AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

RESEARCH ARTICLE

A Previously Undescribed Highly Prevalent Phage Identified in a Danish Enteric Virome Catalog

Lore Van Espen,[a] Emilie Glad Bak,[b] Leen Beller,[a] Lila Close,[a] Ward Deboutte,[a] Helene Bæk Juel,[b] Trine Nielsen,[b] Deniz Sinar,[a] Lander De Coninck,[a] Christine Frithioff-Bøjsøe,[b,c] Cilius Esmann Fonvig,[b,c] Suganya Jacobsen,[d,e] Maria Kjærgaard,[d,e] Maja Thiele,[d,e] Anthony Fullam,[f] Michael Kuhn,[f] Jens-Christian Holm,[b,c,g] Peer Bork,[f,h,i,j] Aleksander Krag,[d,e] Torben Hansen,[b] Manimozhiyan Arumugam,[b] Jelle Matthijnssens[a]

[a]KU Leuven, Department of Microbiology, Immunology, & Transplantation, Rega Institute, Division of Clinical & Epidemiological Virology, Laboratory of Viral Metagenomics, Leuven, Belgium
[b]The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[c]The Children's Obesity Clinic, accredited European Centre for Obesity Management, Department of Paediatrics, Copenhagen University Hospital Holbaek, Holbaek, Denmark
[d]Department of Gastroenterology and Hepatology, Centre for Liver Research, Odense University Hospital, Odense, Denmark
[e]Department of Clinical Research, University of Southern Denmark, Odense, Denmark
[f]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany
[g]Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[h]Max Delbrück Centre for Molecular Medicine, Berlin, Germany
[i]Yonsei Frontier Lab (YFL), Yonsei University, Seoul, South Korea
[j]Department of Bioinformatics, Biocenter, University of Würzburg, Würzburg, Germany

Lore Van Espen and Emilie Glad Bak contributed equally. Author order was decided by the corresponding authors.

ABSTRACT Gut viruses are important, yet often neglected, players in the complex human gut microbial ecosystem. Recently, the number of human gut virome studies has been increasing; however, we are still only scratching the surface of the immense viral diversity. In this study, 254 virus-enriched fecal metagenomes from 204 Danish subjects were used to generate the Danish Enteric Virome Catalog (DEVoC) containing 12,986 nonredundant viral scaffolds, of which the majority was previously undescribed, encoding 190,029 viral genes. The DEVoC was used to compare 91 healthy DEVoC gut viromes from children, adolescents, and adults that were used to create the DEVoC. Gut viromes of healthy Danish subjects were dominated by phages. While most phage genomes (PGs) only occurred in a single subject, indicating large virome individuality, 39 PGs were present in more than 10 healthy subjects. Among these 39 PGs, the prevalences of three PGs were associated with age. To further study the prevalence of these 39 prevalent PGs, 1,880 gut virome data sets of 27 studies from across the world were screened, revealing several age-, geography-, and disease-related prevalence patterns. Two PGs also showed a remarkably high prevalence worldwide—a crAss-like phage (20.6% prevalence), belonging to the tentative *AlphacrAssvirinae* subfamily, and a previously undescribed circular temperate phage infecting *Bacteroides dorei* (14.4% prevalence), called LoVEphage because it encodes lots of viral elements. Due to the LoVEphage's high prevalence and novelty, public data sets in which the LoVEphage was detected were *de novo* assembled, resulting in an additional 18 circular LoVEphage-like genomes (67.9 to 72.4 kb).

IMPORTANCE Through generation of the DEVoC, we added numerous previously uncharacterized viral genomes and genes to the ever-increasing worldwide pool of human gut viromes. The DEVoC, the largest human gut virome catalog generated from consistently processed fecal samples, facilitated the analysis of the 91 healthy Danish gut viromes. Characterizing the biggest cohort of healthy gut viromes from children, adolescents, and adults to date confirmed the previously established high

Citation Van Espen L, Bak EG, Beller L, Close L, Deboutte W, Juel HB, Nielsen T, Sinar D, De Coninck L, Frithioff-Bøjsøe C, Fonvig CE, Jacobsen S, Kjærgaard M, Thiele M, Fullam A, Kuhn M, Holm J-C, Bork P, Krag A, Hansen T, Arumugam M, Matthijnssens J. 2021. A previously undescribed highly prevalent phage identified in a Danish enteric virome catalog. mSystems 6:e00382-21. https://doi.org/10.1128/mSystems.00382-21.

Editor Chaysavanh Manichanh, Vall d'Hebron Research Institute (Ed. Mediterranea)

Ad Hoc Peer Reviewers Adrian Paskey, Reckitt; Ghjuvan Grimaud, Michigan State University

Address correspondence to Torben Hansen, torben.hansen@sund.ku.dk, Manimozhiyan Arumugam, arumugam@sund.ku.dk, or Jelle Matthijnssens, jelle.matthijnssens@kuleuven.be.

Previously undescribed, widely prevalent, Bacteroides-infecting temperate phage discovered during generation of a Danish enteric virome catalog.

Received 29 March 2021
Accepted 2 September 2021
Published 19 October 2021

interindividual variation in human gut viromes and demonstrated that the effect of age on the gut virome composition was limited to the prevalence of specific phage (groups). The identification of a previously undescribed prevalent phage illustrates the usefulness of developing virome catalogs, and we foresee that the DEVoC will benefit future analysis of the roles of gut viruses in human health and disease.

KEYWORDS human gut virome, virome catalog, healthy gut viromes, phages

Gut microbiota, consisting of bacteria, archaea, viruses, fungi, and other eukaryotic microorganisms, play a major role in human health and disease (1, 2). Both structural and functional imbalances of the gut bacteria, called dysbiosis, have been associated with diseases such as obesity (3), diabetes (4, 5), inflammatory bowel disease (IBD) (6), cancer (7), and neurological diseases (8). At the same time, research on human gut viruses, collectively called gut virobiota, is still in its infancy (9), although recent studies demonstrated associations with IBD (10, 11), diabetes (12, 13), liver disease (14, 15), and cancer (16).

Only a minority of the human gut virobiota consists of eukaryotic viruses, infecting human cells, fungal cells, and unicellular eukaryotes residing in the gut or infecting plant or animal cells transiting as part of the diet (17). The vast majority of viruses in the human gut are bacteriophages (phages), which rely on a bacterial host to reproduce (18). The close interplay between phages and bacteria, which are already implicated in numerous diseases, combined with the ability of gut viruses to directly interact with the human host (19, 20), led to gut viruses gaining more interest as potential disease biomarkers (16) and treatments for disease (21, 22). It is therefore important to shed more light on the virobiota, and their collective genomes referred to as the virome, as this will pave the way to unravelling complex interactions within the gut microbiota and their effect on the human host (23).

Recent progress in high-throughput sequencing technologies, viral enrichment procedures, and development of downstream viral bioinformatic tools has facilitated human gut virome studies investigating their association with health (24–29) or disease (30–34), as well as their dynamics (18). However, several significant challenges in studying human gut viromes remain (23). Most importantly, identification of viruses from metagenomes is hampered by incomplete databases (23) and therefore requires specialized viral identification tools, e.g., VirSorter (35), MetaPhinder (36), and DeepVirFinder (37), that do not (only) rely on similarity to known viral genomes/proteins but also look at genome structure to detect viral signatures. High viral mutation rates cause immense viral genetic diversity (38), thereby complicating viral identification based on homology to reference genomes. Even though a few virome databases recently emerged, they are often not gut-specific (IMG/VR [39], Reference Viral DataBase [RVDB; 40] and Earth's virome [41]) or focus only on phages or eukaryotic viruses (Gut Phage Database [GPD; 42], a "circular" phage database [43], and RVDB [40]). However, despite these developments, a large fraction of sequences originating from human gut virome studies cannot be identified as viral because they are not present in databases and are therefore called "viral dark matter" (23). Moreover, taxonomic characterization of human gut viruses is virtually impossible due to the major proportion of viruses being taxonomically unclassified (44, 45), despite ongoing efforts by the International Committee for the Taxonomy of Viruses (ICTV) (46). Thus, viral taxonomic analysis is mostly performed on scaffold level or on artificial taxonomic levels generated by gene-sharing tools, e.g., vConTACT2 (47) or GRAViTy (48). Finally, the lack of host information and functional annotation of proteins complicates the characterization of the phages and their interactions with bacteria in the gut (23, 49). The technical difficulties of identifying and characterizing viruses are numerous. Nevertheless, it is important to make progress in generating human gut virome catalogs and characterizing them to shed light on the viral dark matter. This was exemplified by the discovery of crAssphages from the cross-assembly of human gut metagenomes across publicly available data sets (50). This novel group of phages is now believed to be one of the

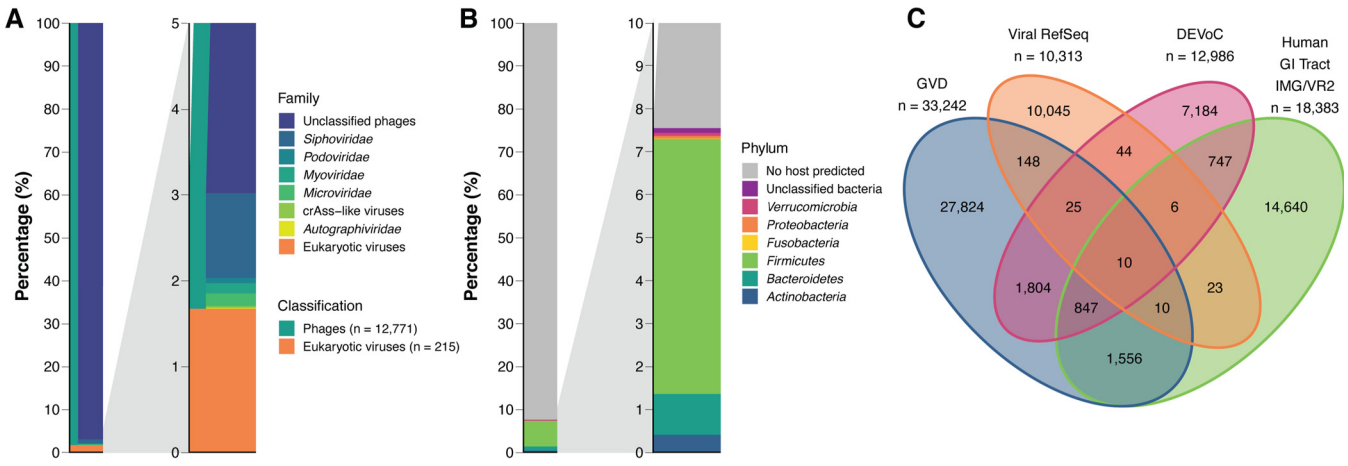

FIG 1 DEVoC mainly consists of undescribed phages. (A) Overview of the DEVoC scaffolds (n = 12,986) by type of virus and phage family. (Breakdown of the eukaryotic viruses into families is visualized in Fig. S3). (B) Overview of the DEVoC phages (n = 12,771) by phylum of the predicted bacterial host. (C) Venn-diagram showing the number of clusters with members of the DEVoC, GVD, IMG/VR2, and ViralRefSeq databases at 95% identity over 80% coverage. Numbers in the Venn diagram do not sum up to the database sizes, as a viral sequence from one database may cluster with multiple partial sequences from a second database; 2,319 sequences in DEVoC, 1,018 in GVD, 544 in IMG/VR2, and 2 in ViralRefSeq were merged in this manner.

most prevalent viruses of the human gut (51, 52). Additionally, a recent study showed that human gut viromes are highly individual (29), emphasizing the importance of cataloging viromes from diverse human populations.

In this study, we characterized 254 fecal viral metagenomes from Danish children and adolescents (6 to 18 years old) and adults (aged 40 to 73 years old), to develop the Danish Enteric Virome Catalog (DEVoC). The DEVoC facilitated assessment of the diversity of the healthy Danish gut viromes, a population in which gut viromes have not been characterized before. Some phage genomes (PGs) were associated with age, while other PGs were present in human gut viromes worldwide. In particular, a previously undescribed PG, which we named LoVEphage, was prevalent in both the healthy Danish subjects and in publicly available human gut viromes. These insights, as well as the DEVoC, will further improve our understanding of the role of viruses in the human gut microbiota and thus human health.

RESULTS

A catalog of 12,986 nonredundant viral scaffolds derived from Danish fecal viromes encoding 190,029 proteins. The Danish Enteric Virome Catalog (DEVoC) was constructed based on 254 Danish fecal viromes (3.86 billion raw reads). The viral scaffolds constituting the DEVoC ranged in size from 1 kb to 191 kb (N_{50}, 16 kb; L_{50}, 1,463 scaffolds), of which 1,867 viral scaffolds (14.4%) were more than 50% complete as estimated by CheckV (53). This small subset of viral scaffolds, however, dominates these Danish fecal viromes, as they represented 87.4% of the total amount of viral reads (Fig. S1A).

Phages represented the vast majority of DEVoC scaffolds (n = 12,771; 98.3%; Fig. 1A) and viral reads (99.2%). The phage scaffolds were clustered using vConTACT2 to generate viral clusters (VCs) as a proxy for viral subfamilies or genera. vConTACT2 formed 1,488 VCs covering 5,222 phage scaffolds (41% of the DEVoC phage scaffolds) representing 73% of the phage reads (Fig. S1B). Merely 176 phage scaffolds (1.4%) could be taxonomically classified based on clustering with RefSeq genomes (30 VCs)—3 crAss-like genomes (1 VC), 19 *Microviridae* genomes (3 VCs), 1 *Autographiviridae* genome (1 VC), 8 *Podoviridae* (3 VCs), 16 *Myoviridae* (6 VCs), and 129 *Siphoviridae* genomes (16 VCs). Bacterial hosts were identified using CRISPR spacers for 963 phage scaffolds (7.5%). At the phylum level, *Firmicutes* (n = 758) and *Bacteroidetes* (n = 121) accounted for the largest fractions of hosts (Fig. 1B), while *Faecalibacterium* (n = 226), *Bacteroides* (n = 62), *Ruminococcus* (n = 58), and *Bifidobacterium* (n = 51) were the most common host genera.

