A previously undescribed highly prevalent phage identified in a Danish enteric virome catalogue

Lore Van Espen
KU Leuven Rega Institute for Medical Research.: Katholieke Universiteit Leuven Rega Institute for Medical Research
Emilie Glad Bak
University of Copenhagen: Kobenhavns Universitet
Leen Beller
KU Leuven Rega Institute for Medical Research.: Katholieke Universiteit Leuven Rega Institute for Medical Research
Lila Close
KU Leuven Rega Institute for Medical Research.: Katholieke Universiteit Leuven Rega Institute for Medical Research
Ward Deboutte
KU Leuven Rega Institute for Medical Research.: Katholieke Universiteit Leuven Rega Institute for Medical Research
Helene Bæk Juel
University of Copenhagen: Kobenhavns Universitet
Trine Nielsen
University of Copenhagen: Kobenhavns Universitet
Deniz Sinar
KU Leuven: Katholieke Universiteit Leuven
Lander De Coninck
KU Leuven Rega Institute for Medical Research.: Katholieke Universiteit Leuven Rega Institute for Medical Research
Christine Frithioff-Bøjsøe
University of Copenhagen: Kobenhavns Universitet
Cilius Esmann Fonvig
University of Copenhagen: Kobenhavns Universitet
Suganya Jacobsen
University of Southern Denmark: Syddansk Universitet
Maria Kjærgaard
University of Southern Denmark: Syddansk Universitet
Maja Thiele
University of Southern Denmark: Syddansk Universitet
Anthony Fullam

EMBL: European Molecular Biology Laboratory

Michael Kuhn

EMBL: European Molecular Biology Laboratory

Jens-Christian Holm

Holbaek Hospital: Holbaek Sygehus

Peer Bork

EMBL: European Molecular Biology Laboratory

Aleksander Krag

University of Southern Denmark: Syddansk Universitet

Torben Hansen

University of Copenhagen: Kobenhavns Universitet

Manimozhiyan Arumugam

University of Copenhagen: Kobenhavns Universitet

Jelle Matthijnssens (jelle.matthijnssens@kuleuven.be)

Katholieke Universiteit Leuven Rega Institute for Medical Research https://orcid.org/0000-0003-1188-9733

Research

Keywords: Human gut virome, virome catalogue, healthy gut viromes, phages

Posted Date: March 15th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-273865/v1

Abstract

Background: Gut viruses are important players in the complex human gut microbial ecosystem. Recently, the number of human gut virome studies is steadily increasing, however we are still only scratching the surface of the immense viral diversity as many wet lab and bio-informatics challenges remain. In this study, 254 virus-enriched faecal metagenomes from 204 Danish subjects were used to generate a Danish Enteric Virome Catalogue (DEVoC) of 12,986 non-redundant viral genome sequences encoding 190,029 viral genes, which formed 67,921 orthologous groups. The DEVoC was used to characterize the composition of the healthy DEVoC gut viromes from 46 children and adolescents (6-18 years old) and 45 adults (40 -73 years old).

Results: The majority of DEVoC viral sequences (67.3 %) and proteins (61.6 %) were not present in other (human gut) viral genome databases. Gut viromes of healthy Danish subjects mostly consisted of phages. While 39 phage genomes (PGs) were present in more than 10 healthy subjects, the degree of viral individuality was high. Among the 39 prevalent PGs, one was significantly more prevalent in the paediatric cohort, whereas two were more prevalent in adults. In 1,880 gut virome samples of 27 studies from across the world, the 39 prevalent PGs reveal 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, genus *I;* and a previously undescribed circular temperate phage (14.4% prevalence), named LoVEphage (because it encodes Lots of Viral Elements). A *de novo* assembly of selected public datasets generated an additional 18 circular LoVEphage-like genomes (67.9-72.4 kb). CRISPR spacer analysis suggested *Bacteroides* as a host genus for the LoVEphage, and a closely related prophage was identified in *Bacteroides dorei*, further confirming the host.

Conclusions: The DEVoC, the largest human gut virome catalogue generated from consistently processed faecal samples, facilitated analysis of healthy Danish human gut viromes and we foresee that it will benefit future analysis on the roles of gut viruses in human health and disease. The identification of a previously undescribed prevalent phage illustrates the usefulness of developing a virome catalogue.

Background

Gut microbiota, consisting of bacteria, archaea, viruses, fungi and other eukaryotic microorganisms, play a major role in human health and disease [1]. In the human gut, viruses and bacteria are highly abundant with an estimated 10^9-10^{12} viral-like particles and 10^{11}-10^{12} bacterial cells per gram of faeces [2]–[4]. While numerous studies have investigated the role of gut bacteria in health and disease [5], research on human gut viruses, collectively called gut virota, is still in its infancy. Only a minority of the human gut virota consists of eukaryotic viruses, most of them infecting human cells, fungal cells and unicellular eukaryotes residing in the gut, while others infecting plants or animals may only be transiting as part of the diet [6]. The vast majority of viruses in the human gut are bacteriophages (phages), which rely on a bacterial host to reproduce [7]. Since bacteria and viruses are the two most abundant components of the human gut microbiota, shedding more light on the virota and their collective genomes referred to as the virome, will pave the way to unravelling complex interactions within the gut microbiota and their effect on the human host.

Recent progress in high throughput sequencing technologies, viral enrichment procedures and development of downstream viral bioinformatic tools has facilitated human gut virome studies investigating the association between gut viromes and health [8]–[13] or disease [14]–[18], as well as fluctuations of specific gut viruses through life [7]. Yet, several significant challenges in studying human gut viromes remain. First, viruses lack a phylogenetic marker gene, thus prohibiting the application of profiling approaches (e.g. 16S rRNA gene amplicon sequencing) and forcing us to employ more complex shotgun metagenomic sequencing approaches. Second, identification of viruses from metagenomes is hampered by incomplete databases and therefore requires specialized viral identification tools e.g. VirSorter [19], MetaPhinder [20], and DeepVirFinder [21]. High viral genomic mutational rates further add to the incompleteness of databases, by creating immense viral genetic diversity [22]. Even though a few virome databases have recently emerged, only one contains gut-specific viral sequences (Gut Virome Database – GVD [7]), while others are either not gut-specific (IMG/VR [23], Reference Viral DataBase – RVDB [24] and Earth's virome [25]) or include only prokaryotic or eukaryotic viruses (gut phage database – GPD [26], a “circular” phage database [27] and RVDB [28]). However, despite these developments, a large fraction of human gut virome studies end up with a significant amount of “viral dark matter”. Third, taxonomic characterization of human gut viruses is virtually impossible due to the major incompleteness of taxonomically classified viruses [29], [30], despite ongoing efforts by the International Committee for the Taxonomy of Viruses (ICTV) to shift towards sequence-based classification of viruses [31]. Thus, viral taxonomic analysis is mostly performed on sequence level or on artificial taxonomic levels generated by gene-sharing tools e.g. vConTACT2 [32] or GRAViTy [33]. 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 [34], [35]. The technical difficulties of identifying and characterizing viruses are numerous, but nevertheless it is important to make progress in generating human gut virome catalogues 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 datasets [36]. This novel group of phages is now believed to be one of the most prevalent viruses of the human gut [37], [38]. Additionally, a recent study showed that human gut viromes are highly individual [13], emphasizing the importance of cataloguing viromes from diverse human populations.

In this study, we characterized 254 faecal viral metagenomes from Danish children and adolescents (6-18 years old) and adults (aged 40-73 years old), to develop the Danish Enteric Virome Catalogue (DEVoC). The DEVoC facilitated assessment of the diversity of the healthy Danish gut viromes. We identified phage genomes (PGs) associated with age in this healthy Danish subset, as well as novel PGs present in human gut metagenomes worldwide. In particular, a previously undescribed PG, 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.

Methods

Subject recruitment and sample collection

The two Danish cohorts involved in this study were included as part of the MicrobLiver project. The paediatric cohort included 50 children and adolescents (6 - 18 years old) with a BMI above the 90th percentile, together

with 50 age- and sex-matched healthy controls[20]. The obese paediatric subjects were enrolled in an obesity treatment, and samples were included at baseline and at the 1-year follow-up. The adult cohort (34 – 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 non-invasive markers of early-stage alcohol-related liver disease. In total, 254 faecal samples were collected from 204 subjects.

Faecal samples were collected at home and kept at -20 °C for 0-4 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 (CryoXtract Instruments)[21]. Faecal samples from the paediatric cohort were aliquoted on ice as the faecal sample sizes were much smaller. All faecal samples were kept at -80 °C until use.

Sample preparation and sequencing

All 254 faecal samples were prepared for high-throughput virome sequencing using the NetoVIR protocol [39]. In short, each faecal aliquot was homogenized in PBS (30 m/v%), centrifuged, filtered (0.8 µm), and subjected to nuclease treatment to enrich for viral-like particles. Next, the QIAamp® Viral RNA Mini kit (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 - 19.4 million) paired-end reads (2x150 bp) were generated.

Development of the Danish Enteric Virome Catalogue

Raw reads were processed as described by Beller *et al*. (in prep.) [40]. In short, reads were quality controlled, after which reads mapping to the "contaminome" and human genome were removed. 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. Abundances were determined by mapping the quality filtered reads to a subset of the cluster representatives containing only those of clusters containing at least one scaffold from that sample. 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. Viral sequences were selected to construct the Danish Enteric Virome Catalogue (DEVoC). These viral sequences were identified by using a combination of homology to known viruses at protein and/or nucleotide level, genome structure (kmer usage and gene content), the presence of viral-specific genes and VirSorter category [19]. Completeness of viral genomes was assessed by CheckV (v0.6.0) [41] and viral sequences were annotated using CenoteTaker2 (v2.0.1, parameters: –prune_prophage False –enforce_start_codon False –hsuite_tool hhsearch) [35].

Taxonomic classification of viral sequences

Eukaryotic viruses were classified based on the lowest common ancestor determined by ktClassifyBLAST (v2.7.1) [42] on DIAMOND protein hits (v0.9.10.111, sensitive mode) [43] and BLASTn nucleotide hits (v2.7.1, e-value 1e-10) [44] (nr and nt databases downloaded from NCBI on 2019/05/03). As taxonomic classifications are unavailable for most phage sequences, vConTACT2 (v0.9.19) was used to create viral clusters (VCs) based on gene-sharing networks that represent genus/subfamily-level taxonomy [32]. If phage sequences 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 faecal samples used to generate the DEVoC (unpublished data) [45]. The predicted CRISPR spacers were blasted against the phage subset of the DEVoC (-evalue 1e -10, task "blastn-short") [44]. Phages required at least two spacer matches with maximum one mismatch for reliable host assignment as the lowest common ancestor of the bacterial matches. Bacterial contigs were mapped against ProGenomes2 [46] and the lowest common ancestor was determined using ktClassifyBLAST [42].

