nature microbiology

Resource

https://doi.org/10.1038/s41564-023-01345-7

Expanding known viral diversity in the healthy infant gut

Received: 10 June 2022

Accepted: 17 February 2023

Published online: 10 April 2023

Check for updates

Shiraz A. Shah[1,11] ✉, Ling Deng[2,11], Jonathan Thorsen[1,3], Anders G. Pedersen[4], Moïra B. Dion[5,6], Josué L. Castro-Mejía[2], Ronalds Silins[2], Fie O. Romme[2], Romain Sausset[7], Leon E. Jessen[1,4], Eric Olo Ndela[8], Mathis Hjelmsø[1], Morten A. Rasmussen[1,2], Tamsin A. Redgwell[1], Cristina Leal Rodríguez[1], Gisle Vestergaard[4], Yichang Zhang[2], Bo Chawes[1], Klaus Bønnelykke[1], Søren J. Sørensen[9], Hans Bisgaard[1,12], Francois Enault[8], Jakob Stokholm[1,2], Sylvain Moineau[5,6,10], Marie-Agnès Petit[7] & Dennis S. Nielsen[2] ✉

The gut microbiome is shaped through infancy and impacts the maturation of the immune system, thus protecting against chronic disease later in life. Phages, or viruses that infect bacteria, modulate bacterial growth by lysis and lysogeny, with the latter being especially prominent in the infant gut. Viral metagenomes (viromes) are difficult to analyse because they span uncharted viral diversity, lacking marker genes and standardized detection methods. Here we systematically resolved the viral diversity in faecal viromes from 647 1-year-olds belonging to Copenhagen Prospective Studies on Asthma in Childhood 2010, an unselected Danish cohort of healthy mother–child pairs. By assembly and curation we uncovered 10,000 viral species from 248 virus family-level clades (VFCs). Most (232 VFCs) were previously unknown, belonging to the *Caudoviricetes* viral class. Hosts were determined for 79% of phage using clustered regularly interspaced short palindromic repeat spacers within bacterial metagenomes from the same children. Typical *Bacteroides*-infecting crAssphages were outnumbered by undescribed phage families infecting *Clostridiales* and *Bifidobacterium*. Phage lifestyles were conserved at the viral family level, with 33 virulent and 118 temperate phage families. Virulent phages were more abundant, while temperate ones were more prevalent and diverse. Together, the viral families found in this study expand existing phage taxonomy and provide a resource aiding future infant gut virome research.

The establishment of the gut microbiome (GM) during the first years of life plays a pivotal role in the maturation of the infant immune system[1,2]. Early-life GM dysbiosis has been linked to a series of chronic diseases occurring later in life, indicative of a lasting effect on immune programming[3–6]. Most existing research has focused on the bacterial component of the GM, but lately it has become evident that viruses are prominent GM members. Recent studies have shown that the transfer of gut viral content from healthy donors can cure recurrent *Clostridioides difficile* infections[7], alleviate diet induced obesity[8] and prevent necrotizing enterocolitis in preterm neonates[9]. The mechanisms are still unclear, but probably involve modulation of GM composition through viral infection, because most gut viruses are bacteriophages (phages) that only infect bacteria[10].

A full list of affiliations appears at the end of the paper. ✉e-mail: shiraz.shah@dbac.dk; dn@food.ku.dk

Phages, like bacteria, appear in the gut during the first months of life following a host-specific pattern[11–14]. Virulent phages undergo the lytic cycle in which they readily multiply and kill their host cell through lysis and release new virions into the ecosystem. Temperate phages can integrate into the bacterial chromosome, thereby becoming prophages. This prophage status postpones the killing of the host until certain environmental conditions induce the prophage to enter the lytic cycle. Some phages can also cause chronic infections leading to continuous shedding of viral particles[15]. Bacteria will defend themselves against these viruses using multiple defence systems[16], including clustered regularly interspaced short palindromic repeat (CRISPR)–Cas systems, an adaptive immune mechanism where DNA records (spacers) of past viral infections are stored on a chromosomal CRISPR array to combat future phage attacks[17].

Phages can alter GM composition and function[8,12], but may also directly elicit immune responses from the human host[18–20], suggesting a tripartite interaction that could modulate host health. The first report on the viral metagenome (virome) composition in the infant gut dates back more than a decade[21], and the infant virome has recently been shown to be influenced by caesarean section and formula milk[22]. Nevertheless large-scale studies establishing the early life virome composition and structure are sparse, and human virome studies in general have been challenged by the large proportion of uncharted viral diversity, which is sometimes referred to as the viral 'dark matter' problem[23].

The latter means that only a small fraction of nucleic acid sequences in a virome can be linked to any known virus. Attempts at de novo virus identification have been limited by the lack of universal viral marker genes, while de novo classification of novel viruses into taxa was hampered by the lack of standardized methods. However, progress has been made in recent years[24–26], leading to several human gut virus databases[27–29], although these are still developing and currently lack viral taxonomies for all the novel viruses they contain. Comprehensive viral taxonomies are important for conducting biologically meaningful statistical analyses against sample metadata.

Traditionally, defining new viral taxa has required laboratory isolation of both virus and host for subsequent characterization[30]. However, the International Committee for the Taxonomy of Viruses (ICTV) has recently made it possible to define new viral taxa on the basis of sequence information alone. This important change is already having major implications as several new taxa are being proposed, particularly among the highly diverse tailed phages (caudoviruses)[31]. Notably, the ICTV established the complete taxonomy of the new *Herelleviridae* family, demonstrating the definition of viral families, subfamilies and genera according to this new paradigm[32]. Subsequently, three new caudoviral families were identified in human gut metagenome data[33]. And recently, the prominent gut phage family *Crassviridae*[34] was elevated into a viral order *Crassvirales*[35], belonging to the new viral class *Caudoviricetes*, which itself is now proposed to encompass caudoviruses[36] as a whole.

In this Resource, we characterized the faecal viromes of 647 infants at 1 year of age enrolled in the Copenhagen Prospective Studies on Asthma in Childhood 2010 (COPSAC2010) cohort[37]. De novo assembly and careful curation allowed us to map out any uncharted viral diversity leading to the identification of hundreds of virus family-level clades (VFCs). In contrast to the adult gut dominated by virulent *Crassvirales*, we found a diverse and largely temperate infant gut virome.

Results

Study population

COPSAC2010 is a population-based mother–child cohort study of 700 Danish children from rural, suburban and urban settings around the greater Copenhagen area (Supplementary Table 1). Participants were recruited in pregnancy with the aim of prospectively studying the causes for chronic inflammatory diseases[37]. Faecal samples were successfully collected and had viromes characterized for 647 children at 1 year of age. Metagenomes were sequenced in parallel[38].

Identifying the viruses and resolving their taxonomies

Virome extractions are known to contain various amounts of bacterial contaminating DNA[39] and uncharted viral diversity[23] makes it difficult to discern novel viruses from contaminants. We resolved this issue by assembly, clustering and successive rounds of manual curation (Methods), to avoid potential selection biases (for details, see Supplementary Information and Supplementary Table 2) in existing tools and criteria such as 'circular contigs'[33] that could have prevented the identification of truly novel viral clades.

In short, the extracted viromes were sequenced to an average of 3 Gbp per infant sample. After assembly and species-level de-duplication, resulting operational taxonomic units (OTUs) were clustered by protein content (Extended Data Fig. 1), visualised (Extended Data Fig. 2) and manually curated. Ultimately, 10,021 manually confirmed viral OTUs (vOTUs) comprised the study's final set of viral species (for details, see Supplementary Information and Online Methods). These vOTUs recruited roughly half of total sequencing reads, with the remaining half mapping mainly to sequence clusters of bacterial contaminating DNA (Supplementary Information and Extended Data Figs. 3–5), which is comparable to other studies[40]. Contaminant sequence clusters were not analysed further.

For determining which vOTUs were parts of existing viral families, we pooled them with 7,705 species-level de-duplicated reference phages[41]. After gene calling, protein alignments were used for defining viral orthologue gene clusters (VOGs) de novo and for constructing an aggregate protein similarity (APS) tree. The tree was rooted and cut at levels reproducing the recent taxonomy for the *Herelleviridae*[32] phage family, thus yielding clusters corresponding to viral families (VFCs), subfamilies and genera covering both vOTUs and reference phages. An additional order-level cutoff was based on the newly proposed caudoviral *Crassvirales* order[35].

The 10,021 species-level vOTUs fell within 248 curated VFCs, including 16 known families (Fig. 1) containing 2,497 vOTUs and 232 previously undescribed VFCs containing 7524 vOTUs. The undescribed VFCs were named after the infants that delivered the faecal samples. The VFCs were additionally grouped into 17 virus order-level clusters (VOCs, Supplementary Table 3), 5 of which were already known (Fig. 1). After estimating the typical complete genome size at the family level (Fig. 1), 56 % of the 10,021 vOTUs were found to be complete or near complete, specifically, 83% of the 2,629 small single-stranded DNA (ssDNA) vOTUs and 46% of the 7,392 larger double-stranded DNA (dsDNA) vOTUs. vOTU DNA sequences and taxonomies along with visualizations of the VFCs (Extended Data Fig. 2) have been made available via an interactive Fig. 1 at http://copsac.com/earlyvir/f1y/fig1.svg.

Infant gut vOTUs are largely absent from gut virus databases

We checked whether any of our 10,021 curated viral species were found within three gut virus databases built mainly on adult faecal metagenomic data. The Gut Virome Database (GVD)[29] contained only 819 of our vOTUs, while the larger and more recent Gut Phage Database (GPD)[28] and Metagenomic Gut Virus catalogue (MGV)[27] covered 2,307 and 2,171 vOTUs, respectively. Combined, 7,046 (70%) of the infant gut vOTUs identified here were not found in any of the three gut virus databases. At the family level, however, most of the 248 VFCs had some representatives in either database, with *Crassvirales* VFCs being particularly well represented in both GPD and MGV. Importantly, the majority of our most species-rich VFCs (for example, candidate family 'Amandaviridae') were poorly represented in all three databases, while the VFCs best covered by the databases were often minor in our data (Fig. 1). In other words, most large gut phage clades in databases are only occasionally found in the infant gut viromes, and vice versa. This pattern suggests that the infant gut is a unique niche harbouring specialized viruses distinct from the adult gut. Alternative explanations for this lack of overlap could be library selection differences (bonafide viromics in our case versus

