K, Chen J, Laughlin A, Grunberg S, Baldassano RN, Lewis JD, Li H *et al*: **Correlation Between Intraluminal Oxygen Gradient and Radial Partitioning of Intestinal Microbiota**. *Gastroenterology* 2014, **147**(5):1055-1063.e1058.

123. Donaldson GP, Lee SM, Mazmanian SK: **Gut biogeography of the bacterial microbiota**. *Nature Reviews Microbiology* 2016, **14**(1):20-32.

Figures

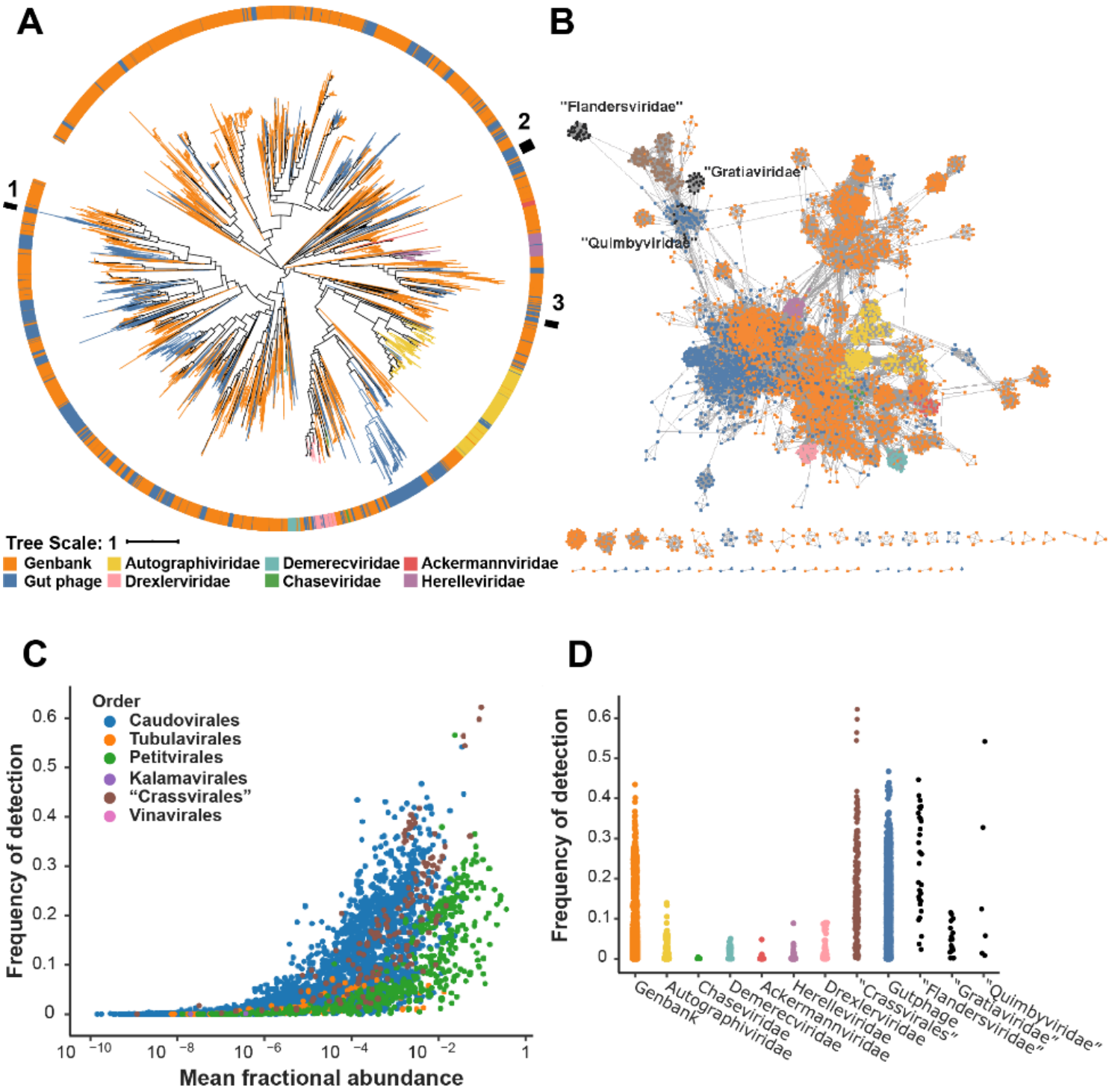

Figure 1

Three candidate families of Caudovirales phages discovered in human gut metagenomes. (A) Phylogenetic tree of the large terminase subunit encoded by Caudovirales phage genomes in GenBank (n = 3,931) and in gut metagenomes (n = 1,298). Branches are colored according to the current ICTV families, except for the Myoviridae, Podoviridae or Siphoviridae, which are in orange. The outermost ring indicates the location of candidate families proposed in this study: 1, “Quimbyviridae” phages; 2, “Flandersviridae” phages; 3, “Gratiaviridae” phages (see main text). (B) Gene sharing network of the Urovicota phages. Phage genomes identified in human gut metagenomes (blue nodes) were compared to