Identification and annotation of DEVoC genes

CenoteTaker2 (v2.0.1) was used to predict and annotate open reading frames (ORFs) on the viral genomes of the DEVoC [35]. CenoteTaker2 predicts ORFs using a combination of Phanotate [47] and Prodigal [48] in metagenomic mode and annotates the predicted ORFs using HMMER [49], RPSBLAST [50] and HHSEARCH [51] 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)[52] with default settings and DIAMOND v0.9.32 [43]. The annotation(s) given to at least 10% of the OG members was assigned to the OG.

Gene prevalence

Due to the short nature of some DEVoC genes (lower cut-off 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 with the gene (and therefore not be assigned) compared to larger genes. Instead, we opted to count a viral gene present when the corresponding viral genome was present (see Beller *et al.* (in prep.)[40] 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 and protein databases including the human Gut Virome Database (GVD, version 2020/07/23) [7], IMG/VR2 (version July 2019) [53], viral RefSeq (v201, version 10/07/2020) and the Human Viral Protein Cluster (HVPC) database [54]. To determine overlap between the different genome databases (GVD, IMG/VR2 and viral RefSeq) and the DEVoC, they were clustered with ClusterGenomes [55] at 95% identity over 80% coverage (using nucmer v3.23) [56]. 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 non-redundant (95% identity over 70% or 100% coverage depending on the type of virus) and

consists of 33,242 viral genomes. The Human Viral Protein Cluster (HVPC) database is generated from 247 viral metagenomes from gut (n = 38), lung (n = 12), mouth (n = 54), oropharynx (n = 117), skin (n = 16), and urine (n = 10) samples. Since the HVPC database contains 926,819 amino acid sequences, the DEVoC genes were translated before they were clustered with the HVPC proteins using CD-HIT (v.4.8.1, parameters: -c 0.6 -G 0 -aS 0.8 -g 1 -n 4)[57].

Prevalence of viral genomes across subjects worldwide

Prevalence of the genomes identified in the DEVoC in other subjects was assessed by mapping publicly available SRA datasets from 26 previously published human gut viral metagenomic studies [8]–[18], [58]–[72] and one unpublished study from our lab to the DEVoC using BWA (v2.0pre2) [73]. Reads were trimmed before mapping using Trimmomatic (v0.63, removing WTA2 and Nextera primers with parameters 30:10:1:true and following quality trimming parameters: HEADCROP:19 LEADING:15 TRAILING:15 SLIDINGWINDOW:4:20 MINLEN:50) [74].

An overview of all included studies is available in **Supplementary Table 1**. As the gut virome is relatively stable over time [13], multiple samples from the same subject were pooled, except for patients undergoing faecal microbiota transplantation (FMT) for which only the baseline sample (before FMT) was included, if available [17], [18], [59]. Two studies sequenced pools of multiple subjects and for further analysis, these pools are regarded as one subject [58], [71]. 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 [8], [9], [71], [72] to 99 [18] years old and originated from different geographical locations (thirteen countries across four continents). Besides healthy subjects, also subjects suffering from inflammatory bowel disease (IBD) [14]–[16], [59], [69], *C. difficile* infection (CDI) [17], [18], diarrhoea [58], malnutrition [75], HIV [68], type 1 diabetes (T1D) [61], [72], colorectal cancer (CRC) [60] and subjects undergoing hematopoietic stem cell transplantation (HSCT) [62] were 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

The prevalent crAss-like phage genome as clustered with the 249 genomes from the crAss-like phage dataset of Guerin *et al.* [76] using ClusterGenomes [77] at 95% identity over 80% coverage to determine to which proposed genus/subfamily the genome belongs.

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 [78] (same settings as before) and assembled using metaSPAdes (v3.11.1, parameters: -k 21,33,55,77) [79]. A BLASTn search of the *de novo* assembled contigs was performed against the reference LoVEphage (e-value: 1e-10) [44]. All contigs larger than 50 kb, which covered the reference genome for at least 70% with a similarity > 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 [80]), resulting in 18 additional complete LoVEphage genomes. CenoteTaker2 (v2.0.1) was used to predict and annotate ORFs on the complete LoVEphage-like genomes (same settings as before) [35]. 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) [81]. The individual protein alignments were concatenated and trimmed using trimAl (v1.4, parameters: -gappyout) [82]. Maximum-likelihood trees were generated using RAxML (v8.2.12, parameters: -f a with 1000 bootstraps and automatic amino acid substitution model selection) [83].

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 less 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) [84]. Phageome analyses were conducted on phage relative abundances, while analyses of the eukaryotic viruses and at 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) [85]. Beta-diversity was analyzed using the phyloseq package (v1.28.0) [84]. 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 chi^2-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) [86], the Venn diagrams using the VennDiagram package (v1.6.20) [87] and the phylogenetic trees using the ggtree package (v1.16.6) [88]. Other figures were generated using the ggplot2 package (v3.3.2) [89].

Results

A catalogue of 12,986 non-redundant viral sequences derived from Danish faecal viromes encoding 190,029 proteins.

The Danish Enteric Virome Catalogue (DEVoC) was constructed based on 254 Danish faecal viromes (3.86 billion raw reads). The viral sequences constituting the DEVoC ranged in size from 1 kb to 191 kb (N50: 16 kb; L50: 1,463 scaffolds) and while CheckV [41] estimated that only 1,867 viral sequences (14.4%) were more than 50% complete, these sequences represented 87.4% of the total amount of viral reads (**Suppl. Fig. 1**).

Phages represented the vast majority of DEVoC sequences (n = 12,771; 98.3%; **Fig. 1A**), representing 99.2% of the viral reads. The phage sequences were clustered using vConTACT2 to generate viral clusters (VCs) as a

proxy for viral subfamilies or genera. vConTACT2 formed 1,488 VCs containing two or more members covering 5,222 phage sequences (41% of the DEVoC phage sequences) representing 73% of the total phage reads (**Suppl. Fig. 2**). Merely 176 phage sequences (1.4%) could be taxonomically classified based on vConTACT2 clustering with RefSeq genomes (30 VCs): 3 crAss-like genomes (1 VC), 19 *Microviridae* genomes (*Petitvirales* order, 3 VCs) and 154 *Caudovirales* genomes (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 sequences (7.5%), and of these, 54% could be predicted up to bacterial genus level. At 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 sequences represented viruses infecting eukaryotes (n = 215; 1.7%). Most eukaryotic viral sequences (65.6%) belonged to the *Picobirnaviridae* family (which is subject to interpretation, as increasing evidence suggests that viruses belonging to this family are phages [90]). The remaining eukaryotic viral genomes belonged to plant-infecting viral families (*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 potentially infecting mammals (*Anelloviridae* (0.9%), *Caliciviridae* (1.4%), *Circoviridae* (5.1%), *Genomoviridae* (2.3%), *Parvoviridae* (0.5%), *Picornaviridae* (2.3%) and *Smacoviridae* (0.5%)) **(Suppl. Fig. 3**).

To understand the functional potential of the viruses in the DEVoC, we predicted viral genes and annotated them using CenoteTaker2 [35], which resulted in 190,029 viral genes (DEVoC genes). These genes ranged in size from 0.06 to 18.2 kb (median: 0.34 kb, IQR: 0.19 – 0.62 kb), and 91.3% were complete. About half of DEVoC genes (n = 102,018; 53.7%) were functionally annotated, with the most common annotations being major capsid protein (> 1.26 %), portal protein (> 1.20 %), large terminase (> 1.16 %), integrase (> 0.91 %) and minor capsid protein (> 0.89 %) – all typical phage functions. The predicted amino acid sequences (DEVoC proteins) were clustered using Proteinortho [52] to form orthologous groups (OGs). This resulted in 18,473 OGs containing up to 360 members (median: 3 members; IQR: 2 – 6 members) covering 140,581 DEVoC proteins (74%), while the remaining 49,448 proteins (26%) remained singletons (regarded as OGs with one member from now on).

Majority of the DEVoC sequences and proteins are previously undescribed.

We compared DEVoC sequences and proteins to existing genome and gene databases to assess their novelty. Viral sequences from the NCBI RefSeq v201 database (ViralRefSeq; n = 10,313), the human Gut Virome Database [7] (GVD; n = 33,242) and the human gastro-intestinal tract subset of the IMG/VR2 database [53] (n = 18,383) were clustered with the DEVoC sequences at 95% identity over 80% coverage (see Methods for details). 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 sequences), followed by IMG/VR2 (n = 1,610 containing 3,096 DEVoC sequences). Only 857 clusters (containing 1,960 DEVoC sequences) 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 sequences) of which 62 were phage clusters and 23 were eukaryotic viral clusters. This limited overlap can be attributed to underrepresentation of phages in ViralRefseq (3,672 phage genomes vs. 9,476 eukaryotic virus genomes). In total, all four databases shared ten viral clusters, including an *uncultured crAssphage*, and members of *Siphoviridae* (*Ceduavirus*, *Limdunavirus*, *Oengusvirus*, *Skunavirus* and *Unaquatrovirus* genera) and *Myoviridae* (*Brigitvirus*, *Lagaffevirus*, *Peduovirus* and *Toutatisvirus* genera) families (**Suppl. Table 2**).

We further evaluated the uniqueness of DEVoC proteins by clustering them with the HVPC database [54], which contains amino acid sequences from 247 human viral metagenomes from different body sites, using CD-HIT at 60% amino acid identity and 80% coverage [57]. Mimicking the comparisons of the viral genome databases, the majority of the DEVoC protein clusters were unique (n = 83,104 containing 117,147 DEVoC proteins; 76.5% of all DEVoC clusters), while 25,536 clusters were shared with the HVPC database (containing 72,882 DEVoC proteins; 23.5% of all DEVoC clusters; **Fig. 1D**). The limited number of shared proteins between both databases supports the idea of highly personalized gut viromes, although the non-gut specificity of the HVPC database and the potential presence of non-viral proteins in the HVPC is likely to contribute to the limited overlap.