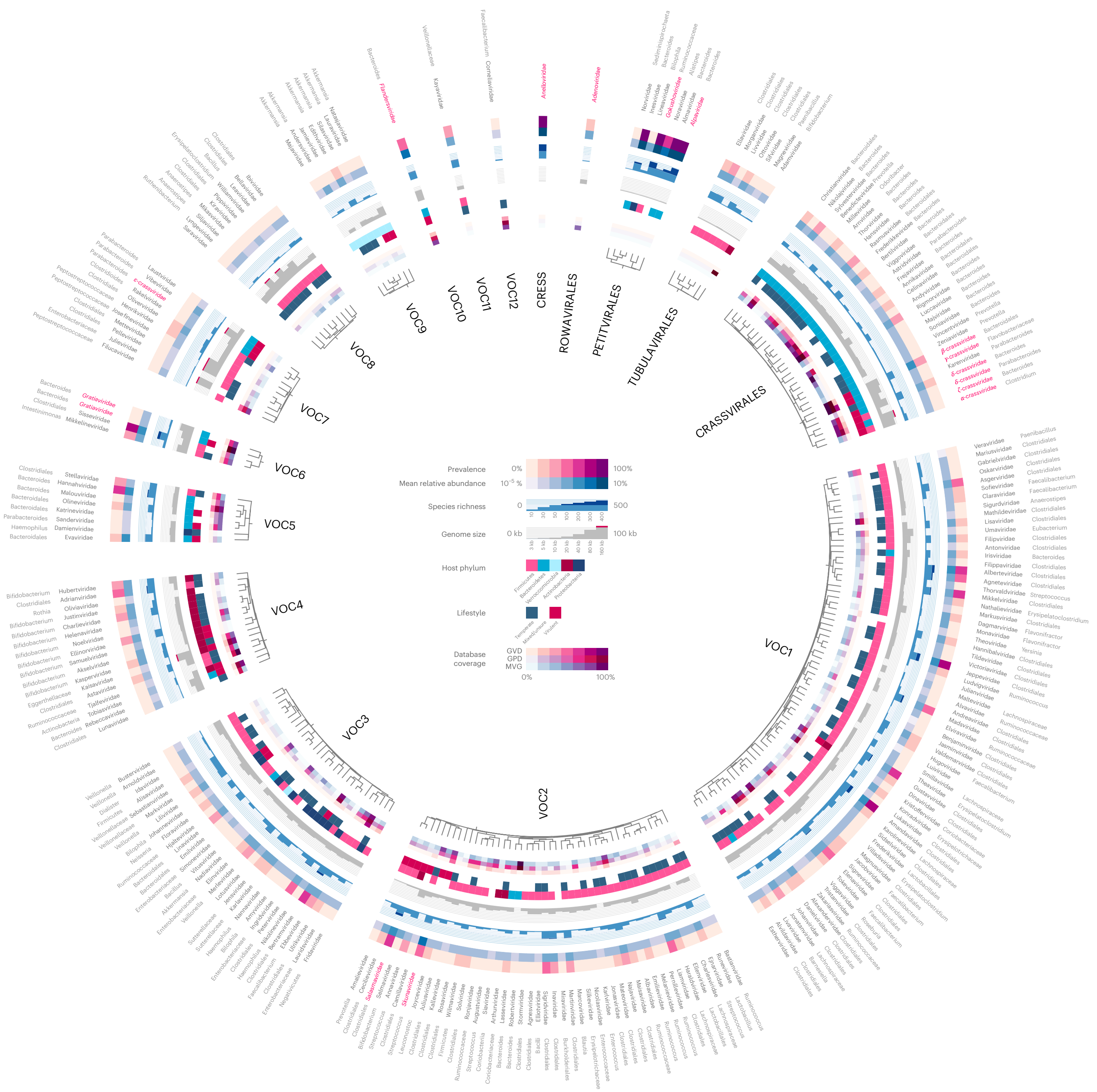


Fig. 1 | An atlas of infant gut DNA virus diversity. Faecal viromes from 647 infants at age 1 year were deeply sequenced, assembled and curated, resulting in the identification of 10,021 viral species falling within 248 VFCs. Predicted host ranges for each VFC are given, and the VFCs have been grouped into 17 VOCs. Trees show how VFCs are interrelated within each VOC, and heat maps and histograms encode their genome size, lifestyle, host range, abundance and prevalence across the cohort as well as in published gut virus databases. For the 16 previously known viral families, names are written in red. An interactive version of the figure with expandable families can be accessed online, for browsing the gene contents and downloading the genome of each virus: http://copsac.com/earlyvir/f1y/fig1.svg.

bulk metagenomics), bioinformatics (curation versus automated detection), limited infant gut sequence diversity (enabling complete assembly of otherwise rare phages) or the fact that gut viromes are extremely individual specific by nature.

Undescribed viral families dominate the infant gut virome

Cutting the APS tree at the family[32] and order level[35] yielded 248 VFCs and 17 VOCs. The family-level cutoff reproduced the recently defined crAssphage families[35] (Fig. 1). The order-level cutoff reproduced five known viral orders (that is, *Petitvirales*, *Tubulavirales*, anelloviruses (CRESS), *Rowavirales* and *Crassvirales*) along with 12 additional strictly caudoviral VOCs (Supplementary Table 3). Even at the family level, 232 out of 248 VFCs were caudoviral, further underlining their diversity. The mean and median VFC size was 40 and 17 species-level vOTUs, respectively, making the typical VFC similar in richness to currently known gut phage families such as *Flandersviridae*[33].

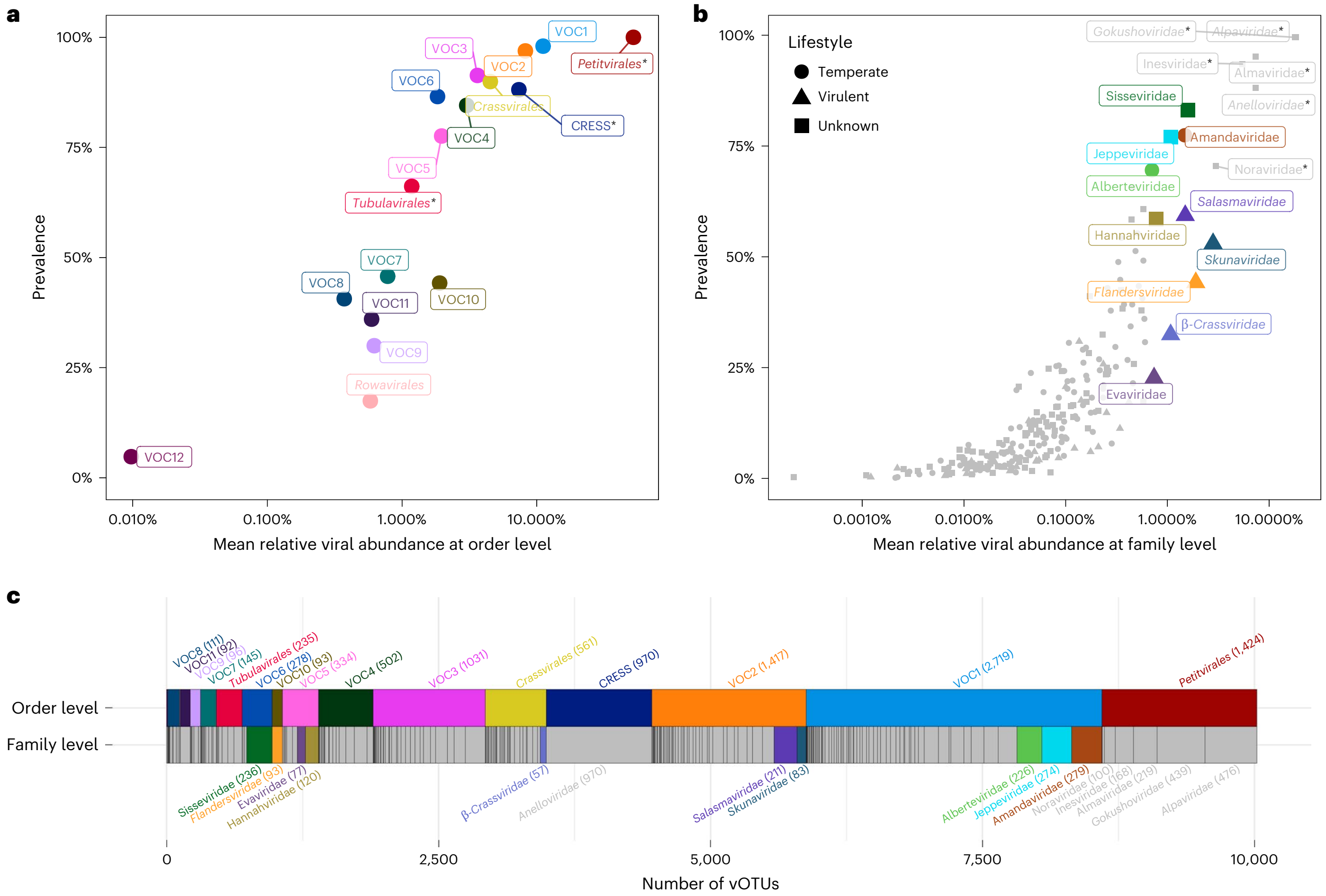


Fig. 2 | Abundance, prevalence and richness of the viral clades in the 1-year-old infant gut. Already-known viral clades are indicated in italics. ssDNA clades have been marked with a star as their abundances may be inflated from amplification bias. **a**, Prevalence and MRA of the 17 VOCs across samples. **b**, Prevalence and MRA of the 248 VFCs. The major VFCs were defined as the ten most abundant caudoviral VFCs in the data, and are coloured and labelled. Minor VFCs as well as ssDNA families are in grey. Predicted lifestyles for the ten major VFCs are indicated by different shapes. **c**, VOCs and VFCs scaled by species richness, ordered by MRA. VOC12 and Rowavirales are not shown due to their small sizes. The VFCs are represented underneath the VOCs they belong to. Clade prevalence, abundance and species richness are highly interrelated, and several previously undescribed clades outnumber crAssphage in the infant gut.

To identify the most predominant viral clades, three measures were calculated: total species richness, prevalence across samples and mean relative abundance (MRA) (Fig. 2). Family- and order-level MRA and prevalence estimates were determined by first mapping sample reads to vOTUs, then aggregating their counts on the basis of taxonomic affiliation. All three measures were highly correlated (Fig. 2 and Extended Data Fig. 6), meaning that the most diverse VFCs and VOCs were also the most widespread and abundant. The correlation between these measures is predicted by the neutral community model, which also applies to bacterial community structures[42,43].

In our data, vertebrate-infecting ssDNA anelloviruses (*Anelloviridae*) and bacterial ssDNA microviruses (*Petitvirales*) were amongst the most abundant viral clades (Fig. 2a and next subsection). These were followed by ten major dsDNA VFCs belonging to the *Caudoviricetes* viral class (Fig. 2b). Four of these are known caudoviral families pending ICTV approval, namely *Skunaviridae*, *Salasmaviridae*, *β-crassviridae* and *Flandersviridae*, while the remaining six comprise new candidate families. Importantly, *Crassvirales*, which are abundant in adult faecal viromes[44], were surpassed by other VOCs in the infant gut (Fig. 2a).

Skunaviridae is a family of virulent phages infecting *Lactococcus* dairy cultures[45]. Possibly originating from the diet, they were the most abundant caudoviral family in our data (2.7% MRA). *Salasmaviridae* is a viral family harbouring around a dozen *Bacillus* podoviral species including phage phi29[46]. Here, we were able to broaden the scope of the *Salasmaviridae* family with over 200 diverse vOTUs spanning more than 20 viral subfamilies, infecting a wide variety of gut-associated *Firmicutes* and *Actinobacteria*. *β-Crassviridae*, a minor *Crassvirales* family in adults, was found in almost a third of the infants ($n = 210$; 647), predicted to infect both *Bacteroides* and *Clostridiales* hosts. The major adult *Crassvirales* family, *α-Crassviridae*[35,47], however, was present in only 5% ($n = 39$) of the infants. *Flandersviridae* is a *Bacteroides*-infecting phage family recently defined on the basis of 30 complete phage genomes[33] from public metagenome assemblies. Found in almost half of the children ($n = 286$), we markedly expand this family with 80 complete species-level vOTUs spanning four subfamilies.