phages in the GenBank database (colored as in Figure 1, with the addition of the crAss-like phages in brown and the new Caudovirales families proposed in this study in black). (C) Abundance of phages across human gut viromes. The x-axis depicts the fractional abundance of a given phage averaged across all viromes (n = 1,258); the y-axis is the fraction of viromes that a given phage recruits at least one read. Each phage genome (n = 7,888 total) is colored at the taxonomic level of Order (C) or Family (Uroviricota families only) (D). The raw data used to generate panel (A) are provided in Additional file 14 and data for panels (C) and (D) are provided in Additional file 15.

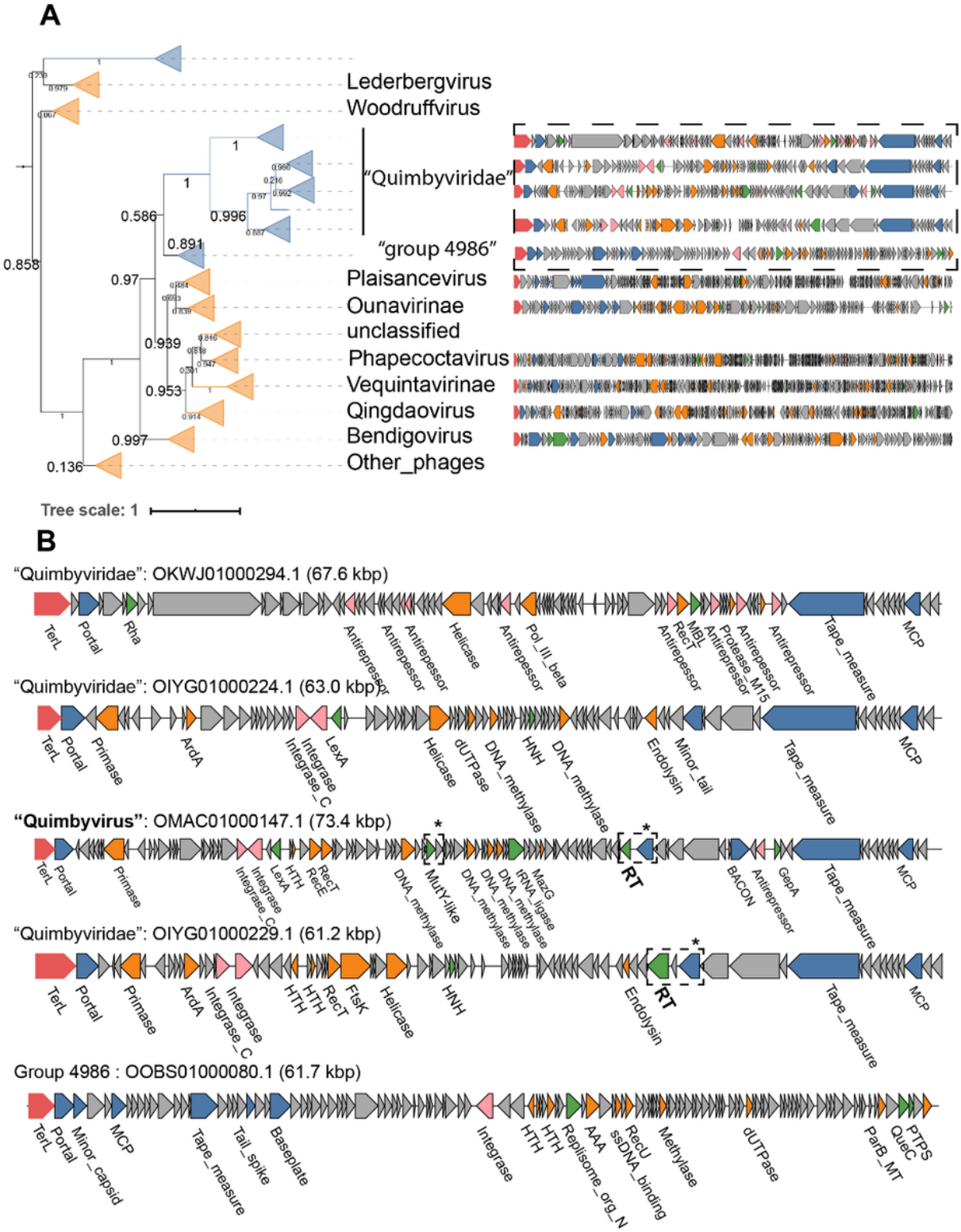

Figure 2

Phylogenetic tree of the large terminase subunit and genome maps of Quimby-like phages. (A) Individual genome maps of Quimby-like phages and ICTV classified phages are shown to the right of each branch. The ORFs are colored according to function: large terminase subunit (red), structural components (blue), DNA replication and repair (orange), lysogeny (pink), general function (green) and unknown (grey). (B) Expansion of four Quimby-like phages and a single gut phage genome from an adjacent branch (“group 4986”). The diversity-generating retroelement and hypervariable ORFs are highlighted with a dashed box and asterisk. The nucleotide scales differ between individual genome maps in both panels.