Healthy Danish gut viromes are highly individual

The remaining analyses solely included gut viromes from 91 healthy Danish subjects, including 46 children and adolescent (6 - 18 years old) from the paediatric cohort and 45 adults (40 – 73 years old). Samples that we did not analyse belong to obese children and adolescents, and ALD patients, which are all part of a larger ongoing study. Characterization of the 91 healthy Danish gut viromes revealed they were dominated by phages (relative abundance vs. all viral reads; median: < 99.9%; IQR: 99.8% – 100%; range: 79.4% – 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 sequences that represented more than 50% of a genome as determined by CheckV [41] (hereafter referred to as Phage Genomes; PGs). This allows us to limit the analysis to maximum one fragment for any given genome. Within the 91 healthy Danish gut viromes, 7,153 phage sequences (56% of the DEVoC phage sequences) 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% - 94.6%; range: 0.83% - 99.6%). The sample in which PGs accounted for 0.83% of phage reads, was dominated by one phage sequence with undetermined completeness (> 99% of viral reads).

The most prevalent PG was a partial *Skunavirus* genome detected in 33% of the subjects. Only a limited number of PGs (n = 39; 3.4%) was found across more than 10 subjects (**Suppl. Table 3**). These included six *Skunavirus*es, two *Eponavirus*es, and one *Limdunavirus, Unaquatrovirus* and crAss-like phage each, while the remaining 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 (vs. all PGs; median: 18.7%; IQR: 14.0% - 24.7%; range: < 0.1% – 40.0%) and their relative abundance (vs. all phage reads; median: 13.5%; IQR: 7.3% – 25.3%; range: < 0.1% – 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% - 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% - 89.0%; range: 0.83% – 99.5%; **Fig. 2B**).

Few eukaryotic viral species were detected in the gut viromes of healthy subjects (n = 33) of which a large fraction (n = 12) were plant viruses and therefore presumably not even stable members of the gut virome, but rather transient passengers. The median eukaryotic viral species richness was barely 1 (IQR: 0-3; range: 0 – 9; **Supp. Fig. 4A**) and most eukaryotic viruses were present in only one or two healthy subjects (**Supp. Fig. 4B**), suggesting eukaryotic viruses are highly individual.

At protein level, all healthy subjects combined harboured 46,620 viral OGs of one or more members. The majority of OGs was 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 – 2,904). A median of 7.9% of the OGs within each subject was unique to that subject (IQR: 5.6% - 10.5%; range: 0.5 – 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 (**Suppl. Table 4**). The five most prevalent OGs (prevalence > 93%), included a recombination protein, a nuclease, a reverse transcriptase, a terminase large subunit and a dUTPase.

Several phage genomes and viral functions associated with age

We investigated if PG diversity differed between the healthy gut phageomes of the paediatric (n = 46) and the adult cohort (n = 45). PG alpha-diversity was not affected by age group (Wilcoxon-test; observed richness: p = 0.89; Shannon's diversity: p = 0.83; **Suppl. Fig. 5**) and although age group was significantly associated with PG beta-diversity measured using Jaccard distance (PERMANOVA; 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 inter-individual diversity in gut phageomes.

To analyse if the occurrence of individual PGs was associated with age group, we compared prevalences between the paediatric and the adult group. Among the 39 most prevalent PGs (present in more than 10 subjects; >12% prevalence), ranging in size from 5.1 to 99.1 kb, PG8 was more common in children and adolescents, while PG7 and PG22 were more prevalent in adults (Chi^2-test; adj. p < 0.05; **Suppl. Table 3**).

The genomic structures of all three age-associated PGs are visualized in **Suppl. Fig. 6**. PG8 encoded proteins involved in the activation or suppression of the lysogenic cycle (including anti-repressor, transposase and site-specific recombinase XerD) indicating that this phage has a temperate lifestyle and can thus exist as a prophage. Temperate phages 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 **Suppl. Table 5**) 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 – 15; range: 0 - 27), representing roughly one-third of a subject's PGs (median: 31.6%, IQR: 21.5% - 39.5%; range: 0 – 66.7%) and accounting for a median of 19.3% of the PG reads (IQR: 8.8% – 42.6%; range: 0% - 95.3%). Among the temperate PGs, the alpha-diversity measures observed (absolute) richness, proportional (vs. all PGs) richness and Shannon diversity was higher in children/adolescents than in adults (Wilcoxon test; p = 0.034, p = 0.0016 and p = 0.018, respectively; **Fig. 3B, Fig. 3C, Fig. 3D**), while we did not observe a difference in relative abundance of temperate PG (vs. all phage reads; p = 0.21; **Suppl**. **Fig. 7**).

We further assessed the association between age group and viral functions represented by OGs. Similar to the previous analysis, observed richness of viral OGs did not differ between age groups (Wilcoxon test; p = 0.11; **Suppl. Fig. 8**). However, age group explained 3% of the beta-diversity between subjects (Jaccard dissimilarity; PERMANOVA; p = 0.001; **Fig. 3E**). Analysis on all OGs of two or more members and present in more than 10 healthy subjects (n = 3,627) identified 29 OGs with a higher prevalence in one of both age groups (Chi^2-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 prevalent in children and adolescents.

Highly prevalent DEVoC phage genomes are detected worldwide

We assessed whether the 39 highly prevalent PGs in the healthy Danish subset (**Suppl. Table 3**) could be recovered worldwide, across age groups and diseases. For this purpose, we obtained 1,880 faecal viral metagenomes from NCBI SRA (denoted SRA viromes hereafter), deriving from 1,181 subjects living in thirteen different countries across Europe, America, Africa and Asia. The subjects ranged from infants (0 – 2 years old) to elderly (> 65 years old) and their health status was extremely diverse (see **Suppl. Table 1** 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 geographical origin (continent) 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 remaining continents (Wilcoxon signed-rank test; vs. America (n = 170): adj. p < 0.0001; vs. Africa (n = 188): adj. p < 0.0001; vs. 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 – 17 years old; n = 12) had the lowest prevalence (Wilcoxon signed-rank test; vs. infants (0 – 2 years old; n = 159): adj. p = 0.0054; vs. adults (18 – 64 years old; n = 231): adj. p < 0.0001; vs. elderly (≥ 65 years old; n = 38): adj. p = 0.0001), followed by infants (Wilcoxon signed-rank test; vs. adults: adj. p < 0.0001, vs. 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 compared to all other disease groups besides the HIV patients (Wilcoxon signed-rank test; vs. IBD (n = 48): adj. p = 0.0044; vs. T1D (n = 29): adj. p = 0.0117; vs. adenoma (n = 28): adj. p = 0.0010; vs. CDI (n = 35): adj. p = 0.0035; vs. CRC (n = 28): adj. p = 0.0056; vs. HSCT (n = 44): adj. p = 0.0004). Furthermore, patients undergoing HSCT (n = 44) had a higher prevalence of the 39 most prevalent PGs compared to T1D (n = 29; adj. p = 0.0077), IBD (n = 48; adj. p = 0.0003) and HIV patients (n = 22; adj. p = 0.0246).

A crAssphage and a previously undescribed phage were highly prevalent in healthy Danish subjects and shared across the world

To investigate globally common phages, we looked for PGs that were highly prevalent in the 1,880 SRA viromes. Among the 39 most prevalent PGs in the healthy DEVoC subset, a 99 kb circular crAss-like phage (PG2) was the most prevalent in SRA viromes (20.6%; **Fig. 5A**). CrAssphages infect *Bacteroidales* sp. and are among the most abundant and globally distributed group of viruses in the human gut [37], [38]. The second most prevalent PG in SRA viromes (PG6; with a prevalence of 14.4%; **Fig. 5B**), was a 71 kb circular phage

without clear homology to previously described phages. Despite lack of clear homology, this PG possessed **L**ots **o**f **V**iral (genetic) **E**lements and was therefore named LoVEphage. The prevalence of these two phages were 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 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; prevalence of 34.8%) and healthy Asians (n = 20; prevalence of 50%) compared to healthy Americans (n = 170; prevalence of 21.1%) (test of equal proportions; adj. $p = 0.02445$ vs. Europeans; adj. $p = 0.02445$ vs. Asians), while less prevalent in healthy Africans (n = 118; prevalence of 3.4%) (test of equal proportions; adj. $p < 0.001$ vs. all other continents). The LoVEphage was more prevalent in healthy Europeans (n = 164; prevalence of 20.7%) and Americans (n = 170; prevalence of 18.8%) compared to healthy Africans (n = 118; prevalence of 2.5%) (test of equal proportions; vs. Europeans: adj. $p = 0.00011$; vs. Americans: adj. $p = 0.00035$). Asians (n = 20) had a prevalence of 15% for the LoVEphage (PG6). In addition, we found that the prevalence of the crAss-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; prevalence of 6.3%) than in CRC patients (n = 28; prevalence of 39.3%) (test of equal proportions; adj. $p = 0.029$), while other diseases did not affect its presence.

The crAss-like phage (PG2) had a circular genome of 99 kb encoding 99 proteins, of which 32 (32.3%) were functionally annotated (**Fig. 5A**) and showed ≥ 95 % identity and > 95 % coverage to sequenced crAssphages (**Suppl. Table 6**). PG2 was classified as candidate genus *I* (*AlphacrAssvirinae* subfamily), which is currently known to be one of the most prevalent gut viruses in Western subjects independent of age. Typical of a crAssphage, PG2 was clearly subdivided into two regions with opposite gene orientation, with one region encoding structural proteins and proteins involved in host interaction, and the other region encoding proteins 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 [91]. No gene was annotated as RNA polymerase, however, we suspect that one of the large unknown genes of PG2 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 [92]. 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 compared to 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). We suspect that the LoVEphage is a temperate phage as it encoded two integrase proteins, a repressor protein, a prophage protein. Furthermore, the genome was highly similar to *Bacteroides dorei* strain CL03T12C01 (CP011531.1) (95.6% nucleotide identity over 96% of its length) indicating that a LoVEphage-like 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), we attempted to reconstruct additional complete LoVEphage-like genomes from the DEVoC samples as well as from the SRA viromes. This resulted in reconstruction of 18 additional complete LoVEphage-like genomes. Each complete genome (67.9 - 72.4 kb) encoded between 122 and 131 genes, of which 61 conserved proteins were selected for phylogenetic analysis based on concatenated protein alignment (see Methods for details). **Fig. 5C** shows their phylogeny and **Fig. 5D** their genomic organizations. Two large phylogenetic clusters can be distinguished. The largest cluster contains 12 genomes mainly obtained from healthy adults, while the smaller cluster contains seven 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, with few insertions/deletions. The largest gene in these genomes has a conserved position, but has one of three annotations (and is assigned to three different functional groups with three different colours in **Fig. 5D**). Four proteins, including the one from the reference, are annotated as “mu-like prophage protein” (indicative of temperate phages; green), while 11 proteins are annotated as “tail tape measure protein” (involved in 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). The latter are > 99% identical at the amino acid level to the proteins annotated as “tail tape measure protein” and therefore we assume that “reticulocyte binding protein rhoptry” is likely a misannotation in one of the databases used. The “mu-like prophage protein” plays a role in viral tail assembly, as do the “tail tape measure proteins” to which they showed > 88% similarity at amino acid level.