A small subset of the DEVoC scaffolds represented viruses infecting eukaryotes (n = 215; 1.7%). Most putative eukaryotic viral scaffolds (65.6%) belonged to the

Picobirnaviridae family (subject to interpretation, as increasing evidence suggests that viruses belonging to this family are phages [54]). The remaining putative eukaryotic viral genomes belonged to plant-infecting viral families probably originating from the diet (*Alphaflexiviridae* [0.9%], *Betaflexiviridae* [1.4%], *Bromoviridae* [1.4%], *Partitiviridae* [7.0%], *Tombusviridae* [0.5%], *Tymoviridae* [0.5%], and *Virgaviridae* [5.1%]), fungi-infecting viral families (*Chrysoviridae* [1.4%] and *Totiviridae* [2.8%]), and viral families that are known or hypothesized to infect mammals (*Anelloviridae* [0.9%], *Caliciviridae* [1.4%], *Circoviridae* [5.1%], *Genomoviridae* [2.3%], *Parvoviridae* [0.5%], *Picornaviridae* [2.3%], and *Smacoviridae* [0.5%]) (Fig. S2A).

To understand the functional potential of the viruses in the DEVoC, we predicted viral genes and annotated them using Cenote-Taker 2 (49). The 190,029 DEVoC genes ranged in size from 0.06 to 18.2 kb (median, 0.34 kb; interquartile range [IQR], 0.19 to 0.62 kb), and 91.3% were complete. About half of the DEVoC genes (n = 102,018; 53.7%) were functionally annotated, with the most common predicted annotations being major capsid protein, portal protein, large terminase, integrase, and minor capsid protein, all typical phage functions. The DEVoC proteins were clustered using Proteinortho (55) and formed 18,473 orthologous groups (OGs), containing up to 360 members (median, 3 members; IQR, 2 to 6 members) covering 140,581 DEVoC proteins (74%). The remaining 49,448 proteins (26%) remained singletons (regarded as OGs with one member from now on).

The majority of the DEVoC scaffolds are previously undescribed. We compared DEVoC scaffolds to existing viral genome databases to assess their novelty. Viral scaffolds from the NCBI RefSeq v201 database (n = 10,313), the human Gut Virome Database (18) (GVD; n = 33,242), and the human gastrointestinal tract subset of the IMG/VR2 database (56) (n = 18,383) were clustered with the DEVoC scaffolds at 95% identity over 80% coverage. Each of the databases contained a remarkably large set of previously undescribed viral sequence clusters (DEVoC, 67.3%; GVD, 86.3%; IMG/VR, 82.1%; ViralRefSeq, 97.4%; Fig. 1C). DEVoC shared the largest number of clusters with the GVD (n = 2,686 containing 4,583 DEVoC scaffolds), followed by IMG/VR2 (n = 1,610 containing 3,096 DEVoC scaffolds). Only 857 clusters (containing 1,960 DEVoC scaffolds) were shared among all three human gut-specific viral genome databases, and these were all phage clusters. This small overlap between the databases reflects the high interpersonal, potentially cross-regional, age-spanning variation of the human gut virome that metagenomic research has merely begun to uncover. A minor fraction of the DEVoC clusters was shared with ViralRefSeq (n = 85 containing 222 DEVoC scaffolds), 62 phage and 23 eukaryotic viral clusters. This limited overlap can be attributed to the underrepresentation of phages in ViralRefSeq (3,672 phage genomes versus 9,476 eukaryotic virus genomes). In total, all four databases shared 10 viral clusters, including an uncultured crAssphage, and members of the *Siphoviridae* (*Ceduavirus*, *Limdunavirus*, *Oengusvirus*, *Skunavirus*, and *Unaquatrovirus* genera) and *Myoviridae* (*Brigitvirus*, *Lagaffevirus*, *Peduovirus*, and *Toutatisvirus* genera) families (Table S1).

Healthy Danish gut viromes are highly individual. The remaining analyses solely included gut viromes from 91 healthy Danish subjects, including 46 children and adolescents (6 to 18 years old) and 45 adults (40 to 73 years old). Samples that we did not analyze belong to obese children and adolescents and alcoholic liver disease (ALD) patients, which are all part of a larger ongoing study. The 91 healthy Danish gut viromes were dominated by phages (relative abundance versus all viral reads; median, >99.9%; IQR, 99.8% to 100%; range, 79.4% to 100%). As multiple fragments from the same genome can hamper phage community-level analysis when they are treated as separate viruses, we restricted the analysis to phage scaffolds that represented more than 50% of a genome as determined by CheckV (53) (here referred to as phage genomes [PGs]). This allows us to limit the analysis to a maximum of one fragment for any given genome. Within the 91 healthy Danish gut viromes, 7,153 phage scaffolds (56% of the DEVoC phage scaffolds) were detected, and 1,162 of these were PGs (62.2% of DEVoC PGs). The PGs recruited a median of 90.2% of the phage reads per sample (IQR, 78.8% to 94.6%; range, 0.83% to 99.6%). The sample in which PGs

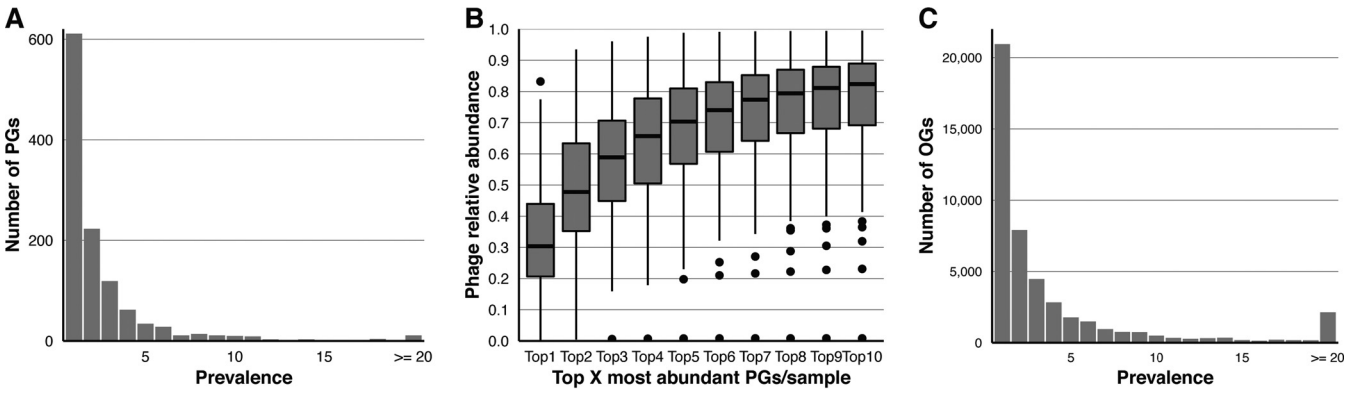

FIG 2 Danish gut viromes are highly individual and dominated by a limited number of phages. (A) Bar plot of the prevalence of PGs (n = 1,162) in healthy Danish subjects (n = 91). PGs occurring in 20 or more subjects are grouped. (B) Boxplots of the fraction of all phage reads taken up by the most dominant PGs in different healthy Danish subjects (n = 91). (C) Bar plot of the prevalence of the viral OGs (n = 46,620) in healthy Danish subjects (n = 91). OGs occurring in 20 or more subjects are grouped.

accounted for 0.83% of phage reads was dominated by one phage scaffold with undetermined completeness (>99% of viral reads).

The most prevalent PG was a partial *Skunavirus* genome detected in 30 subjects (33% prevalence). Including this PG, only 39 PGs (3.4% of all PGs) occurred in more than 10 subjects (>12% prevalence; Table S2). This subset of 39 highly prevalent PGs will be further looked into in the next sections and included six skunaviruses, two eponaviruses, and one *Limdunavirus*, *Unaquatrovirus*, and crAss-like phage each, while the remaining 18 highly prevalent PGs remained unclassified. In contrast, more than half of the PGs were subject-specific (n = 611; 52.6%; Fig. 2A), suggesting that the healthy gut phageome is highly individual. Within each subject's phageome, the proportion of subject-specific PGs (versus all PGs; median, 18.7%; IQR, 14.0% to 24.7%; range, <0.1% to 40.0%) and their relative abundance (versus all phage reads; median, 13.5%; IQR, 7.3% to 25.3%; range, <0.1% to 83.7%) varied greatly. The most abundant PG within each subject recruited between 0.24% and 83.2% of the phage reads (median, 30.4%; IQR, 20.7% to 44.0%; Fig. 2B), while the 10 most abundant PGs represented the majority of the phage reads in most subjects (median, 82.4%; IQR, 69.2% to 89.0%; range, 0.83% to 99.5%; Fig. 2B). This suggests that the overall diversity of the phageome can be captured by the 10 most abundant PGs in most samples.

Few eukaryotic viral species were detected in the gut viromes of healthy subjects (n = 33). The majority (n = 12) were plant viruses and therefore presumably not stable members of the gut virome but, rather, transient passengers. The median observed eukaryotic viral species richness was barely 1 (IQR, 0 to 3; range, 0 to 9; Fig. S2B), and most eukaryotic viruses were present in only one or two healthy subjects (Fig. S2C), suggesting that eukaryotic viruses are highly individual.

At the protein level, all healthy subjects combined harbored 46,620 viral OGs (68.6% of DEVoC OGs). The majority of OGs were present in only one or two healthy subjects (Fig. 2C), and the number of OGs in healthy subjects ranged from 282 to 6,397 (median, 2,270; IQR, 1,584 to 2,904). A median of 7.9% of the OGs within each subject were unique to that subject (IQR, 5.6% to 10.5%; range, 0.5 to 23.8%). Notably, the most prevalent OG was recovered in almost all subjects (n = 88; 96.7%), and 51 OGs were found across more than 80% of the subjects (Table S3). The five most prevalent OGs (prevalence, >93%) were predicted to encode a recombination protein, a nuclease, a reverse transcriptase, a terminase large subunit, and a dUTPase.

Several phage genomes and viral functions are associated with age. We investigated if the virome composition differed between the healthy gut phageomes of the pediatric (n = 46) and the adult cohort (n = 45) cohorts. PG alpha diversity was not affected by age group (Wilcoxon-test; observed richness, P = 0.89; Shannon's diversity, P = 0.83; Fig. S3A and B), and although age group was significantly associated with PG

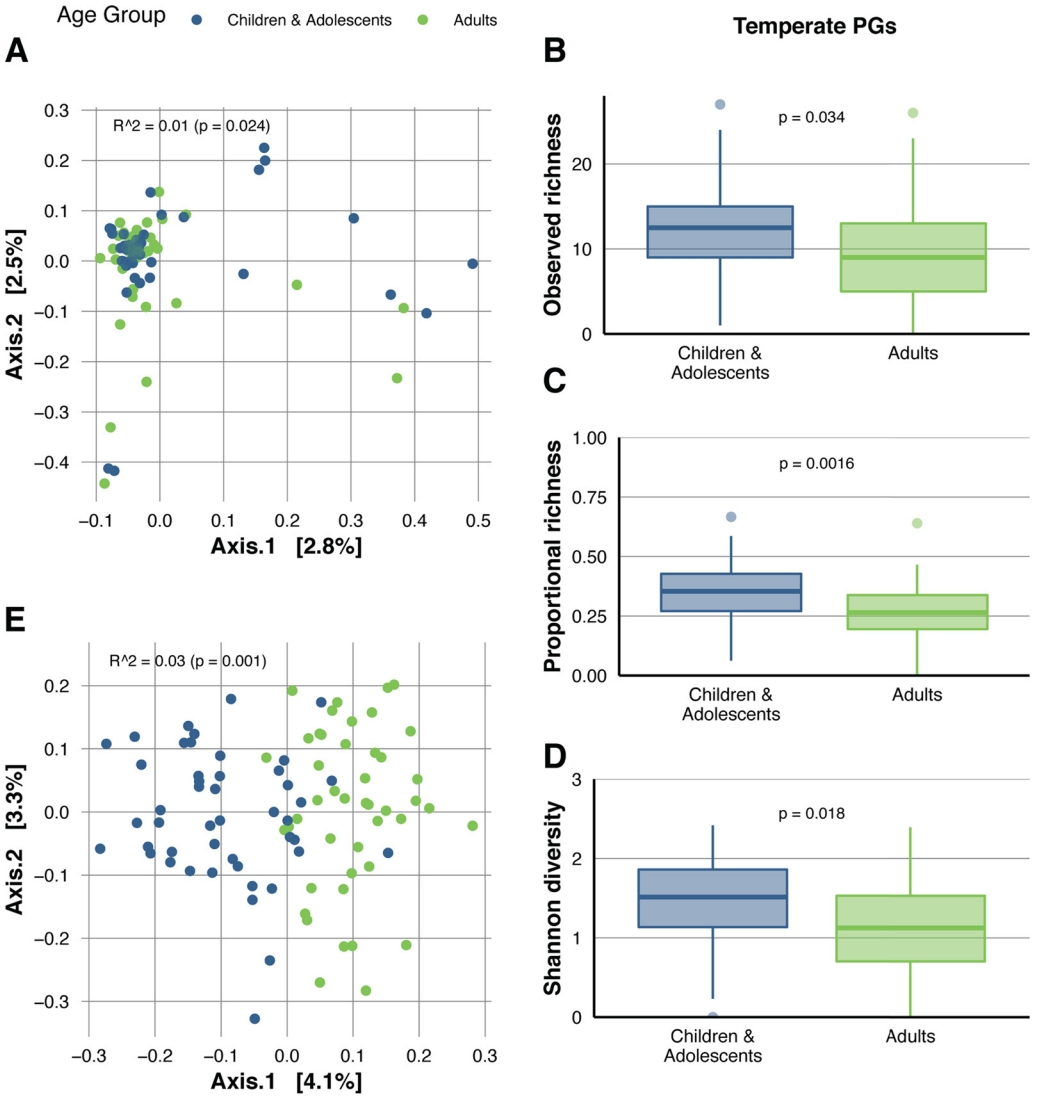

FIG 3 Age group-associated virome patterns in healthy Danish subjects. (A) Principal-coordinate analysis on Jaccard dissimilarities between healthy Danish subjects at the PG level (PERMANOVA of age group; R^2 = 0.01; P = 0.024). Subjects are colored by age group. (B) Boxplots of the number of temperate PGs in healthy Danish children and adolescents (n = 46) and adults (n = 45) (Wilcoxon test; P = 0.034). (C) Boxplots of the proportional richness of temperate PGs (number of temperate PGs versus total number of PGs) in healthy Danish children and adolescents (n = 46) and adults (n = 45) (Wilcoxon test; P = 0.0016). (D) Boxplots of Shannon's diversity of temperate PGs in healthy Danish children and adolescents (n = 46) and adults (n = 45) (Wilcoxon test; P = 0.018). (E) Principal-coordinate analysis on Jaccard dissimilarities between healthy Danish subjects at the OG level (PERMANOVA of age group; R^2 = 0.03; P = 0.001). Subjects are colored by age group. All analyses are performed on 46 children/adolescents and 45 adults.

beta diversity (permutational multivariate analysis of variance [PERMANOVA]; Jaccard dissimilarity; P = 0.024), it only explained 1% of the variance and might hence not be biologically relevant (Fig. 3A). Low percentages of explained variability by the first two principal components indicated a large interindividual diversity in gut phageomes.

