Diversity and Ecology of Caudoviricetes Phages with Genome Terminal Repeats in Fecal Metagenomes from Four Dutch Cohorts

Gulyaeva, Anastasia; Garmaeva, Sanzhima; Kurilshikov, Alexander; Vich Vila, Arnau; Riksen, Niels P.; Netea, Mihai G.; Weersma, Rinse K.; Fu, Jingyuan; Zhernakova, Alexandra

Published in:
Viruses

DOI:
10.3390/v14102305

Publication date:
2022

viruses

MDPI

Article

Diversity and Ecology of *Caudoviricetes* Phages with Genome Terminal Repeats in Fecal Metagenomes from Four Dutch Cohorts

Anastasia Gulyaeva [1,*], Sanzhima Garmaeva [1], Alexander Kurilshikov [1], Arnau Vich Vila [1,2], Niels P. Riksen [3], Mihai G. Netea [3], Rinse K. Weersma [1,2], Jingyuan Fu [1,4] and Alexandra Zhernakova [1,*]

1 Department of Genetics, University of Groningen, University Medical Center Groningen, 9713GZ Groningen, The Netherlands
2 Department of Gastroenterology and Hepatology, University Medical Center Groningen, 9713GZ Groningen, The Netherlands
3 Department of Internal Medicine, Radboud University Medical Center, 6525GA Nijmegen, The Netherlands
4 Department of Pediatrics, University of Groningen, University Medical Center Groningen, 9713GZ Groningen, The Netherlands
* Correspondence: a.gulyaeva@umcg.nl (A.G.); a.zhernakova@umcg.nl (A.Z.)

Abstract: The human gut harbors numerous viruses infecting the human host, microbes, and other inhabitants of the gastrointestinal tract. Most of these viruses remain undiscovered, and their influence on human health is unknown. Here, we characterize viral genomes in gut metagenomic data from 1950 individuals from four population and patient cohorts. We focus on a subset of viruses that is highly abundant in the gut, remains largely uncharacterized, and allows confident complete genome identification—phages that belong to the class *Caudoviricetes* and possess genome terminal repeats. We detect 1899 species-level units belonging to this subset, 19% of which do not have complete representative genomes in major public gut virome databases. These units display diverse genomic features, are predicted to infect a wide range of microbial hosts, and on average account for <1% of metagenomic reads. Analysis of longitudinal data from 338 individuals shows that the composition of this fraction of the virome remained relatively stable over a period of 4 years. We also demonstrate that 54 species-level units are highly prevalent (detected in >5% of individuals in a cohort). Finally, we find 34 associations between highly prevalent phages and human phenotypes, 24 of which can be explained by the relative abundance of potential hosts.

Keywords: human gut metagenome; *Caudoviricetes*; human phenotypes

Citation: Gulyaeva, A.; Garmaeva, S.; Kurilshikov, A.; Vich Vila, A.; Riksen, N.P.; Netea, M.G.; Weersma, R.K.; Fu, J.; Zhernakova, A. Diversity and Ecology of *Caudoviricetes* Phages with Genome Terminal Repeats in Fecal Metagenomes from Four Dutch Cohorts. *Viruses* **2022**, *14*, 2305. https://doi.org/10.3390/v14102305

Academic Editor: Jennifer Mahony

Received: 22 September 2022
Accepted: 17 October 2022
Published: 20 October 2022

1. Introduction

The human gut harbors a large and diverse collection of viruses. These include viruses that infect human cells, viruses that infect archaea and bacteria inhabiting the gut (phages), viruses that infect protists and parasites, and viruses that pass through the intestinal tract with food. The viruses of the human gut belong to diverse lineages and possess different types of genomes: single- or double-stranded (ss or ds) DNA or RNA [1,2]. The diversity of the human gut virome is actually so large that, despite the unprecedented attention to the human gut ecosystem in recent years, saturation in the number of known species of human gut viruses has not been reached [3], and a multitude of unanswered questions about their biology and links to human health remain.

One of the most abundant and diverse groups of viruses in the human gut are the tailed phages unified in the class *Caudoviricetes* [4,5]. These phages possess dsDNA genomes that encode a distinctive major capsid protein (MCP) with the HK97 fold and a distinctive packaging enzyme, terminase, consisting of a small and a large subunit (TerL) [6,7]. Phages belonging to class *Caudoviricetes* employ a wide range of replication mechanisms, which is reflected in their genome termini type. The virion-packaged genome molecules of all

Caudoviricetes phages are believed to be linear. Those that replicate via a circular intermediate possess cohesive ends or direct terminal repeats (DTR) when packaged; if sequenced during replication, these genomes are also likely to produce contigs with DTR [2,8,9]. Genomes that replicate by transposition are flanked by random host genome fragments when packaged [8]. Genomes that employ a protein-primed replication mechanism remain linear during replication and possess inverted terminal repeats (ITR) [10]. Importantly, the presence of DTR or ITR at the phage contig termini can be used in bioinformatics analysis as an indicator of phage genome sequencing completion [11]. Depending on their lifestyle, many *Caudoviricetes* phages can be referred to as virulent or temperate. Virulent phages enter a lytic state upon genome injection into a host cell: replicate, then lyse the cell to release viral progeny. Temperate phages can enter a lysogenic state (becoming dormant, for example, by integrating into the host genome as a prophage), and subsequently, switch to a lytic state [12]. Accumulated mutations can render a prophage incapable of switching to a lytic state, turning it into a cryptic prophage [13]. The taxonomic structure of class *Caudoviricetes* is currently undergoing a major overhaul to produce a primarily genome-based classification that accurately reflects the evolutionary relationships between member phages, and order *Caudovirales* and families *Myoviridae*, *Siphoviridae, Podoviridae* were recently abolished as a part of this effort [5,14].

During the last decade, metagenomics—the analysis of nucleic acid sequences extracted from an entire ecological community—has become the primary method for studying the diversity of the human gut virome. The two popular approaches are the sequencing of nucleic acid isolated either from the entire human gut community (total metagenome) or from virus-like particles (virus-enriched metagenome). Both approaches can produce a biased representation of the virome composition. For example, virus-enriched metagenomes do not include the genomes of phages in the lysogenic state, while the preparation of total metagenome libraries usually does not include the steps required to sequence the genomes of RNA and ssDNA viruses [15]. As metagenomics rapidly increases the amount of sequencing data available, bioinformatics analysis often becomes the bottleneck of virome discovery [11]. One promising approach to meeting this challenge is to use protein markers for virus identification (proteins encoded by viruses but not by cellular organisms, e.g., viral structural proteins) and taxonomic assignment (proteins uniquely encoded by viruses belonging to a specific lineage). Despite the challenges, recent metagenomics studies were able to uncover virome signatures associated with a number of diseases including inflammatory bowel disease (IBD), colorectal cancer, and type 1 diabetes [4]. Nonetheless, the role of gut phages in relation to human diseases is still underexplored.

To improve our understanding of the human gut virome, we analyzed viruses in total fecal metagenomes (i.e., metagenomes that were not enriched for viruses) from four cohorts collected in The Netherlands: the population cohorts Lifelines-DEEP (LLD) and LLD follow-up [16–18], a cohort of overweight and obese individuals with BMI > 27 kg/m^2 (300OB) [19,20], and a cohort of patients with IBD [21,22]. We relied on protein markers for virus identification and taxonomic assignment. The analysis was focused primarily on viruses with genome terminal repeats belonging to the class *Caudoviricetes*, and examined their diversity, abundance, stability, predicted hosts, and links to human phenotypes.

2. Materials and Methods

2.1. Virus Detection in Metagenomes

Total metagenome sequencing data from 2291 samples from four Dutch cohorts were assembled into contigs using metaSPAdes 3.14.1 [23], as described in [24]. To identify viral genomes, contigs from each sample were screened using Cenote-Taker 2 version 2.1.3, program *unlimited_breadsticks.py* with the following parameters: "–virus_domain_db 'virion' –minimum_length_circular 3000 –minimum_length_linear 10000 –circ_minimum_hallmark_genes 1 –lin_minimum_hallmark_genes 2 –prune_prophage True –filter_out_plasmids True" [25]. Identified virus-like contigs were then screened for the presence of ribosomal RNA (rRNA) genes using a BLASTN 2.10.1+ [26] search in the SILVA 138.1 NR99 rRNA genes database [27]

with an E-value threshold of 0.001. An rRNA gene was considered to be detected in a contig if the gene and the contig produced a hit covering > 50% of the gene length. The eight contigs with detected rRNA genes were excluded from further consideration. Importantly, 520 IBD cohort samples were screened for the presence of viral genomes, but we excluded 62 samples from individuals with stoma and ileoanal pouches, 1 duplicated sample, and 2 samples without metadata from all subsequent analyses, bringing the number of IBD samples under consideration down to 455.

2.2. Nucleotide Sequence Characterization

A nucleotide sequence was considered to contain a DTR or ITR if identical terminal repeats ≥ 20 nt were detected [28]. Nucleotide content, GC- and AT-skew were calculated using a 1001 nt window sliding along the genome sequence with a 200 nt step, as described in [24]. Prediction of tRNA genes was conducted for individual genome sequences using tRNAscan-SE 2.0.9 with the "-B" parameter [29]. To search for nucleotide repeats, genome sequences were compared to themselves using BLASTN 2.12.0+ with the "-task 'blastn' -evalue 0.001" parameters [26], and only hits with alignment length ≥ 100 and identity ≥ 80% were considered.

2.3. Identification of Potential Plasmids

Nucleotide sequences with DTR characterized by a PlasX score > 0.9 were considered potential plasmids [30].

2.4. Genetic Code Prediction

Open reading frames (ORFs) were predicted using Prodigal 2.6.3 [31]. For sequences shorter than 20 kb, the prediction was made in the "meta" mode using standard bacterial genetic code 11. For each individual sequence ≥ 20 kb, the prediction was made in the "single" mode using standard bacterial genetic code 11 and alternative genetic codes 4 (TGA codon encodes tryptophan) and 15 (TAG codon encodes glutamine). If the sum of the ORF coding potential scores was higher under an alternative genetic code and exceeded the sum under the standard genetic code by 10%, the alternative genetic code was assigned to the viral genome sequence [3,32]. We tested this method on 378 crAss-like phage genome sequences for which the genetic code had been previously predicted using manual ORF analysis [24] and obtained an identical prediction in 97% of cases.

2.5. Proteome Annotation

Individual proteomes, predicted as described above, were compared to Pfam 35.0 profiles [33] using the HMMER 3.3.2 program *hmmsearch* with the "–max -E 0.001" parameters (http://hmmer.org/). Only protein–profile pairs where ≥100 amino acid residues of the protein were covered by hit(s) to the profile were considered, the profile providing maximal coverage was used for annotation. Coverage was measured in HMMER envelope coordinates combined using the R package *IRanges* 2.22.2 in case of overlap [34]. Multiple sequence alignment (MSA) of proteins annotated as reverse transcriptases was generated by adding them to the Pfam 35.0 seed PF00078 alignment with the help of the R package *seqinr* 3.6-1 and MAFFT 7.453 with an "–add" parameter [35,36].