Discussion

Human gut viruses represent a major pool of diverse and relatively underexplored microbes that, together with other gut microbiome components, are believed to impact human health and disease [93]. Currently, the number of studies exploring the human gut viruses is expanding significantly, and subsequently is the cataloguing of viral genomes and genes, which collectively will advance the virome field.

In this study, a human enteric virome catalogue (the DEVoC) containing 12,986 viral sequences and encoding 190,029 genes was generated from 254 faecal viromes from Danish children, adolescents and adults. The majority of the DEVoC sequences were of phage origin, mostly taxonomically unclassified (**Fig. 1A**) and without assigned bacterial host (**Fig. 1B**), as described in other human gut virome databases [7]. 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 sequences formed a viral cluster (VC) with RefSeq representatives of these families. Phages of the *Caudovirales* and *Petitvirales* orders that were identified in DEVoC have all been commonly described in human gut viromes [7], [13]. The predicted bacterial hosts of these phages are well known gut bacteria including members of the *Firmicutes* and *Bacteroidetes* phyla [94]. Recent human gut virome studies have all concluded gut viromes to be highly individual [11], [13] and therefore the majority of the identified phages were novel. Individual-differentiating factors likely include geographical origin [95], age [7], diet [10] and health status [72], [96], [97]. Each of the compared viral databases contained a rather unique set of viruses, 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 sequences 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, we clustered proteins into OGs and used these clusters as proxy for function. The sparse overlap with proteins from the HVPC database can – in part – be explained by the high level of uniqueness of the original DEVoC sequences, although the limited number of faecal samples used to generate the HVPC database, as well as the apparent lack of viral selection and consequential inclusion of non-viral proteins could contribute as well [54] (**Fig. 1D**).

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, which was further reflected by the fact that the majority of the PGs (phage genomes predicted to be at least 50% complete) were found only in a single healthy subject (**Fig. 2A**). It should be noted that in each healthy subject the majority of the viral reads belonged to only a handful of phage genomes (**Fig. 2B**). Despite this individuality and our stringent selection criteria, we identified 39 PGs in more than 10 healthy subjects (> 12% prevalence) with a maximum prevalence of 33% (30 subjects) (**Suppl. Table 3**). This finding refutes the existence of a “core” virome (phages present in > 50% of subjects) [64] – at least at genome level – similar to previous studies [7]. On the other hand, OGs were much more prevalent and could be detected in up to 97% of the healthy subjects (**Suppl. Table 4**) – most of these are involved in typical phage functions. However, similar to previous findings [98], the majority of the OGs remained specific to only one subject (**Fig. 2C**).

Gregory *et al.* reported an age-dependent virome diversity using publicly available data [7]. They included studies produced with varying wetlab procedures and sequencing depths, as well as age groups with unequal age ranges and sample sizes (infants (< 3 years): n = 27 vs. children/adolescents (3-18 years): n = 11 vs. adults (18-65 years): n = 93 vs. elderly (> 65 years): n = 20). Our study could not confirm the former as the PG richness and Shannon diversity did not differ across age groups (**Suppl. Fig 5**). While our study has the advantage of consistently processed samples of different age groups (range 6-73 years), we lacked data from infants and young children (< 6 years old) as well as young adults (19 – 39 years old) to make associations with age as a continuous variable. Beta-diversity at PG and at OG level were associated with age group, although the biological importance of this effect is probably limited (**Fig. 3A,E**). OG richness was, similar to PG richness, not different between age groups (**Suppl. Fig. 8**). Interestingly, at the level of individual OGs and PGs, 45 OGs and 3 PGs had different prevalences across age groups. One OG and two PGs, were more common in adults, while the others were more prevalent in children/adolescents (**Table 1** and **Suppl. Table 3**). The presence of age-associated PGs may indicate that some more common (or even core) phages might exist in smaller, more homogeneous, populations, although core phages do not exist for the general healthy human population.

We observed a clear decrease in the number and proportion of temperate PGs in our healthy adult population (**Fig. 3B, Fig. 3C**). This is in accordance with the finding from Beller *et al.* [40] that demonstrates a decrease in

the proportion of temperate phages across the first year of life in infants, and our findings suggest 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 lifecycle 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. 4BCD**). However, the conclusions should be interpreted cautiously, as subsets consisted of heterogeneous sample sizes (Kruskal-Wallis tests do not take into account the number of subjects within each subset). Especially, 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 with regard to 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 geographic proximity and cultural similarities. 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 part of our paediatric cohort, and thus not included in the development of the DEVoC. Prevalences in the 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. Due to the same lack of samples from healthy children and adolescents, 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 [37], [76], [99], [100]. 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 and the LoVEphages across SRA viromes are 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 genes typical for phages could be identified in the LoVEphage, besides a majority of genes with unknown function (**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 [101]. However, such patterns may also become apparent when more LoVEphage-like genomes are included/investigated.

Conclusion

The human gut virome catalogue 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.

Declarations

Ethical approval and consent to participate

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 (paediatric cohort). All participants or their legal guardians gave consent to participate in this study.

Consent for publication

Not applicable.

Availability of data and materials

The virome sequencing reads supporting the conclusions of this article will be made available at the Sequenced Read Archive (SRA) under projectID XXX. The DEVoC and its encoded genes will be made available at https://zenodo.org/XXX. The LoVEphage genome is submitted to GenBank (accession XXXXXX). The scripts used to perform the analysis and make figures starting from the abundance table will be made available at https://github.com/Matthijnssenslab/ViromeCatalogue.

Competing interests

The authors declare that they have no competing interests.

Funding

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; the European Union’s Horizon 2020 research and innovation programme (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 Economy, Science and Innovation.

Author contributions

The study was conceived by EGB, LVE, AK, PB, TH, MA and JM. CFB, CEF, SJ, MKj, MJ, HBJ, TN, JCH and AK handled the collection and management of faecal samples. LVE and EGB managed the project. EGB and LCC carried out the viral DNA/RNA extraction, amplifications and library preparation. LVE performed the bioinformatic processing of the reads, generated the catalogue and performed the statistical analysis in close collaboration with EGB, LB, WD, MA and JM. AF and MKu predicted CRISPR spacers in bacterial metagenomes. LVE performed SRA screening with assistance of DS and LDC. LVE, EGB, MA and JM drafted the manuscript. All authors critically revised the article and approved the final version for publication.

Acknowledgements

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

References

[1] S. V. Lynch and O. Pedersen, “The Human Intestinal Microbiome in Health and Disease,” *N. Engl. J. Med.*, vol. 375, no. 24, pp. 2369–2379, 2016.

[2] M. S. Kim, E. J. Park, S. W. Roh, and J. W. Bae, “Diversity and abundance of single-stranded DNA viruses in human feces,” *Appl. Environ. Microbiol.*, vol. 77, no. 22, pp. 8062–8070, 2011.

[3] J. L. Castro-Mejía *et al.*, “Optimizing protocols for extraction of bacteriophages prior to metagenomic analyses of phage communities in the human gut,” *Microbiome*, vol. 3, p. 64, 2015.

[4] L. Hoyles *et al.*, “Characterization of virus-like particles associated with the human faecal and caecal microbiota,” *Res. Microbiol.*, vol. 165, no. 10, pp. 803–812, 2014.

[5] J. C. Clemente, L. K. Ursell, L. W. Parfrey, and R. Knight, “The impact of the gut microbiota on human health: An integrative view,” *Cell*, vol. 148, no. 6, pp. 1258–1270, 2012.

[6] S. R. Carding, N. Davis, and L. Hoyles, “Review article: the human intestinal virome in health and disease,” *Aliment. Pharmacol. Ther.*, vol. 46, no. 9, pp. 800–815, 2017.

[7] A. C. Gregory, O. Zablocki, A. A. Zayed, A. Howell, B. Bolduc, and M. B. Sullivan, “The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut,” *Cell Host Microbe*, vol. 28, no. 5, pp. 724-740.e8, 2020.

[8] E. S. Lim *et al.*, “Early life dynamics of the human gut virome and bacterial microbiome in infants,” *Nat. Med.*, vol. 21, no. 10, pp. 1228–1234, 2015.

[9] R. Maqsood *et al.*, “Discordant transmission of bacteria and viruses from mothers to babies at birth,” *Microbiome*, vol. 7, no. 1, p. 156, Dec. 2019.

[10] S. Minot *et al.*, “The human gut virome: Inter-individual variation and dynamic response to diet,” *Genome Res.*, vol. 21, no. 10, pp. 1616–1625, 2011.

[11] J. L. Moreno-Gallego *et al.*, “Virome Diversity Correlates with Intestinal Microbiome Diversity in Adult Monozygotic Twins,” *Cell Host Microbe*, vol. 25, no. 2, pp. 261-272.e5, 2019.

[12] S. R. Stockdale, F. J. Ryan, A. McCann, M. Dalmasso, and C. Hill, “Viral dark matter in the gut virome of elderly humans,” *Preprints*, Jul. 2018.

[13] A. N. Shkoporov *et al.*, “The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific,” *Cell Host Microbe*, vol. 26, no. 4, pp. 527-541.e5, 2019.

[14] M. A. Fernandes *et al.*, “Enteric Virome and Bacterial Microbiota in Children with Ulcerative Colitis and Crohn Disease,” *J. Pediatr. Gastroenterol. Nutr.*, vol. 68, no. 1, pp. 30–36, 2019.

[15] V. Pérez-Brocal, R. García-López, P. Nos, B. Beltrán, I. Moret, and A. Moya, “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.*, vol. 21, no. 11, pp. 2515–2532, 2015.