To analyze if the occurrence of individual PGs was associated with age group, we compared prevalences between the pediatric and the adult group. Among the subset of the 39 most prevalent PGs (present in more than 10 subjects; >12% prevalence), PG8 was more common in children and adolescents, while PG7 and PG22 were more prevalent in adults (Chi-square test; adjusted [adj.] $P < 0.05$; Table S2).

The genomic structures of all three age-associated PGs are visualized in Fig. S4. PG8 is predicted to encode proteins involved in the activation or suppression of the lysogenic cycle, indicating that this phage has a temperate lifestyle. Temperate phages

TABLE 1 Orthologous groups with age-associated absence/presence profiles

Orthologous group identifier	Size	Annotation	Function	Prevalence in the healthy subset [*n* (%)]			Chi2 test adjusted *P* value[a]
				All (*n* = 91)	Pediatric cohort (*n* = 46)[b]	Adult cohort (*n* = 45)[b]	
OG_17005	180	Hypothetical protein	Unknown	63 (69.2)	**42 (91.3)**	21 (46.7)	0.0011
OG_17212	205	Carlavirus endopeptidase	Assembly	63 (69.2)	**42 (91.3)**	21 (46.7)	0.0011
OG_116	149	Tail assembly chaperone protein	Assembly	58 (63.7)	**40 (87)**	18 (40)	0.0008
OG_17197	159	Hypothetical protein	Unknown	56 (61.5)	**42 (91.3)**	14 (31.1)	< 0.0001
OG_17367	149	Major capsid/head protein	Structural	54 (59.3)	**41 (89.1)**	13 (28.9)	< 0.0001
OG_17685	118	Minor structural protein	Structural	54 (59.3)	**39 (84.8)**	15 (33.3)	0.0002
OG_863	115	Hypothetical protein	Unknown	49 (53.8)	**36 (78.3)**	13 (28.9)	0.0006
OG_16990	86	Hypothetical protein	Unknown	46 (50.5)	**35 (76.1)**	11 (24.4)	0.0002
OG_1899	95	Tail completion protein	Assembly	44 (48.4)	**35 (76.1)**	9 (20)	< 0.0001
OG_2871	96	Portal protein	Packaging	43 (47.3)	**33 (71.7)**	10 (22.2)	0.0006
OG_3146	86	Hypothetical protein	Unknown	41 (45.1)	**32 (69.6)**	9 (20)	0.0005
OG_3199	20	Polysaccharide export protein	Other	37 (40.7)	**31 (67.4)**	6 (13.3)	< 0.0001
OG_16319	71	tRNA synthase	Translation	35 (38.5)	**29 (63)**	6 (13.3)	0.0003
OG_2045	48	Hypothetical protein	Unknown	34 (37.4)	**28 (60.9)**	6 (13.3)	0.0007
OG_2076	6	Putative metallopeptidase	Other	33 (36.3)	5 (10.9)	**28 (62.2)**	0.0001
OG_752	45	Hypothetical protein	Unknown	33 (36.3)	**29 (63)**	4 (8.9)	< 0.0001
OG_16058	4	Hypothetical protein	Unknown	32 (35.2)	**28 (60.9)**	4 (8.9)	0.0001
OG_17591	68	Hypothetical protein	Unknown	31 (34.1)	**27 (58.7)**	4 (8.9)	0.0002
OG_2749	34	Hypothetical protein	Unknown	31 (34.1)	**26 (56.5)**	5 (11.1)	0.0012
OG_1811	3	Plasmid recombination enzyme	Recombination	30 (33)	**26 (56.5)**	4 (8.9)	0.0004
OG_16535	37	LytR response regulator	Other	29 (31.9)	**25 (54.3)**	4 (8.9)	0.0009
OG_17358	59	Hypothetical protein	Unknown	27 (29.7)	**25 (54.3)**	2 (4.4)	0.0001
OG_17353	63	Bromodomain RACK7-like subfamily	Other	26 (28.6)	**24 (52.2)**	2 (4.4)	0.0001
OG_18041	45	Head tail connector protein	Structural	26 (28.6)	**24 (52.2)**	2 (4.4)	0.0001
OG_18129	44	Minor structural protein	Structural	24 (26.4)	**22 (47.8)**	2 (4.4)	0.0008
OG_2613	38	Hypothetical protein	Unknown	22 (24.2)	**21 (45.7)**	1 (2.2)	0.0004
OG_3028	35	Hypothetical protein	Unknown	22 (24.2)	**21 (45.7)**	1 (2.2)	0.0004
OG_2406	37	DNA binding protein	Other	20 (22)	**20 (43.5)**	0 (0)	0.0002
OG_2150	27	Hypothetical protein	Unknown	19 (20.9)	**19 (41.3)**	0 (0)	0.0004

[a]Bonferroni-adjusted *P* values of chi-squared test on prevalences.
[b]Prevalences in bold indicate the cohort with the highest prevalence.

have the potential to alter the bacterial host phenotype and shift the dynamics of the complex gut microbial network. Therefore, we identified lysogeny-associated genes (listed in Table S4) in the PGs and classified 345 temperate PGs in the healthy Danish subjects (29.7%). Each subject had a median of 11 different temperate PGs (IQR, 6.5 to 15; range, 0 to 27), representing roughly one-third of a subject's PGs (median, 31.6%, IQR, 21.5% to 39.5%; range, 0 to 66.7%) and accounting for a median of 19.3% of the PG reads (IQR, 8.8% to 42.6%; range, 0% to 95.3%). Among the temperate PGs, the alpha-diversity measures observed (absolute) richness, proportional (versus all PGs) richness, and Shannon diversity were higher in children/adolescents than in adults (Wilcoxon test; $P = 0.034$, $P = 0.0016$ and $P = 0.018$, respectively; Fig. 3B to D), while we did not observe a difference in the relative abundance of temperate PG (versus all phage reads; $P = 0.21$; Fig. S3C).

We further assessed the association between age group and viral functions represented by OGs. Similar to the previous analysis, the observed richness of viral OGs did not differ between age groups (Wilcoxon test; $P = 0.11$; Fig. S3D). However, age group explained 3% of the beta diversity between subjects (Jaccard dissimilarity; PERMANOVA; $P = 0.001$; Fig. 3E). Analysis of all OGs containing two or more members and present in more than 10 healthy subjects ($n = 3{,}627$) identified 29 OGs with a higher prevalence in one or both age groups (Chi-squared test; adj. $P < 0.05$; Table 1). Interestingly, only one OG (a putative metallopeptidase) was detected more often in adults, while the remaining OGs were more common in the pediatric cohort.

Highly prevalent DEVoC phage genomes are detected worldwide. We further examined whether the subset of 39 highly prevalent PGs defined earlier in the healthy Danish subjects (Table S2) could be recovered worldwide, across age groups and

diseases. For this purpose, we obtained 1,880 fecal viral metagenomes from NCBI SRA (denoted SRA viromes here), deriving from 1,181 subjects (see Table S5 for an overview of the included studies). The highly prevalent DEVoC PGs were widely detected in SRA viromes (Fig. 4A). The prevalence of these 39 PGs was significantly associated with the geographical region (continent of sample collection) of the SRA viromes (Kruskal-Wallis test; $P < 0.0001$; Fig. 4B). Our prevalent PGs were found more often in Europeans ($n = 164$) than in subjects from the other continents (Wilcoxon signed-rank test; versus America ($n = 170$), adj. $P < 0.0001$; versus Africa ($n = 188$), adj. $P < 0.0001$; versus Asia ($n = 20$), adj. $P = 0.038$). Moreover, they exhibited higher prevalence in Americans than Africans (Wilcoxon signed-rank test; adj. $P < 0.0001$). Age groups were also significantly associated with the prevalence of these PGs (Kruskal-Wallis test; $P < 0.0001$; Fig. 4C). Children and adolescents (3 to 17 years old; $n = 12$) had the lowest prevalence (Wilcoxon signed-rank test; versus infants [0 to 2 years old; $n = 159$], adj. $P = 0.0054$; versus adults [18 to 64 years old; $n = 231$], adj. $P < 0.0001$; versus elderly [≥65 years old; $n = 38$], adj. $P = 0.0001$), followed by infants (Wilcoxon signed-rank test; versus adults, adj. $P < 0.0001$; versus elderly, adj. $P = 0.0025$). We did not observe a significant association between healthy ($n = 472$) and all diseased ($n = 247$) subjects (Wilcoxon rank sum test; $P = 0.13$). The type of disease did, however, have an effect (Kruskal-Wallis test; $P < 0.0001$; Fig. 4D). Remarkably, malnourished Malawian infants ($n = 12$) lacked all 39 highly prevalent PGs, and consequently, prevalence was significantly lower in this group than in all other disease groups besides the HIV patients (Wilcoxon signed-rank test; versus inflammatory bowel disease [IBD, $n = 48$], adj. $P = 0.0044$; versus type 1 diabetes [T1D, $n = 29$], adj. $P = 0.0117$; versus adenoma [$n = 28$], adj. $P = 0.0010$; versus *C. difficile* infection [CDI, $n = 35$], adj. $P = 0.0035$; versus colorectal carcinoma [CRC, $n = 28$], adj. $P = 0.0056$; versus hematopoetic stem cell transplantation [HSCT, $n = 44$], adj. $P = 0.0004$). Furthermore, patients undergoing HSCT ($n = 44$) had a higher prevalence than T1D ($n = 29$; adj. $P = 0.0077$), IBD ($n = 48$; adj. $P = 0.0003$), and HIV patients ($n = 22$; adj. $P = 0.0246$).

A crAss-like phage and a previously undescribed phage were highly prevalent in healthy Danish subjects and shared across the world. Among the 39 most prevalent PGs in the healthy DEVoC subset, two were widely distributed in SRA viromes (Fig. 4A). A 99-kb circular crAss-like phage (PG2) was the most prevalent in SRA viromes (20.6% prevalence; Fig. 5A). CrAssphages infect *Bacteroidales* sp. and are among the most abundant and globally distributed group of viruses in the human gut (51, 52). The second most prevalent PG in SRA viromes (PG6; 14.4% prevalence; Fig. 5B), was a 71-kb circular phage without clear homology to previously described phages. Despite the lack of clear homology, this PG possessed lots of viral (genetic) elements and was therefore named LoVEphage. The prevalence of these two phages was associated with age group and geographical location (test of equal proportions between multiple groups; $P < 0.001$ for both age group and geographical location for both PG2 and PG6). None of the two PGs were detected in healthy children/adolescents from other studies ($n = 12$), although they were detected in the DEVoC healthy children/adolescents. While they occurred in, respectively, 7.5% and 5% of the infants ($n = 159$), their prevalence significantly increased to 32.5% and 20.8% in adulthood ($n = 231$; test of equal proportions; PG2, adj. $P < 0.00001$; PG6, $P = 0.00015$) and to 42.1% and 28.9% in the elderly ($n = 38$; test of equal proportions; PG2, adj. $P < 0.0001$; PG6, $P = 0.00015$). The crAss-like phage was significantly more prevalent in healthy Europeans ($n = 164$; 34.8% prevalence) and healthy Asians ($n = 20$; 50% prevalence) than in healthy Americans ($n = 170$; 21.1% prevalence; test of equal proportions; adj. $P = 0.02445$ versus Europeans; adj. $P = 0.02445$ versus Asians) while less prevalent in healthy Africans ($n = 118$; 3.4% prevalence; test of equal proportions; adj. $P < 0.001$ versus all other continents). The LoVEphage was more prevalent in healthy Europeans ($n = 164$; 20.7% prevalence) and Americans ($n = 170$; 18.8% prevalence) than in healthy Africans ($n = 118$; 2.5% prevalence; test of equal proportions; versus Europeans, adj. $P = 0.00011$; versus Americans, adj. $P = 0.00035$). Asians ($n = 20$) had a prevalence of 15% for the LoVEphage (PG6). Additionally, we found that the prevalence of the crAss-