2.6. Taxonomic Assignment Based on Marker Genes

Taxonomic assignment was performed using the proteomes predicted as described above. Proteomes were compared to profiles of marker proteins of seven dsDNA virus groups using the HMMER 3.3.2 program *hmmsearch* with the "–max -E 0.001" parameters. The following MSAs were used to generate marker profiles (Table 1): class *Caudoviricetes*—alignment of 5130 TerL sequences from [37], alignment of 823 TerL sequences from [38], TerL MSA VOG00461 from the VOG 207 database [39]; family *Herpesviridae*—MCP MSA PF03122 from the Pfam 35.0 database [33]; family *Papillomaviridae*—capsid protein L1 MSA VOG05075 [39]; family *Polyomaviridae*—coat protein MSA PF00718 [33]; family *Adenoviri-*

dae—hexon protein MSA VOG05391 [39]; class *Tectiliviricetes*, excluding the *Adenoviridae* family—6 MCP MSAs corresponding to 6 prokaryotic virus groups from [40]; and phylum *Nucleocytoviricota*—7 marker MSAs (MCP, DNA-directed RNA polymerase alpha and beta subunits, DNA polymerase family B, transcription initiation factor IIB, DNA topoisomerase II, poxvirus late transcription factor VLTF3) from [41]. The profiles were visualized using Skylign [42]. For class *Caudoviricetes*, alignments between the TerL queries and a target protein were required to span the following TerL profile residues: (1) the second conserved acidic residue of the TerL Walker B motif, (2) the first conserved acidic residue of the TerL nuclease motif I, and (3) the conserved acidic residue of the TerL nuclease motif II; if the genome sequence encoding the target protein also encoded a protein producing a hit with the *Herpesviridae* MCP, taxonomic assignment was based on the latter. In the case of the phylum *Nucleocytoviricota*, genome sequences encoding proteins hit by any of the seven marker profiles were analyzed by ViralRecall with the "–contiglevel –evalue 1e-3" parameters [43], and an assignment to this phylum was made only if the ViralRecall score was >2 and ViralRecall was able to detect at least one marker protein gene.

Table 1. Benchmarking of virus detection and taxonomic assignment [a].

Taxonomic Group	Viral RefSeq Genome Sequences Recognized by Cenote-Taker 2, %	Taxonomic Assignment		
		Marker Protein Profile(s) [b]	Sensitivity, %	Specificity, % [d]
Class *Caudoviricetes*	99.61	Terminase large subunit (TerL): VOG00461 and alignments from Yutin et al., 2021 [38] and Benler et al., 2021 [37]. Target proteins were required to include hits to the TerL Walker B motif and the TerL nuclease motifs I and II. Target genomes encoding the *Herpesviridae* marker protein were discarded.	94.29	99.98
Family *Herpesviridae*	82.46	Major capsid protein (MCP): PF03122	79.82	100
Family *Papillomaviridae*	0	Capsid protein L1: VOG05075	99.52	100
Family *Polyomaviridae*	0.76	Coat protein: PF00718	99.24	100
Family *Adenoviridae*	91.89	Hexon protein: VOG05391	91.89	100
Class *Tectiliviricetes* (excluding the *Adenoviridae* family)	89.47	Double Jelly Roll MCP alignments corresponding to six prokaryotic virus groups [c] from Yutin et al., 2018 [40]	100	99.99
Phylum *Nucleocytoviricota*	74.04	Seven marker protein alignments from Aylward et al., 2021 [41]; each genome hit by a marker protein profile was further analyzed using ViralRecall (see Section 2)	70.19	99.98

[a] See Text S1. [b] Marker protein alignment identifiers beginning with "VOG" and "PF" refer to the VOG and Pfam database entries, respectively. [c] Two of the groups, Odin and FLiP, are divergent from the recognized members of the class *Tectiliviricetes*. [d] Unclassified viruses were excluded from consideration when calculating the number of true negatives and false positives.

2.7. Species-Level Clustering

Viral nucleotide sequences were clustered into virus operational taxonomic units (vOTUs) using the CheckV 0.7.0 script *aniclust.py* with the "–min_ani 95 –min_qcov 0 –

min_tcov 85" parameters [44,45]. If a vOTU contained genomes with terminal repeats, the median length genome with terminal repeats was selected as a vOTU representative. Otherwise, the longest genome without terminal repeats (encoding TerL, if available) was selected.

2.8. Read Mapping

Sequencing reads of each individual sample from the four Dutch cohorts and from the collection of 254 Danish fecal viromes [46], filtered and quality-trimmed as described in [24], were competitively mapped to a database of 30,461 virus-like genome sequences representing vOTUs using Bowtie2 2.4.4 with a "–very-sensitive" parameter [47]. The breadth of genome coverage by reads was calculated using the BEDTools 2.29.2 command *coverage* [48]. The depth of genome coverage by reads was calculated using the SAMtools 1.10 command *depth* [49]. The abundance of a vOTU in a sample was considered to be zero if the breadth of the representative genome coverage by reads was below 75% [50]; otherwise, it was estimated as $(N \cdot 10^6)/(L \cdot S)$, where N is the number of reads mapped to a genome, L is the length of a genome and S is the number of sample reads after filtering and quality trimming.

2.9. Building TerL MSA

We used the HH-suite 3.3.0 command *hhalign* with the "-M 50 -mact 0 -all" parameters and script *reformat.pl* [51] to first combine TerL alignment VOG00461 from VOG 207 [39] and the alignment of 5130 TerL sequences from [37], and then added the alignment of 823 TerL sequences from [38] to the MSA. Next, the TerL protein sequences (detected by hits to TerL profiles, as described above) from the *Caudoviricetes* genomes with terminal repeats representing vOTUs in this study were added to the alignment using MAFFT 7.453 with an "–add" parameter [36]. If a genome was predicted to encode multiple copies of TerL, the one with the highest number of hit TerL motifs was used. If there were multiple candidates with an equal number of hit motifs, we used a single protein characterized by the maximal length of the alignment with TerL profiles (measured in HMMER envelope coordinates combined using the R package *IRanges* 2.22.2 in case of overlap). Finally, the MSA was inspected with the help of Jalview 2.11.2.2 [52], and only the sequences containing acidic residues in the following three alignment positions were preserved in the MSA: (1) the second conserved acidic residue of the TerL Walker B motif, (2) the first conserved acidic residue of the TerL nuclease motif I, and (3) the conserved acidic residue of the TerL nuclease motif II. MSA columns with $\geq$50% gaps were excluded from consideration using R package *Bio3D* 2.4-1 [53]. MSA conservation was estimated using the *Bio3D* 2.4-1 function *conserv* with the "method = 'similarity', sub.matrix = 'blosum62'" parameters.

2.10. Building the Phylogenetic Tree

The phylogenetic tree was reconstructed based on the TerL MSA using IQ-TREE 2.0.3, 1000 replicates of ultrafast bootstrap, WAG amino acid replacement matrix [54–56]. The tree was midpoint-rooted using the R package *phangorn* 2.5.5 [57].

2.11. The Gene-Sharing Network Reconstruction

The gene-sharing network was reconstructed by vConTACT2 0.11.3 with the "–db 'None'" parameter based on proteomes of the 1899 *Caudoviricetes* genomes with terminal repeats under consideration in this study (proteomes were predicted as described above), proteomes corresponding to 4167 complete *Caudoviricetes* and 92 complete *Herviviricetes* genomes from the Viral RefSeq 209 (proteomes were extracted from Viral RefSeq), and proteomes of the 111 recently discovered *Mirusviricota* viruses [58,59]. The resulting graph was visualized with the help of R packages *igraph* 1.2.4.2, *ggraph* 2.0.6 and *ggtext* 0.1.2.

2.12. Virome Stability Estimation

Bray–Curtis dissimilarities between samples were calculated using the function *vegdist* from the R package *vegan* 2.5-7. Data about the relative abundance of vOTUs in 338 corresponding LLD and LLD follow-up samples were utilized. vOTUs that were not detected in any sample and samples without detected vOTUs were excluded from consideration. The significance of the difference between intra- and inter-individual Bray–Curtis dissimilarities was assessed using a permutation test with 10,000 iterations [18]. On each iteration, real Bray–Curtis dissimilarity values were randomly reassigned between pairs of samples, and a Wilcoxon signed-rank test comparing intra- and inter-individual dissimilarities was performed using the R function *wilcox.test* with the "alternative = 'two.sided', paired = F" parameters [60]. The empirical *p*-value was calculated as the proportion of *p*-values that were obtained based on permuted data and lower than the *p*-value obtained based on real data.

2.13. Prophage-Based Host Prediction

The analysis was based on contigs from the four Dutch cohorts that were predicted to contain both host and prophage genome fragments by Cenote-Taker 2. Host fragment(s) of the contig $\geq$ 1000 nt were extracted for the analysis. If the number of extracted host fragments per vOTU exceeded 100, we considered a subset of 100 randomly selected host fragments. Host fragments of each vOTU were compared to the bacterial and archaeal genome sequences from the NCBI "nt" database (downloaded on 23 May 2022) using BLASTN 2.12.0+ with the "-task 'blastn' -perc_identity 95" parameters [26,61]. Only hits characterized by alignment length $\geq$ 1000 were considered; if multiple targets produced hits with a query, the one associated with the maximal query coverage by BLASTN alignments was used for taxonomic assignment. When multiple predictions made for a vOTU were incompatible at the host phylum level, they were disregarded, but this only occurred in one case.

2.14. CRISPR-Based Host Prediction

The database of CRISPR spacers published in [62] and the CRISPR-Cas++ spacers database (21 January 2021) [63] were independently compared to genomes representing vOTUs using BLASTN 2.12.0+ with the parameters "-task blastn-short -dust no -evalue 1 -max_target_seqs 1000000" [26]. A phage genome was linked to a host if there was a spacer–protospacer match characterized by $\geq$95% identity over the length of the spacer, or multiple spacer–protospacer matches characterized by $\geq$80% identity over the length of each spacer [64]. If multiple spacers of a host matched exactly the same region of a phage genome, or if multiple regions of a phage genome matched the same spacer of a host, a single spacer–protospacer match characterized by the highest bit-score was considered. Host taxonomy was retrieved from GenBank using the EFetch utility [61]. When multiple predictions made for a vOTU were incompatible at the host phylum level, they were disregarded (observed in 10 cases).

2.15. Co-Abundance-Based Host Prediction

The relative abundance of microbial taxa was estimated using MetaPhlAn 3.0.7 [65]. Correlations between the relative abundances of microbial taxa (from kingdoms to species) and vOTUs were assessed using the R function *cor.test* with the "method = 'spearman'" parameter [60] for each cohort independently. Only taxa present in >10 samples in a given cohort were considered. Meta-analysis of the results obtained for the independent LLD, 300OB and IBD cohorts was conducted using the R package *meta* 5.1-1 [66], function *metacor* with the "sm = 'ZCOR', method.tau = 'SJ'" parameters. Multiple testing correction was performed by the R function *p.adjust* using the Benjamini-Hochberg procedure [67]. The host of each vOTU was predicted based on a correlation characterized by the minimal false discovery rate (FDR) obtained for this vOTU in meta-analysis.