[16] A. N. Shkoporov *et al.*, “Reproducible protocols for metagenomic analysis of human faecal phageomes,” *Microbiome*, vol. 6, no. 1, pp. 1–17, 2018.

[17] L. A. Draper *et al.*, “Long-term colonisation with donor bacteriophages following successful faecal microbial transplantation,” *Microbiome*, vol. 6, no. 1, p. 220, Dec. 2018.

[18] T. Zuo *et al.*, “Bacteriophage transfer during faecal microbiota transplantation in Clostridium difficile infection is associated with treatment outcome,” *Gut*, vol. 67, no. 4, pp. 634–643, 2018.

[19] S. Roux, F. Enault, B. L. Hurwitz, and M. B. Sullivan, “VirSorter: mining viral signal from microbial genomic data,” *PeerJ*, vol. 3, p. e985, 2015.

[20] V. I. Jurtz, J. Villarroel, O. Lund, M. Voldby Larsen, and M. Nielsen, “MetaPhinder - Identifying bacteriophage sequences in metagenomic data sets,” *PLoS One*, vol. 11, no. 9, pp. 1–14, 2016.

[21] J. Ren *et al.*, “Identifying viruses from metagenomic data using deep learning,” *Quant. Biol.*, vol. 8, no. 1, pp. 64–77, 2020.

[22] L. Fancello, D. Raoult, and C. Desnues, “Computational tools for viral metagenomics and their application in clinical research,” *Virology*, vol. 434, no. 2, pp. 162–174, 2012.

[23] D. Paez-Espino *et al.*, “IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses,” *Nucleic Acids Res.*, vol. 45, 2016.

[24] N. Goodacre, A. Aljanahi, S. Nandakumar, M. Mikailov, and A. S. Khan, “A Reference Viral Database (RVDB) To Enhance Bioinformatics Analysis of High-Throughput Sequencing for Novel Virus Detection.”

[25] D. Paez-espino *et al.*, “Uncovering Earth ’ s virome,” *Nat. Publ. Gr.*, vol. 536, no. 7617, pp. 425–430, 2016.

[26] L. F. Camarillo-guerrero *et al.*, “Massive expansion of human gut bacteriophage diversity,” *Cell*, vol. 184, no. 4, pp. 1098-1109.e9, 2021.

[27] S. Benler *et al.*, “Thousands of previously unknown phages discovered in whole-community human gut metagenomes,” *bioRxiv*, 2020.

[28] E. Science, N. Goodacre, A. Aljanahi, S. Nandakumar, M. Mikailov, and A. S. Khan, “crossm A Reference Viral Database (RVDB) To Enhance Bioinformatics Analysis of High-Throughput Sequencing for Novel Virus,”

vol. 3, no. 2, pp. 1–18, 2018.

[29] R. Sausset, M. A. Petit, V. Gaboriau-Routhiau, and M. De Paepe, “New insights into intestinal phages,” *Mucosal Immunol.*, vol. 13, no. 2, pp. 205–215, 2020.

[30] A. N. Shkoporov and C. Hill, “Bacteriophages of the Human Gut: The ‘Known Unknown’ of the Microbiome,” *Cell Host Microbe*, vol. 25, no. 2, pp. 195–209, 2019.

[31] S. Roux *et al.*, “Minimum information about an uncultivated virus genome (MIUVIG),” *Nat. Biotechnol.*, vol. 37, no. 1, pp. 29–37, Jan. 2019.

[32] H. Bin Jang *et al.*, “Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks,” *Nat. Biotechnol.*, vol. 37, no. 6, pp. 632–639, Jun. 2019.

[33] P. Aiewsakun and P. Simmonds, “The genomic underpinnings of eukaryotic virus taxonomy: Creating a sequence-based framework for family-level virus classification,” *Microbiome*, vol. 6, no. 1, pp. 1–24, 2018.

[34] T. D. S. Sutton and C. Hill, “Gut Bacteriophage: Current Understanding and Challenges,” *Front. Endocrinol. (Lausanne).*, vol. 10, p. 784, Nov. 2019.

[35] M. J. Tisza, A. K. Belford, G. Dominguez-Huerta, B. Bolduc, and C. B. Buck, “Cenote-Taker 2 Democratizes Virus Discovery And Sequence Annotation,” *Virus Evol.*, Dec. 2020.

[36] B. E. Dutilh *et al.*, “A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes,” *Nat. Commun.*, vol. 5, pp. 1–11, 2014.

[37] A. N. Shkoporov *et al.*, “ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis,” *Nat. Commun.*, vol. 9, no. 1, pp. 1–8, 2018.

[38] N. Yutin *et al.*, “Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut,” *Nat. Microbiol.*, vol. 3, no. 1, pp. 38–46, 2018.

[39] N. Conceição-Neto *et al.*, “Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis,” *Nat. Publ. Gr.*, 2015.

[40] L. Beller *et al.*, “THE VIROME AND ITS TRANSKINGDOM INTERACTIONS IN THE HEALTHY INFANT GUT,” 2021.

[41] S. Nayfach, A. P. Camargo, F. Schulz, E. Eloe-Fadrosh, S. Roux, and N. C. Kyrpides, “CheckV assesses the quality and completeness of metagenome-assembled viral genomes,” *Nat. Biotechnol.*, pp. 1–8, Dec. 2020.

[42] B. D. Ondov, N. H. Bergman, and A. M. Phillippy, “Interactive metagenomic visualization in a Web browser,” *BMC Bioinformatics*, vol. 12, no. 1, p. 385, Dec. 2011.

[43] B. Buchfink, C. Xie, and D. H. Huson, “Fast and sensitive protein alignment using DIAMOND,” *Nat. Methods*, vol. 12, no. 1, pp. 59–60, 2014.

[44] C. Camacho *et al.*, “BLAST+: architecture and applications,” *BMC Bioinformatics*, vol. 10, no. 1, p. 421, Dec. 2009.

[45] C. Bland *et al.*, “CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats,” *BMC Bioinformatics*, vol. 8, no. 1, p. 209, Dec. 2007.

[46] D. R. Mende *et al.*, “proGenomes2: an improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes,” *Nucleic Acids Res.*, vol. 48, no. D1, pp. D621–D625, Oct. 2019.

[47] K. Mcnair, C. Zhou, E. A. Dinsdale, B. Souza, and R. A. Edwards, “PHANOTATE: A novel approach to gene identification in phage genomes.”

[48] D. Hyatt, G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, and L. J. Hauser, “Prodigal: prokaryotic gene recognition and translation initiation site identification,” *BMC Bioinformatics*, vol. 11, no. 1, p. 119, Mar. 2010.

[49] S. R. Eddy, “A new generation of homology search tools based on probabilistic inference.,” *Genome Inform.*, vol. 23, no. 1, pp. 205–211, 2009.

[50] A. Marchler-Bauer *et al.*, “CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D200–D203, Jan. 2017.

[51] A. Meier and J. Söding, “Automatic Prediction of Protein 3D Structures by Probabilistic Multi-template Homology Modeling,” *PLoS Comput. Biol.*, vol. 11, no. 10, 2015.

[52] M. Lechner, S. Findeiß, L. Steiner, M. Marz, P. F. Stadler, and S. J. Prohaska, “Proteinortho: Detection of (Co-)orthologs in large-scale analysis,” 2011.

[53] D. Paez-Espino *et al.*, “IMG/VR v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D678–D686, Jan. 2019.

[54] A. Shankar Bhattacharjee *et al.*, “The Human Virome Protein Cluster Database (HVPC): A Human Viral Metagenomic Database for Diversity and Function Annotation,” 2018.

[55] S. Roux and B. Bolduc, “ClusterGenomes.” .

[56] S. Kurtz *et al.*, “Versatile and open software for comparing large genomes,” *Genome Biol.*, vol. 5, no. 2, p. R12, Jan. 2004.

[57] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, “CD-HIT: accelerated for clustering the next-generation sequencing data,” *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, Dec. 2012.

[58] K. Aiemjoy *et al.*, “Viral species richness and composition in young children with loose or watery stool in Ethiopia,” *BMC Infect. Dis.*, vol. 19, no. 1, pp. 1–10, Jan. 2019.

[59] C. Chehoud *et al.*, “Transfer of Viral Communities between Human Individuals during,” *MBio*, vol. 7, no. 2, pp. 1–8, 2016.

[60] G. D. Hannigan, M. B. Duhaime, M. T. Ruffin, C. C. Koumpouras, and P. D. Schloss, “Diagnostic potential and interactive dynamics of the colorectal cancer virome,” *MBio*, vol. 9, no. 6, pp. 1–13, 2018.

[61] L. Kramná *et al.*, “Gut virome sequencing in children with early islet autoimmunity,” *Diabetes Care*, vol. 38, no. 5, pp. 930–933, May 2015.

[62] J. Legoff *et al.*, “The eukaryotic gut virome in hematopoietic stem cell transplantation: New clues in enteric graft-versus-host disease,” *Nat. Med.*, vol. 23, no. 9, pp. 1080–1085, Sep. 2017.

[63] M. Ly *et al.*, “Transmission of viruses via our microbiomes,” *Microbiome*, vol. 4, no. 1, p. 64, 2016.

[64] P. Manrique, B. Bolduc, S. T. Walk, J. Der Van Oost, W. M. De Vos, and M. J. Young, “Healthy human gut phageome,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 113, no. 37, pp. 10400–10405, 2016.

[65] A. McCann *et al.*, “Viromes of one year old infants reveal the impact of birth mode on microbiome diversity,” *PeerJ*, vol. 2018, no. 5, pp. 1–13, 2018.

[66] S. Minot, S. Grunberg, G. D. Wu, J. D. Lewis, and F. D. Bushman, “Hypervariable loci in the human gut virome,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 10, pp. 3962–3966, 2012.

[67] S. Minot, A. Bryson, C. Chehoud, G. D. Wu, J. D. Lewis, and F. D. Bushman, “Rapid evolution of the human gut virome,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 30, pp. 12450–12455, 2013.

[68] C. L. Monaco *et al.*, “Altered Virome and Bacterial Microbiome in Human Immunodeficiency Virus-Associated Acquired Immunodeficiency Syndrome,” *Cell Host Microbe*, vol. 19, no. 3, pp. 311–322, 2016.