Apart from these four known virulent viral families, six previously undescribed candidate families were found to be highly abundant, prevalent and diverse. The prevalence and richness estimates for these candidate families indicate that they are at least as predominant in the infant gut ecosystem as crAssphage is in adults[44]. Candidate family 'Sisseviridae', almost universally present in the infants (80%), harbours the highly prevalent *Faecalibacterium* phage Oengus[48] and encompasses a wide range of both temperate and virulent vOTUs infecting diverse *Firmicutes* and *Actinobacteria*. The temperate candidate families 'Amandaviridae', 'Jeppeviridae' and 'Alberteviridae' are related, belonging to the major VOC1. These candidate families were present in 70% of the infants, containing between 200 and 300 viral species each, infecting *Clostridiales* genera such as *Ruminococcus*, *Blautia*, *Anaerostipes* and *Hungatella*. Apart from a few unclassified *Clostridium* and *Brevibacillus* reference phage species that co-cluster within them, these expansive clades are largely unexplored. Finally, 'Evaviridae' and 'Hannahviridae' comprise two related candidate families of *Bacteroides*-infecting phages containing around 200 species in total. The former appears strictly virulent while the latter harbours separate subfamilies that are either virulent or temperate. 'Hannahviridae' includes the recently described *Bacteroides* phage 'Hankyphage'[49] known for its diversity-generating retroelements, and it has been extensively described in a parallel provirome study performed on the same samples[50].

Clades of ssDNA viruses in the infant gut

ssDNA vOTUs recruited around a third of the sequencing reads, but after normalizing for their short genome sizes, they accounted for 60% of the MRA (Extended Data Fig. 3). The short multiple-displacement amplification (sMDA) protocol used to detect the ssDNA viruses could have inflated their counts[51]. However, the families did still display canonical positioning along the neutral community model (Figs. 2b and 3f) so we infer that any artificial inflation would have been limited. The ssDNA families fell within three separate viral classes, *Malgrandeviricetes*, CRESS viruses and *Faserviricetes*, each harbouring a single viral order.

Microviruses of the *Petitvirales* viral order (class *Malgrandeviricetes*) are ubiquitous small icosahedral ssDNA phages and were the most prevalent and abundant group of viruses in our viromes, making up 52% of the MRA. Twenty-one per cent of all CRISPR spacer matches from the metagenome targeted microviruses (http://copsac.com/earlyvir/f1y/taxtable.html), underlining their importance. vOTUs from the two major families *Gokushoviridae* and *Alpaviridae* (currently subfamilies *Gokushovirinae* and *Alpavirinae*) in our data are predicted to infect *Clostridiales* and *Bacteroidales,* respectively, but other minor VFCs were also detected (Fig. 1).

Anelloviruses from the CRESS class of ssDNA viruses, also known as Torque Teno viruses, comprise a single family (*Anelloviridae*) of small 3 kb ssDNA viruses that infect vertebrate cells. They cause chronic asymptomatic infections in healthy humans, with elevated titre in immunocompromised patients[52]. The immature infant immune response may explain their abundance in our samples (7% of the MRA). They comprise by far the richest single family with 970 species-level vOTUs. On average, each infant harboured ten species of *Anelloviridae*, consistent with earlier findings[13]. Unsurprisingly, no CRISPR spacer matched any *Anelloviridae* vOTUs.

Inoviruses from the *Tubulavirales* order (class *Faserviricetes*) are a ubiquitous and diverse group of filamentous phages with small ssDNA genomes[53]. Some integrate into their host genomes using integrases while others cause chronic non-lethal infections that result in the continuous shedding of viral particles[15]. Although they were diverse in our data, distributed among seven families like the *Petitvirales*, their species richness was lower at 235 vOTUs, and abundances were correspondingly lower totalling 1% MRA. Most of the inoviral families found were predicted to infect *Clostridiales*, although members of the VFC 'Adamviridae', appear to specifically infect *Bifidobacterium* (Fig. 1).

Virus lifestyle determines both abundance and prevalence

Most of the ten major caudoviral VFCs lacked integrases, otherwise commonly found in less abundant VFCs. Since an integrase is an indicator of a temperate lifestyle, we investigated whether a virulent lifestyle was linked to higher abundances. First, the typical complete genome size per VFC was determined for 228 VFCs by examining the size distribution of their constituent vOTUs. The median (interquartile range (IQR)) complete genome size for the VFCs was 35 kb (30–50 kb). Using the determined minimum complete size limit per viral family (Fig. 1), 5,608 vOTUs with complete and near-complete genomes were screened for integrases (Methods). Phage lifestyles were mostly homogeneous at the family level and a total of 118 VFCs were deemed temperate, while only 33 were found to be virulent. The remaining 97 VFCs exhibited either a mixed lifestyle pattern or were uncertain due to an insufficient number of complete genomes.

Family-level abundance was not significantly linked to phage lifestyle (two-sided Wilcoxon test, $P = 0.90$; Fig. 3a), but temperate VFCs were significantly more prevalent than virulent VFCs ($P = 0.048$; Fig. 3b). Temperate phages have been shown to be more genetically diverse than their virulent counterparts[54], so we compared the amount of unique branch length (as a fraction of total branch length) in virulent versus temperate family-level APS subtrees. Indeed, temperate caudoviral VFCs were more genetically diverse ($P = 0.021$; Fig. 3c) than virulent VFCs. *Clostridiales* hosts were particularly enriched in temperate VFCs, whereas most virulent VFCs were predicted to infect *Bacteroidales* (Fig. 1). According to our CRISPR spacer mappings, and in line with other studies[28,55], some vOTUs appeared to infect multiple host species, genera or even families of bacteria. We checked whether spacers targeted virulent phages more often than temperate phages, or whether a virulent lifestyle was associated with a broader host range. This was not the case as both temperate and virulent families exhibited similar mean host ranges ($P = 0.2$; Fig. 3d) and numbers of targeting spacers ($P = 0.097$; Fig. 3e).

Finally, plotting the abundance and prevalence of the virulent and temperate VFCs against each other (Fig. 3f) suggested that virulent VFCs had elevated titre despite being found in fewer children. We tested this hypothesis systematically using the neutral community model (Fig. 3g), which describes the community relationship between abundance and prevalence[56]. After fitting the model on all of our VFC abundances, virulent VFCs had significantly lower residuals against it than temperate VFCs (two-sided Wilcoxon test, $P = 2.1 \times 10^{-5}$; Fig. 3h), confirming that they were both less prevalent and more abundant than temperate VFCs.

Phage–host abundances are linked in spite of virus lifestyle

Bacterial hosts for the vOTUs were predicted using 317,968 CRISPR spacers from our metagenome assembled genomes (MAGs)[38], 11 million spacers from the CRISPR spacer database[57] and using WIsH[58]. These predictions were merged by their last common ancestor. Bacterial host genera were predicted for 63% of the vOTUs, with 77% being covered at the order level (Fig. 4a), and 79% at the host phylum level. *Bacteroides* was by far the most commonly predicted host genus followed by *Faecalibacterium* and *Bifidobacterium*. At the order level, however, approximately half of the annotated vOTUs had *Clostridiales* as hosts, with *Bacteroidales* covering just one-quarter (Fig. 4a). This mirrors the corresponding pattern for the bacterial taxa in the metagenomes, where *Bacteroides* was the most abundant genus, while *Clostridiales* were more diverse (Fig. 4b).

MRAs of bacterial host genera in the metagenomes were strongly correlated with cognate phage MRAs in the viromes (Spearman's $\rho = 0.76$, $P < 1.45 \times 10^{-17}$; Fig. 4c) supporting both the accuracies of host predictions and viral abundance estimations. Overall, in the infant gut, both virulent and temperate phages correlate positively with the abundances of their hosts (Extended Data Figs. 7 and 8). Although virulent phages lyse their hosts, cross-sectionally they still act as positive markers for their presence.

Discussion

The recent publication of several large and curated gut virus databases illustrates the massive diversity of the human gut viral community[27–29]. Yet, significant parts of this ecological niche remain uncharacterized. A thorough description of the gut viruses is essential to understand their roles, especially if one aims at modulating the GM for prevention and treatment of chronic disease. We deeply sequenced 647 infant gut viromes and mapped the uncharted viral diversity by de novo