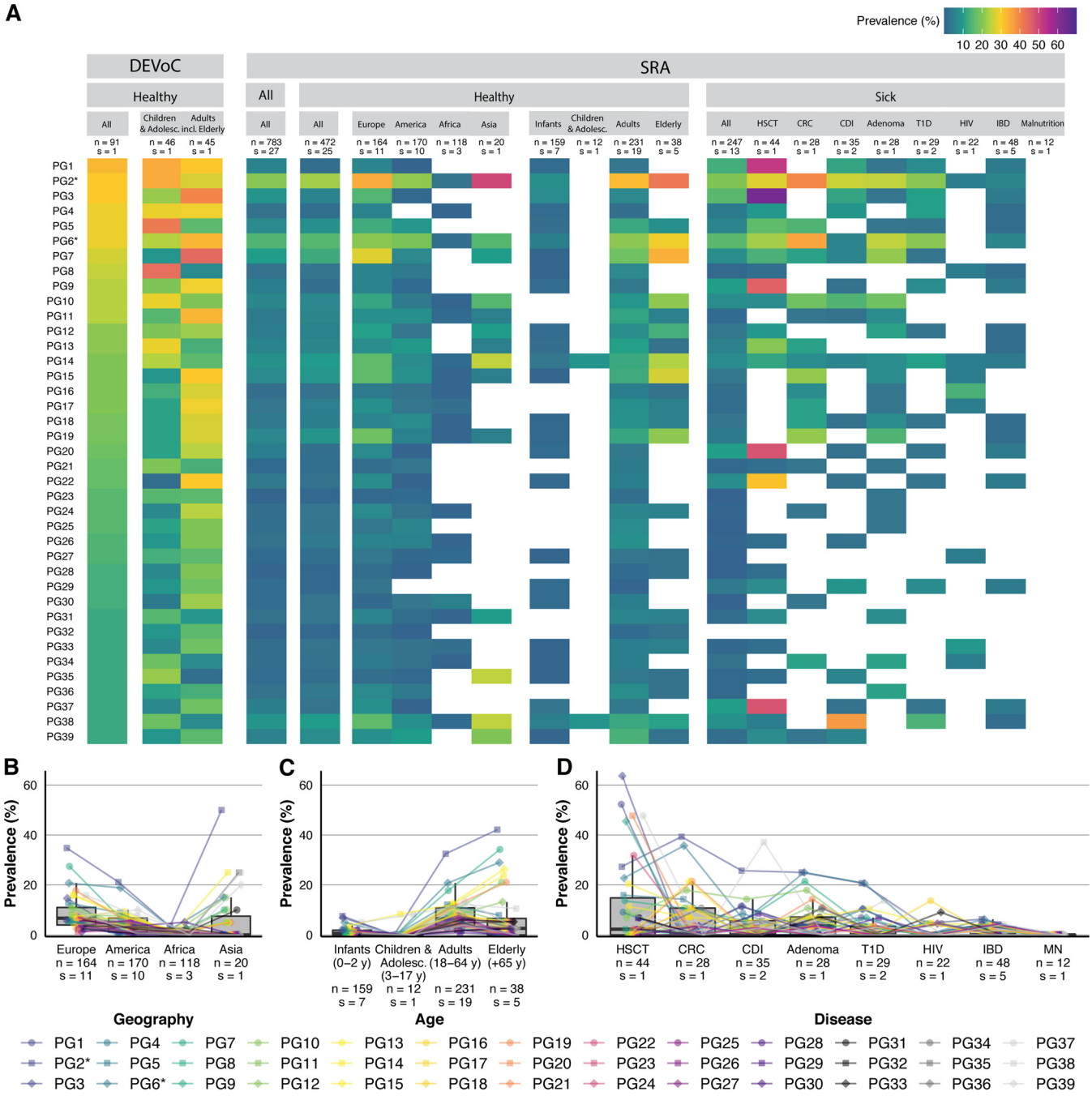

FIG 4 Worldwide prevalence of the 39 most prevalent healthy Danish PGs. (A) Heatmap of the prevalence of the top 39 most prevalent PGs (rows) in different subsets of subjects (columns) from this study's healthy Danish population (columns 1 to 3) and from other human gut virome studies (columns 4 to 19). The first four columns represent the prevalence in healthy Danish subjects, all healthy Danish children and adolescents (6 to 18 years old), and all healthy Danish adults, including the elderly (40 to 73 years old) from the DEVoC cohort. The fourth column shows the overall prevalence in all subjects from all the other studies combined. Columns 5 to 13 represent the prevalence in the healthy subjects (column 5), separated by continent (column 6 to 9) and age group (columns 10 to 13). Columns 14 to 22 represent the prevalence in disease subjects (column 14) in different diseases (column 15 to 22). The numbers of included subjects (*n*) and studies (*s*) are indicated on top of each column. PGs not detected in a specific subset are marked by a blank square. (B) Boxplots showing the prevalence of the top 39 PGs in different continents. (C) Boxplots showing the prevalence of the top 39 PGs in different age groups. (D) Boxplots showing the prevalence of the top 39 PGs in different diseases. Prevalences are indicated by different shapes and colors by PG and connected across boxplots in panels B, C, and D, and the numbers of subjects (*n*) and studies (*s*) included in each subgroup are indicated below each boxplot. PGs with an asterisk are further discussed in Fig. 5. HSCT, hematopoietic stem cell transplantation; CRC, colorectal cancer; CDI, *Clostridium difficile* infection; T1D, type 1 diabetes; HIV, human immunodeficiency virus; IBD, inflammatory bowel disease; MN, malnutrition.

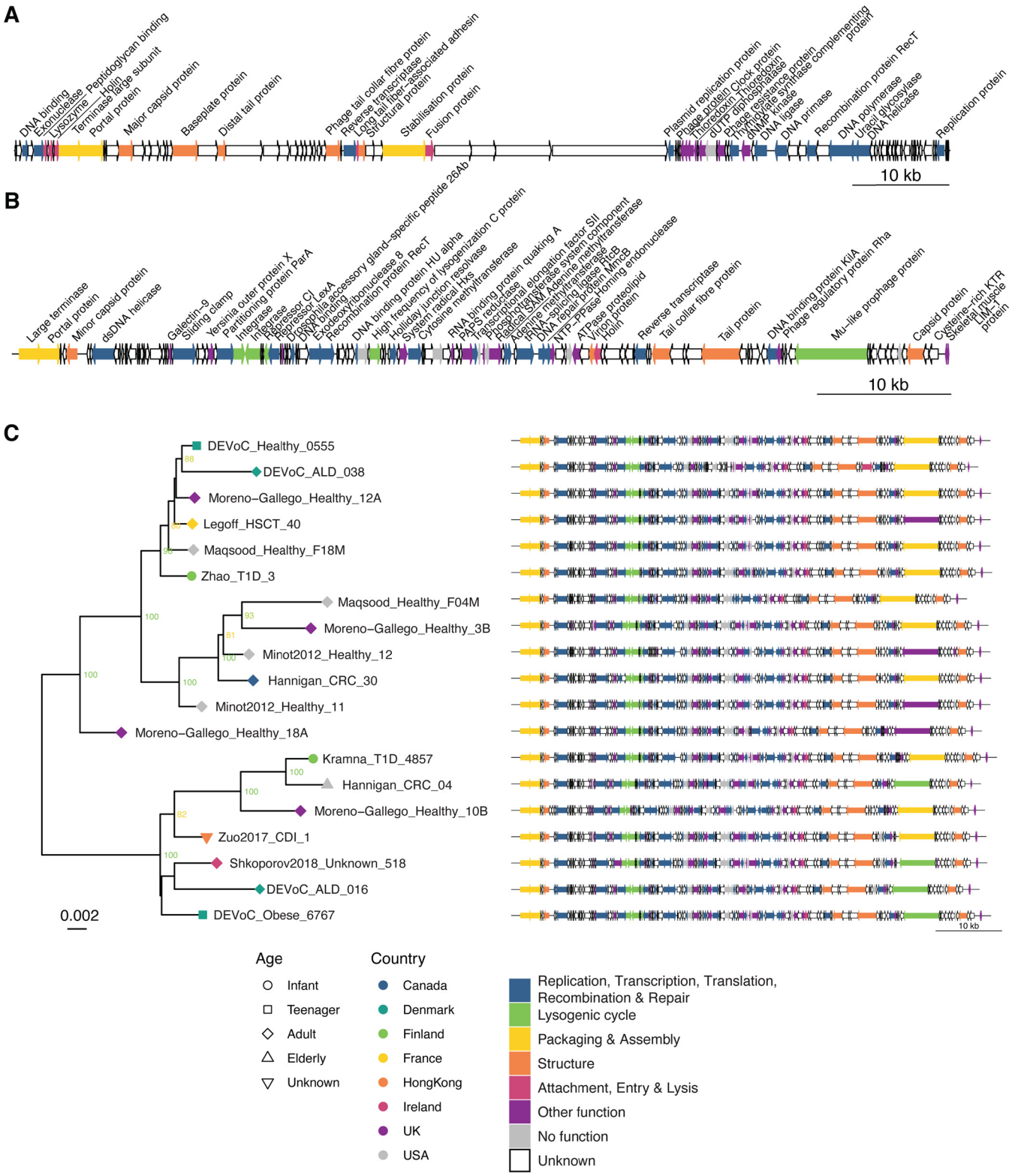

FIG 5 Genomic structure of the identified crAss-like and LoVEphage genomes. (A) The genomic structure of a crAss-like phage (PG2). (B) The genomic structure of the LoVEphage (PG6). (C) Maximum likelihood trees of concatenated protein alignments (n = 61) of 19 LoVEphage-like genomes recovered from Danish subjects and SRA viromes. Only bootstrap values higher than 70 are shown. Tip symbol and color are indicative of, respectively, the age group and country of origin of the individual from which the LoVEphage-like genome was assembled. Disease states are indicated in the tip labels. Full-genome structures of the 19 LoVEphage-like genomes are visualized next to the phylogenetic tree. ALD, alcohol liver disease; CRC, colorectal cancer; T1D, type 1 diabetes; HSCT, hematopoietic stem cell transplantation. All genomes in panels A, B, and C are represented linearly for clarity, although all have a circular genome. Arrows indicate ORFs, and annotations for known ORFs are given. Unknown proteins indicate ORFs with no hit to any of the databases. ORFs without function indicate proteins with hits to hypothetical proteins.

like phage was affected by disease (test of equal proportions between multiple groups; $P = 0.004$). Remarkably, its prevalence was significantly lower in IBD patients ($n = 48$; 6.3% prevalence) than in CRC patients ($n = 28$; 39.3% prevalence; test of equal proportions; adj. $P = 0.029$), while other diseases did not affect its presence.

The circular crAss-like phage (PG2) genome of 99 kb encoded 99 proteins, of which 32 (32.3%) were functionally annotated (Fig. 5A). PG2 was classified as the *AlphacrAssvirinae* subfamily, genus *I*, one of the most prevalent gut viruses in Western subjects independent of age. Typical of a crAssphage, PG2 was subdivided into two regions with opposite gene orientation, one region encoding proteins predicted to be involved in host interaction and phage structure and the other region encoding proteins predicted to be involved in DNA replication, recombination, and nucleotide metabolism. Downstream of the tail collar fiber protein we observed a reverse transcriptase—indicative of a diversity-generating retroelement previously described in crAssphages (57). No gene was annotated as RNA polymerase; however, we suspect that one of the large unknown genes may encode a divergent RNA polymerase subunit, as large unannotated proteins in crAss-like phages often contain an amino acid motif typical for RNA polymerases (58). PG2 had no tRNA genes, otherwise commonly found in genus *I*, *II*, and *IV AlphacrAssvirinae*.

The LoVEphage (PG6) had a circular 71-kb genome encoding 130 proteins, of which 45 (34.6%) were functionally annotated. Nine tRNA genes were identified in the LoVEphage, and the orientation of the genes was more random than that of the crAss-like phage. In this genome, we also observed a tail collar fiber protein, located upstream of a reverse transcriptase, similar to what we observe in the crAss-like phage (PG2). The predicted presence of two integrase proteins, a repressor protein, and a prophage protein suggest that the LoVEphage is a temperate phage. Furthermore, the genome was highly similar to *Bacteroides dorei* strain CL03T12C01 (GenBank accession number CP011531.1) (95.6% nucleotide identity and 96% coverage), indicating that a LoVEphage-like phage has occurred as prophage in this bacterial genome. Additionally, the *Bacteroides* genus was also predicted to be the host for the LoVEphage based on matches with CRISPR spacers.

To investigate the genetic diversity of the LoVEphage (PG6), 18 additional complete LoVEphage-like genomes were reconstructed from the DEVoC (3/18) and SRA (15/18) viromes. Each complete genome (67.9 to 72.4 kb) encoded between 122 and 131 genes, of which 61 conserved proteins were selected for phylogenetic analysis based on concatenated protein alignment. Two large phylogenetic clusters can be distinguished (Fig. 5C). The largest cluster contains 12 genomes mainly obtained from healthy adults, while the smaller cluster contains 7 genomes from subjects with variable ages and disease states. However, no distinct clustering based on geography, age, or health status was observed. All 19 genomes show remarkable conservation of synteny (Fig. 5C). The largest gene in these genomes has a conserved position but has one of three annotations. Four proteins, including the protein from the reference, are annotated as "mu-like prophage protein" (indicative of temperate phages and involved in tail assembly; green), while 11 proteins are annotated as "tail tape measure protein" (involved in tail assembly; yellow) and four proteins, as "reticulocyte binding protein rhoptry" (a protein involved in the entry of the malaria parasite in red blood cells; purple). Proteins from the latter two groups show >88% amino acid identity to the proteins annotated as "mu-like prophage protein," while they show >99% pairwise amino acid identity. Therefore, we assume that "reticulocyte binding protein rhoptry" is likely a misannotation in one of the databases used.

DISCUSSION

Human gut viruses represent a major pool of diverse and relatively underexplored microbes that, together with other gut microbiota, are believed to impact human health and disease (59). Currently, the number of studies exploring the human gut

viruses and cataloging their viral genomes and genes is expanding significantly, collectively advancing the virome field.

In this study, a human enteric virome catalog (the DEVoC), containing 12,986 viral scaffolds and encoding 190,029 genes, was generated from 254 fecal viromes from Danish children, adolescents, and adults. The majority of the DEVoC scaffolds originated from unclassified phages (Fig. 1A) without an assigned bacterial host (Fig. 1B), as described in other human gut virome databases (18). Even though the viral RefSeq version used during vConTACT2 clustering contained the most recently established phage families *Ackermannviridae*, *Herelleviridae*, *Chaseviridae*, *Dexlerviridae*, and *Demerecviridae*, none of the DEVoC scaffolds formed a viral cluster (VC) with these families. Although less stringent taxonomical classification approaches could increase the number of phage genomes with assigned taxonomy, a large fraction of phage scaffolds would remain unclassified nonetheless, hampering potential subsequent analyses at the family/genus level. Hence, further analyses were conducted at the individual scaffold level, thereby also avoiding having the results become outdated due to the constantly evolving phage taxonomy. DEVoC phages (*Caudovirales* and *Petitvirales* orders) and their bacterial hosts (*Firmicutes* and *Bacteroidetes* phyla) have all been commonly described in the human gut (18, 29, 60, 61).