2.16. Finding Similar Extensively Characterized Phages

To identify extensively characterized genome sequences similar to genomes representing vOTUs, each genome representing a vOTU was compared to the viral genome sequences from the NCBI "nt" database (downloaded on 23 May 2022) using BLASTN 2.12.0+ with the "-task 'blastn' -evalue 0.001 -perc_identity 50" parameters [26,61]. Only query–target pairs characterized by ≥10% query and ≥50% target length coverage by the query–target BLASTN alignments were considered. Coverage was calculated with the help of R package *IRanges* 2.22.2 [34]. Information about the publications associated with the target sequences was obtained using the EFetch utility [61]. Sequence similarity within the query–target pairs was visualized as dot plots. The underlying matching words data were generated using the EMBOSS 6.6.0 function *polydot* with the "-wordsize 12" parameter [68].

2.17. Associations with Human Phenotypes

Association analysis was carried out based on 1135 LLD cohort samples and 207 phenotypes (missing values imputed). Prevalence of vOTUs was compared between the following groups: (1) LLD vs. IBD, (2) LLD vs. 300OB, (3) within the IBD cohort: Crohn's disease (CD) vs. ulcerative colitis (UC), (4) within the IBD cohort: exclusively colonic vs. ileal-inclusive disease location, and (5) within the 300OB cohort: absence vs. presence of the metabolic syndrome. All analyses were conducted using logistic regression adjusted for (1) the age and sex of the cohort participants and (2) the age and sex of the participants and the log-transformed abundance of the host predicted as described above. The corresponding phenotype was used as a predictor, and the detection of a phage was an outcome of the logistic regression. Logistic regression was fitted using the R 4.0.3 function *glm* with the "family = 'binomial'" parameter [60]. Multiple testing correction was conducted using the R function *p.adjust* employing the Benjamini-Hochberg procedure [67]. A significance threshold of FDR < 0.05 was used.

2.18. Visualization

The Sankey diagram was prepared using R package *alluvial* v0.1-2. Sequence logos were constructed using R package *ggseqlogo* 0.1 [69]. The phylogenetic tree was visualized using R package *ape* 5.4-1 [70]. Boxplots were plotted using R package *vioplot* 0.3.7. Colors designating host phyla were selected using R package *RColorBrewer* 1.1-2. Genome annotation labels in Material S1 were positioned with the help of R package *TeachingDemos* 2.12. Reverse transcriptase MSA was visualized using ESPript 3.0 [71].

3. Results

3.1. Viral Fraction of Total Fecal Metagenomes

To identify the virus-like fraction of the total fecal metagenomes from the LLD (n = 1135), LLD follow-up (n = 338), 300OB (n = 298), and IBD (n = 520) cohorts, we used Cenote-Taker 2, a tool relying on detection of virus marker genes for virus discovery in sequencing data [25]. We set the software to recognize contigs encoding virion proteins and to cleave off fragments of microbial genomes from these contigs (see Section 2). Of the 58,776 virus-like contigs detected (Figure S1), 45% originated from LLD, 21% from LLD follow-up, 15% from 300OB, and 19% from IBD. Microbial genome fragments were cleaved off from 15,570 contigs, suggesting that these contigs represent prophages. A total of 1613 contigs had terminal repeats: 97% had DTR and 3% had ITR. Additionally, 5706 and 21 contigs were predicted to employ alternative genetic codes with TAG and TGA stop codons recoded to amino acids, respectively.

Next, we explored the taxonomic composition of the detected virus-like contigs. Predicted proteomes of the contigs were compared to profiles of marker proteins selected to identify seven dsDNA virus groups (Table 1, Text S1). As a result, 39,752 contigs were assigned to class *Caudoviricetes*, one contig was predicted to belong to family *Adenoviridae* (99.9% nucleotide identity to *Human mastadenovirus D* HM770721.2 over the entire 17.4 kb contig) and the remaining 19,023 contigs did not receive any taxonomic assignment.

In order to represent the gut virome as comprehensively as possible, we incorporated genomes from existing viral databases [3,37,46,72–78] (Table 2) into the analysis. To avoid redundancy, both the viral genomes from the databases and the virus-like contigs identified in the four Dutch cohorts were clustered into vOTUs together. As a result, 30,461 vOTUs were delineated (Figure S1). One sequence per vOTU (with terminal repeats if available) was selected as a vOTU representative.

Table 2. Virus genomes from databases included in the analysis.

Database	Viral Genomes Included	
	Number	Description
Viral RefSeq 209	6049	Genome sequences ≥ 3000 nt, realm *Riboviria* excluded.
Human Virome Database (HuVirDB)	1660	Genome sequences ≥ 5000 nt with terminal repeats belonging [a] to the class *Caudoviricetes*.
Gut Virome Database (GVD)	936	
Gut Phage Database (GPD)	10,870	
Metagenomic Gut Virus (MGV) catalog	19,694	
Danish Enteric Virome Catalog (DEVoC)	137	
Devoto et al., 2019 [76] Al-Shayeb et al., 2020 [77] Borges et al., 2022 [b] [78]	458	
Benler et al., 2021 [37]	1480	Genome sequences belonging to the phylum *Uroviricota*.

[a] Based on TerL gene detection, see Table 1 and Section 2. [b] The version of the dataset accompanying the preprint was used (https://doi.org/10.5281/zenodo.5275335; accessed on 9 October 2021).

To estimate the relative abundance of the vOTUs in metagenomic samples, sequencing reads from individual samples were mapped to the genome sequences representing the vOTUs. A vOTU was considered detected if ≥75% of its representative sequence length was covered by reads. In total, 15,196 vOTUs were detected in at least one sample from the four Dutch cohorts (Figure S1). Based on the congruent taxonomy of their members, 69% of these vOTUs were assigned to class *Caudoviricetes*, 31% did not receive any taxonomic assignment, and the remaining 7 vOTUs included ssDNA prokaryotic viruses from the family *Microviridae* (likely sequenced while in a DNA duplex state during replication) and dsDNA human viruses from families *Papillomaviridae*, *Polyomaviridae*, and *Adenoviridae*.

Analyzing the entire set of virus-like sequences detected in total metagenome sequencing data (Figure 1) poses a number of challenges. When analyzing virus-like genomes that lack terminal repeats, it is difficult to estimate their completeness and to distinguish between prophages that can excise from the host genome and enter a lytic state and cryptic prophages that have lost this ability. When considering virus-like genomes with DTR that did not receive any taxonomic assignment, it can be difficult to distinguish these from plasmids. Notably, 20% of the taxonomically unassigned vOTUs represented by sequences with DTR were predicted to be plasmids with high confidence (PlasX score > 0.9 [30]), whereas the same was the case for only 2% of the *Caudoviricetes* vOTUs represented by sequences with DTR.

We, therefore, decided to focus on *Caudoviricetes* vOTUs represented by sequences with terminal repeats (Figure 1). Initially, there were 2106 such vOTUs, but after excluding a highly prevalent vOTU represented by a chimeric nucleotide sequence (NL_vir005341) and vOTUs with an undetected or dubious TerL gene in the representative genome (see Section 2, Figure S2), 1899 vOTUs remained (Table S1). Below, we refer to genome sequences belonging to these vOTUs as the *Caudoviricetes* Genomes with Terminal Repeats (CGTR1899) database.

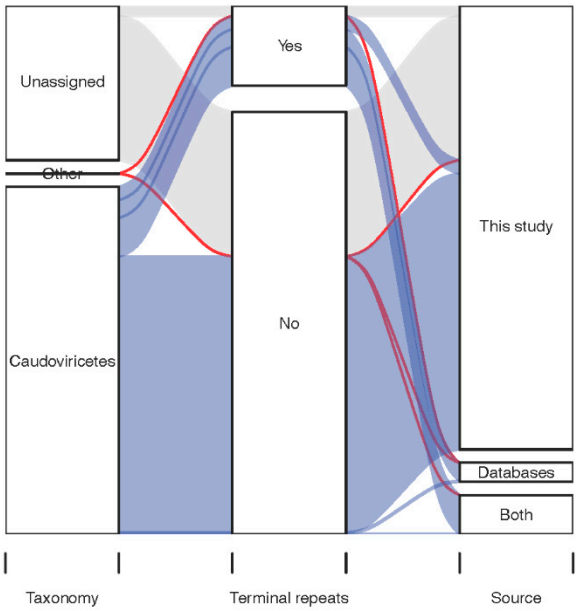

Figure 1. Properties of the vOTUs detected in the four Dutch cohorts. Sankey diagram shows the relationships between taxonomy, detection of terminal repeats in representative genome, and source of the 15,196 vOTUs.

Importantly, the overall proteome composition of the CGTR1899 genomes was similar to that of recognized *Caudoviricetes* phages, but not to that of other viruses that possess evolutionary-related terminase genes: in the gene-sharing network reconstructed by vConTACT2, 52% of CGTR1899 phages were directly connected to *Caudoviricetes* phages, but none were directly connected to viruses belonging to the class *Herviviricetes* or the recently discovered phylum *Mirusviricota* (Figure S3) [58,59].

3.2. Diversity of Caudoviricetes Phages with Genome Terminal Repeats

We next aimed to characterize the diversity, abundance, and long-term stability of the *Caudoviricetes* phages with genome terminal repeats represented by the CGTR1899 database in the four Dutch cohorts. To compare the abundance of phages in the total metagenomes from the four Dutch cohorts to estimates based on virus-enriched metagenomes, we explored a collection of 254 Danish fecal viromes [46].

The CGTR1899 database constitutes only 12% of the vOTUs detected in the four Dutch cohorts based on read alignment but encompasses 29% of all virus-like contigs assembled for these four cohorts. The CGTR1899 database includes vOTUs composed entirely of sequences from the four cohorts (19%), entirely of sequences from the databases (25%), or of a mixture of both (56%). A total of 404 CGTR1899 vOTUs were detected in the Danish fecal viromes, providing additional confirmation for the viral nature of these vOTUs (Table S2).

We measured the abundance of CGTR1899 phages per sample both as the number of viruses detected and the percentage of recruited reads. The mean number of CGTR1899 vOTUs detected in a sample was seven for LLD, 10 for LLD follow-up, nine for 300OB and five for IBD. On average, genomes representing CGTR1899 vOTUs recruited 0.68% (LLD), 0.75% (LLD follow-up), 0.74% (300OB), and 0.60% (IBD) reads per sample (Figure 2A). When we compared the CGTR1899 phage abundances in the four Dutch cohorts to those in the Danish healthy adult viromes (n = 52), the mean number of vOTUs detected was similar (six) but the mean read recruitment rate for the Danish samples was considerably higher (6.92%), as could be expected for these virus-enriched samples.