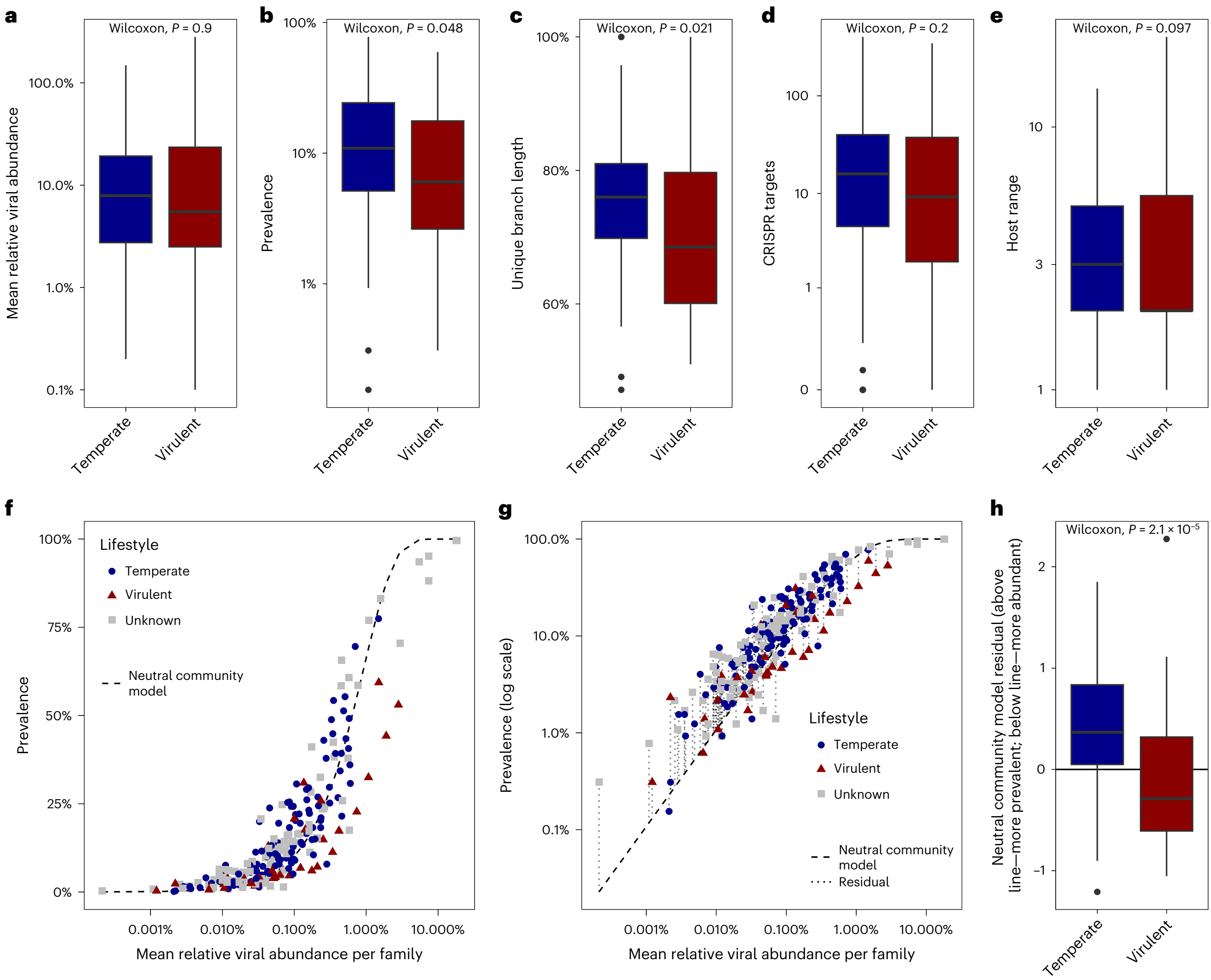
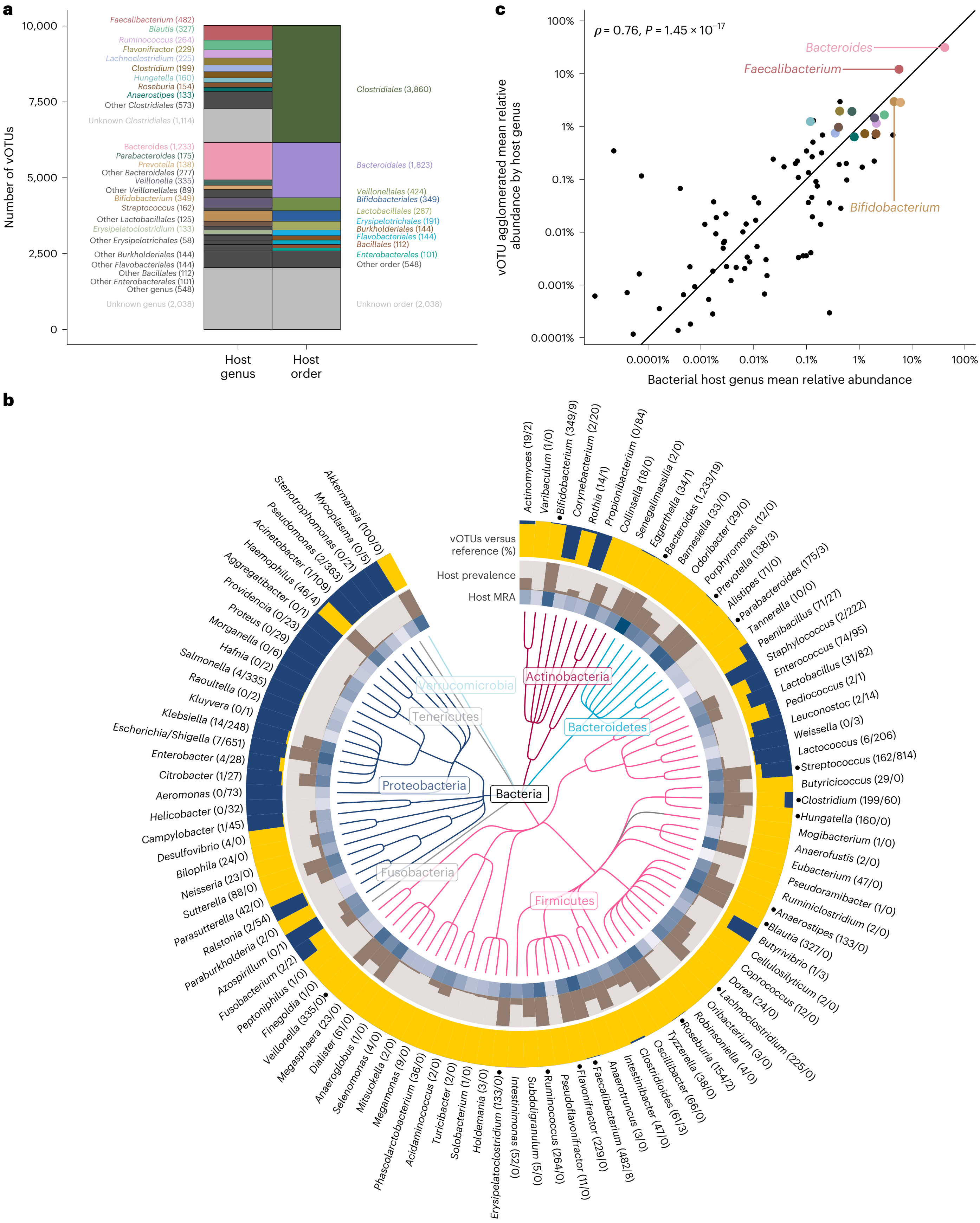


Fig. 3 | Temperate versus virulent viral families in the infant gut. **a**–**e**, Characteristics of temperate versus virulent VFCs in the data in terms of MRA (**a**), prevalence (**b**), genetic diversity as measured by unique branch length (**c**), number of metagenomic CRISPR spacer matches (**d**) and host range (number of host species) (**e**). **f**, Fit of the neutral community model, on the VFCs from Fig. 2b. **g**, Deriving neutral community model residuals from the log-transformed prevalences. **h**, Comparison of neutral community model residuals, showing that temperate VFCs tend to have positive residuals, whereas virulent VFCs tend towards negative residuals, indicating that temperate phages are present in lower abundance despite being found in more children, as compared with virulent phages. For **a**–**e** and **h**, $n = 151$ (118 temperate + 33 virulent). Box plot elements: centre line, median; box limits, upper and lower quartiles; whiskers, 1.5× IQR; points, outliers. Two-sided Wilcoxon test P values reported. For **f** and **g**, $n = 248$ (118 temperate + 33 virulent + 97 unknown).

assembly and classification. The approach led to the uncovering of 248 VFCs, 232 of which were previously unknown, and most of which belonged to the *Caudoviricetes* viral class. Temperate phages dominate the 1-year-old gut virome, and crAssphage is overshadowed by several previously undescribed viral clades. Such comprehensive taxonomic resolution of virome data allows for biologically meaningful statistical analyses against sample metadata, aiding future research in translational viromics.

Systematically resolving the uncharted viral diversity ('dark matter') left only 7% of the virome sequences unaccounted (Extended Data Fig. 3) and the VFCs that were uncovered in the process represent a major expansion of current phage taxonomy. Resolution of phage lifestyles showed that most phages in the infant gut ecosystem are temperate, even though the less diverse virulent phages can be more abundant. This echoes recent findings on the neonatal gut[12], also dominated by temperate phages, and it is in contrast to adults where virulent phages dominate[29].

In addition to the six major candidate families described, numerous additional predominant caudoviral VFCs can be browsed online (Fig. 1). In general, *Bacteroides*-infecting VFCs were more often virulent and host specific, while VFCs infecting *Clostridiales* featured wider host ranges and were overwhelmingly temperate. Multiple VFCs were often specialized for a single host genus, for example, seven *Akkermansia*-specific VFCs (Fig. 1). Others were more agnostic, having multiple host genera, for example, *Clostidiales*-infecting VFCs such as 'Amandaviridae'. Some vOTUs were even predicted to infect multiple bacterial families within the same order. Such features underscore the rapid rates of speciation that caudoviruses attain both horizontally across hosts, but also vertically within tight host niches. Phylogenetically distinct hosts such as *Bacteroides* and *Akkermansia*[59] probably present greater barriers for



host switching, making their phage families more host specific in a human gut context. This is in contrast to *Clostridiales* genera, dozens of which often co-exist, encouraging host flexibility. Overall, we found caudoviral richness to exceed host richness by an order of magnitude, both at the species and genus levels (for example, 2,858 caudoviral genera versus 203 host genera in the metagenome).

Fig. 4 | Phages and their bacterial hosts in the 1-year-old infant gut. Prediction of bacterial hosts for the 10,021 vOTUs found in the infant gut virome shows that ***Bacteroides, Faecalibacterium and Bifidobacterium*** are the three most prominent host genera. **a**, Distribution of virus host predictions collapsed to bacterial order and genus levels, respectively. Numbers in parentheses denote the number of vOTUs with a given host genus or order, respectively. **b**, The top 100 gut bacterial genera found in gut metagenomes from the same infant faecal samples, as represented by a taxonomic tree. The MRA of each bacterial genus is shown in the blue heat map, while the fraction of the 647 infants harbouring the host genus (that is its prevalence) is shown with the brown bar plot. The outer ring displays per bacterial genus, the proportion of infant gut vOTUs (yellow) relative to reference phage species with known hosts[41] (dark blue). Numbers behind each genus name denote the total number of vOTUs versus reference phage species per bacterial host genus. The 16 major host genera from **a** are indicated by a dot in front of their names in **b**. **c**, Each dot represents a genus from **b**, by its MRA in the metagenome against the aggregate MRA of all its vOTUs in the virome. Host abundances correlated strongly with corresponding phage abundances as tested by a Spearman's rank test (two-sided *P* value).

Most virome studies so far amplified the extracted DNA with MDA before sequencing, which may bias sequence composition towards ssDNA viruses[60,61] in addition to compromising quantitative analyses overall. However, the largest meta-analysis of virome studies[29] did not find differences between non-MDA and standard 2 h MDA gut viromes. Furthermore, in a recent gut virome study using a different DNA library kit enabling ssDNA detection supposedly without biases[61], microviruses outnumbered caudoviruses in a third of the samples[62]. Here, we used a 30 min sMDA step to enable ssDNA detection while limiting biases. We found the opposite trend; that microviruses outnumbered caudoviruses in two-thirds of the infants. But we also showed strong co-abundances between phages and their hosts. Moreover, we made a thorough comparison linking plaque forming units to virome abundances (Extended Data Fig. 9). We conclude that our results on viral abundances are relevant in quantitative terms despite using sMDA, at least for dsDNA viruses.

Skunaviridae, our most abundant caudoviral family, comprised only eight complete vOTUs in the dataset. This is atypical considering the hundreds of vOTUs in most of the other abundant viral families. All reference phages belonging to the family infect *Lactococcus* while our vOTUs were predicted to infect *Streptococcus*, but this could be an artefact caused by the lack of CRISPR-Cas systems in *Lactococcus*[63]. *Streptococcus*, although very prevalent in the children, may not have been abundant enough to support the high counts of virulent *Skunaviridae*. We also did not find any strong correlation between *Skunaviridae* and *Streptococcus* or *Lactococcus* in the data. Thus, it remains a possibility that these strictly virulent phages were ingested via fermented dairy products where they naturally occur, as previously proposed[64].