[69] V. Pérez-Brocal *et al.*, “Study of the Viral and Microbial Communities Associated With Crohn’s Disease: A Metagenomic Approach,” *Clin. Transl. Gastroenterol.*, vol. 4, no. 6, p. e36, Jun. 2013.

[70] A. Reyes *et al.*, “Viruses in the faecal microbiota of monozygotic twins and their mothers,” *Nature*, vol. 466, no. 7304, pp. 334–338, 2010.

[71] C. K. Yinda *et al.*, “Gut Virome Analysis of Cameroonians Reveals High Diversity of Enteric Viruses, Including Potential Interspecies Transmitted Viruses,” *mSphere*, vol. 4, no. 1, 2019.

[72] G. Zhao *et al.*, “Intestinal virome changes precede autoimmunity in type I diabetes-susceptible children,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 114, no. 30, pp. E6166–E6175, Jul. 2017.

[73] H. Li and R. Durbin, “Fast and accurate short read alignment with Burrows-Wheeler transform,” vol. 25, no. 14, pp. 1754–1760, 2009.

[74] A. M. Bolger, M. Lohse, and B. Usadel, “Trimmomatic: A flexible trimmer for Illumina sequence data,” *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, 2014.

[75] A. Reyes *et al.*, “Gut DNA viromes of Malawian twins discordant for severe acute malnutrition,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 112, no. 38, pp. 11941–11946, 2015.

[76] E. Guerin, A. Shkoporov, S. R. Stockdale, E. Gonzalez-Tortuero, R. P. Ross, and C. Hill, “Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut,” *Cell Host Microbe*, vol. 24, pp. 653–664, 2018.

[77] S. Roux and B. Bolduc, “ClusterGenomes.”

[78] A. M. Bolger, M. Lohse, and B. Usadel, “Trimmomatic: A flexible trimmer for Illumina sequence data,” *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, 2014.

[79] A. Bankevich *et al.*, “SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing,” *J. Comput. Biol.*, vol. 19, no. 5, pp. 455–477, 2012.

[80] H. Li and R. Durbin, “Fast and accurate short read alignment with Burrows-Wheeler transform,” *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[81] K. Katoh and D. M. Standley, “MAFFT multiple sequence alignment software version 7: Improvements in performance and usability,” *Mol. Biol. Evol.*, vol. 30, no. 4, pp. 772–780, 2013.

[82] S. Capella-Gutiérrez, J. M. Silla-Martínez, and T. Gabaldón, “trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses,” *Bioinformatics*, vol. 25, no. 15, pp. 1972–1973, Aug. 2009.

[83] A. Stamatakis, “RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies,” *Bioinformatics*, vol. 30, no. 9, pp. 1312–1313, May 2014.

[84] P. J. McMurdie and S. Holmes, “phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data.,” *PLoS One*, vol. 8, no. 4, p. e61217, 2013.

[85] J. Oksanen *et al.*, “Package ‘vegan’: Community Ecology Package Version 2.5-7.” 2020.

[86] L. Guy, J. R. Kultima, S. G. E. Andersson, and J. Quackenbush, “GenoPlotR: comparative gene and genome visualization in R,” in *Bioinformatics*, 2011, vol. 27, no. 13, pp. 2334–2335.

[87] H. Chen and P. C. Boutros, “VennDiagram: A package for the generation of highly-customizable Venn and Euler diagrams in R,” *BMC Bioinformatics*, vol. 12, no. 1, p. 35, Jan. 2011.

[88] G. Yu, D. K. Smith, H. Zhu, Y. Guan, and T. T. Y. Lam, “ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data,” *Methods Ecol. Evol.*, vol. 8, no. 1, pp. 28–36, Jan. 2017.

[89] H. Wickham, “ggplot2: Elegant Graphics for Data Analysis.” 2009.

[90] S. R. Krishnamurthy and D. Wang, “Extensive conservation of prokaryotic ribosomal binding sites in known and novel picobirnaviruses,” *Virology*, vol. 516, pp. 108–114, Mar. 2018.

[91] V. Morozova, M. Fofanov, N. Tikunova, I. Babkin, V. V Morozov, and A. Tikunov, “First crAss-Like Phage Genome Encoding the Diversity-Generating Retroelement (DGR),” *Viruses*, vol. 12, no. 573, 2020.

[92] E. V Koonin and N. Yutin, “Special Issue: Infection Biology in the Age of the Microbiome The crAss-like Phage Group: How Metagenomics Reshaped the Human Virome Trends in Microbiology,” *Trends Microbiol.*, vol. 28, no. 5, 2020.

[93] H. W. Virgin, “The Virome in Mammalian Physiology and Disease,” *Cell*, vol. 157, pp. 142–150, 2014.

[94] M. Arumugam *et al.*, “Enterotypes of the human gut microbiome,” 2011.

[95] S. Rampelli *et al.*, “Characterization of the human DNA gut virome across populations with different subsistence strategies and geographical origin,” *Environ. Microbiol.*, vol. 19, no. 11, pp. 4728–4735, Nov. 2017.

[96] J. M. Norman *et al.*, “Disease-specific Alterations in the Enteric Virome in Inflammatory Bowel Disease,” *NIH Public Access*, vol. 160, no. 3, pp. 447–460, 2015.

[97] A. G. Clooney *et al.*, “Whole-Virome Analysis Sheds Light on Viral Dark Matter in Inflammatory Bowel Disease,” *Cell Host Microbe*, vol. 26, no. 6, pp. 764-778.e5, Dec. 2019.

[98] D. M. Kristensen, A. S. Waller, T. Yamada, P. Bork, A. R. Mushegian, and E. V. Koonin, “Orthologous gene clusters and taxon signature genes for viruses of prokaryotes,” *J. Bacteriol.*, vol. 195, no. 5, pp. 941–950, 2013.

[99] T. P. Honap, K. Sankaranarayanan, S. L. Schnorr, A. T. Ozga, C. Warinner, and C. M. Lewis, “Biogeographic study of human gut-associated crAssphage suggests impacts from industrialization and recent expansion,” *PLoS One*, vol. 15, no. 1, p. e0226930, Jan. 2020.

[100] B. A. Siranosian, F. B. Tamburini, G. Sherlock, and A. S. Bhatt, “Acquisition, transmission and strain diversity of human gut-colonizing crAss-like phages,” *Nat. Commun.*, vol. 11, no. 1, p. 280, Dec. 2020.

[101] R. Edwards *et al.*, “Global phylogeography and ancient evolution of the widespread human gut virus crAssphage,” *bioRxiv*, p. 527796, 2019.

Table

Table 1: Orthologous groups with age-associated absence/presence profiles.

Orthologous Group	Size	Annotation	Function	Prevalence in healthy subset			Chi^2 test adj. p-value
				All (n = 91)	Paediatric cohort (n = 46)	Adult cohort (n = 45)	
17005	180	Hypothetical protein	Unknown	63 (69.2%)	**42 (91.3%)**	21 (46.7%)	p = 0.0011
17212	205	Carlavirus endopeptidase	Assembly	63 (69.2%)	**42 (91.3%)**	21 (46.7%)	p = 0.0011
116	149	Tail assembly chaperone protein	Assembly	58 (63.7%)	**40 (87%)**	18 (40%)	p = 0.0008
17197	159	Hypothetical protein	Unknown	56 (61.5%)	**42 (91.3%)**	14 (31.1%)	p < 0.0001
17367	149	Major capsid/head protein	Structural	54 (59.3%)	**41 (89.1%)**	13 (28.9%)	p < 0.0001
17685	118	Minor structural protein	Structural	54 (59.3%)	**39 (84.8%)**	15 (33.3%)	p = 0.0002
863	115	Hypothetical protein	Unknown	49 (53.8%)	**36 (78.3%)**	13 (28.9%)	p = 0.0006
16990	86	Hypothetical protein	Unknown	46 (50.5%)	**35 (76.1%)**	11 (24.4%)	p = 0.0002
1899	95	Tail completion protein	Assembly	44 (48.4%)	**35 (76.1%)**	9 (20%)	p < 0.0001
2871	96	Portal protein	Packaging	43 (47.3%)	**33 (71.7%)**	10 (22.2%)	p = 0.0006
3146	86	Hypothetical protein	Unknown	41 (45.1%)	**32 (69.6%)**	9 (20%)	p = 0.0005
3199	20	Polysaccharide export protein	Other	37 (40.7%)	**31 (67.4%)**	6 (13.3%)	p < 0.0001
16319	71	tRNA synthase	Translation	35 (38.5%)	**29 (63%)**	6 (13.3%)	p = 0.0003
2045	48	Hypothetical protein	Unknown	34 (37.4%)	**28 (60.9%)**	6 (13.3%)	p = 0.0007
2076	6	Putative metallopeptidasee	Other	33 (36.3%)	5 (10.9%)	**28 (62.2%)**	p = 0.0001
752	45	Hypothetical protein	Unknown	33 (36.3%)	**29 (63%)**	4 (8.9%)	p < 0.0001
16058	4	Hypothetical protein	Unknown	32 (35.2%)	**28 (60.9%)**	4 (8.9%)	p = 0.0001
17591	68	Hypothetical	Unknown	31	**27**	4	p =

		protein		(34.1%)	**(58.7%)**	(8.9%)	0.0002
2749	34	Hypothetical protein	Unknown	31 (34.1%)	**26 (56.5%)**	5 (11.1%)	p = 0.0012
1811	3	Plasmid recombination enzyme	Recombination	30 (33%)	**26 (56.5%)**	4 (8.9%)	p = 0.0004
16535	37	LytR response regulator	Other	29 (31.9%)	**25 (54.3%)**	4 (8.9%)	p = 0.0009
17358	59	Hypothetical protein	Unknown	27 (29.7%)	**25 (54.3%)**	2 (4.4%)	p = 0.0001
17353	63	Bromodomain RACK7 like subfamily	Other	26 (28.6%)	**24 (52.2%)**	2 (4.4%)	p = 0.0001
18041	45	Head tail connector protein	Structural	26 (28.6%)	**24 (52.2%)**	2 (4.4%)	p = 0.0001
18129	44	Minor structural protein	Structural	24 (26.4%)	**22 (47.8%)**	2 (4.4%)	p = 0.0008
2613	38	Hypothetical protein	Unknown	22 (24.2%)	**21 (45.7%)**	1 (2.2%)	p = 0.0004
3028	35	Hypothetical protein	Unknown	22 (24.2%)	**21 (45.7%)**	1 (2.2%)	p = 0.0004
2406	37	DNA binding protein	Other	20 (22%)	**20 (43.5%)**	0 (0%)	p = 0.0002
2150	27	Hypothetical protein	Unknown	19 (20.9%)	**19 (41.3%)**	0 (0%)	p = 0.0004

Bonferroni-adjusted p-values of chi^2 test on prevalences. Prevalences in bold indicate cohort with highest prevalence.