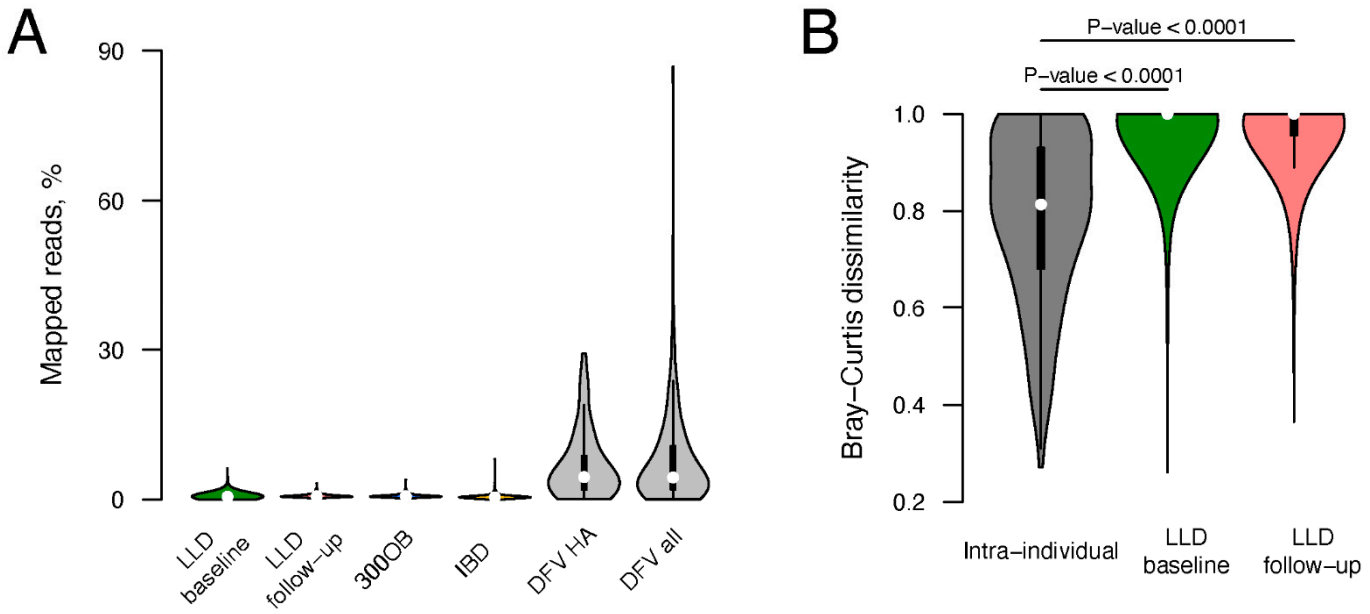

Figure 2. Abundance and stability of the CGTR1899 phages. (**A**) Violin plots show the percent of sample reads mapping to the genomes representing CGTR1899 vOTUs for the four Dutch cohorts (LLD baseline, LLD follow-up, 300OB and IBD), healthy adult Danish fecal viromes (DFV HA) and all Danish fecal viromes (DFV all). (**B**) Bray–Curtis dissimilarities between pairs of samples collected from the same individual 4 years apart (gray) and from different individuals at the same time point (green and red). Empirical *p*-values are indicated above the violin plots. Data from the 338 individuals sampled as part of the LLD and LLD follow-up cohorts were utilized.

We estimated the stability of the CGTR1899 fraction of the virome based on longitudinal data from the LLD and LLD follow-up cohorts. The Bray–Curtis dissimilarity between samples collected from the same individual 4 years apart was significantly lower than that between samples collected from different individuals at the same timepoint (p-value < 0.0001, Figure 2B), indicating relative stability. Importantly, as virus detection was based on total metagenome data, it was impossible to distinguish between phages in lytic and lysogenic states, and thus, stability was estimated for phages in all states taken together.

The diversity of the CGTR1899 phages was assessed via characteristics of the genomes representing vOTUs such as genome length, GC content, and genetic code. Notably, genomes with similar characteristics tended to cluster on a TerL-based phylogenetic tree (Figure 3, Table S1). The length of the genomes representing vOTUs ranged from 5061 to 352,502 nt. GC content varied from 25% to 69%. A minority of the genomes representing vOTUs (90, 5%) possessed ITR at their termini; the rest possessed DTR. A total of 1608 genomes representing vOTUs were predicted to employ standard bacterial genetic code, with the remaining 285, and six predicted to employ alternative genetic codes with a TAG or TGA stop codon recoded to an amino acid, respectively. Interestingly, transfer RNA (tRNA) genes with an anticodon matching TAG were detected in just 13% of the 285 genomes predicted to be TAG-recoded. There were 395 vOTUs containing phage genome sequences originating from prophage contigs (i.e., contigs including phage and host genome fragments) identified in the four Dutch cohorts (Table S1).

Taxonomic assignment of the CGTR1889 phages proved to be very sparse. The majority of the CGTR1889 vOTUs could not be assigned to an established monophyletic group at the level of family or order. Only 7% of the vOTUs were assigned to such groups based on the presence of previously classified genomes within the vOTUs (Figure 3, Table S1). Extending assignment by finding the most recent common ancestor (MRCA) of all vOTUs assigned to a group on the TerL-based phylogenetic tree, and then placing all descendants of the

MRCA into the group, resulted in 11% of vOTUs being taxonomically assigned: family *Autographiviridae* (4 vOTUs), family *Drexlerviridae* (2 vOTUs), Flandersviridae (also known as Gubaphages, 32 vOTUs) [2,37,75], Gratiaviridae (16 vOTUs) [37], crAss-like phages (135 vOTUs), group Topaz (18 vOTUs), and group Garnet (1 vOTU) [78].

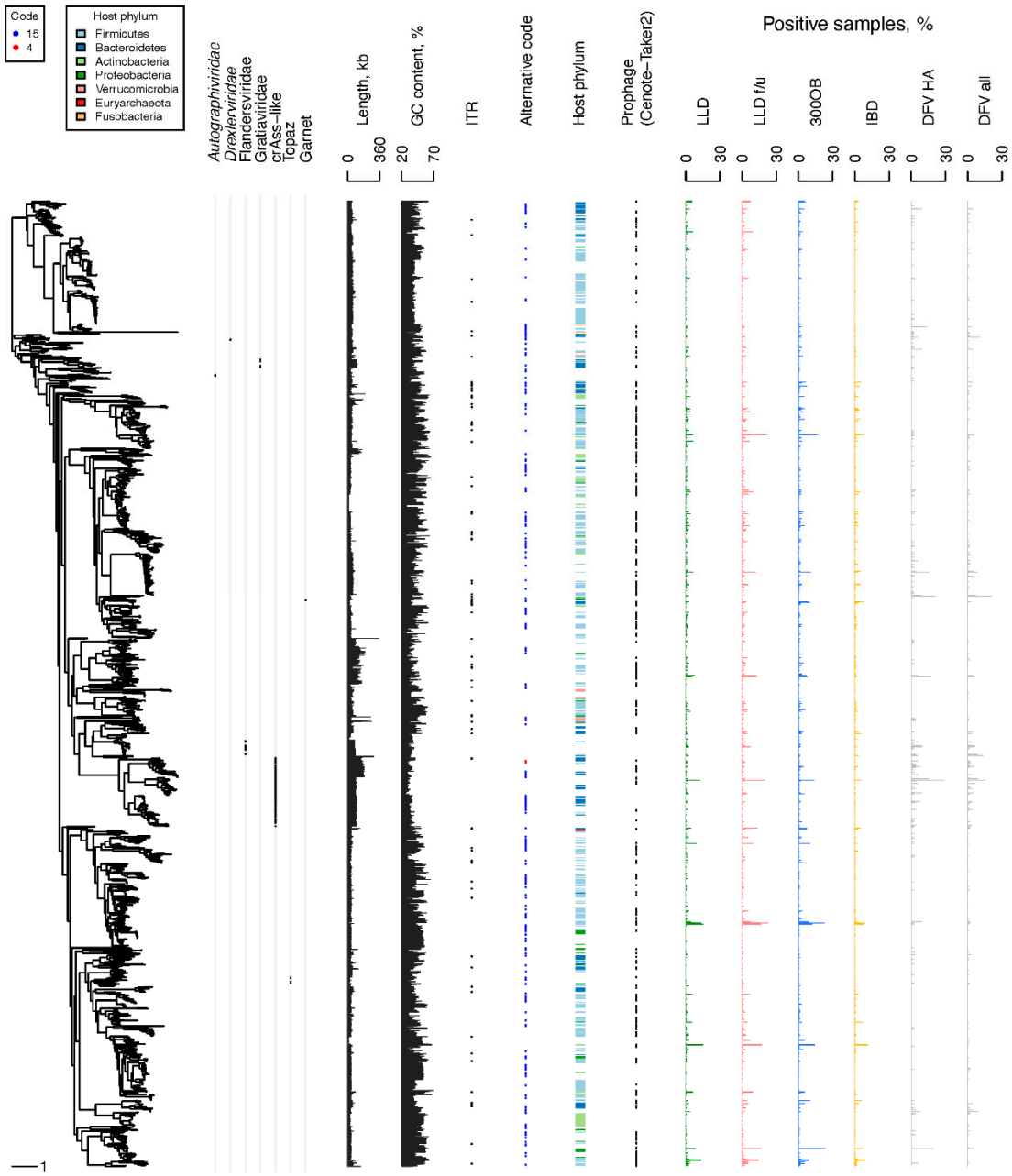

Figure 3. Properties of the CGTR1899 phages. Left, a phylogenetic tree reconstructed based on the TerL proteins of phages representing the CGTR1899 vOTUs. From left to right, the following genome properties are depicted per tree tip: assignment to order- or family-level taxonomic groups, length, GC content, terminal repeats (presence or absence of a black dot indicates ITR or DTR, respectively), predicted genetic code (blue dot for code 15, red dot for code 4 or empty space for standard code 11), predicted host phyla designated by colored bars (empty space if prediction is unavailable), presence of vOTU members derived from prophage contigs identified by Cenote-Taker 2, and prevalence in the four Dutch cohorts (LLD, LLD follow-up, 300OB and IBD), healthy adult Danish fecal viromes (DFV HA) and all Danish fecal viromes (DFV all).

Next, we predicted the hosts of the CGTR1899 phages using two methods: (1) analysis of the host genome fragments attached to prophage contigs identified in the four Dutch cohorts and (2) detection of sequence similarity between phage genomes and microbial CRISPR spacers (see Section 2). The first approach yielded predictions for 228 vOTUs. The second approach yielded predictions for 578 vOTUs. In total, predictions were made for 713 vOTUs (Table S3). Predictions were made by both methods for 93 vOTUs. In 88 cases, the predictions made by both methods were identical at the host phylum level. In the remaining five cases, the predictions were incompatible at the host phylum level. In these cases, we prioritized the prophage-based predictions. Predicted hosts belonged to the phyla Firmicutes (453 vOTUs), Bacteroidetes (138 vOTUs), Actinobacteria (60 vOTUs), Proteobacteria (47 vOTUs), Verrucomicrobia (13 vOTUs), Fusobacteria (1 vOTU) and Euryarchaeota (1 vOTU).

The prevalence of the individual CGTR1899 phages in the four Dutch cohorts varied depending on the phage. Some were detected in a single sample, whereas others were detected in dozens of samples (Figures 3 and 4, Table S2). The most prevalent vOTUs per cohort were NL_vir026707 in LLD (detected in 16% of samples), NL_vir053139 in LLD follow-up (detected in 22% of samples), MGV-GENOME-0279285 in 300OB (detected in 23% of samples) and OLXK01000549.1 in IBD (detected in 11% of samples). Unsurprisingly, the prevalence of the CGTR1899 phages in the Danish fecal viromes was often drastically different from that in the four Dutch cohorts. For example, the second most prevalent CGTR1899 vOTU among the Danish healthy adult viromes was MGV-GENOME-0193745 (21% positive samples), which was detected in <1% samples in every Dutch cohort. This vOTU includes genomes of the virulent *Leuconostoc* phages ΦLN03, ΦLN04, and ΦLN12 [79,80] that could be expected to be abundant among DNA sequences isolated from virus-like particles.