In a previous study on *Escherichia coli* phages isolated from the same samples[65], virulent coliphages were less prevalent but more abundant and had broader host ranges than temperate coliphages. Here we found the same pattern on a more global scale. Virulent phage families across diverse hosts were more abundant but less prevalent than temperate phage families. Although we found no difference in host ranges, temperate phage families were more genetically diverse compared with the virulent ones. The higher prevalence and lower abundance of temperate phages probably reflects frequent prophage induction, as shown in mouse models[66–68], and that induced virions do not readily re-infect and multiply. In viromics, this would appear as a stable background of diverse temperate phages on top of which virulent blooms would stochastically appear from random phage–host encounters. For our infant samples, this temperate background was intense enough to overshadow the diversity of virulent phages. Possibly, in adult viromes where the GM and host immunity have stabilized, the bacteria are less stressed and the temperate virome, in turn, less dominant. This notion is consistent with how a virulent phage core is linked to adult gut health[69], as well as the paucity of crAssphage in infant viromes[29].

Methods

The COPSAC2010 cohort

The study was embedded in the Danish population-based COPSAC2010 prospective mother–child cohort of 736 women and their children followed from week 24 of pregnancy, with the aim of studying the mechanisms underlying chronic inflammatory diseases[37] (Supplementary Table 1). The study was conducted in accordance with the guiding principles of the Declaration of Helsinki and was approved by The National Committee on Health Research Ethics (H-B-2008-093) and the Danish Data Protection Agency (2015-41-3696). Both parents gave written informed consent before enrollment. Faecal samples were collected for 660 participants at age 1 year.

Virome extraction

Each sample was mixed with 10% vol/vol glycerol and stored at −80 °C until DNA extraction for metagenomes[38] and virome extraction. Extraction and sequencing of viromes were done using previously described protocols[70]. Briefly, DNA from faecal filtrates enriched in viral particles was extracted and subjected to short (30 min) MDA amplification and libraries were prepared following the manufacturer's procedures for the Illumina Nextera XT kit (FC-131-1096). For epiflorescence virus-like particle (VLP) estimations, 10 µl of a virome sample was diluted 100-fold, fixed and deposited on a 0.02 µm filter, dried and stained with SYBR-Gold (200×), then visualized with an epifluorescence microscope using a 475 nm laser. VLPs were counted in eight to ten fields and multiplied over the remaining filter surface area.

Sequencing, assembly and decontamination

Virome libraries were sequenced on the Illumina HiSeq X platform to an average depth of 3 Gb per sample with paired-end 2× 150 bp reads. Satisfactory sequencing results were obtained for 647 out of 660 samples. Virome reads were quality filtered and trimmed using Fastq Quality Trimmer/Filter v0.0.14 (options -Q 33 -t 13 -l 32 -p 90 -q 13), and residual Illumina adaptors were removed using cutadapt (v2.0). Trimmed reads were de-replicated using the VSEARCH[71] (v2.4.3) derep_prefix and then assembled with Spades[72] (v3.10.1) using the meta flag while disabling read error correction. Decontamination clusters were generated by reducing redundancy by de-duplicating the 1.5 M contigs above 1 kb in size into 267k 90% ANI representatives using a previously published pipeline[73] then calling genes using Prodigal[74] (v2.6.3) and aligning proteins all-against-all using FASTA[75] (v36.3.6f) for building an APS tree[76] using custom code (https://github.com/shiraz-shah/VFCs). The tree was cut close to the root to obtain the decontamination clusters. Bacterial MAGs from the same samples[38] were mined for CRISPR spacers using CRISPRDetect[77] (v2.2), and the virome decontamination clusters were ranked by their extent of CRISPR targeting multiplied by sample prevalence. The protein alignment results were passed through an orthology filter[78] (https://github.com/shiraz-shah/VFCs) and clustered using Markov clustering[79] (v14-137) to obtain VOGs de novo. VOGs were used to visualize the gene contents of contigs within each decontamination cluster. The top 400 ranked clusters were inspected visually for two viral signatures, namely conservation of contig sizes and of gene content. There were diminishing returns beyond the top 400 mark and the remaining decontamination clusters were assumed to represent contaminants.

OTU delineation and protein annotation

Species-level (95% ANI) de-duplication of contigs into OTUs was done using BLAT[80] and custom code for clustering (https://github.

com/shiraz-shah/VFCs). Reference phages were de-duplicated to the species-level using the same strategy. Comparisons of the vOTUs to the GVD, GPD and MGV were also performed similarly. Decontaminated vOTUs and reference phage species[41] were pooled and the APS tree and VOGs were recomputed. Multiple sequence alignments (MSAs) of VOGs were constructed with MUSCLE[81] v3.8.425. VOG MSAs were aligned against MSAs from Pfam[82], the Conserved Domains Database[83], the Clusters of Orthologous Groups of proteins database[84] and TIGRFAMs[85] using HH-suite3 (ref. 86) v3.0-beta.3 to gain functional annotations.

Resolution of viral taxonomy

We first used FigTree (v1.4.4) to root the APS tree by selecting an outgroup that branched out directly from the stem of the tree. Next we used phylotreelib and treetool (https://github.com/agormp/phylotreelib) to generate viral genera, subfamilies, VFCs and VOCs as follows. First, treetool's cladeinfo option was used to retrieve the distances from the root to the branch points corresponding to existing phage genera, subfamilies, families and orders[32,35]. Next, treetool.py's–clustcut option was used to cut the rooted APS tree at the above distances in order to obtain clades of both vOTUs and reference phages corresponding to viral genera, subfamilies, families and orders. The distances we used to cut the tree were 0.250, 0.125, 0.04 and 0.025, respectively, corresponding to average amino-acid identity (AAI) and coverage thresholds of 70%, 50%, 28% and 22% for each respective taxonomic level.

Curation of VFCs

Viral families from above were visualized (Extended Data Fig. 2) to (1) further curate each individual member vOTU to separate confirmable viruses that had structural VOGs, from subclades of vOTUs representing various virus-related MGEs, such as satellites, that did not harbour genes coding for typical structural proteins. (2) The OTU length distribution within each family was inspected and then plotted in a histogram with 5 kb steps to locate the right-most size peak. The 5 kb step immediately preceding this peak was set as the lower size bound for a complete or near-complete genome. (3) The family visualizations were inspected to manually remove families that were dominated by reference phages, so as to avoid interference with ongoing classification efforts. Weak families composed mainly of MGEs or fragments, having fewer than five vOTUs or fewer than two complete vOTUs were also removed. For the final version of the family visualizations available online, VOG MSAs were realigned against MSAs from PHROGs[87] because this database was more informative than Pfam, Conserved Domains Database, Clusters of Orthologous Groups of proteins database and TIGRAMs.

Host prediction

MAG spacers, along with spacers from CRISPRopenDB[57] and WIsH[57] (v1.0) were used to generate separate host predictions for each vOTU. The three predictions were integrated using the last common ancestor of the two most closely matching predictions, as an error-correction strategy, since all three methods would occasionally mispredict. Bacterial genus abundances in the metagenome were derived by running mOTUs[88] (v2) on the reads from each sample followed by aggregating mOTU abundances at the genus-level in R (v4.0.2) using phyloseq[89] (v1.41.1).

Abundance estimation

Bacterial contamination was estimated for each virome sample using ViromeQC[40] (v1.0) along with a custom approach where we leveraged the metagenomes cognate to each virome: Reads were mapped from both fractions against the 16S rRNA gene[90] and *cpn60* (ref. 91) and the degree of contamination was calculated as the ratio between the two fractions. Abundances of vOTUs in each sample were determined by mapping sample reads to sample contigs using the Burrows–Wheeler Aligner[92] (v0.7.17-r1188) with the option mem -a, then using the msamtools (v0.9.6) profile to determine depth and length-normalized relative abundances with iterative redistribution of ambiguously mapped reads proportionally to uniquely mapped reads (https://github.com/arumugamlab/msamtools). The obtained contig abundances were then aggregated at the OTU level using custom code (https://github.com/shiraz-shah/VFCs) to obtain vOTU abundances per sample. vOTU abundances were aggregated at the family and order levels in R (v4.0.2) using phyloseq[89] (v1.41.1) to obtain the statistics used for Figs. 2 and 3.

Phage lifestyle prediction

A list of VOGs matching to integrase and large serine recombinase protein families was first curated, then used to predict whether complete vOTUs within viral families were temperate or virulent. Families where more than 95% of complete vOTUs did not harbour an integrase were deemed virulent, whereas for temperate families at least 50% of both complete and incomplete vOTUs were required to carry an integrase.

Benchmarking

The versions of virus discovery tools used for benchmarking (Supplementary Table 2) were DeepVirFinder (v1.0), VIBRANT (v1.2.1), VIRSorter (1.0.6), VIRSorter2 (v2.0 commit 22f6a7d), Seeker (commit 9ae1488), PPR-Meta (v1.1) and CheckV (v.0.7.0). The random prediction was created by randomly sampling the 362,668 OTUs 12,500 times without replacement. The number 12,500 was chosen because it was reasonably close to our own positive set and the number of positives generated by most tools.

Figures and statistical analysis

Figure 1 was drawn by first collating data at the family level using phyloseq[89] then using Circos v0.69-8 (ref. 93) for rendering. Figures 2–4, Extended Data Figs. 4–8 and corresponding statistical analyses were generated using the statistical software R and the tidyverse suite, including ggplot2 (ref. 94) and related add-on packages ggraph[95], ggforce[96], ggpubr[97], ggrepel[98], ggstance[99] and patchwork[100]. For deriving unique branch lengths (Fig. 4), we used the function pd.calc from the caper package[101]. The neutral.fit function from the MicEco R library (https://github.com/Russel88/MicEco) was used for fitting the family-level abundances to the neutral community model.

Availability of unique biological materials

Access upon request of the infant faecal samples to third parties is not part of the consent granted by the parents upon enrollment into the COPSAC2010 cohort. Nor is such access compliant with Danish or EU regulations for safeguarding rights of underage human research participants. Materials might however be obtained as part of a scientific collaboration agreement with COPSAC, and queries for such may be sent to the COPSAC Data Protection Officer, Ulrik Ralkiaer, PhD (administration@dbac.dk).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Viral genome sequences, taxonomy and host predictions and VOGs for all viruses are available through the online version of Fig. 1 on http://copsac.com/earlyvir/f1y/fig1.svg as well as the FigShare repository https://doi.org/10.6084/m9.figshare.21102805. Benchmarking data including the non-viral sequence clusters is also available through the above as well as via http://copsac.com/earlyvir/f1y/benchmark.tsv. Sequencing FASTQ files can be accessed through the European Nucleotide Archive (ebi.ac.uk) using the project number PRJEB46943. Reference phages were obtained from the INPHARED database on millardlab.org. Reference Bacterial *cpn60* sequences were obtained from cpndb.ca.

Code availability

Data analyses were carried out using free and open source software as specified in the Online Methods. Custom code was also used and is available on GitHub (https://github.com/shiraz-shah/VFCs).