Figures

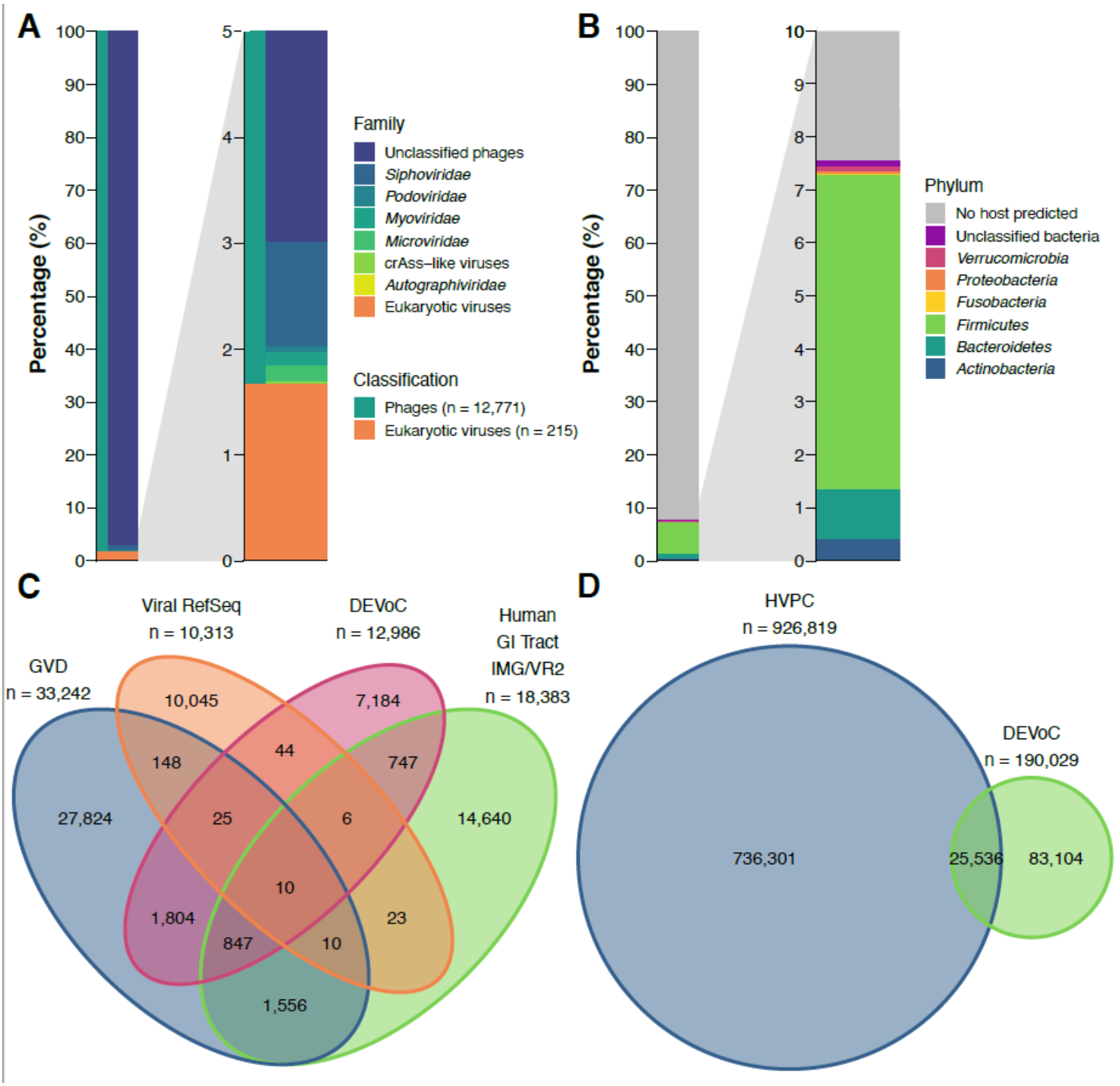

Figure 1

DEVoC mainly consists of undescribed phages. (A) Overview of the DEVoC sequences (n = 12,986) by type of virus and phage family. Breakdown of the eukaryotic viruses into families is visualized in Suppl. Fig. 3). (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. (D) Venn-diagram showing the number of clusters with members of the DEVoC and HVPC proteins at 60% amino

acid identity over 80% coverage. Numbers in the Venn diagram do not sum up to the database sizes, as both databases were redundant before clustering.

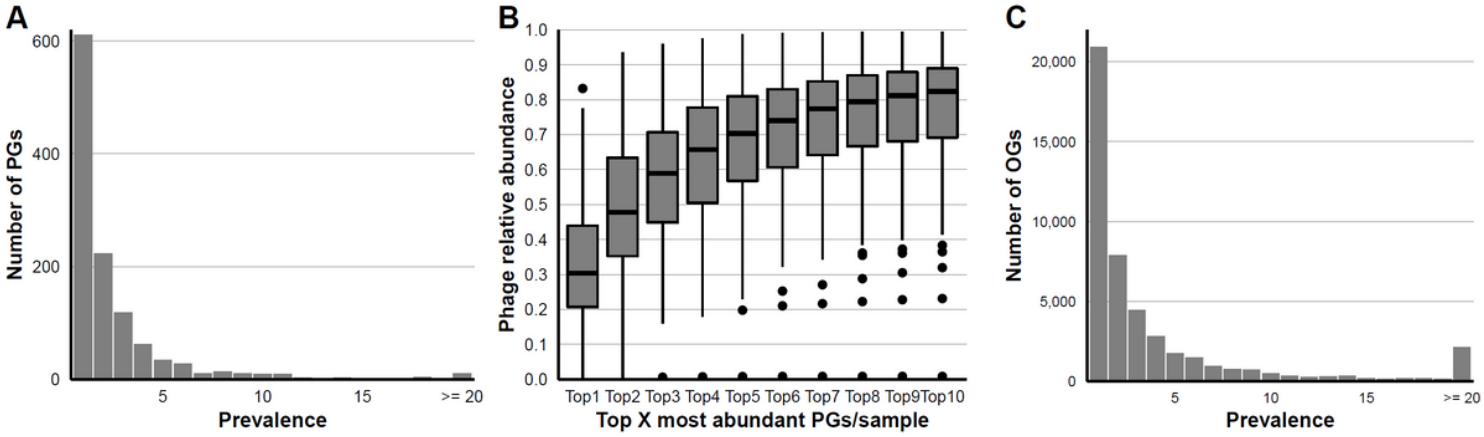

Figure 2

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 together. (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 together.

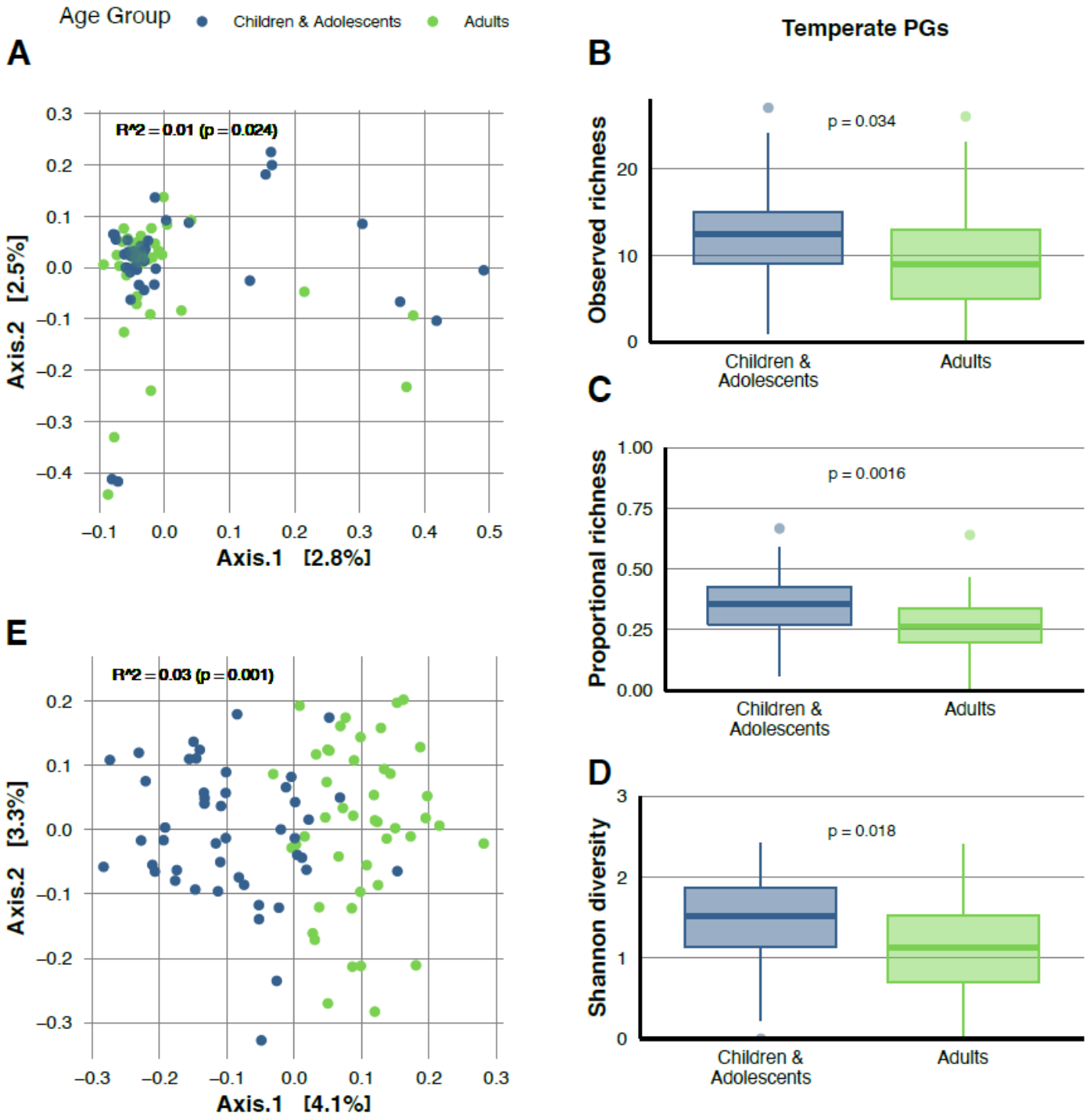

Figure 3

Age group-associated virome patterns in healthy Danish subjects. (A) Principal Coordinate Analysis on Jaccard dissimilarities between healthy Danish subjects at PG level (PERMANOVA of age group; R2 = 0.01; p = 0.024). Subjects are coloured by age group. (B) Boxplots of number of temperate PGs in healthy Danish children and adolescents (n = 46) and adults (n = 45) (Wilcoxon test; p = 0.034). (C) Boxplots of proportional richness of temperate PGs (number of temperate PGs vs. 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 OG level (PERMANOVA of age group; R2 = 0.03; p = 0.001). Subjects are coloured by age group. All analyses are performed on 46 children/adolescents and 45 adults.