3.3. Diversity of the Most Prevalent Caudoviricetes Phages with Genome Terminal Repeats

In the final part of the analysis, we focused on the most prevalent CGTR1899 phages: 54 vOTUs detected in >5% samples in at least one of the four Dutch cohorts (and referred to as the CGTR54 database below). We explored their diversity, searched for known closely related viruses, and analyzed associations between the prevalence of these phages and human phenotypes. Importantly, since only total metagenome sequencing data are available for the four Dutch cohorts, we could not determine if each detected phage was in a lytic or lysogenic state in a particular sample. However, it is worth noting that 35 of the 54 vOTUs were detected in the Danish fecal viromes (Figure 4), suggesting that these vOTUs can exist in the form of virus particles.

The diversity of the CGTR54 phages can be assessed through the characteristics of their representative genomes. The lengths of the representative genomes varied from 5061 to 140,662 nt, and their GC content varied from 25% to 67%. A total of 44 of the CGTR54 vOTUs were represented by genomes with DTR. The remaining 10 vOTUs were represented by genomes with ITR. Four representative genomes were predicted to use an alternative genetic code with TAG stop codons recoded to an amino acid. Finally, 38 vOTUs included sequences originating from prophage contigs (Figure 4).

The diversity of the CGTR54 phages was also reflected in the organization of their representative genomes (Material S1). There were representative genomes with the majority of ORFs positioned on a single strand, as well as genomes with ORFs occupying both the forward and reverse strands. The 54 vOTUs were represented by genomes with various shapes of AT- and GC-skew curves, including the V-shaped cumulative GC-skew curve, previously suggested to be associated with bidirectional replication [81]. Some of the 54 genomes were evenly covered by reads, while in others we observed an anomaly where a specific region receives almost no coverage in a fraction of samples (Material S1), which may be a consequence of recombination [24]. Inspection of the genome maps also revealed that forward and reverse strand sequences of one of the genomes (MGV-GENOME-0281541)

are completely identical, indicating that this genome might be a sequencing or assembly artifact (Material S1).

Despite the diversity of the 54 genomes, there were some shared characteristics. Genes with similar functions, such as tRNA genes or genes encoding structural proteins and proteins implicated in assembly of virus particles, tended to form clusters (Material S1). Genes encoding integrases were detected in multiple genomes. Potential diversity-generating retroelements—a reverse transcriptase (RT) gene and a pair of nucleotide repeats with at least one repeat positioned in close proximity to the RT gene [82]—were detected in 21 genomes (Material S1). Interestingly, in multiple genomes, there was a pair of repeats with both repeats close to the RT gene and a pair of repeats with one repeat close to the RT gene and another separated from the RT gene by a large distance (Material S1, S2).

Taxonomic assignment was only possible for 4 CGTR54 vOTUs (Figure 4). MGV-GENOME-0359371 and uvig_191910 vOTUs were recognized as crAss-like phages because of the presence of previously described crAss-like phage genomes in these vOTUs (Table S1) [24]. The MGV-GENOME-0359371 vOTU belongs to the same clade of crAss-like phages as the crAssphage *sensu stricto*—clade alpha—although, the nucleotide sequence similarity between them is relatively low (only 27% of the crAssphage *sensu stricto* genome was covered by hits when compared to the MGV-GENOME-0359371 genome using BLASTN with E-value threshold 0.05) [83]. The uvig_191910 vOTU belongs to the gamma clade of crAss-like phages [38,84]. For the other two CGTR54 vOTUs, the MGV-GENOME-0351725 vOTU was recognized as belonging to the Flandersviridae group because it contains flandersvirus OJML01000036 [37] and the MGV-GENOME-0305083 vOTU was recognized as *Faecalibacterium* phage Lagaffe because it contains the genome of this phage, NC_047911 [85].

To investigate if any of the CGTR54 vOTUs had already been extensively characterized, we searched the viral fraction of GenBank for similar sequences (see Section 2). We found that five viral contigs that had been revealed to contain diversity-generating retroelements in [86] exhibited various degrees of similarity to six genomes representing CGTR54 vOTUs. In addition, eight phage sequences obtained in a project where single-cell viral tagging was used to identify unknown host–phage pairs [87] demonstrated similarity to five genomes representing CGTR54 vOTUs. Finally, contig71, which was subjected to PCR amplification and Sanger sequencing in [88], displayed similarity to a fragment of the CGTR54 flandersvirus (Figures 4 and S4, Table S4).

Hosts of the CGTR54 phages predicted based on the analysis of prophage contigs and CRISPR spacers included phyla Firmicutes (20 vOTUs), Bacteroidetes (10 vOTU) and Actinobacteria (2 vOTUs) (Figure 4). We also predicted the hosts of the CGTR54 phages based on co-abundance with microbial taxa. The potential host of each vOTU was predicted as the microbial taxon demonstrating the most reliable (minimal FDR in meta-analysis) relative abundance correlation with the vOTU (Table S3). Notably, there were 32 vOTUs with a prophage-based and/or CRISPR-based prediction available in addition to the co-abundance-based prediction, and we observed an agreement between all available predictions at host phylum level for 91% of these vOTUs (Table S3).

3.4. Associations with Human Phenotypes

The availability of phenotypic data for participants of the four Dutch cohorts provides an opportunity to explore associations between the detection of gut phages and human phenotypes. We explored the associations of the CGTR54 vOTUs using logistic regression where the presence of the phage represents the outcome and the phenotype represents the predictor, while adjusting for the age and sex of cohort participants. Subsequently, we conducted an additional analysis where the logistic regression was additionally adjusted for the abundance of the potential host predicted by the co-abundance analysis. Associations were considered significant at a FDR < 0.05.

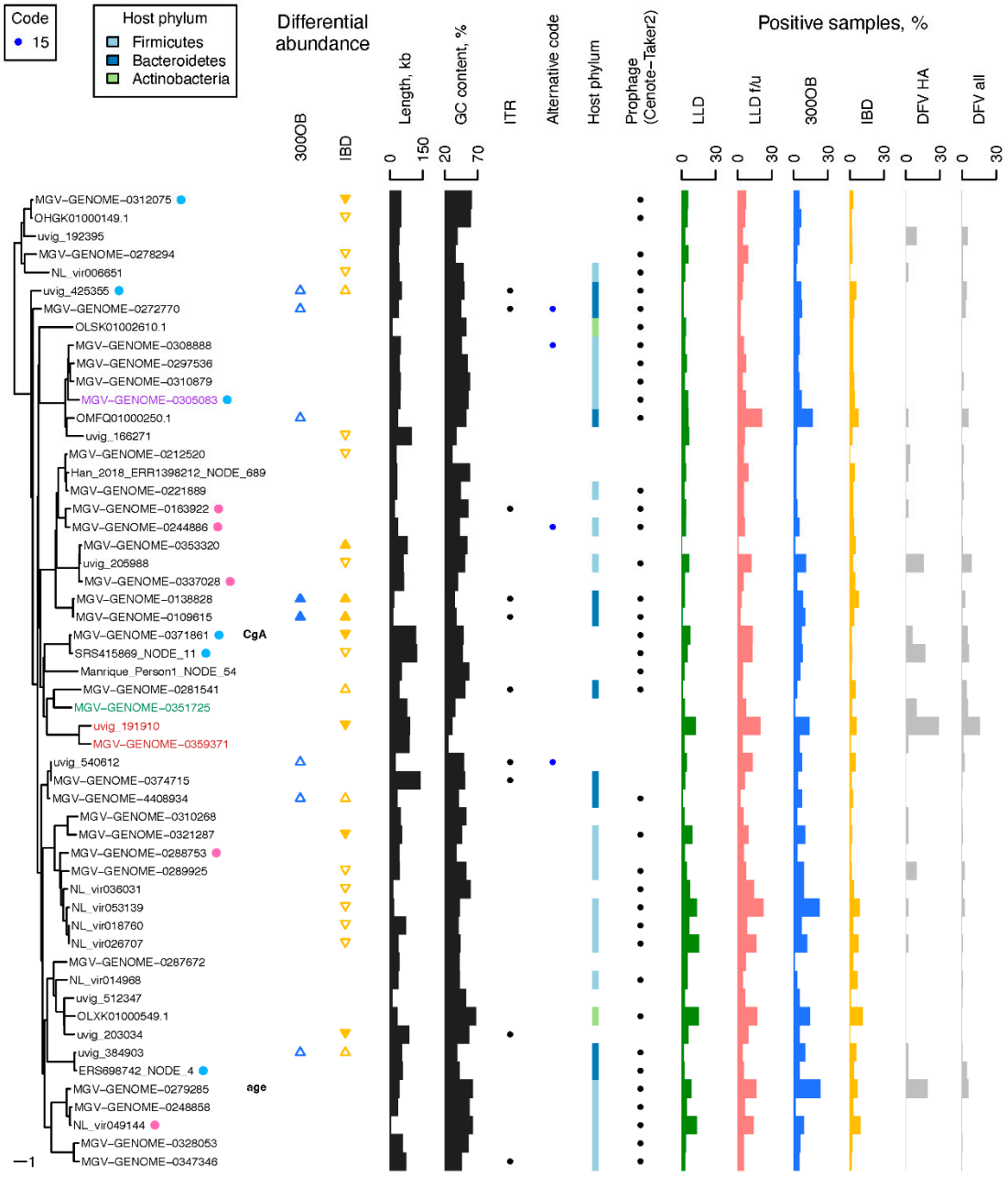

Figure 4. Properties of the CGTR54 phages. This figure shows a subset of the data presented in Figure 3 including only information about the 54 vOTUs detected in >5% samples of a Dutch cohort. See legend of Figure 3 for details. The crAss-like phage, flandersvirus, and *Faecalibacterium* phage Lagaffe vOTU names are written in red, green, and violet font, respectively. A dot next to a vOTU indicates that a phage with a similar genome sequence was described in Minot et al., 2012 [86] (pink) or Dzunkova et al., 2019 [87] (blue) (see Figure S4). A name of a phenotype ("CgA", "age") next to a name of a vOTU indicates a phage–phenotype association in the LLD cohort. Statistically significant differences in prevalence of vOTUs between the population cohort LLD and patient cohort 300OB (IBD) are indicated by blue (yellow) triangles. If a vOTU was overrepresented in a patient cohort, the triangle points upward. If it was underrepresented, the triangle points downward. If the association was significant after the logistic regression was adjusted for relative abundance of the predicted host, the triangle is filled, otherwise the triangle is empty. Contig length and coverage are omitted from the phage names for brevity, where applicable.

The association analysis based on the LLD cohort data revealed a negative association between the fecal level of the secretory protein chromogranin A (CgA) and detection of the MGV-GENOME-0371861 vOTU, and a positive association between the age of human subjects and detection of the MGV-GENOME-0279285 vOTU (Figure 4, Table S5). However, both associations were no longer significant after the adjustment for the abundance of the predicted hosts.

Eight vOTUs were found to be significantly more prevalent in overweight and obese individuals (300OB cohort, BMI > 27 kg/m^2) compared to the general population (LLD cohort). After adjustment for the abundance of the predicted hosts, only two of these associations remained statistically significant (Figure 4, Table S5).