References

1. Moeller, A. H. et al. Cospeciation of gut microbiota with hominids. *Science* **353**, 380–382 (2016).
2. Milani, C. et al. The first microbial colonizers of the human gut: composition, activities, and health implications of the infant gut microbiota. *Microbiol. Mol. Biol. Rev.* **81**, e00036-17 (2017).
3. Johnson, C. C. & Ownby, D. R. The infant gut bacterial microbiota and risk of pediatric asthma and allergic diseases. *Transl. Res.* **179**, 60–70 (2017).
4. Kalliomäki, M., Collado, M. C., Salminen, S. & Isolauri, E. Early differences in fecal microbiota composition in children may predict overweight. *Am. J. Clin. Nutr.* **87**, 534–538 (2008).
5. Stokholm, J. et al. Maturation of the gut microbiome and risk of asthma in childhood. *Nat. Commun.* **9**, 141 (2018).
6. Bisgaard, H. et al. Reduced diversity of the intestinal microbiota during infancy is associated with increased risk of allergic disease at school age. *J. Allergy Clin. Immunol.* **128**, 646–52.e1–5 (2011).
7. Ott, S. J. et al. Efficacy of sterile fecal filtrate transfer for treating patients with *Clostridioides difficile* infection. *Gastroenterology* **152**, 799–811.e7 (2017).
8. Rasmussen, T. S. et al. Faecal virome transplantation decreases symptoms of type 2 diabetes and obesity in a murine model. *Gut* **69**, 2122–2130 (2020).
9. Brunse, A. et al. Fecal filtrate transplantation protects against necrotizing enterocolitis. *ISME J* **16**, 686–694 (2022).
10. Shkoporov, A. N. & Hill, C. Bacteriophages of the human gut: the 'known unknown' of the microbiome. *Cell Host Microbe* **25**, 195–209 (2019).
11. Siljander, H., Honkanen, J. & Knip, M. Microbiome and type 1 diabetes. *EBioMedicine* **46**, 512–521 (2019).
12. Liang, G. et al. The stepwise assembly of the neonatal virome is modulated by breastfeeding. *Nature* **581**, 470–474 (2020).
13. Lim, E. S. et al. Early life dynamics of the human gut virome and bacterial microbiome in infants. *Nat. Med.* **21**, 1228–1234 (2015).
14. Zhao, G. et al. Intestinal virome changes precede autoimmunity in type I diabetes-susceptible children. *Proc. Natl Acad. Sci. USA* **114**, E6166–E6175 (2017).
15. Hobbs, Z. & Abedon, S. T. Diversity of phage infection types and associated terminology: the problem with 'Lytic or lysogenic'. *FEMS Microbiol. Lett.* https://doi.org/10.1093/femsle/fnw047 (2016).
16. Bernheim, A. & Sorek, R. The pan-immune system of bacteria: antiviral defence as a community resource. *Nat. Rev. Microbiol.* **18**, 113–119 (2020).
17. Sorek, R., Kunin, V. & Hugenholtz, P. CRISPR—a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat. Rev. Microbiol.* **6**, 181–186 (2008).
18. Fluckiger, A. et al. Cross-reactivity between tumor MHC class I-restricted antigens and an enterococcal bacteriophage. *Science* **369**, 936–942 (2020).
19. Górski, A. et al. Perspectives of phage therapy in non-bacterial infections. *Front. Microbiol.* **9**, 3306 (2018).
20. Dufour, N., Delattre, R., Chevallereau, A., Ricard, J.-D. & Debarbieux, L. Phage therapy of pneumonia is not associated with an overstimulation of the inflammatory response compared to antibiotic treatment in mice. *Antimicrob. Agents Chemother.* **63**, e00379–19 (2019).
21. Breitbart, M. et al. Viral diversity and dynamics in an infant gut. *Res. Microbiol.* **159**, 367–373 (2008).
22. McCann, A. et al. Viromes of one year old infants reveal the impact of birth mode on microbiome diversity. *PeerJ* **6**, e4694 (2018).
23. Aggarwala, V., Liang, G. & Bushman, F. D. Viral communities of the human gut: metagenomic analysis of composition and dynamics. *Mob. DNA* **8**, 12 (2017).
24. Bin Jang, H. et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* **37**, 632–639 (2019).
25. Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **8**, 90 (2020).
26. Nayfach, S. et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2020).
27. Nayfach, S. et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat. Microbiol* **6**, 960–970 (2021).
28. Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D. & Lawley, T. D. Massive expansion of human gut bacteriophage diversity. *Cell* **184**, 1098–1109.e9 (2021).
29. Gregory, A. C. et al. The gut virome database reveals age-dependent patterns of virome diversity in the human gut. *Cell Host Microbe* **28**, 724–740.e8 (2020).
30. Simmonds, P. et al. Consensus statement: virus taxonomy in the age of metagenomics. *Nat. Rev. Microbiol.* **15**, 161–168 (2017).
31. Low, S. J., Džunková, M., Chaumeil, P.-A., Parks, D. H. & Hugenholtz, P. Evaluation of a concatenated protein phylogeny for classification of tailed double-stranded DNA viruses belonging to the order Caudovirales. *Nat. Microbiol* **4**, 1306–1315 (2019).
32. Barylski, J. et al. Analysis of spounaviruses as a case study for the overdue reclassification of tailed phages. *Syst. Biol.* **69**, 110–123 (2020).
33. Benler, S. et al. Thousands of previously unknown phages discovered in whole-community human gut metagenomes. *Microbiome* **9**, 78 (2021).
34. Yutin, N. et al. Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nat. Microbiol.* **3**, 38–46 (2018).
35. Yutin, N. et al. Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features. *Nat. Commun.* **12**, 1044 (2021).
36. Koonin, E. V. et al. Global organization and proposed megataxonomy of the virus world. *Microbiol. Mol. Biol. Rev.* **84**, e00061–19 (2020).
37. Bisgaard, H. et al. Deep phenotyping of the unselected COPSAC 2010 birth cohort study. *Clin. Exp. Allergy* **43**, 1384–1394 (2013).
38. Li, X. et al. The infant gut resistome associates with *E. coli*, environmental exposures, gut microbiome maturity, and asthma-associated bacterial composition. *Cell Host Microbe* https://doi.org/10.1016/j.chom.2021.03.017 (2021).
39. Roux, S., Krupovic, M., Debroas, D., Forterre, P. & Enault, F. Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. *Open Biol.* **3**, 130160 (2013).
40. Zolfo, M. et al. Detecting contamination in viromes using ViromeQC. *Nat. Biotechnol.* **37**, 1408–1412 (2019).
41. Cook, R. et al. INfrastructure for a PHAge REference database: identification of large-scale biases in the current collection of cultured phage genomes. *Phage* **2**, 214–223 (2021).
42. Roguet, A. et al. Neutral community model explains the bacterial community assembly in freshwater lakes. *FEMS Microbiol. Ecol.* **91**, fiv125 (2015).

43. Venkataraman, A. et al. Application of a neutral community model to assess structuring of the human lung microbiome. *mBio* **6**, e02284–14 (2015).
44. Dutilh, B. E. et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* **5**, 4498 (2014).
45. Murphy, J. et al. Comparative genomics and functional analysis of the 936 group of lactococcal Siphoviridae phages. *Sci. Rep.* **6**, 21345 (2016).
46. Stanton, C. R., Rice, D. T. F., Beer, M., Batinovic, S. & Petrovski, S. Isolation and characterisation of the genus and phylogenetic investigation of the bacteriophages. *Viruses* **13**, 1557 (2021).
47. Guerin, E. et al. Biology and taxonomy of crAss-like bacteriophages, the most abundant virus in the human gut. *Cell Host Microbe* **24**, 653–664.e6 (2018).
48. Cornuault, J. K. et al. Phages infecting *Faecalibacterium prausnitzii* belong to novel viral genera that help to decipher intestinal viromes. *Microbiome* **6**, 65 (2018).
49. Benler, S. et al. A diversity-generating retroelement encoded by a globally ubiquitous *Bacteroides* phage. *Microbiome* **6**, 191 (2018).
50. Redgwell, T. A. et al. Prophages in the infant gut are largely induced, and may be functionally relevant to their hosts. Preprint at *bioRxiv* https://doi.org/10.1101/2021.06.25.449885 (2021).
51. Petit, M.-A. et al. An alternative method to multiple displacement amplification for preparing virome DNA in a way adapted for the sequencing of both double-strand and single-strand DNA viruses. Preprint at *bioRxiv* https://doi.org/10.1101/2022.12.12.520144 (2022).
52. Freer, G. et al. The virome and its major component, anellovirus, a convoluted system molding human immune defenses and possibly affecting the development of asthma and respiratory diseases in childhood. *Front. Microbiol.* **9**, 686 (2018).
53. Roux, S. et al. Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. *Nat. Microbiol* **4**, 1895–1906 (2019).
54. Mavrich, T. N. & Hatfull, G. F. Bacteriophage evolution differs by host, lifestyle and genome. *Nat. Microbiol* **2**, 17112 (2017).
55. Paez-Espino, D. et al. Uncovering Earth's virome. *Nature* **536**, 425–430 (2016).
56. Hubbell, S. P. & Borda-de-Água, L. The unified neutral theory of biodiversity and biogeography: reply. *Ecology* **85**, 3175–3178 (2004).
57. Dion, M. B. et al. Streamlining CRISPR spacer-based bacterial host predictions to decipher the viral dark matter. *Nucleic Acids Res.* **49**, 3127–3138 (2021).
58. Galiez, C., Siebert, M., Enault, F., Vincent, J. & Söding, J. WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics* **33**, 3113–3114 (2017).
59. Karcher, N. et al. Genomic diversity and ecology of human-associated *Akkermansia* species in the gut microbiome revealed by extensive metagenomic assembly. *Genome Biol.* **22**, 209 (2021).
60. Kim, K.-H. & Bae, J.-W. Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Appl. Environ. Microbiol.* **77**, 7663–7668 (2011).
61. Roux, S. et al. Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. *PeerJ* **4**, e2777 (2016).
62. Garmaeva, S. et al. Stability of the human gut virome and effect of gluten-free diet. *Cell Rep.* **35**, 109132 (2021).
63. Millen, A. M. et al. *Lactococcus lactis* type III-A CRISPR–Cas system cleaves bacteriophage RNA. *RNA Biol.* **16**, 461–468 (2019).
64. Waller, A. S. et al. Classification and quantification of bacteriophage taxa in human gut metagenomes. *ISME J.* **8**, 1391–1402 (2014).
65. Mathieu, A. et al. Virulent coliphages in 1-year-old children fecal samples are fewer, but more infectious than temperate coliphages. *Nat. Commun.* **11**, 378 (2020).
66. De Paepe, M. et al. Carriage of λ latent virus is costly for its bacterial host due to frequent reactivation in monoxenic mouse intestine. *PLoS Genet.* **12**, e1005861 (2016).
67. Cornuault, J. K. et al. The enemy from within: a prophage of *Roseburia intestinalis* systematically turns lytic in the mouse gut, driving bacterial adaptation by CRISPR spacer acquisition. *ISME J.* **14**, 771–787 (2020).
68. Tyler, J. S. et al. Prophage induction is enhanced and required for renal disease and lethality in an EHEC mouse model. *PLoS Pathog.* **9**, e1003236 (2013).
69. Clooney, A. G. et al. Whole-virome analysis sheds light on viral dark matter in inflammatory bowel disease. *Cell Host Microbe* **26**, 764–778.e5 (2019).
70. Deng, L. et al. A protocol for extraction of infective viromes suitable for metagenomics sequencing from low volume fecal samples. *Viruses* **11**, 667 (2019).
71. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).
72. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
73. Moreno-Gallego, J. L. et al. Virome diversity correlates with intestinal microbiome diversity in adult monozygotic twins. *Cell Host Microbe* **25**, 261–272.e5 (2019).
74. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **11**, 119 (2010).
75. Pearson, W. R. Finding protein and nucleotide similarities with FASTA. *Curr. Protoc. Bioinform.* **53**, 3.9.1–3.9.25 (2016).
76. Vestergaard, G., Garrett, R. A. & Shah, S. A. CRISPR adaptive immune systems of Archaea. *RNA Biol.* **11**, 156–167 (2014).
77. Biswas, A., Staals, R. H. J., Morales, S. E., Fineran, P. C. & Brown, C. M. CRISPRDetect: a flexible algorithm to define CRISPR arrays. *BMC Genomics* **17**, 356 (2016).
78. Shah, S. A. et al. Comprehensive search for accessory proteins encoded with archaeal and bacterial type III CRISPR–cas gene cassettes reveals 39 new cas gene families. *RNA Biol.* **16**, 530–542 (2019).
79. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
80. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
81. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinform.* **5**, 113 (2004).
82. Mistry, J. et al. Pfam: the protein families database in 2021. *Nucleic Acids Res* **49**, D412–D419 (2021).
83. Lu, S. et al. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.* **48**, D265–D268 (2020).
84. Tatusov, R. L. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**, 33–36 (2000).
85. Haft, D. H. The TIGRFAMs database of protein families. *Nucleic Acids Res* **31**, 371–373 (2003).
86. Steinegger, M. et al. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinform.* **20**, 473 (2019).
87. Terzian, P. et al. PHROG: families of prokaryotic virus proteins clustered using remote homology. *NAR Genom. Bioinform* **3**, lqab067 (2021).