Recent human gut virome studies concluded gut viromes of healthy Western adults to be highly individual (27, 29, 62). Individual-differentiating factors likely include geographical origin (63), age (18), diet (26, 62), and health status (11, 14, 15, 64, 65). Hence, our findings of large individuality of the gut viromes in healthy Danish adults is expected. The virome composition of healthy children (>3 years) and adolescents has not been studied before but is expected to show similar subject specificity since gut virome individuality has also been observed in infants (25). Due to this high virome individuality, it is not surprising that the majority of the identified viral genomes were not previously described, indicating that we are only scratching the surface of the viral diversity in the human gut microbiota worldwide (Fig. 1C). Notably, the very limited overlap of DEVoC with viral RefSeq indicates the clear underrepresentation of gut phages in the RefSeq database.

The DEVoC scaffolds encoded 190,029 genes, of which 53.7% could be annotated. However, exact estimates of functions were impeded, as multiple descriptions of the same function exist. To overcome this issue and the problem of unannotated proteins in general, proteins were clustered into OGs which were used as a proxy for function. As there is currently no database cataloguing the proteins encoded by gut viral genomes, the DEVoC encoded proteins could serve as a starting point to study the functional capacities of human gut viromes.

We further characterized the gut viromes in a subset of Danish healthy children, adolescents, and adults (n = 91) used to develop the DEVoC. The substantial number of previously undescribed DEVoC viral genomes is a clear indication of high individuality of human gut viromes and was further reflected by the low prevalence of most PGs (phage genomes predicted to be at least 50% complete) (Fig. 2A). It should be noted that in each healthy subject the majority of the viral reads belonged to only a few phage genomes (Fig. 2B). Despite this virome individuality, we identified 39 PGs present in more than 10 healthy subjects (>12% prevalence) with a maximum prevalence of 33% (30 subjects) (Table S2). This finding refutes the existence of a "core" virome (phages present in >50% of subjects) (66)—at least at genome level—in line with previous studies (18). OGs, on the other hand, were much more prevalent and could be detected in up to 97% of the healthy subjects (Table S3); most of these are involved in typical phage functions. However, similar to previous findings (67), the majority of the OGs remained specific to only one subject (Fig. 2C).

Gregory et al. (2020) reported an age-dependent virome diversity using publicly available data (18). They included studies produced with various wet-lab procedures and sequencing depths, as well as age groups with unequal age ranges and sample sizes (infants [<3 years], n = 27, versus children/adolescents [3 to 18 years], n = 11, versus adults [18 to 65 years], n = 93 versus elderly [>65 years], n = 20). Our study could

not confirm the former finding, as the PG richness and Shannon diversity did not differ across age groups (Fig. S3A and B). While our study has the advantage of consistently processed samples of different age groups (range, 6 to 73 years), we lacked data from infants, young children (<6 years old), and young adults (19 to 39 years old) to make associations with age as a continuous variable. OG richness was, similar to PG richness, not different between age groups (Fig. S8). Beta diversity at the PG and OG levels were associated with age group, although the biological importance of this effect is probably limited (Fig. 3A and E). Interestingly, at the level of individual OGs and PGs, 45 OGs and 3 PGs had different prevalences across age groups (Table 1 and Table S2). The presence of age-associated PGs may indicate that some "common" (prevalence between 20 and 50%) or even "core" (prevalence higher than 50%) phages (66) might exist in smaller, more homogeneous, populations, although core phages do not exist for the general healthy human population. Moreover, while age does not seem to affect overall diversity of the gut virome, age seems to affect the presence of certain viruses. The association of specific phages with age group might be linked to the gut microbiota with the human host, affecting human host metabolism and immune response.

We observed a clear decrease in the number and proportion of temperate PGs in our healthy adult population (Fig. 3B and C). This is accordant with the finding from L. Beller, W. Deboutte, S. Vieira-Silva, G. Falony, R. Yhossef Tito, L. Rymenans, C. Kwe Yinda, B. Vanmechelen, L. Van Espen, D. Jansen, C. Shi, M. Zeller, P. Maes, K. Faust, M. Van Ranst, J. Raes, and J. Matthijnssens (unpublished data) that demonstrates a decrease in the proportion of temperate phages across the first year of life in infants and suggests that this decrease continues during childhood into adulthood. It should, however, be noted that the identification of phage genomes with the potential to enter the lysogenic life cycle will be underestimated, as not all lysogeny-associated genes are currently known, and genes could also be encoded on the missing fragments of partial phage genomes.

Finally, we investigated the prevalence of the 39 most prevalent healthy Danish PGs in worldwide gut virome studies (Fig. 4A). Geography, age, and disease were all associated with the prevalence of the top 39 PGs (Fig. 4B to D). However, the conclusions should be interpreted cautiously, as subsets consisted of heterogeneous sample sizes. Some patient subsets showed a remarkably high prevalence (e.g., CRC or HSCT patients) or complete absence (malnourished Malawian infants) of the top 39 PGs. These striking differences are possibly confounded, as they often consist of a limited number of samples from only a single study, which can cause a severe bias regarding sample preparation, sequencing depth, or study setup. The top 39 PGs were, nonetheless, most commonly found in other European adults, which could be expected given the demographic similarity to our cohorts. The top 39 PGs were less commonly observed in infants, which are known to have a more distinct gut virome composition, and this age group was not included in the development of the DEVoC. Prevalences from healthy children and adolescents should be interpreted cautiously, as this SRA subset contained very few subjects due to the limited availability of these samples. For the same reason, the age-specific PGs could not be confirmed within the SRA viromes.

The group of crAss-like phages and their high prevalence and abundance across human gut viromes have been described extensively (51, 68–70). Although not as widespread as the crAss-like phages, the newly discovered LoVEphage seems to be rather common as well, with a prevalence of 28.6% in the healthy Danish subjects and 14.4% in the SRA viromes (Fig. 4A). However, the prevalence of crAss-like phages and the LoVEphages across SRA viromes is probably an underestimate due to the stringent criteria used and the low sequencing depths of some samples. Despite not having clear homology to previously described phages, numerous typical phage genes were identified in the LoVEphage (Fig. 5B). Phylogenetic analysis of 19 LoVEphage genomes did not reveal any clustering based on age, geography, or disease status (Fig. 5C), in contrast to the crAss-like phages, which seem to have some level of local geographic

clustering (71). However, such patterns may become apparent when more LoVEphage-like genomes are included/investigated. Future studies should experimentally determine the host range and morphology of the LoVEphage, as well as their broad genetic diversity in the general population, to uncover the potential associations of variants with disease. It could be worthwhile to also investigate whether this phage is specific to humans or is also found in nonhuman primates and other mammals as is the case for crAss-like phages (71, 72).

In conclusion, the human gut virome catalog DEVoC and its encoded genes generated from Danish children, adolescents, and adults assisted in the characterization of the healthy gut virome and will prove very helpful in investigating the role of the gut virome in human health and disease in the future. Furthermore, by investigating the presence of the top healthy Danish PGs in other human gut virome studies, we identified a previously undescribed phage, called LoVEphage, with a high worldwide prevalence.

MATERIALS AND METHODS

Subject recruitment and sample collection. The two Danish cohorts involved in this study were included as part of the MicrobLiver project. The pediatric cohort included 50 children and adolescents (6 to 18 years old) with a BMI above the 90th percentile, together with 50 age- and sex-matched healthy controls (73). The obese pediatric subjects were enrolled in an obesity treatment, and samples were included at baseline and the 1-year follow-up. The adult cohort (34 to 76 years old) included 52 patients with alcohol-related liver disease (ALD) and 52 sex, BMI, and age-matched healthy controls. They represent a selection of participants from a study aimed to develop noninvasive markers of early-stage alcohol-related liver disease. In total, 254 fecal samples were collected from 204 subjects.

Fecal samples were collected at home and kept at −20°C for 0 to 4 days, after which they were brought to the clinic (frozen) and stored at −80°C. For the adult cohort, samples were aliquoted at −120°C with the CryoXtract CXT350 device (CryoXtract Instruments) (74). Fecal samples from the pediatric cohort were aliquoted on ice, as the fecal sample sizes were much smaller. All fecal samples were kept at −80°C until use.

Sample preparation and sequencing. All 254 fecal samples were prepared for high-throughput virome sequencing using the NetoVIR protocol (75). In short, each fecal aliquot was homogenized in phosphate-buffered saline (PBS) (30 mass/volume percentage), centrifuged, filtered (0.8 μm), and subjected to nuclease treatment to enrich for viral-like particles. Next, the QIAamp viral RNA minikit (Qiagen) without carrier RNA was used to extract both RNA and DNA. The extracts were reverse transcribed and randomly amplified (17 cycles) using a modified WTA2 kit (Sigma-Aldrich). Sequencing libraries were prepared with the Nextera XT DNA library preparation kit (Illumina) and sequenced on the NextSeq 500 high-throughput Illumina platform (Nucleomics Core facility, KU Leuven, Belgium). Per sample, a median of 12.1 million (IQR, 6.9 million to 19.4 million) paired-end reads (2 × 150 bp) were generated.

Development of the Danish Enteric Virome Catalog. Raw reads were processed as described by L. Beller et al. (submitted for publication). In short, reads were quality controlled using Trimmomatic (v0.36) (76), after which reads mapping to the "contaminome" and human genome were removed using Bowtie2 (v2.3.4.1) in "very-sensitive" mode (77). Quality-filtered reads were *de novo* assembled, and all scaffolds longer than 1 kb were clustered at 95% identity over 80% coverage to remove redundancy in line with Roux et al. (78). Instead of calculating abundances by mapping the quality-filtered reads to the complete set of nonredundant scaffolds, reads were only mapped against to the representatives of the clusters containing a scaffold from that sample to avoid false-positive detection of closely related sequences. A scaffold was assumed to be present if 70% of its length was covered by reads. Scaffolds representing less than 0.00001% of the total amount of mapped reads across all samples were removed to reduce background noise. Viral scaffolds were selected to construct the Danish Enteric Virome Catalog (DEVoC). These viral scaffolds were identified by using a combination of homology to known viruses at the protein and/or nucleotide level, genome structure (kmer usage and gene content), the presence of virus-specific genes, and VirSorter category (35). The completeness of viral genomes was assessed with CheckV (v0.6.0) (53), and viral scaffolds were annotated using Cenote-Taker 2 (v2.0.1; parameters –prune_prophage False –enforce_start_codon False –hsuite_tool hhsearch) (49).

Taxonomic classification of viral scaffolds. Eukaryotic viruses were classified based on the lowest common ancestor determined using ktClassifyBLAST (v2.7.1) (79) on DIAMOND protein hits (v0.9.10.111, sensitive mode) (80) and BLASTn nucleotide hits (v2.7.1; E value, 1e-10) (81) (nonredundant [nr] and nucleotide [nt] databases downloaded from NCBI on 3 May 2019). As taxonomic classifications are unavailable for most phage scaffolds, vConTACT2 (v0.9.19) was used to create viral clusters (VCs) based on gene-sharing networks that represent genus/subfamily-level taxonomy (47). If phage scaffolds clustered with a RefSeq phage genome (v201), the taxonomy of the RefSeq phage genome(s) was assigned to the other members of the VC up to genus level.

Phage host prediction. CRISPR spacers were predicted using MinCED (v0.4.2) on the bacterial contigs assembled from shotgun metagenomic sequencing data of the same 254 fecal samples used to generate the DEVoC (unpublished data) (82). The predicted CRISPR spacers were submitted to a BLAST search against the phage subset of the DEVoC (-evalue, 1e-10, task, "blastn-short") (81). Phages required

at least two spacer matches with a maximum of one mismatch for reliable host assignment as the lowest common ancestor of the bacterial matches. Bacterial contigs were mapped against ProGenomes2 (83), and the lowest common ancestor was determined using ktClassifyBLAST (79).

Identification and annotation of DEVoC genes. Cenote-Taker 2 (v2.0.1) was used to predict and annotate open reading frames (ORFs) on the viral genomes of the DEVoC (49). Cenote-Taker 2 predicts ORFs using a combination of PHANOTATE (84) and Prodigal (85) in metagenomic mode and annotates the predicted ORFs using HMMER (86), RPSBLAST, (87) and HHSEARCH (88) searches against custom viral HMM, CDD, Pfam, and PDB databases. Next, amino acid sequences of the predicted ORFs were clustered into orthologous groups (OGs) using Proteinortho (v6.0.18) (55) with default settings and DIAMOND v0.9.32 (80). The annotation(s) given to at least 10% of the protein members of a specific OG was assigned to that OG of interest (manually, to overcome differences in spelling, capitalization, and abbreviations, as well as synonyms of the same protein, due to the use of different databases). This 10% threshold was introduced to avoid spurious annotations of OGs.

Gene prevalence. Due to the short nature of some DEVoC genes (the lower cutoff length for ORF identification was 20 amino acids) compared to the read length, we decided against a mapping approach to determine gene presence per sample. This is because short genes would be underrepresented, as a substantial fraction of the reads would only partially overlap the gene (and therefore not be assigned) compared to larger genes. Instead, we opted to count a viral gene as present when the corresponding viral genome was present (see L. Beller et al. [submitted for publication] for genome abundances). The presence of orthologous groups (OGs) was determined by grouping the prevalence information of all genes within the specific OG.