Seven vOTUs were found to be significantly more prevalent among IBD cohort participants compared to the general population (LLD cohort), while 16 vOTUs were found to be significantly less prevalent. After the adjustment for the abundance of the predicted hosts, these numbers changed to 3 and 5 vOTUs, respectively (Figure 4, Table S5).

Overall, these results indicate that in many cases the driving force behind the association may not be the phage itself but rather its microbial host.

4. Discussion

We used a marker-based bioinformatics approach to identify and classify viral genomes in total fecal metagenomes from four Dutch cohorts: two population cohorts, a cohort of overweight and obese individuals, and a cohort of IBD patients. Detected viruses included those belonging to class *Caudoviricetes* and families *Microviridae*, *Papillomaviridae*, *Polyomaviridae,* and *Adenoviridae*, and we further focused specifically on *Caudoviricetes* phages with genome terminal repeats. We estimated the proportion of their nucleic acid in the human gut metagenomes (<1% on average), noted the relative stability of this virome fraction over a period of 4 years and described the diversity of these viruses, including their genome characteristics, predicted hosts and prevalence in human gut metagenomes. A small fraction of the *Caudoviricetes* phages with genome terminal repeats were highly prevalent (detected in >5% of Dutch cohort samples), allowing us to conduct a statistical analysis that identified associations between the prevalence of these phages and human phenotypes including age, fecal levels of CgA, obesity, and IBD diagnosis.

Metagenomics is a powerful approach to studying viruses that allows us to see the big picture of the human gut virome. However, technical challenges on each step of the study, from sample collection to bioinformatics analysis, may influence the results. Working with fecal samples, as opposed to intestinal wall biopsy samples, may affect the ratio of the number of viruses infecting microbes and human cells. Conducting total metagenome sequencing, as opposed to virus-enriched metagenome sequencing, means that viruses with dsDNA genomes will be included in the analysis even if they are in a lysogenic state, while viruses with ssDNA and RNA genomes will be excluded. Subsequent bioinformatics analysis carries its own limitations. Metagenomically assembled contigs may include sequencing and assembly artifacts, which might be difficult to distinguish from genuine biological phenomena based solely on metagenomics data. Properties of some CGTR1899 vOTUs may indicate the presence of artifacts, for example, the MGV-Genome-0281541 vOTU is represented by a sequence with completely identical forward and reverse strands, the MGV-GENOME-0370088 vOTU is one of 87 vOTUs that include both sequences with DTR and ITR, vOTUs represented by sequences with ITR are more widespread than reported for thoroughly described *Caudovirites* phages with ITR [10], the NL_vir049144 vOTU is represented by a 5061 nt sequence that is shorter than most known *Caudovirites* genomes. The marker-based approach that we used to identify virus genomes and tentatively assign them to taxonomic groups is designed to be very specific as it relies on the presence of protein genes uniquely associated with a particular group of viruses. However, it also has limitations. Marker proteins can only be identified based on the known virosphere, so it is always possible that a seemingly unique association between a protein gene and a group of viruses will be disproven with the discovery of novel viruses.

Likewise, the definitions of virus taxonomic groups may be shifting with the expansion of the known virosphere [5]. Another important aspect of our bioinformatics analysis was the identification of genomes with DTR and ITR, which also carries several potential pitfalls. It is possible to overlook terminal repeats if they are shorter than the threshold of 20 nt or contain a sequencing error, making the 5′- and the 3′-terminal repeat sequences non-identical. On the other hand, a circular plasmid with an integrated prophage or a partial viral genome flanked by repeats can be mistakenly identified as a complete viral genome with terminal repeats. Predicting ORFs is yet another virus bioinformatics challenge that we encountered: some phages employ alternative genetic codes with stop codon reassignment, and thus their ORFs cannot be correctly predicted by standard tools [32,38,78]. We solved this challenge by applying an approach similar to the one described in [3,32], which worked well on a crAss-like phage test dataset (see Section 2). Notably, the percent of viral contigs from the four Dutch cohorts predicted to employ an alternative genetic code (9.74%) was higher than reported in the literature based on the gut microbiomes of people consuming a westernized diet (2.25%) [78]. Furthermore, different regions of the same phage genome may employ different genetic codes [38,89]. Finally, the sensitivity and selectivity of virus detection is strongly influenced by the selection of the breadth of coverage threshold [50,90]. In this study, we considered a vOTU detected if the breadth of representative sequence coverage by reads reached 75%.

This study was specifically focused on phages that belong to class *Caudoviricetes* and possess genome terminal repeats. We used *Caudoviricetes* TerL detection as an indicator that a virus belongs to this group, while requiring that the genome in question does not encode the marker of the family *Herpesviridae*, as herpesviruses possess an evolutionarily-related TerL [6]. Detection of TerL was required to involve three TerL motifs: the Walker B motif belonging to the adenosine triphosphatase domain and motifs I and II belonging to the nuclease domain (Figure S2). This approach performed well in benchmarking, reaching a sensitivity of 94.3% and a specificity of 99.9%, although with the caveat that the Viral RefSeq database used for benchmarking contained sequences employed in the TerL detection procedure, leading to a potential overestimation of the approach's robustness (Text S1). Notably, the sensitivity did not reach 100%, and there may be several reasons for that. Not all phages belonging to the class *Caudoviricetes* encode TerL: for example, a satellite phage may lack terminase genes and employ a terminase enzyme of a helper phage instead [91]. Alternatively, a phage might possess a packaging enzyme that differs from the terminase of the majority of known *Caudoviricetes* phages [92]. The TerL gene may also be interrupted by an intron [93]. Finally, a TerL protein encoded by a divergent *Caudoviricetes* phage may fail to be recognized based on sequence similarity. On the other hand, given the diversity and mosaicism of viral genomes, it is impossible to exclude that uncharacterized viruses outside of class *Caudoviricetes* may encode close homologs of the *Caudoviricetes* TerL.

One of the most intriguing aspects of the human gut virome is its potential role in human health and disease. In this study, we found a positive correlation between the detection of a phage and human age, and a negative correlation between the detection of a phage and fecal levels of CgA, which is a precursor to peptides with regulatory and antimicrobial activities [94]. We also found that multiple phages are underrepresented or overrepresented in metagenomes of overweight and obese people and in patients with IBD. However, the interpretation of these results is not straightforward. Although we do not know the true host of each of the viruses in question, when we adjusted our statistical model for the relative abundance of their predicted hosts, most of the associations became statistically insignificant. This strongly suggests that the microbial host, and not its phage, may be the driving force behind many of the associations we detected. This seems especially logical because we were working with total metagenome data and thus could not differentiate between the detection of a phage in an actively replicating lytic state versus one in a dormant lysogenic state.

To summarize, while the big picture of the human gut virome painted with the help of metagenomics is very informative, further research and improvement of the analysis techniques in the future can help resolve remaining uncertainties.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/v14102305/s1, Figure S1: Workflow leading to delineation of the CGTR1899 and CGTR54 databases; Figure S2: TerL alignment conservation; Figure S3: Gene-sharing network of viruses with evolutionary related terminase genes; Figure S4: Sequence similarity between genomes representing the CGTR54 vOTUs and extensively characterized phages; Table S1: Properties of the CGTR1899 phage genome sequences; Table S2: Properties of the CGTR1899 vOTU representatives; Table S3: Predicted phage hosts; Table S4: Extensively characterized phage sequences similar to the CGTR54 vOTU representatives; Table S5: Associations with human phenotypes; Material S1: Characteristics of the CGTR54 genomes; Material S2. MSA of RTs from the CGTR54 genomes representing vOTUs; Text S1: Benchmarking of virus detection and taxonomic assignment.

Author Contributions: Conceptualization, methodology, software, formal analysis, investigation, visualization, and writing—original draft preparation, A.G.; validation, S.G., A.K. and A.Z.; writing—review and editing, all authors; resources, A.V.V., N.P.R., M.G.N., R.K.W., J.F. and A.Z.; supervision and project administration, A.Z.; funding acquisition, N.P.R., M.G.N., R.K.W., J.F. and A.Z. All authors have read and agreed to the published version of the manuscript.

Funding: S.G. holds a scholarship from the Graduate School of Medical Sciences, University of Groningen. A.Z. is supported by European Research Council (ERC) Starting Grant 715772, Netherlands Organization for Scientific Research (NWO) VIDI grant 016.178.056 and NWO Gravitation grant ExposomeNL 024.004.017. J.F. is supported by NWO Gravitation grant Netherlands Organ-on-Chip Initiative 024.003.001, ERC Consolidator grant 101001678 and NWO VICI grant VI.C.202.022. N.P.R., M.G.N., J.F. and A.Z. are supported by The Netherlands Heart Foundation CVON grant 2018-27. R.K.W. is supported by the Seerave Foundation and the Dutch Digestive Foundation (16-14).

Institutional Review Board Statement: The LifeLines-DEEP baseline and follow-up cohort projects were approved by the institutional ethics review board of the University Medical Center Groningen (ref. M12.113965). The 300OB cohort project was approved by the Ethical Committee of the Radboud University (nr. 46846.091.13). The 1000IBD cohort project was approved by the institutional review board (IRB) of the University Medical Center Groningen (IRB number 2008.338).

Informed Consent Statement: All participants signed an informed consent form prior to cohort enrolment.

Data Availability Statement: Phage genomes and genome fragments identified in this study and belonging to the CGTR1899 were published in the Figshare repository https://doi.org/10.6084/m9.figshare.20747248. The code used to conduct the analysis was deposited to the GitHub repository https://github.com/aag1/NL_vir_analysis/.