88. Milanese, A. et al. Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat. Commun.* **10**, 1014 (2019).
89. McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **8**, e61217 (2013).
90. Kalvari, I. et al. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.* **49**, D192–D200 (2021).
91. Hill, J. E., Penny, S. L., Crowell, K. G., Goh, S. H. & Hemmingsen, S. M.cpnDB: a chaperonin sequence database. *Genome Res.* **14**, 1669–1675 (2004).
92. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
93. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
94. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. *Springer* (2016).
95. Pedersen, T. L. ggraph: an implementation of grammar of graphics for graphs and networks. *CRAN* https://CRAN.R-project.org/package=ggraph (2019).
96. Pedersen, T. L. ggforce: accelerating 'ggplot2'. *CRAN* https://CRAN.R-project.org/package=ggforce (2019).
97. Kassambara, A. ggpubr: 'ggplot2' based publication ready plots. *CRAN* https://CRAN.R-project.org/package=ggpubr (2020).
98. Slowikowski, K. ggrepel: automatically position non-overlapping text labels with'ggplot2'. *CRAN* https://CRAN.R-project.org/package=ggrepel (2019).
99. Henry, L., Wickham, H. & Chang, W. ggstance: horizontal 'ggplot2' components. *CRAN* https://CRAN.R-project.org/package=ggstance (2020).
100. Pedersen, T. L. patchwork: the composer of plots. *CRAN* https://CRAN.R-project.org/package=patchwork (2020).
101. Orme, D. et al. caper: comparative analyses of phylogenetics and evolution in R. *CRAN* https://CRAN.R-project.org/package=caper (2018).

Acknowledgements

We express our deepest gratitude to the children and families of the COPSAC2010 cohort study for all their support and commitment. We acknowledge and appreciate the unique efforts of the COPSAC research team. This work is supported by the Joint Programming Initiative 'Healthy Diet for a Healthy Life', specifically here, the Danish Agency for Science and Higher Education, Institut National de la Recherche pour l'Agriculture, l'Alimentation et l'Environnement (INRAE), and the Canadian Institutes of Health Research (Team grant on Intestinal Microbiomics, Institute of Nutrition, Metabolism, and Diabetes, grant number 143924). M.B.D. is recipient of graduate scholarships from the Fonds de Recherche du Québec—Nature et Technologies (259257) as well as Sentinel North and is a recipient of the Goran-Enhorning Graduate Student Research Award from the Canadian Allergy, Asthma and Immunology Foundation. J.T. is supported by the BRIDGE Translational Excellence Program (bridge.ku.dk) at the Faculty of Health and Medical Sciences, University of Copenhagen, funded by the Novo Nordisk Foundation (grant no. NNF18SA0034956). S.M. holds the Tier 1 Canada Research Chair in Bacteriophages (950-232136). S.A.S. is a recipient of a Novo Nordisk Foundation project grant in basic bioscience (NNF18OC0052965), along with M.A.R., J.S. and D.S.N. are recipients of Novo Nordisk Foundation grant NNF20OC0061029.

Author contributions

S.A.S. performed all bioinformatics analyses, prepared Fig. 1, the online resources and wrote the manuscript. L.D. extracted viromes and prepared sequencing libraries. J.T. prepared Figs. 2–4, performed statistics and assisted with writing. A.G.P. wrote code for cutting the APS tree at required taxonomic levels. M.B.D. and T.A.R. performed bacterial host predictions. J.L.C.-M. and G.V. assisted with bioinformatics. R. Sausset performed viral counts. R. Silins and F.O.R. assisted with virome extractions and sequencing library preparation. E.O.N. and F.E. cross-referenced microviral taxonomy. M.H., C.L.-R., M.A.R., Y.Z., B.C., K.B., S.J.S. and F.E. assisted with writing. H.B., J.S., S.M., M.-A.P. and D.S.N. conceived the study, aided in data interpretation and assisted with writing. All authors read, revised and approved the manuscript.

Competing interests

S.A.S. is a consultant for profluent.bio on a matter that is unrelated to the present study. D.S.N. has functioned as a consultant for the companies Pfizer and Sniprbiome on scientific matters not related to the present study. All remaining authors declare no conflicts of interest.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41564-023-01345-7.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-023-01345-7.

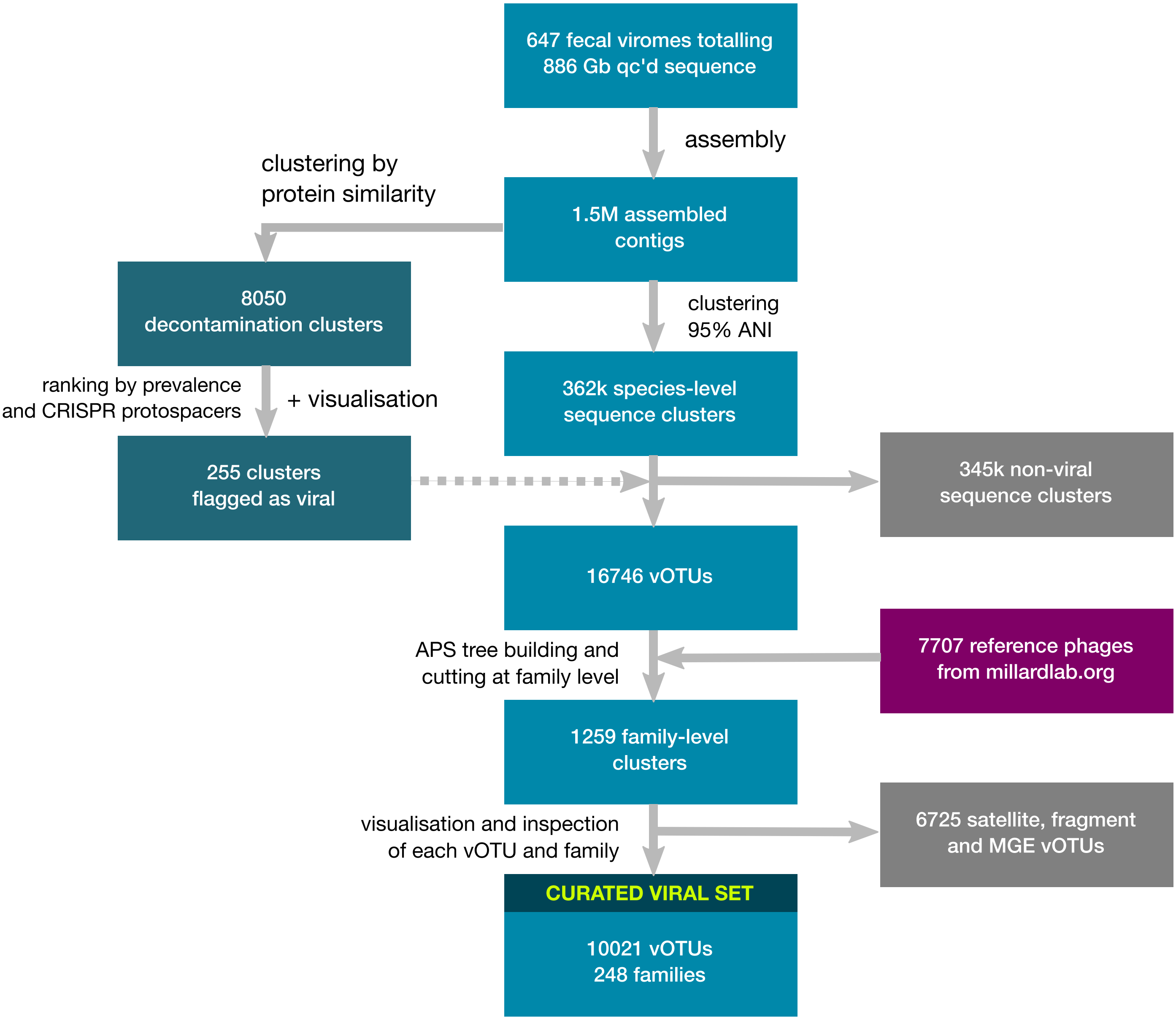
Correspondence and requests for materials should be addressed to Shiraz A. Shah or Dennis S. Nielsen.

Peer review information *Nature Microbiology* thanks the anonymous reviewers for their contribution to the peer review of this work.