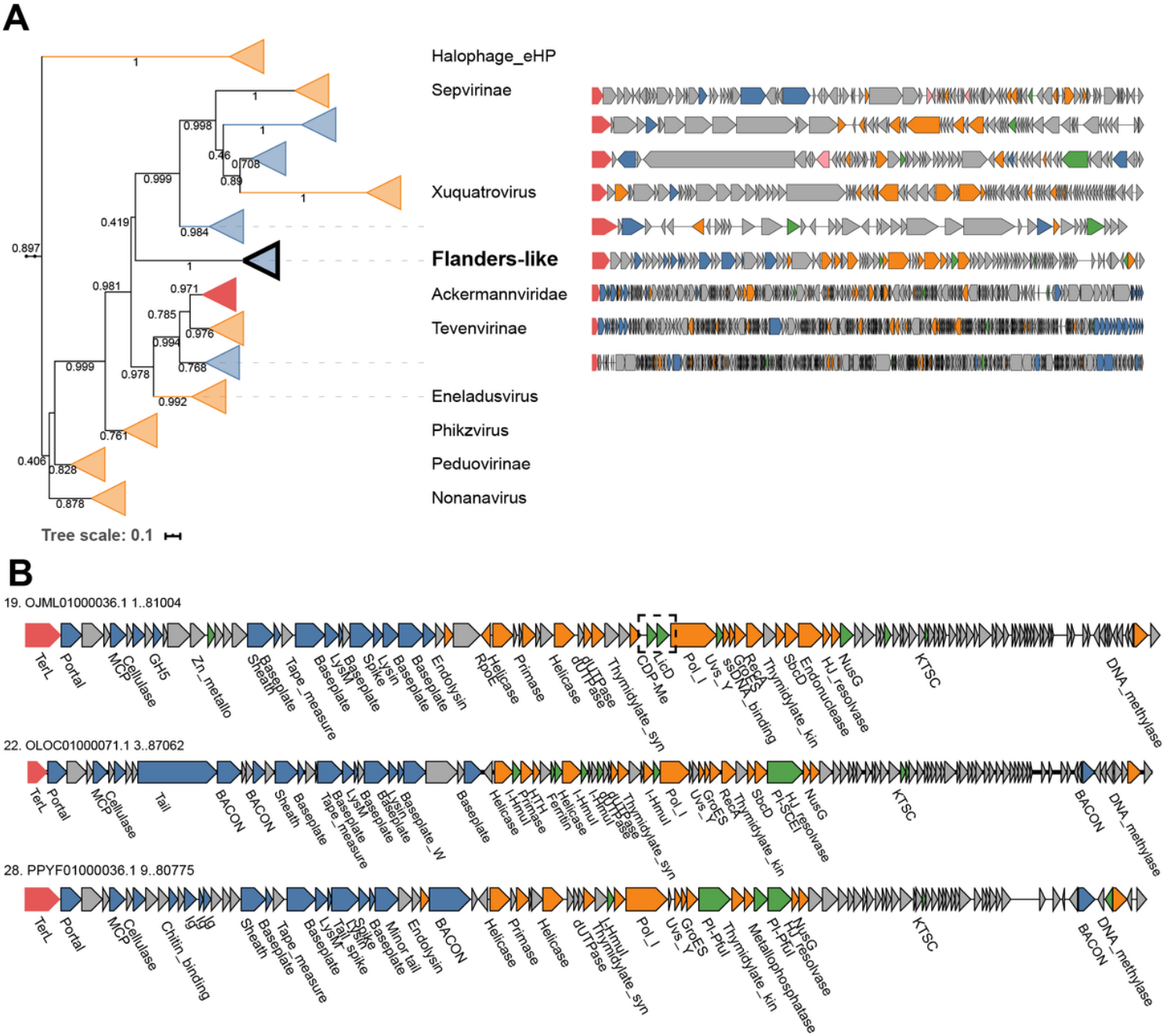

Figure 3

Phylogenetic tree of the large terminase subunit and complete genome maps for “Flandersviridae”. (A) Genome maps of members of the “Flandersviridae and selected ICTV-classified phages were constructed and colored as in Figure 3. (B) Genome maps of three genera from the “Flandersviridae” family. The dashed box highlights the insertion of licD- and ispD-family enzymes in the replication module of one “Flandersviridae” phage.