Comparison to existing databases. The DEVoC and its encoded genes were compared against existing (human gut) viral genome databases, including the human Gut Virome Database (GVD, version 2020/07/23) (18), IMG/VR2 (version July 2019) (56), and viral RefSeq (v201, version 10/07/2020). To determine overlap between the different genome databases (GVD, IMG/VR2, and viral RefSeq) and the DEVoC, they were clustered with ClusterGenomes (89) at 95% identity over 80% coverage (using nucmer v3.23) (90). Only IMG/VR2 sequences originating from human digestive tract samples (n = 78,016) were selected for comparison, and they were clustered in advance to remove redundancy (resulting in n = 18,383). Likewise, also viral RefSeq sequences larger than 1 kb (n = 12,681) were clustered in advance (resulting in n = 10,313). The GVD is already nonredundant (95% identity over 70% or 100% coverage depending on the type of virus) and consists of 33,242 viral genomes.

Prevalence of viral genomes across subjects worldwide. The prevalence of the genomes identified in the DEVoC in other subjects was assessed by mapping publicly available SRA data sets from 26 previously published human gut viral metagenomic studies (24–34, 65, 66, 91–103) and one unpublished study from our lab to the DEVoC using the Burrows-Wheeler Aligner (BWA) (v2.0pre2) (104). Reads were trimmed before mapping using Trimmomatic (v0.63; removing WTA2 and Nextera primers with the parameters 30:10:1:true and the following quality trimming parameters: HEADCROP:19 LEADING:15 TRAILING:15 SLIDINGWINDOW:4:20 MINLEN:50) (76).

An overview of all included studies is available in Table S5. Studies were selected based on a PubMed search in December 2019 searching for "human gut/fecal/enteric viromes." Studies using targeted sequencing or not using viral enrichment were excluded, as well as studies for which raw sequencing reads and/or metadata (subject ID, age group, health status, and geographic region of inclusion) were unavailable. The metadata were curated by screening the original article's subject recruitment section, supplementary tables, and/or the information association with the BioSample/SRA entry. As the gut virome is relatively stable over time (29), multiple samples from the same subject were pooled, except for patients undergoing fecal microbiota transplantation (FMT), for which only the baseline sample (before FMT) was included, if available (33, 34, 92). Two studies sequenced pools of multiple subjects, and for further analysis, these pools are regarded as one subject (91, 103). In total, 1,880 samples from 1,181 subjects (of which 490 were sequenced in 92 pools) were assessed. The subjects ranged in age from 0 (24, 25, 65, 103) to 99 (34) years old and originated from different geographical locations (13 countries across 4 continents). Besides healthy subjects, subjects suffering from inflammatory bowel disease (IBD) (30–32, 92, 101), *C. difficile* infection (CDI) (33, 34), diarrhea (91), malnutrition (105), HIV (100), type 1 diabetes (T1D) (65, 94), and colorectal cancer (CRC) (93) and subjects undergoing hematopoietic stem cell transplantation (HSCT) (95) were also included. A viral sequence was considered present in a subject if it was covered for more than 70% of its length by reads from the subject.

CrAss-like phage genome. To determine to which proposed genus/subfamily the prevalent crAss-like phage genome belongs (PG2), its genome is clustered with the 249 genomes from the crAss-like phage data set of Guerin et al. (70) using ClusterGenomes (89) at 95% identity over 80% coverage.

LoVEphage genome. To investigate the genetic diversity of the LoVEphage, we attempted to retrieve (near-)complete genomes from samples in which the LoVEphage was present. For the Danish samples, scaffolds longer than 50 kb clustering together with the LoVEphage (see above) were selected. All SRAs of subjects in which the LoVEphage was covered by reads for at least 70% of its length were quality-trimmed using Trimmomatic (76) (same settings as before) and assembled using metaSPAdes (v3.11.1; parameters -k 21,33,55,77) (106). A BLASTn search of the *de novo* assembled contigs was performed against the reference LoVEphage (E value, 1e-10) (81). All contigs larger than 50 kb, which covered the reference genome for at least 70% with a similarity of >70%, were selected. Incomplete genomes were completed using additional smaller scaffolds and/or individual quality-filtered reads (after mapping to the reference LoVEphage using BWA [104]), resulting in 18 additional complete LoVEphage genomes. Cenote-Taker 2 (v2.0.1) was used to predict and annotate ORFs on the complete LoVEphage-like genomes (same settings as before) (49). All 61 proteins that showed more than 70%

identity over 70% coverage and were present in all 19 LoVEphage-like genomes were aligned individually using MAFFT (v7.464; with automatic alignment strategy selection) (107). The individual protein alignments were concatenated and trimmed using trimAl (v1.4; parameter, -gappyout) (108). Maximum-likelihood trees were generated using RAxML (v8.2.12; parameters, -f a with 1,000 bootstraps and automatic amino acid substitution model selection) (109).

Ecological analyses, statistical analyses, and visualization. All ecological and statistical analyses, as well as visualizations, were done in R (http://www.R-project.org; v3.6.0). Viral reads were subsampled to a depth of 176,256 viral reads/sample, removing 21 samples with fewer viral reads, to allow unbiased characterization of the gut virome across the samples, as virome sequencing depth is equal. Random subsampling was done using the "rarefy_even_depth" function of the phyloseq package (v1.28.0) (110). Phageome analyses were conducted on phage relative abundances, while analyses of the eukaryotic viruses and at the protein level were performed on absence/presence profiles. Alpha-diversity indices (observed richness and Shannon's diversity) were calculated using the vegan package (v2.6-7) (111). Beta diversity was analyzed using the phyloseq package (v1.28.0) (110). Principal-coordinate analysis (PCoA) was used to visualize Jaccard distance. PERMANOVA was calculated using the "adonis" function from the vegan package. Medians of two groups were compared using the Wilcoxon test. Proportions of two groups were compared using the chi-squared test corrected for multiple testing using the Bonferroni method. Prevalences of the selected phage genomes across multiple sample subsets were compared using the Kruskal-Wallis test, after which *post hoc* Wilcoxon signed-rank tests (paired) were performed on each pair of groups corrected for multiple testing using the Holm method. Multiple proportions were compared using the test of equal proportions (prop.test in R), followed by *post hoc* tests of equal proportions on each pair of groups corrected for multiple testing using the Holm method. The genomic structure of individual phage genomes was visualized using the GenoPlotR package (v0.8.9) (112), the Venn diagrams using the VennDiagram package (v1.6.20) (113), and the phylogenetic trees using the ggtree package (v1.16.6) (114). Other figures were generated using the ggplot2 package (v3.3.2) (115).

Data availability. The virome sequencing reads supporting the conclusions of this article are in the Sequence Read Archive (SRA) with accession numbers PRJNA723467 (pediatric cohort) and PRJNA722819 (adult cohort). The DEVoC and its encoded genes, annotations, and normalized counts are available at https://doi.org/10.5281/zenodo.5173012. The LoVEphage genomes assembled from the above-mentioned BioProjects are available in GenBank under accession no. MW660583, MZ919976, MZ919981, and MZ919987. The scripts used to perform the analysis and make figures starting from the abundance table are available at https://github.com/Matthijnssenslab/ViromeCatalogue.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, EPS file, 1.7 MB.

FIG S2, PDF file, 0.4 MB.

FIG S3, EPS file, 1.5 MB.

FIG S4, PDF file, 0.2 MB.

TABLE S1, XLSX file, 0.01 MB.

TABLE S2, XLSX file, 0.01 MB.

TABLE S3, XLSX file, 0.01 MB.

TABLE S4, XLSX file, 0.01 MB.

TABLE S5, XLSX file, 0.01 MB.

ACKNOWLEDGMENTS

This research was supported by the Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark (grant number NNF18CC0034900), the Challenge Grant "MicrobLiver" (grant number NNF15OC0016692), and grant number NNF15OC0016544 from the Novo Nordisk Foundation; the Innovation Fund Denmark (TARGET: grant number 0603-00484B), the Region Zealand Health Scientific Research Foundation; and the European Union's Horizon 2020 research and innovation program (GALAXY: grant number 668031); the "Fonds Wetenschappelijk Onderzoek" (FWO, Research Foundation Flanders) (Lore Van Espen: 1S25720N, Leen Beller: 1S61618N).

The computational resources were provided by the Flemish Supercomputer Center (VSC) and funded by FWO and the Flemish Government Department of Economy, Science, and Innovation.

The study was conceptualized by E.G.B., L.V.E., A.K., P.B., T.H., M.A., and J.M. C.F.-B., C.E.F., S.J., M. Kjærgaard, M.T. H.B.J., T.N., J.-C.H., and A.K. handled the collection and management of fecal samples. L.V.E. and E.G.B. managed the project. E.G.B. and L.C.

carried out the viral DNA/RNA extraction, amplifications, and library preparation. L.V.E. performed the bioinformatic processing of the reads, generated the catalog, and performed the statistical analysis in close collaboration with E.G.B., L.B., W.D., M.A., and J.M. A.F. and M. Kuhn predicted CRISPR spacers in bacterial metagenomes. L.V.E. performed SRA screening with the assistance of D.S. and L.D.C., L.V.E., E.G.B., M.A., and J.M. drafted the manuscript. All authors critically revised and approved the final version for publication.

We declare that we have no competing interests. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

The study was approved by the Ethical Committees for the Region of Southern Denmark with reference numbers S-20120071, S-20160021, and S-20170087 (adult cohort) and by the Ethical Committees for Region Zeeland with reference number REG-043-2013 (pediatric cohort). All participants or their legal guardians gave consent to participate in this study.

We would like to thank the participants in The Danish Childhood Obesity Data and Biobank and the GALAXY study.

REFERENCES

1. Lynch SV, Pedersen O. 2016. The human intestinal microbiome in health and disease. N Engl J Med 375:2369–2379. https://doi.org/10.1056/NEJMra1600266.
2. Clemente JC, Ursell LK, Parfrey LW, Knight R. 2012. The impact of the gut microbiota on human health: an integrative view. Cell 148:1258–1270. https://doi.org/10.1016/j.cell.2012.01.035.
3. Gérard P. 2016. Gut microbiota and obesity. Cell Mol Life Sci 73:147–162. https://doi.org/10.1007/s00018-015-2061-5.
4. Gurung M, Li Z, You H, Rodrigues R, Jump DB, Morgun A, Shulzhenko N. 2020. Role of gut microbiota in type 2 diabetes pathophysiology. EBioMedicine 51:102590. https://doi.org/10.1016/j.ebiom.2019.11.051.
5. Zheng P, Li Z, Zhou Z. 2018. Gut microbiome in type 1 diabetes: a comprehensive review. Diabetes Metab Res Rev 34:e3043. https://doi.org/10.1002/dmrr.3043.
6. Nishida A, Inoue R, Inatomi O, Bamba S, Naito Y, Andoh A. 2018. Gut microbiota in the pathogenesis of inflammatory bowel disease. Clin J Gastroenterol 11:1–10. https://doi.org/10.1007/s12328-017-0813-5.
7. Cheng WY, Wu C-Y, Yu J. 2020. The role of gut microbiota in cancer treatment: friend or foe? Gut 69:1867–1876. https://doi.org/10.1136/gutjnl-2020-321153.
8. Ma Q, Xing C, Long W, Wang HY, Liu Q, Wang R-F. 2019. Impact of microbiota on central nervous system and neurological diseases: the gut-brain axis. J Neuroinflammation 16:53. https://doi.org/10.1186/s12974-019-1434-3.
9. García-López R, Pérez-Brocal V, Moya A. 2019. Beyond cells: the virome in the human holobiont. Microb Cell 6:373–396. https://doi.org/10.15698/mic2019.09.689.
10. Liang G, Conrad MA, Kelsen JR, Kessler LR, Breton J, Albenberg LG, Marakos S, Galgano A, Devas N, Erlichman J, Zhang H, Mattei L, Bittinger K, Baldassano RN, Bushman FD. 2020. Dynamics of the stool virome in very early- onset inflammatory bowel disease. J Crohns Colitis 14:1600–1610. https://doi.org/10.1093/ecco-jcc/jjaa094.
11. Clooney AG, Sutton TDS, Shkoporov AN, Holohan RK, Daly KM, O'Regan O, Ryan FJ, Draper LA, Plevy SE, Ross RP, Hill C. 2019. Whole-virome analysis sheds light on viral dark matter in inflammatory bowel disease. Cell Host Microbe 26:764–778.e5. https://doi.org/10.1016/j.chom.2019.10.009.
12. Chen Q, Ma X, Li C, Shen Y, Zhu W, Zhang Y, Guo X, Zhou J, Liu C. 2020. Enteric phageome alterations in patients with type 2 diabetes. Front Cell Infect Microbiol 10:575084.
13. Vehik K, Lynch KF, Wong MC, Tian X, Ross MC, Gibbs RA, Ajami NJ, Petrosino JF, Rewers M, Toppari J, Ziegler AG, She JX, Lernmark A, Akolkar B, Hagopian WA, Schatz DA, Krischer JP, Hyöty H, Lloyd RE, TEDDY Study Group. 2019. Prospective virome analyses in young children at increased genetic risk for type 1 diabetes. Nat Med 25:1865–1872. https://doi.org/10.1038/s41591-019-0667-0.
14. Lang S, Demir M, Martin A, Jiang L, Zhang X, Duan Y, Gao B, Wisplinghoff H, Kasper P, Roderburg C, Tacke F, Steffen H-M, Goeser T, Abraldes JG, Tu XM, Loomba R, Stärkel P, Pride D, Fouts DE, Schnabl B. 2020. Intestinal virome signature associated with severity of nonalcoholic fatty liver disease. Gastroenterology 159:1839–1852. https://doi.org/10.1053/j.gastro.2020.07.005.
15. Jiang L, Lang S, Duan Y, Zhang X, Gao B, Chopyk J, Schwanemann LK, Ventura-Cots M, Bataller R, Bosques-Padilla F, Verna EC, Abraldes JG, Brown RS, Vargas V, Altamirano J, Caballería J, Shawcross DL, Ho SB, Louvet A, Lucey MR, Mathurin P, Garcia-Tsao G, Kisseleva T, Brenner DA, Tu XM, Stärkel P, Pride D, Fouts DE, Schnabl B. 2020. Intestinal virome in patients with alcoholic hepatitis. Hepatology 72:2182–2196. https://doi.org/10.1002/hep.31459.
16. Nakatsu G, Zhou H, Wu WKK, Wong SH, Coker OO, Dai Z, Li X, Szeto C-H, Sugimura N, Lam TY-T, Yu AC-S, Wang X, Chen Z, Wong MC-S, Ng SC, Chan MTV, Chan PKS, Chan FKL, Sung JJ-Y, Yu J. 2018. Alterations in enteric virome are associated with colorectal cancer and survival outcomes. Gastroenterology 155:529–541.e5. https://doi.org/10.1053/j.gastro.2018.04.018.
17. Carding SR, Davis N, Hoyles L. 2017. The human intestinal virome in health and disease. Aliment Pharmacol Ther 46:800–815. https://doi.org/10.1111/apt.14280.
18. Gregory AC, Zablocki O, Zayed AA, Howell A, Bolduc B, Sullivan MB. 2020. The gut virome database reveals age-dependent patterns of virome diversity in the human gut. Cell Host Microbe 28:724–740.e8. https://doi.org/10.1016/j.chom.2020.08.003.
19. Ogilvie LA, Jones BV. 2015. The human gut virome: a multifaceted majority. Front Microbiol 6:918.
20. Tetz G, Tetz V. 2018. Bacteriophages as new human viral pathogens. Microorganisms 6:54. https://doi.org/10.3390/microorganisms6020054.
21. Duan Y, Llorente C, Lang S, Brandl K, Chu H, Jiang L, White RC, Clarke TH, Nguyen K, Torralba M, Shao Y, Liu J, Hernandez-Morales A, Lessor L, Rahman IR, Miyamoto Y, Ly M, Gao B, Sun W, Kiesel R, Hutmacher F, Lee S, Ventura-Cots M, Bosques-Padilla F, Verna EC, Abraldes JG, Brown RS, Vargas V, Altamirano J, Caballería J, Shawcross DL, Ho SB, Louvet A, Lucey MR, Mathurin P, Garcia-Tsao G, Bataller R, Tu XM, Eckmann L, Van Der Donk WA, Young R, Lawley TD, Stärkel P, Pride D, Fouts DE, Schnabl B. 2019. Bacteriophage targeting of gut bacterium attenuates alcoholic liver disease. Nature 575:505–511. https://doi.org/10.1038/s41586-019-1742-x.
22. Ott SJ, Waetzig GH, Rehman A, Moltzau-Anderson J, Bharti R, Grasis JA, Cassidy L, Tholey A, Fickenscher H, Seegert D, Rosenstiel P, Schreiber S. 2017. Efficacy of sterile fecal filtrate transfer for treating patients with Clostridium difficile infection. Gastroenterology 152:799–811.e7. https://doi.org/10.1053/j.gastro.2016.11.010.
23. Sutton TDS, Hill C. 2019. Gut bacteriophage: current understanding and challenges. Front Endocrinol (Lausanne) 10:784. https://doi.org/10.3389/fendo.2019.00784.
24. Lim ES, Zhou Y, Zhao G, Bauer IK, Droit L, Ndao IM, Warner BB, Tarr PI, Wang D, Holtz LR. 2015. Early life dynamics of the human gut virome