Acknowledgments: We would like to thank Kate Mc Intyre for editing the manuscript, the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster and the volunteers of the LifeLines-DEEP, 300OB and IBD cohorts for their participation.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Koonin, E.V.; Dolja, V.V.; Krupovic, M. The healthy human virome: From virus-host symbiosis to disease. *Curr. Opin. Virol.* **2021**, *47*, 86–94. [CrossRef] [PubMed]
2. Benler, S.; Koonin, E.V. Fishing for phages in metagenomes: What do we catch, what do we miss? *Curr. Opin. Virol.* **2021**, *49*, 142–150. [CrossRef] [PubMed]
3. Nayfach, S.; Paez-Espino, D.; Call, L.; Low, S.J.; Sberro, H.; Ivanova, N.N.; Proal, A.D.; Fischbach, M.A.; Bhatt, A.S.; Hugenholtz, P.; et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat. Microbiol.* **2021**, *6*, 960–970. [CrossRef] [PubMed]
4. Liang, G.; Bushman, F.D. The human virome: Assembly, composition and host interactions. *Nat. Rev. Microbiol.* **2021**, *19*, 514–527. [CrossRef] [PubMed]

5. Turner, D.; Kropinski, A.M.; Adriaenssens, E.M. A Roadmap for Genome-Based Phage Taxonomy. *Viruses* **2021**, *13*, 506. [CrossRef] [PubMed]
6. Iranzo, J.; Krupovic, M.; Koonin, E.V. The Double-Stranded DNA Virosphere as a Modular Hierarchical Network of Gene Sharing. *mBio* **2016**, *7*, e00978-16. [CrossRef] [PubMed]
7. Adriaenssens, E.M. Phage Diversity in the Human Gut Microbiome: A Taxonomist's Perspective. *mSystems* **2021**, *6*, e0079921. [CrossRef] [PubMed]
8. Casjens, S.R.; Gilcrease, E.B. Determining DNA packaging strategy by analysis of the termini of the chromosomes in tailed-bacteriophage virions. *Methods Mol. Biol.* **2009**, *502*, 91–111. [CrossRef] [PubMed]
9. Merrill, B.D.; Ward, A.T.; Grose, J.H.; Hope, S. Software-based analysis of bacteriophage genomes, physical ends, and packaging strategies. *BMC Genom.* **2016**, *17*, 679. [CrossRef] [PubMed]
10. Meijer, W.J.; Horcajadas, J.A.; Salas, M. Phi29 family of phages. *Microbiol. Mol. Biol. Rev.* **2001**, *65*, 261–287. [CrossRef] [PubMed]
11. Kieft, K.; Anantharaman, K. Virus genomics: What is being overlooked? *Curr. Opin. Virol.* **2022**, *53*, 101200. [CrossRef] [PubMed]
12. Mantynen, S.; Laanto, E.; Oksanen, H.M.; Poranen, M.M.; Diaz-Munoz, S.L. Black box of phage-bacterium interactions: Exploring alternative phage infection strategies. *Open Biol.* **2021**, *11*, 210188. [CrossRef] [PubMed]
13. Howard-Varona, C.; Hargreaves, K.R.; Abedon, S.T.; Sullivan, M.B. Lysogeny in nature: Mechanisms, impact and ecology of temperate phages. *ISME J.* **2017**, *11*, 1511–1520. [CrossRef] [PubMed]
14. Walker, P.J.; Siddell, S.G.; Lefkowitz, E.J.; Mushegian, A.R.; Adriaenssens, E.M.; Alfenas-Zerbini, P.; Dempsey, D.M.; Dutilh, B.E.; Garcia, M.L.; Curtis Hendrickson, R.; et al. Recent changes to virus taxonomy ratified by the International Committee on Taxonomy of Viruses (2022). *Arch. Virol.* **2022**, *167*, 2429–2440. [CrossRef] [PubMed]
15. Garmaeva, S.; Sinha, T.; Kurilshikov, A.; Fu, J.; Wijmenga, C.; Zhernakova, A. Studying the gut virome in the metagenomic era: Challenges and perspectives. *BMC Biol.* **2019**, *17*, 84. [CrossRef]
16. Tigchelaar, E.F.; Zhernakova, A.; Dekens, J.A.; Hermes, G.; Baranska, A.; Mujagic, Z.; Swertz, M.A.; Munoz, A.M.; Deelen, P.; Cenit, M.C.; et al. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: Study design and baseline characteristics. *BMJ Open* **2015**, *5*, e006772. [CrossRef] [PubMed]
17. Zhernakova, A.; Kurilshikov, A.; Bonder, M.J.; Tigchelaar, E.F.; Schirmer, M.; Vatanen, T.; Mujagic, Z.; Vila, A.V.; Falony, G.; Vieira-Silva, S.; et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **2016**, *352*, 565–569. [CrossRef] [PubMed]
18. Chen, L.; Wang, D.; Garmaeva, S.; Kurilshikov, A.; Vich Vila, A.; Gacesa, R.; Sinha, T.; Lifelines Cohort, S.; Segal, E.; Weersma, R.K.; et al. The long-term genetic stability and individual specificity of the human gut microbiome. *Cell* **2021**, *184*, 2302–2315.e12. [CrossRef]
19. Ter Horst, R.; van den Munckhof, I.C.L.; Schraa, K.; Aguirre-Gamboa, R.; Jaeger, M.; Smeekens, S.P.; Brand, T.; Lemmers, H.; Dijkstra, H.; Galesloot, T.E.; et al. Sex-Specific Regulation of Inflammation and Metabolic Syndrome in Obesity. *Arter. Thromb. Vasc Biol.* **2020**, *40*, 1787–1800. [CrossRef] [PubMed]
20. Kurilshikov, A.; van den Munckhof, I.C.L.; Chen, L.; Bonder, M.J.; Schraa, K.; Rutten, J.H.W.; Riksen, N.P.; de Graaf, J.; Oosting, M.; Sanna, S.; et al. Gut Microbial Associations to Plasma Metabolites Linked to Cardiovascular Phenotypes and Risk. *Circ. Res.* **2019**, *124*, 1808–1820. [CrossRef]
21. Vich Vila, A.; Imhann, F.; Collij, V.; Jankipersadsing, S.A.; Gurry, T.; Mujagic, Z.; Kurilshikov, A.; Bonder, M.J.; Jiang, X.; Tigchelaar, E.F.; et al. Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome. *Sci. Transl. Med.* **2018**, *10*, eaap8914. [CrossRef] [PubMed]
22. Imhann, F.; Van der Velde, K.J.; Barbieri, R.; Alberts, R.; Voskuil, M.D.; Vich Vila, A.; Collij, V.; Spekhorst, L.M.; Van der Sloot, K.W.J.; Peters, V.; et al. The 1000IBD project: Multi-omics data of 1000 inflammatory bowel disease patients; data release 1. *BMC Gastroenterol.* **2019**, *19*, 5. [CrossRef]
23. Nurk, S.; Meleshko, D.; Korobeynikov, A.; Pevzner, P.A. metaSPAdes: A new versatile metagenomic assembler. *Genome Res.* **2017**, *27*, 824–834. [CrossRef] [PubMed]
24. Gulyaeva, A.; Garmaeva, S.; Ruigrok, R.; Wang, D.; Riksen, N.P.; Netea, M.G.; Wijmenga, C.; Weersma, R.K.; Fu, J.; Vila, A.V.; et al. Discovery, diversity, and functional associations of crAss-like phages in human gut metagenomes from four Dutch cohorts. *Cell Rep.* **2022**, *38*, 110204. [CrossRef]
25. Tisza, M.J.; Belford, A.K.; Dominguez-Huerta, G.; Bolduc, B.; Buck, C.B. Cenote-Taker 2 democratizes virus discovery and sequence annotation. *Virus Evol.* **2021**, *7*, veaa100. [CrossRef] [PubMed]
26. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
27. Quast, C.; Pruesse, E.; Yilmaz, P.; Gerken, J.; Schweer, T.; Yarza, P.; Peplies, J.; Glockner, F.O. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* **2013**, *41*, D590–D596. [CrossRef] [PubMed]
28. Roux, S.; Enault, F.; Hurwitz, B.L.; Sullivan, M.B. VirSorter: Mining viral signal from microbial genomic data. *PeerJ* **2015**, *3*, e985. [CrossRef] [PubMed]
29. Chan, P.P.; Lin, B.Y.; Mak, A.J.; Lowe, T.M. tRNAscan-SE 2.0: Improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.* **2021**, *49*, 9077–9096. [CrossRef] [PubMed]
30. Yu, M.K.; Fogarty, E.C.; Eren, A.M. The genetic and ecological landscape of plasmids in the human gut. *bioRxiv* **2022**. [CrossRef]

31. Hyatt, D.; Chen, G.L.; Locascio, P.F.; Land, M.L.; Larimer, F.W.; Hauser, L.J. Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **2010**, *11*, 119. [CrossRef] [PubMed]
32. Ivanova, N.N.; Schwientek, P.; Tripp, H.J.; Rinke, C.; Pati, A.; Huntemann, M.; Visel, A.; Woyke, T.; Kyrpides, N.C.; Rubin, E.M. Stop codon reassignments in the wild. *Science* **2014**, *344*, 909–913. [CrossRef]
33. Mistry, J.; Chuguransky, S.; Williams, L.; Qureshi, M.; Salazar, G.A.; Sonnhammer, E.L.L.; Tosatto, S.C.E.; Paladin, L.; Raj, S.; Richardson, L.J.; et al. Pfam: The protein families database in 2021. *Nucleic Acids Res.* **2021**, *49*, D412–D419. [CrossRef] [PubMed]
34. Lawrence, M.; Huber, W.; Pages, H.; Aboyoun, P.; Carlson, M.; Gentleman, R.; Morgan, M.T.; Carey, V.J. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **2013**, *9*, e1003118. [CrossRef] [PubMed]
35. Charif, D.; Lobry, J.R. SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. In *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations*; Bastolla, U., Porto, M., Roman, H.E., Vendruscolo, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2007; pp. 207–232.
36. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [CrossRef]
37. Benler, S.; Yutin, N.; Antipov, D.; Rayko, M.; Shmakov, S.; Gussow, A.B.; Pevzner, P.; Koonin, E.V. Thousands of previously unknown phages discovered in whole-community human gut metagenomes. *Microbiome* **2021**, *9*, 78. [CrossRef]
38. Yutin, N.; Benler, S.; Shmakov, S.A.; Wolf, Y.I.; Tolstoy, I.; Rayko, M.; Antipov, D.; Pevzner, P.A.; Koonin, E.V. Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features. *Nat. Commun.* **2021**, *12*, 1044. [CrossRef]
39. Grazziotin, A.L.; Koonin, E.V.; Kristensen, D.M. Prokaryotic Virus Orthologous Groups (pVOGs): A resource for comparative genomics and protein family annotation. *Nucleic Acids Res.* **2017**, *45*, D491–D498. [CrossRef]
40. Yutin, N.; Backstrom, D.; Ettema, T.J.G.; Krupovic, M.; Koonin, E.V. Vast diversity of prokaryotic virus genomes encoding double jelly-roll major capsid proteins uncovered by genomic and metagenomic sequence analysis. *Virol. J.* **2018**, *15*, 67. [CrossRef]
41. Aylward, F.O.; Moniruzzaman, M.; Ha, A.D.; Koonin, E.V. A phylogenomic framework for charting the diversity and evolution of giant viruses. *PLoS Biol.* **2021**, *19*, e3001430. [CrossRef]
42. Wheeler, T.J.; Clements, J.; Finn, R.D. Skylign: A tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. *BMC Bioinform.* **2014**, *15*, 7. [CrossRef] [PubMed]
43. Aylward, F.O.; Moniruzzaman, M. ViralRecall-A Flexible Command-Line Tool for the Detection of Giant Virus Signatures in 'Omic Data. *Viruses* **2021**, *13*, 150. [CrossRef] [PubMed]
44. Nayfach, S.; Camargo, A.P.; Schulz, F.; Eloe-Fadrosh, E.; Roux, S.; Kyrpides, N.C. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **2020**, *39*, 578–585. [CrossRef] [PubMed]
45. Roux, S.; Adriaenssens, E.M.; Dutilh, B.E.; Koonin, E.V.; Kropinski, A.M.; Krupovic, M.; Kuhn, J.H.; Lavigne, R.; Brister, J.R.; Varsani, A.; et al. Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nat. Biotechnol.* **2019**, *37*, 29–37. [CrossRef]
46. Van Espen, L.; Bak, E.G.; Beller, L.; Close, L.; Deboutte, W.; Juel, H.B.; Nielsen, T.; Sinar, D.; De Coninck, L.; Frithioff-Bojsoe, C.; et al. A Previously Undescribed Highly Prevalent Phage Identified in a Danish Enteric Virome Catalog. *mSystems* **2021**, *6*, e0038221. [CrossRef]
47. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [CrossRef]
48. Quinlan, A.R.; Hall, I.M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **2010**, *26*, 841–842. [CrossRef]
49. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; Genome Project Data Processing, S. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef]
50. Roux, S.; Emerson, J.B.; Eloe-Fadrosh, E.A.; Sullivan, M.B. Benchmarking viromics: An in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ* **2017**, *5*, e3817. [CrossRef]
51. Steinegger, M.; Meier, M.; Mirdita, M.; Vohringer, H.; Haunsberger, S.J.; Soding, J. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinform.* **2019**, *20*, 473. [CrossRef]
52. Waterhouse, A.M.; Procter, J.B.; Martin, D.M.; Clamp, M.; Barton, G.J. Jalview Version 2—A multiple sequence alignment editor and analysis workbench. *Bioinformatics* **2009**, *25*, 1189–1191. [CrossRef] [PubMed]
53. Grant, B.J.; Rodrigues, A.P.; ElSawy, K.M.; McCammon, J.A.; Caves, L.S. Bio3d: An R package for the comparative analysis of protein structures. *Bioinformatics* **2006**, *22*, 2695–2696. [CrossRef] [PubMed]
54. Minh, B.Q.; Schmidt, H.A.; Chernomor, O.; Schrempf, D.; Woodhams, M.D.; von Haeseler, A.; Lanfear, R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **2020**, *37*, 1530–1534. [CrossRef] [PubMed]
55. Hoang, D.T.; Chernomor, O.; von Haeseler, A.; Minh, B.Q.; Vinh, L.S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **2018**, *35*, 518–522. [CrossRef]
56. Whelan, S.; Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **2001**, *18*, 691–699. [CrossRef]
57. Schliep, K.P. phangorn: Phylogenetic analysis in R. *Bioinformatics* **2011**, *27*, 592–593. [CrossRef]
58. Gaïa, M.; Meng, L.; Pelletier, E.; Forterre, P.; Vanni, C.; Fernandez-Guerra, A.; Jaillon, O.; Wincker, P.; Ogata, H.; Krupovic, M.; et al. Plankton-infecting relatives of herpesviruses clarify the evolutionary trajectory of giant viruses. *bioRxiv* **2022**. [CrossRef]