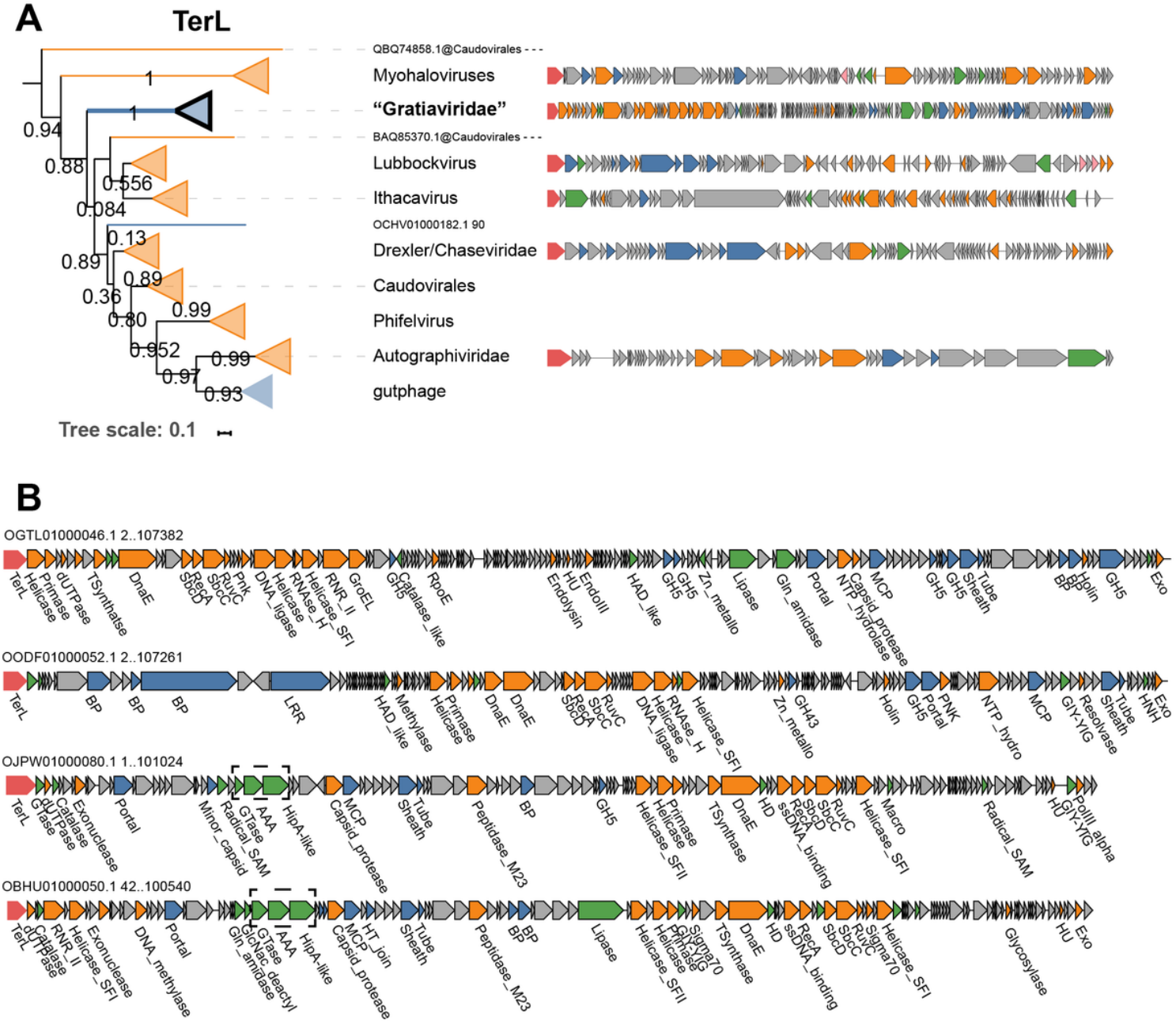

Figure 4

Phylogenetic tree of the large terminase subunit and genome maps of the “Gratiaviridae” phages. (A) Genome maps of ICTV-classified phages were constructed and colored as in Figure 3. (B) Genome maps of four genera from the Gratiaviridae family. The dashed box highlights a HipA-family kinase domain-containing protein, AAA-family ATPase and glycosyltransferase (see main text).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.csv
- Additionalfile2.csv
- Additionalfile3.pdf
- Additionalfile4.csv
- Additionalfile5.csv
- Additionalfile6.pdf
- Additionalfile7.pdf
- Additionalfile8.pdf
- Additionalfile9.pdf
- Additionalfile10.docx
- Additionalfile11.pdf
- Additionalfile12.pdf
- Additionalfile13.pdf