and bacterial microbiome in infants. Nat Med 21:1228–1234. https://doi.org/10.1038/nm.3950.
25. Maqsood R, Rodgers R, Rodriguez C, Handley SA, Ndao IM, Tarr PI, Warner BB, Lim ES, Holtz LR. 2019. Discordant transmission of bacteria and viruses from mothers to babies at birth. Microbiome 7:156. https://doi.org/10.1186/s40168-019-0766-7.
26. Minot S, Sinha R, Chen J, Li H, Keilbaugh S. a, Wu GD, Lewis JD, Bushman FD. 2011. The human gut virome: inter-individual variation and dynamic response to diet. Genome Res 21:1616–1625. https://doi.org/10.1101/gr.122705.111.
27. Moreno-Gallego JL, Chou S-P, Di Rienzi SC, Goodrich JK, Spector TD, Bell JT, Youngblut ND, Hewson I, Reyes A, Ley RE. 2019. Virome diversity correlates with intestinal microbiome diversity in adult monozygotic twins. Cell Host Microbe 25:261–272.e5. https://doi.org/10.1016/j.chom.2019.01.019.
28. Stockdale SR, Ryan FJ, McCann A, Dalmasso M, Hill C. 2018. Viral dark matter in the gut virome of elderly humans. https://doi.org/10.20944/preprints201807.0128.v1.
29. Shkoporov AN, Clooney AG, Sutton TDS, Ryan FJ, Daly KM, Nolan JA, McDonnell SA, Khokhlova EV, Draper LA, Forde A, Guerin E, Velayudhan V, Ross RP, Hill C. 2019. The human gut virome is highly diverse, stable, and individual specific. Cell Host Microbe 26:527–541.e5. https://doi.org/10.1016/j.chom.2019.09.009.
30. Fernandes MA, Verstraete SG, Phan TG, Deng X, Stekol E, LaMere B, Lynch SV, Heyman MB, Delwart E. 2019. Enteric virome and bacterial microbiota in children with ulcerative colitis and Crohn disease. J Pediatr Gastroenterol Nutr 68:30–36. https://doi.org/10.1097/MPG.0000000000002140.
31. Pérez-Brocal V, García-López R, Nos P, Beltrán B, Moret I, Moya A. 2015. Metagenomic analysis of Crohn's disease patients identifies changes in the virome and microbiome related to disease status and therapy, and detects potential interactions and biomarkers. Inflamm Bowel Dis 21:2515–2532. https://doi.org/10.1097/MIB.0000000000000549.
32. Shkoporov AN, Ryan FJ, Draper LA, Forde A, Stockdale SR, Daly KM, McDonnell SA, Nolan JA, Sutton TDS, Dalmasso M, Mccann A, Ross RP, Hill C. 2018. Reproducible protocols for metagenomic analysis of human faecal phageomes. Microbiome 6:68. https://doi.org/10.1186/s40168-018-0446-z.
33. Draper LA, Ryan FJ, Smith MK, Jalanka J, Mattila E, Arkkila PA, Ross RP, Satokari R, Hill C. 2018. Long-term colonisation with donor bacteriophages following successful faecal microbial transplantation. Microbiome 6:220. https://doi.org/10.1186/s40168-018-0598-x.
34. Zuo T, Wong SH, Lam K, Lui R, Cheung K, Tang W, Ching JYL, Chan PKS, Chan MCW, Wu JCY, Chan FKL, Yu J, Sung JJY, Ng SC. 2018. Bacteriophage transfer during faecal microbiota transplantation in Clostridium difficile infection is associated with treatment outcome. Gut 67:634–643. https://doi.org/10.1136/gutjnl-2017-313952.
35. Roux S, Enault F, Hurwitz BL, Sullivan MB. 2015. VirSorter: mining viral signal from microbial genomic data. PeerJ 3:e985. https://doi.org/10.7717/peerj.985.
36. Jurtz VI, Villarroel J, Lund O, Voldby Larsen M, Nielsen M. 2016. MetaPhinder: identifying bacteriophage sequences in metagenomic data sets. PLoS One 11:e0163111-14. https://doi.org/10.1371/journal.pone.0163111.
37. Ren J, Song K, Deng C, Ahlgren NA, Fuhrman JA, Li Y, Xie X, Poplin R, Sun F. 2020. Identifying viruses from metagenomic data using deep learning. Quant Biol 8:64–77. https://doi.org/10.1007/s40484-019-0187-4.
38. Fancello L, Raoult D, Desnues C. 2012. Computational tools for viral metagenomics and their application in clinical research. Virology 434:162–174. https://doi.org/10.1016/j.virol.2012.09.025.
39. Paez-Espino D, Chen I-MA, Palaniappan K, Ratner A, Chu K, Szeto E, Pillay M, Huang J, Markowitz VM, Nielsen T, Huntemann M, Reddy TBK, Pavlopoulos GA, Sullivan MB, Campbell BJ, Chen F, Mcmahon K, Hallam SJ, Denef V, Cavicchioli R, Caffrey SM, Streit WR, Webster J, Handley KM, Salekdeh GH, Tsesmetzis N, Setubal JC, Pope PB, Liu W-T, Rivers AR, Ivanova NN, Kyrpides NC. 2017. IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses. Nucleic Acids Res 45:D457–D465.
40. Goodacre N, Aljanahi A, Nandakumar S, Mikailov M, Khan AS. 2018. A reference viral database (RVDB) to enhance bioinformatics analysis of high-throughput sequencing for novel virus detection. mSphere 3:e00069-18. https://doi.org/10.1128/mSphereDirect.00069-18.
41. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, Rubin E, Ivanova NN, Kyrpides NC. 2016. Uncovering Earth's virome. Nature 536:425–430. https://doi.org/10.1038/nature19094.
42. Camarillo-Guerrero LF, Almeida A, Rangel-Pineros G, Finn RD, Lawley TD. 2021. Massive expansion of human gut bacteriophage diversity. Cell 184:1098–1109.e9. https://doi.org/10.1016/j.cell.2021.01.029.
43. Benler S, Yutin N, Antipov D, Raykov M, Shmakov S, Pevzner P, Koonin EV. 2020. Thousands of previously unknown phages discovered in whole-community human gut metagenomes. bioRxiv doi:https://doi.org/10.1101/2020.10.07.330464.
44. Sausset R, Petit MA, Gaboriau-Routhiau V, De Paepe M. 2020. New insights into intestinal phages. Mucosal Immunol 13:559. https://doi.org/10.1038/s41385-020-0260-3.
45. Shkoporov AN, Hill C. 2019. Bacteriophages of the human gut: the "known unknown" of the microbiome. Cell Host Microbe 25:195–209. https://doi.org/10.1016/j.chom.2019.01.017.
46. Roux S, Adriaenssens EM, Dutilh BE, Koonin EV, Kropinski AM, Krupovic M, Kuhn JH, Lavigne R, Brister JR, Varsani A, Amid C, Aziz RK, Bordenstein SR, Bork P, Breitbart M, Cochrane GR, Daly RA, Desnues C, Duhaime MB, Emerson JB, Enault F, Fuhrman JA, Hingamp P, Hugenholtz P, Hurwitz BL, Ivanova NN, Labonté JM, Lee KB, Malmstrom RR, Martinez-Garcia M, Mizrachi IK, Ogata H, Páez-Espino D, Petit MA, Putonti C, Rattei T, Reyes A, Rodriguez-Valera F, Rosario K, Schriml L, Schulz F, Steward GF, Sullivan MB, Sunagawa S, Suttle CA, Temperton B, Tringe SG, Thurber RV, Webster NS, Whiteson KL, Wilhelm SW, Wommack KE, Woyke T, Wrighton KC, et al. 2019. Minimum information about an uncultivated virus genome (MIUVIG). Nat Biotechnol 37:29–37. https://doi.org/10.1038/nbt.4306.
47. Bin Jang H, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, Brister JR, Kropinski AM, Krupovic M, Lavigne R, Turner D, Sullivan MB. 2019. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat Biotechnol 37:632–639. https://doi.org/10.1038/s41587-019-0100-8.
48. Aiewsakun P, Simmonds P. 2018. The genomic underpinnings of eukaryotic virus taxonomy: creating a sequence-based framework for family-level virus classification. Microbiome 6:38–24. https://doi.org/10.1186/s40168-018-0422-7.
49. Tisza MJ, Belford AK, Dominguez-Huerta G, Bolduc B, Buck CB. 2021. Cenote-Taker 2 democratizes virus discovery and sequence annotation. Virus Evol 7:veaa100. https://doi.org/10.1093/ve/veaa100.
50. Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GGZ, Boling L, Barr JJ, Speth DR, Seguritan V, Aziz RK, Felts B, Dinsdale EA, Mokili JL, Edwards RA. 2014. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nat Commun 5:4498–4411. https://doi.org/10.1038/ncomms5498.
51. Shkoporov AN, Khokhlova EV, Fitzgerald CB, Stockdale SR, Draper LA, Ross RP, Hill C. 2018. ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. Nat Commun 9:4781–4788. https://doi.org/10.1038/s41467-018-07225-7.
52. Yutin N, Makarova KS, Gussow AB, Krupovic M, Segall A, Edwards RA, Koonin EV. 2018. Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. Nat Microbiol 3:38–46. https://doi.org/10.1038/s41564-017-0053-y.
53. Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides NC. 2021. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol 39:578–578. https://doi.org/10.1038/s41587-020-00774-7.
54. Krishnamurthy SR, Wang D. 2018. Extensive conservation of prokaryotic ribosomal binding sites in known and novel picobirnaviruses. Virology 516:108–114. https://doi.org/10.1016/j.virol.2018.01.006.
55. Lechner M, Findeiß S, Steiner L, Marz M, Stadler PF, Prohaska SJ. 2011. Proteinortho: detection of (co-)orthologs in large-scale analysis. BMC Bioinformatics 12:124. https://doi.org/10.1186/1471-2105-12-124.
56. Paez-Espino D, Roux S, Chen I-MA, Palaniappan K, Ratner A, Chu K, Huntemann M, Reddy TBK, Pons JC, Llabrés M, Eloe-Fadrosh EA, Ivanova NN, Kyrpides NC. 2019. IMG/VR v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes. Nucleic Acids Res 47:D678–D686. https://doi.org/10.1093/nar/gky1127.
57. Morozova V, Fofanov M, Tikunova N, Babkin I, Morozov VV, Tikunov A. 2020. First crAss-like phage genome encoding the diversity-generating retroelement (DGR). Viruses 12:573. https://doi.org/10.3390/v12050573.
58. Koonin EV, Yutin N. 2020. The crAss-like phage group: how metagenomics reshaped the human virome. Trends Microbiol 28:349–359. https://doi.org/10.1016/j.tim.2020.01.010.
59. Virgin HW. 2014. The virome in mammalian physiology and disease. Cell 157:142–150. https://doi.org/10.1016/j.cell.2014.02.032.
60. Koliada A, Moseiko V, Romanenko M, Lushchak O, Kryzhanovska N, Guryanov V, Vaiserman A. 2021. Sex differences in the phylum-level