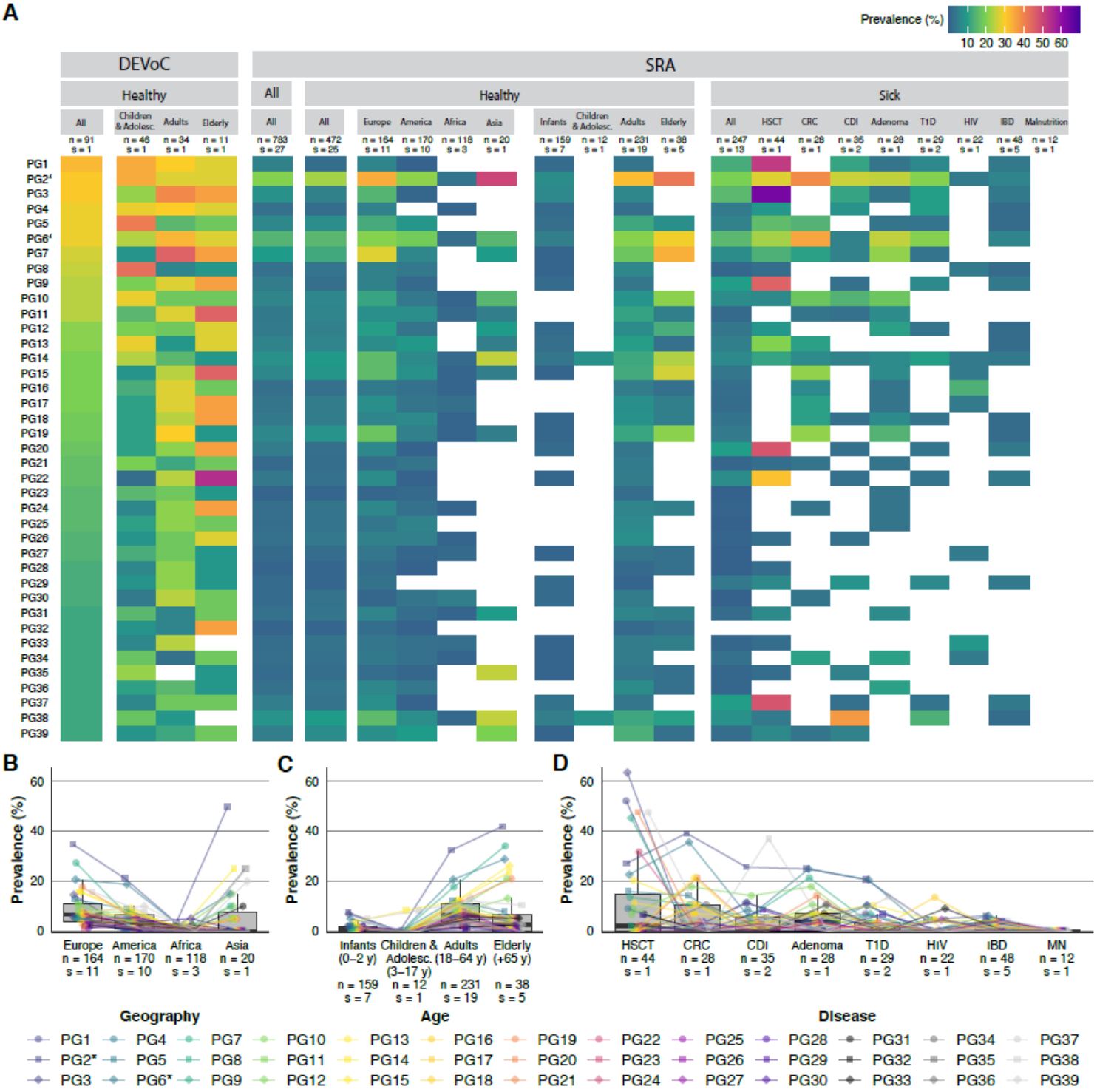

Figure 4

Worldwide prevalence of 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 – 4) and from other human gut virome studies (columns 5 – 23). The first four columns represent the prevalence in healthy Danish subjects, all healthy Danish children and adolescents (6 -

18 years old), all healthy Danish adults (40 - 65 years old) and all healthy elderly (≥ 65 years old) from the DEVoC cohort. The fifth column shows the overall prevalence in all subjects from all the other studies combined. Columns 6-14 represent the prevalence in the healthy subjects (column 6), separated by continent (column 7-10) and age group (columns 11-14). Columns 15-23 represent the prevalence in disease subjects (column 15) in different diseases (column 16-23). 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 indicate by different shapes and colours by PG and connected across boxplots in panels B, C and D and the number of subjects (n) and studies (s) included in each subgroup is indicated below each boxplot. PGs with 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.

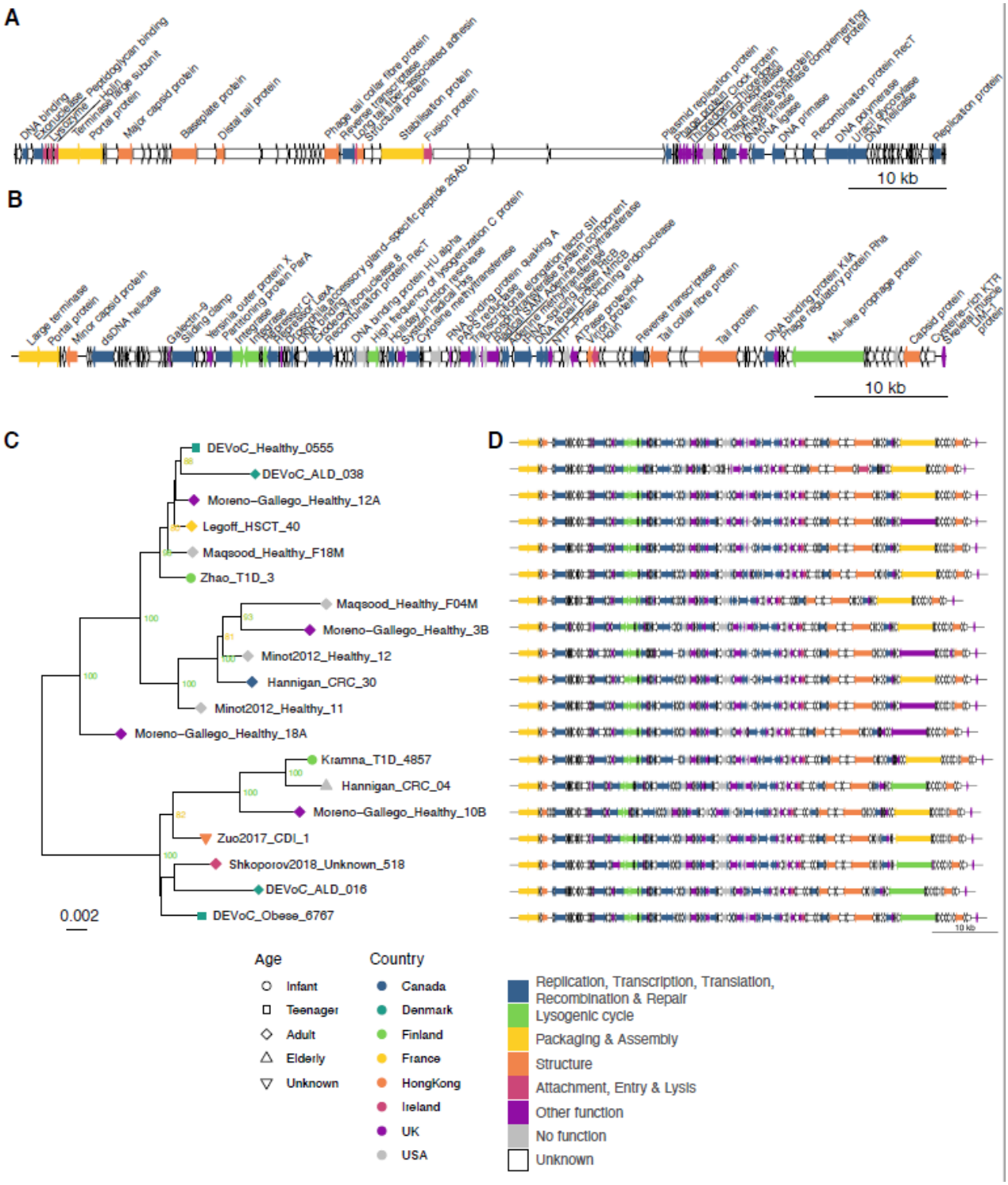

Figure 5

Genomic stucture of the identified crAss-like and LoVEphage genomes. (A) Genomic structure of a crAss-like phage (PG2) commonly found in the healthy Danish subjects and in other virome studies. (B) Genomic structure of the LoVEphage (PG6) commonly found in the healthy Danish subjects and in other virome studies. (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 colour 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. ALD =

alcohol liver disease; CRC = colorectal cancer; T1D = type 1 diabetes; HSCT = hematopoietic stem cell transplantation. (D) Genome structure of the 19 LoVEphage-like genomes shown in panel C. All genomes in panels A, B and D 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.

Supplementary Files

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

- SupplFig1.pdf
- SupplFig2.pdf
- SupplFig3.pdf
- SupplFig4.pdf
- SupplFig5.pdf
- SupplFig6.pdf
- SupplFig7.pdf
- SupplFig8.pdf
- SupplementaryMaterials.docx