59. Bin Jang, H.; Bolduc, B.; Zablocki, O.; Kuhn, J.H.; Roux, S.; Adriaenssens, E.M.; Brister, J.R.; Kropinski, A.M.; Krupovic, M.; Lavigne, R.; et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* **2019**, *37*, 632–639. [CrossRef]
60. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020.
61. Sayers, E.W.; Cavanaugh, M.; Clark, K.; Pruitt, K.D.; Schoch, C.L.; Sherry, S.T.; Karsch-Mizrachi, I. GenBank. *Nucleic Acids Res.* **2022**, *50*, D161–D164. [CrossRef]
62. Shmakov, S.A.; Sitnik, V.; Makarova, K.S.; Wolf, Y.I.; Severinov, K.V.; Koonin, E.V. The CRISPR Spacer Space Is Dominated by Sequences from Species-Specific Mobilomes. *mBio* **2017**, *8*, e01397-17. [CrossRef]
63. Pourcel, C.; Touchon, M.; Villeriot, N.; Vernadet, J.P.; Couvin, D.; Toffano-Nioche, C.; Vergnaud, G. CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays and cas genes from complete genome sequences, and tools to download and query lists of repeats and spacers. *Nucleic Acids Res.* **2020**, *48*, D535–D544. [CrossRef] [PubMed]
64. Roux, S.; Paez-Espino, D.; Chen, I.A.; Palaniappan, K.; Ratner, A.; Chu, K.; Reddy, T.B.K.; Nayfach, S.; Schulz, F.; Call, L.; et al. IMG/VR v3: An integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Res.* **2021**, *49*, D764–D775. [CrossRef] [PubMed]
65. Beghini, F.; McIver, L.J.; Blanco-Miguez, A.; Dubois, L.; Asnicar, F.; Maharjan, S.; Mailyan, A.; Manghi, P.; Scholz, M.; Thomas, A.M.; et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife* **2021**, *10*, e65088. [CrossRef] [PubMed]
66. Balduzzi, S.; Rücker, G.; Schwarzer, G. How to perform a meta-analysis with R: A practical tutorial. *Evid. Based Ment. Health* **2019**, *22*, 153–160. [CrossRef]
67. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Society. Ser. B (Methodol.)* **1995**, *57*, 289–300.
68. Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **2000**, *16*, 276–277. [CrossRef]
69. Wagih, O. ggseqlogo: A versatile R package for drawing sequence logos. *Bioinformatics* **2017**, *33*, 3645–3647. [CrossRef]
70. Paradis, E.; Schliep, K. ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **2019**, *35*, 526–528. [CrossRef]
71. Robert, X.; Gouet, P. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* **2014**, *42*, W320–W324. [CrossRef]
72. Brister, J.R.; Ako-Adjei, D.; Bao, Y.; Blinkova, O. NCBI viral genomes resource. *Nucleic Acids Res.* **2015**, *43*, D571–D577. [CrossRef]
73. Soto-Perez, P.; Bisanz, J.E.; Berry, J.D.; Lam, K.N.; Bondy-Denomy, J.; Turnbaugh, P.J. CRISPR-Cas System of a Prevalent Human Gut Bacterium Reveals Hyper-targeting against Phages in a Human Virome Catalog. *Cell Host Microbe* **2019**, *26*, 325–335.e325. [CrossRef] [PubMed]
74. Gregory, A.C.; Zablocki, O.; Zayed, A.A.; Howell, A.; Bolduc, B.; Sullivan, M.B. The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut. *Cell Host Microbe* **2020**, *28*, 724–740.e8. [CrossRef] [PubMed]
75. Camarillo-Guerrero, L.F.; Almeida, A.; Rangel-Pineros, G.; Finn, R.D.; Lawley, T.D. Massive expansion of human gut bacteriophage diversity. *Cell* **2021**, *184*, 1098–1109.e1099. [CrossRef] [PubMed]
76. Devoto, A.E.; Santini, J.M.; Olm, M.R.; Anantharaman, K.; Munk, P.; Tung, J.; Archie, E.A.; Turnbaugh, P.J.; Seed, K.D.; Blekhman, R.; et al. Megaphages infect Prevotella and variants are widespread in gut microbiomes. *Nat. Microbiol.* **2019**, *4*, 693–700. [CrossRef]
77. Al-Shayeb, B.; Sachdeva, R.; Chen, L.X.; Ward, F.; Munk, P.; Devoto, A.; Castelle, C.J.; Olm, M.R.; Bouma-Gregson, K.; Amano, Y.; et al. Clades of huge phages from across Earth's ecosystems. *Nature* **2020**, *578*, 425–431. [CrossRef]
78. Borges, A.L.; Lou, Y.C.; Sachdeva, R.; Al-Shayeb, B.; Penev, P.I.; Jaffe, A.L.; Lei, S.; Santini, J.M.; Banfield, J.F. Widespread stop-codon recoding in bacteriophages may regulate translation of lytic genes. *Nat. Microbiol.* **2022**, *7*, 918–927. [CrossRef]
79. Kot, W.; Hammer, K.; Neve, H.; Vogensen, F.K. Identification of the receptor-binding protein in lytic Leuconostoc pseudomesenteroides bacteriophages. *Appl. Env. Microbiol.* **2013**, *79*, 3311–3314. [CrossRef]
80. Kot, W.; Hansen, L.H.; Neve, H.; Hammer, K.; Jacobsen, S.; Pedersen, P.D.; Sorensen, S.J.; Heller, K.J.; Vogensen, F.K. Sequence and comparative analysis of Leuconostoc dairy bacteriophages. *Int. J. Food Microbiol.* **2014**, *176*, 29–37. [CrossRef]
81. Grigoriev, A. Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res.* **1998**, *26*, 2286–2290. [CrossRef]
82. Liu, M.; Deora, R.; Doulatov, S.R.; Gingery, M.; Eiserling, F.A.; Preston, A.; Maskell, D.J.; Simons, R.W.; Cotter, P.A.; Parkhill, J.; et al. Reverse transcriptase-mediated tropism switching in Bordetella bacteriophage. *Science* **2002**, *295*, 2091–2094. [CrossRef]
83. Dutilh, B.E.; Cassman, N.; McNair, K.; Sanchez, S.E.; Silva, G.G.; Boling, L.; Barr, J.J.; Speth, D.R.; Seguritan, V.; Aziz, R.K.; et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* **2014**, *5*, 4498. [CrossRef] [PubMed]
84. 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.; Hill, C. Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. *Cell Host Microbe* **2018**, *24*, 653–664.e6. [CrossRef] [PubMed]

85. Cornuault, J.K.; 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*, 65. [CrossRef] [PubMed]
86. Minot, S.; Grunberg, S.; Wu, G.D.; Lewis, J.D.; Bushman, F.D. Hypervariable loci in the human gut virome. *Proc. Natl Acad Sci. USA* **2012**, *109*, 3962–3966. [CrossRef]
87. Dzunkova, M.; Low, S.J.; Daly, J.N.; Deng, L.; Rinke, C.; Hugenholtz, P. Defining the human gut host-phage network through single-cell viral tagging. *Nat. Microbiol.* **2019**, *4*, 2192–2203. [CrossRef] [PubMed]
88. Ly, M.; Jones, M.B.; Abeles, S.R.; Santiago-Rodriguez, T.M.; Gao, J.; Chan, I.C.; Ghose, C.; Pride, D.T. Transmission of viruses via our microbiomes. *Microbiome* **2016**, *4*, 64. [CrossRef] [PubMed]
89. Pfennig, A.; Lomsadze, A.; Borodovsky, M. Annotation of Phage Genomes with Multiple Genetic Codes. *bioRxiv* **2022**. [CrossRef]
90. Weinheimer, A.R.; Aylward, F.O. Infection strategy and biogeography distinguish cosmopolitan groups of marine jumbo bacteriophages. *ISME J.* **2022**, *16*, 1657–1667. [CrossRef] [PubMed]
91. Christie, G.E.; Dokland, T. Pirates of the Caudovirales. *Virology* **2012**, *434*, 210–221. [CrossRef] [PubMed]
92. Mao, H.; Saha, M.; Reyes-Aldrete, E.; Sherman, M.B.; Woodson, M.; Atz, R.; Grimes, S.; Jardine, P.J.; Morais, M.C. Structural and Molecular Basis for Coordination in a Viral DNA Packaging Motor. *Cell Rep.* **2016**, *14*, 2017–2029. [CrossRef] [PubMed]
93. Mikkonen, M.; Alatossava, T. A group I intron in the terminase gene of Lactobacillus delbrueckii subsp. lactis phage LL-H. *Microbiology* **1995**, *141 Pt 9*, 2183–2190. [CrossRef] [PubMed]
94. Bartolomucci, A.; Possenti, R.; Mahata, S.K.; Fischer-Colbrie, R.; Loh, Y.P.; Salton, S.R. The extended granin family: Structure, function, and biomedical implications. *Endocr. Rev.* **2011**, *32*, 755–797. [CrossRef] [PubMed]