human gut microbiota composition. BMC Microbiol 21:131. https://doi.org/10.1186/s12866-021-02198-y.
61. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto J-M, Bertalan M, Borruel N, Casellas F, Fernandez L, Gautier L, Hansen T, Hattori M, Hayashi T, Kleerebezem M, Kurokawa K, Leclerc M, Levenez F, Manichanh C, Bjørn Nielsen H, Nielsen T, Pons N, Poulain J, Qin J, Sicheritz-Ponten T, Tims S, Torrents D, Ugarte E, Zoetendal EG, Wang J, Guarner F, Pedersen O, De Vos WM, Brunak S, Doré J, MetaHIT Consortium, Weissenbach J, Dusko Ehrlich S. 2011. Enterotypes of the human gut microbiome. Nature 473:174–180. doi:https://doi.org/10.1038/nature09944.
62. Garmaeva S, Gulyaeva A, Sinha T, Shkoporov AN, Clooney AG, Stockdale SR, Spreckels JE, Sutton TDS, Draper LA, Dutilh BE, Wijmenga C, Kurilshikov A, Fu J, Hill C, Zhernakova A. 2021. Stability of the human gut virome and effect of gluten-free diet. Cell Rep 35:109132. https://doi.org/10.1016/j.celrep.2021.109132.
63. Rampelli S, Turroni S, Schnorr SL, Soverini M, Quercia S, Barone M, Castagnetti A, Biagi E, Gallinella G, Brigidi P, Candela M. 2017. Characterization of the human DNA gut virome across populations with different subsistence strategies and geographical origin. Environ Microbiol 19:4728–4735. https://doi.org/10.1111/1462-2920.13938.
64. Norman JM, Handley SA, Baldridge MT, Droit L, Liu CY, Keller BC, Kambal A, Monaco CL, Zhao G, Fleshner P, Stappenbeck TS, McGovern DPB, Keshavarzian A, Mutlu EA, Sauk J, Gevers D, Xavier RJ, Wang D, Parkes M, Virgin HW. 2015. Disease-specific alterations in the enteric virome in inflammatory bowel disease. Cell 160:447–460. https://doi.org/10.1016/j.cell.2015.01.002.
65. Zhao G, Vatanen T, Droit L, Park A, Kostic AD, Poon TW, Vlamakis H, Siljander H, Härkönen T, Hämäläinen A-M, Peet A, Tillmann V, Ilonen J, Wang D, Knip M, Xavier RJ, Virgin HW. 2017. Intestinal virome changes precede autoimmunity in type I diabetes-susceptible children. Proc Natl Acad Sci U S A 114:E6166–E6175. https://doi.org/10.1073/pnas.1706359114.
66. Manrique P, Bolduc B, Walk ST, Van Der Oost J, De Vos WM, Young MJ. 2016. Healthy human gut phageome. Proc Natl Acad Sci U S A 113:10400–10405. https://doi.org/10.1073/pnas.1601060113.
67. Kristensen DM, Waller AS, Yamada T, Bork P, Mushegian AR, Koonin EV. 2013. Orthologous gene clusters and taxon signature genes for viruses of prokaryotes. J Bacteriol 195:941–950. https://doi.org/10.1128/JB.01801-12.
68. Honap TP, Sankaranarayanan K, Schnorr SL, Ozga AT, Warinner C, Lewis CM. 2020. Biogeographic study of human gut-associated crAssphage suggests impacts from industrialization and recent expansion. PLoS One 15:e0226930. https://doi.org/10.1371/journal.pone.0226930.
69. Siranosian BA, Tamburini FB, Sherlock G, Bhatt AS. 2020. Acquisition, transmission and strain diversity of human gut-colonizing crAss-like phages. Nat Commun 11:280. https://doi.org/10.1038/s41467-019-14103-3.
70. Guerin E, Shkoporov A, Stockdale SR, Clooney AG, Ryan FJ, Sutton TDS, Draper LA, Gonzalez-Tortuero E, Ross RP, Hill C. 2018. Biology and taxonomy of crAss-like bacteriophages, the most abundant virus in the human gut. Cell Host Microbe 24:653–664. https://doi.org/10.1016/j.chom.2018.10.002.
71. Edwards RA, Vega AA, Norman HM, Ohaeri M, Levi K, Dinsdale EA, Cinek O, Aziz RK, McNair K, Barr JJ, Bibby K, Brouns SJJ, Cazares A, de Jonge PA, Desnues C, Díaz Muñoz SL, Fineran PC, Kurilshikov A, Lavigne R, Mazankova K, McCarthy DT, Nobrega FL, Reyes Muñoz A, Tapia G, Trefault N, Tyakht AV, Vinuesa P, Wagemans J, Zhernakova A, Aarestrup FM, Ahmadov G, Alassaf A, Anton J, Asangba A, Billings EK, Cantu VA, Carlton JM, Cazares D, Cho G-S, Condeff T, Cortés P, Cranfield M, Cuevas DA, De la Iglesia R, Decewicz P, Doane MP, Dominy NJ, Dziewit L, Elwasila BM, Eren AM, et al. 2019. Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. Nat Microbiol 4:1727–1736. https://doi.org/10.1038/s41564-019-0494-6.
72. Li Y, Gordon E, Shean RC, Idle A, Deng X, Greninger AL, Delwart E. 2021. CrAssphage and its bacterial host in cat feces. Sci Rep 11:815. https://doi.org/10.1038/s41598-020-80076-9.
73. Patton HM, Sirlin C, Behling C, Middleton M, Schwimmer JB, Lavine JE. 2006. Pediatric nonalcoholic fatty liver disease: a critical appraisal of current data and implications for future research. J Pediatr Gastroenterol Nutr 43:413–427. https://doi.org/10.1097/01.mpg.0000239995.58388.56.
74. Mathieson W, Sanchez I, Mommaerts K, Frasquilho S, Betsou F. 2016. An independent evaluation of the CryoXtract instruments' CXT350 frozen sample aliquotter using tissue and fecal biospecimens. Biopreserv Biobank 14:2–8. https://doi.org/10.1089/bio.2015.0016.
75. Conceição-Neto N, Zeller M, Lefrère H, De Bruyn P, Beller L, Deboutte W, Kwe Yinda C, Lavigne R, Maes P, Van Ranst M, Heylen E, Matthijnssens J. 2015. Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. Sci Rep 5:16532. https://doi.org/10.1038/srep16532.
76. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170.
77. Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25. https://doi.org/10.1186/gb-2009-10-3-r25.
78. Roux S, Sullivan MB, Emerson JB, Eloe-Fadrosh EA, Sullivan MB. 2017. Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. PeerJ 5:e3817. https://doi.org/10.7717/peerj.3817.
79. Ondov BD, Bergman NH, Phillippy AM. 2011. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics 12:385. https://doi.org/10.1186/1471-210s5-12-385.
80. Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59–60. https://doi.org/10.1038/nmeth.3176.
81. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421.
82. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, Hugenholtz P. 2007. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics 8:209. https://doi.org/10.1186/1471-2105-8-209.
83. Mende DR, Letunic I, Maistrenko OM, Schmidt TSB, Milanese A, Paoli L, Hernández-Plaza A, Orakov AN, Forslund SK, Sunagawa S, Zeller G, Huerta-Cepas J, Coelho LP, Bork P. 2019. ProGenomes2: an improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes. Nucleic Acids Res 48:D621–D625. https://doi.org/10.1093/nar/gkz1002.
84. Mcnair K, Zhou C, Dinsdale EA, Souza B, Edwards RA. 2019. PHANOTATE: a novel approach to gene identification in phage genomes. Bioinformatics 35:4537–4542. https://doi.org/10.1093/bioinformatics/btz265.
85. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119. https://doi.org/10.1186/1471-2105-11-119.
86. Eddy SR. 2009. A new generation of homology search tools based on probabilistic inference. Genome Inform 23:205–211.
87. Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Geer LY, Bryant SH. 2017. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res 45:D200–D203. https://doi.org/10.1093/nar/gkw1129.
88. Meier A, Söding J. 2015. Automatic prediction of protein 3D structures by probabilistic multi-template homology modeling. PLoS Comput Biol 11:e1004343. https://doi.org/10.1371/journal.pcbi.1004343.
89. Roux S, Bolduc B. ClusterGenomes. https://bitbucket.org/MAVERICLab/stampede-clustergenomes/src.
90. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. Genome Biol 5:R12. https://doi.org/10.1186/gb-2004-5-2-r12.
91. Aiemjoy K, Altan E, Aragie S, Fry DM, Phan TG, Deng X, Chanyalew M, Tadesse Z, Callahan EK, Delwart E, Keenan JD. 2019. Viral species richness and composition in young children with loose or watery stool in Ethiopia. BMC Infect Dis 19:53–10. https://doi.org/10.1186/s12879-019-3674-3.
92. Chehoud C, Dryga A, Hwang Y, Nagy-Szakal D, Hollister EB, Luna RA, Versalovic J, Kellermayer R, Bushman FD. 2016. Transfer of viral communities between human individuals during fecal microbiota transplantation. mBio 7:e00322-16. https://doi.org/10.1128/mBio.00322-16.
93. Hannigan GD, Duhaime MB, Ruffin MT, Koumpouras CC, Schloss PD. 2018. Diagnostic potential and interactive dynamics of the colorectal cancer virome. mBio 9:1–13. https://doi.org/10.1128/mBio.02248-18.
94. Kramná L, Kolářová K, Oikarinen S, Pursiheimo J-P, Ilonen J, Simell O, Knip M, Veijola R, Hyöty H, Cinek O. 2015. Gut virome sequencing in children with early islet autoimmunity. Diabetes Care 38:930–933. https://doi.org/10.2337/dc14-2490.
95. Legoff J, Resche-Rigon M, Bouquet J, Robin M, Naccache SN, Mercier-Delarue S, Federman S, Samayoa E, Rousseau C, Piron P, Kapel N, Simon F, Socié G, Chiu CY. 2017. The eukaryotic gut virome in hematopoietic

stem cell transplantation: new clues in enteric graft-versus-host disease. Nat Med 23:1080–1085. https://doi.org/10.1038/nm.4380.
96. Ly M, Jones MB, Abeles SR, Santiago-Rodriguez TM, Gao J, Chan IC, Ghose C, Pride DT. 2016. Transmission of viruses via our microbiomes. Microbiome 4:64. https://doi.org/10.1186/s40168-016-0212-z.
97. McCann A, Ryan FJ, Stockdale SR, Dalmasso M, Blake T, Anthony Ryan C, Stanton C, Mills S, Ross PR, Hill C. 2018. Viromes of one year old infants reveal the impact of birth mode on microbiome diversity. PeerJ 6:e4694-13. https://doi.org/10.7717/peerj.4694.
98. Minot S, Grunberg S, Wu GD, Lewis JD, Bushman FD. 2012. Hypervariable loci in the human gut virome. Proc Natl Acad Sci U S A 109:3962–3966. https://doi.org/10.1073/pnas.1119061109.
99. Minot S, Bryson A, Chehoud C, Wu GD, Lewis JD, Bushman FD. 2013. Rapid evolution of the human gut virome. Proc Natl Acad Sci U S A 110:12450–12455. https://doi.org/10.1073/pnas.1300833110.
100. Monaco CL, Gootenberg DB, Zhao G, Handley SA, Ghebremichael MS, Lim ES, Lankowski A, Baldridge MT, Wilen CB, Flagg M, Norman JM, Keller BC, Luévano JM, Wang D, Boum Y, Martin JN, Hunt PW, Bangsberg DR, Siedner MJ, Kwon DS, Virgin HW. 2016. Altered virome and bacterial microbiome in human immunodeficiency virus-associated acquired immunodeficiency syndrome. Cell Host Microbe 19:311–322. https://doi.org/10.1016/j.chom.2016.02.011.
101. Pérez-Brocal V, García-López R, Vázquez-Castellanos JF, Nos P, Beltrán B, Latorre A, Moya A. 2013. Study of the viral and microbial communities associated with Crohn's disease: a metagenomic approach. Clin Transl Gastroenterol 4:e36. https://doi.org/10.1038/ctg.2013.9.
102. Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, Gordon JI. 2010. Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature 466:334–338. https://doi.org/10.1038/nature09199.
103. Yinda CK, Vanhulle E, Conceição-Neto N, Beller L, Deboutte W, Shi C, Ghogomu SM, Maes P, Van Ranst M, Matthijnssens J. 2019. Gut virome analysis of Cameroonians reveals high diversity of enteric viruses, including potential interspecies transmitted viruses. mSphere 4:e00585-18. https://doi.org/10.1128/mSphere.00585-18.
104. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760. https://doi.org/10.1093/bioinformatics/btp324.
105. Reyes A, Blanton LV, Cao S, Zhao G, Manary M, Trehan I, Smith MI, Wang D, Virgin HW, Rohwer F, Gordon JI. 2015. Gut DNA viromes of Malawian twins discordant for severe acute malnutrition. Proc Natl Acad Sci U S A 112:11941–11946. https://doi.org/10.1073/pnas.1514285112.
106. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021.
107. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010.
108. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973. https://doi.org/10.1093/bioinformatics/btp348.
109. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033.
110. McMurdie PJ, Holmes S. 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 8:e61217. https://doi.org/10.1371/journal.pone.0061217.
111. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, Mcglinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Henry M, Stevens H, Szoecs E, Maintainer HW. 2020. Package "vegan": Community Ecology Package Version 2.5-7.
112. Guy L, Kultima JR, Andersson SGE. 2010. GenoPlotR: comparative gene and genome visualization in R. Bioinformatics 26:2334–2335. https://doi.org/10.1093/bioinformatics/btq413.
113. Chen H, Boutros PC. 2011. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. BMC Bioinformatics 12:35. https://doi.org/10.1186/1471-2105-12-35.
114. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. 2017. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol 8:28–36. https://doi.org/10.1111/2041-210X.12628.
115. Wickham H. 2009. ggplot2: elegant graphics for data analysis. Springer-Verlag, New York, NY. https://ggplot2.tidyverse.org.