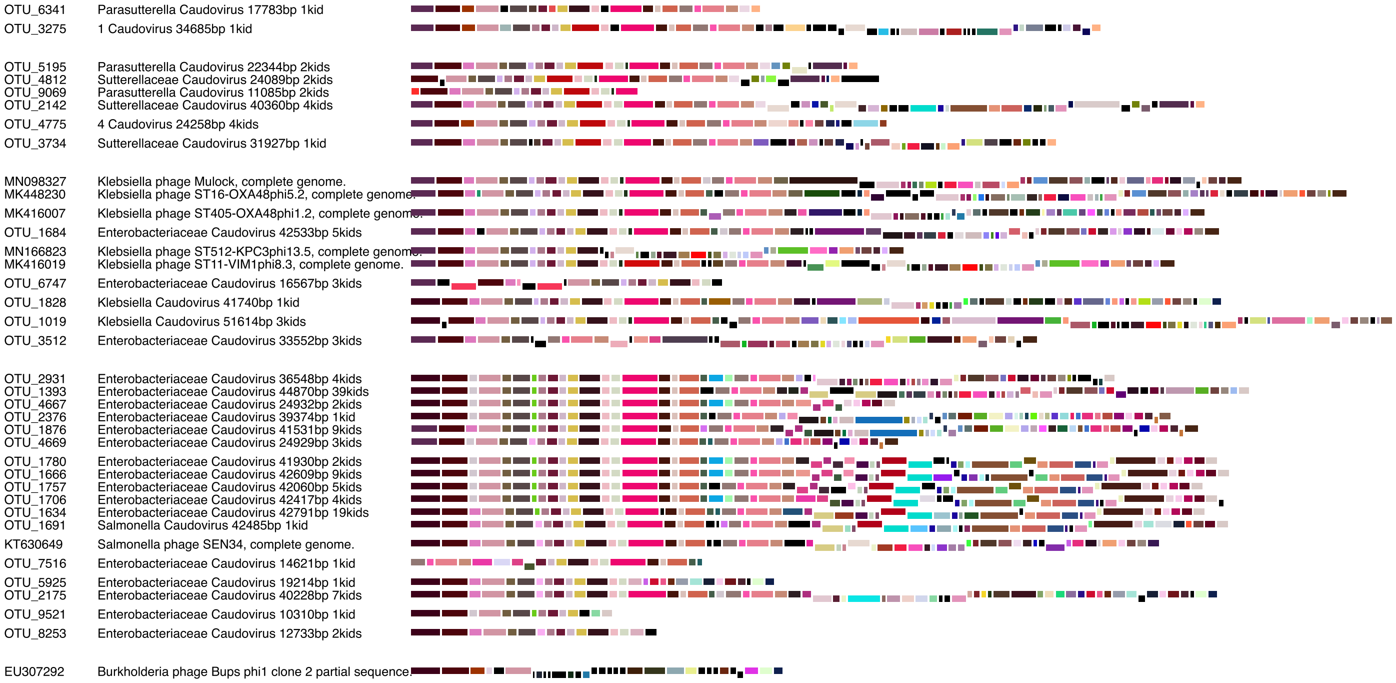
Reprints and permissions information is available at www.nature.com/reprints.

[1]Copenhagen Prospective Studies on Asthma in Childhood, Copenhagen University Hospital, Herlev-Gentofte, Gentofte, Denmark. [2]Department of Food Science, University of Copenhagen, Copenhagen, Denmark. [3]Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [4]Department of Health Technology, Technical University of Denmark, Lyngby, Denmark. [5]Département de biochimie, de microbiologie, et de bio-informatique, Faculté des sciences et de génie, Université

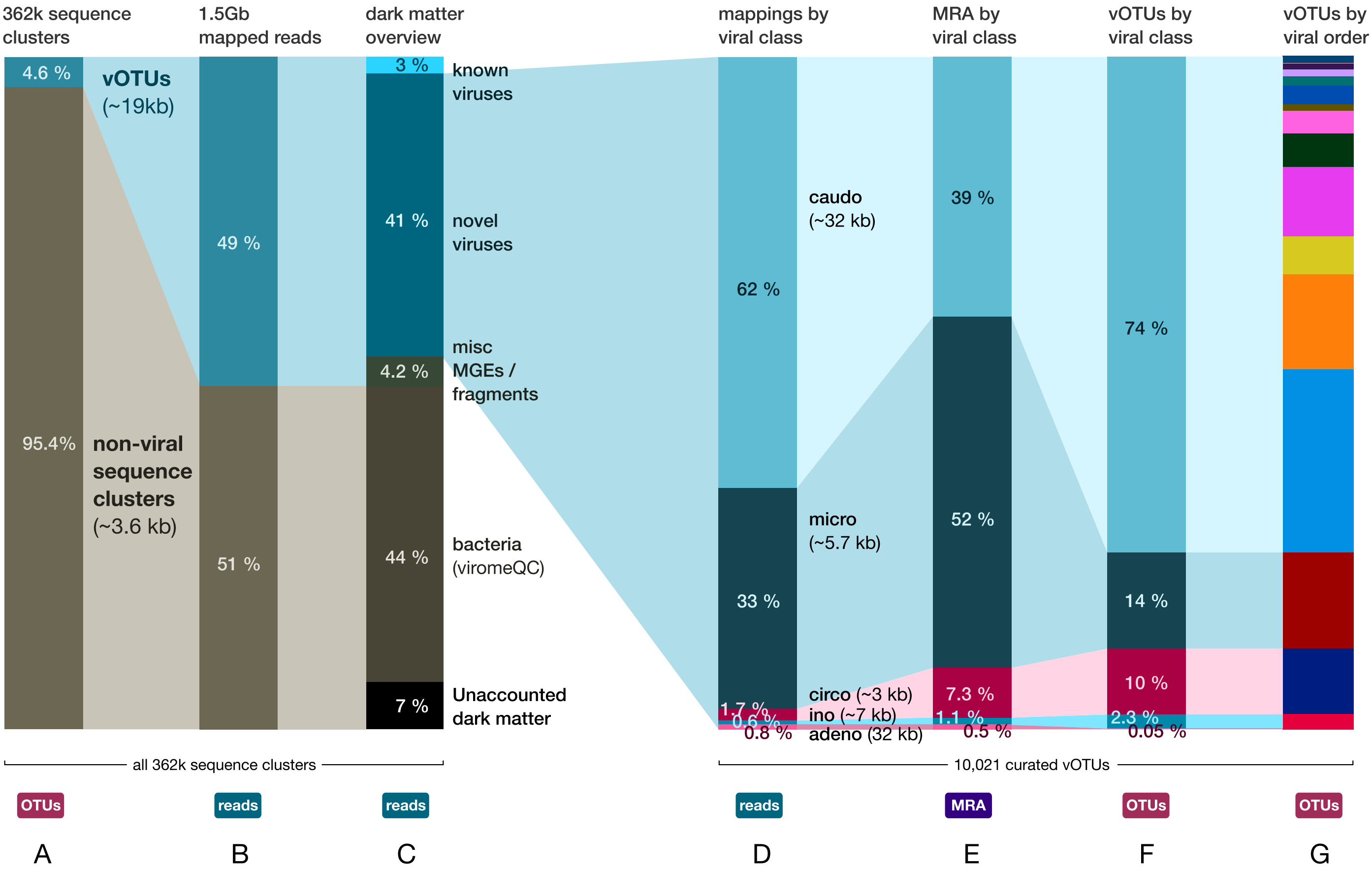
Laval, Québec City, Quebec, Canada. [6]Groupe de recherche en écologie buccale, Faculté de médecine dentaire, Université Laval, Québec City, Quebec, Canada. [7]Université Paris-Saclay, INRAE, Agroparistech, Micalis institute, Jouy-en-Josas, France. [8]Lab de Microorganismes: Génome et Environnement, Université Clermont Auvergne, Clermont-Ferrand, France. [9]Department of Biology, University of Copenhagen, Copenhagen, Denmark. [10]Félix d'Hérelle Reference Center for Bacterial Viruses, Université Laval, Québec City, Quebec, Canada. [11]These authors contributed equally: Shiraz A. Shah, Ling Deng. [12]Deceased: Hans Bisgaard. ✉e-mail: shiraz.shah@dbac.dk; dn@food.ku.dk


647 fecal viromes totalling 886 Gb qc'd sequence
assembly
1.5M assembled contigs
clustering by protein similarity
8050 decontamination clusters
ranking by prevalence and CRISPR protospacers
+ visualisation
255 clusters flagged as viral
clustering 95% ANI
362k species-level sequence clusters
345k non-viral sequence clusters
16746 vOTUs
APS tree building and cutting at family level
7707 reference phages from millardlab.org
1259 family-level clusters
visualisation and inspection of each vOTU and family
6725 satellite, fragment and MGE vOTUs
CURATED VIRAL SET
10021 vOTUs
248 families

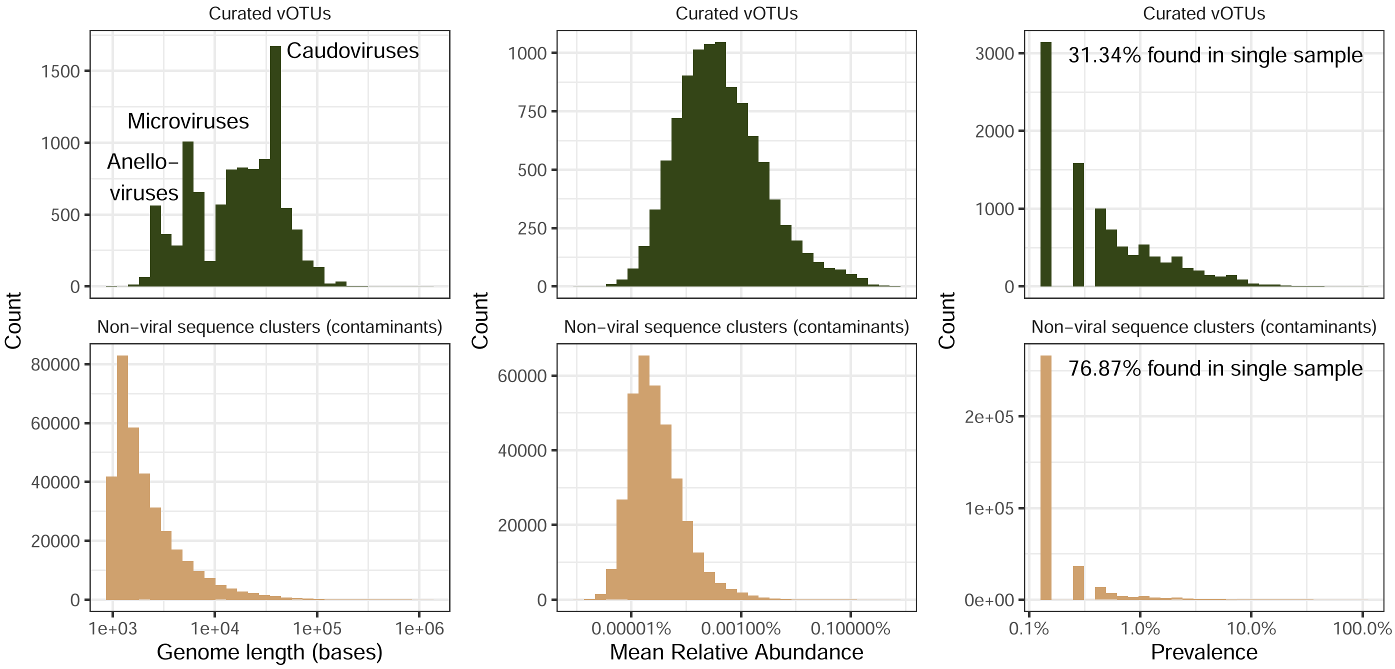
Extended Data Fig. 1 | Overview of decontamination and curation procedure.


OTU_6341 Parasutterella Caudovirus 17783bp 1kid
OTU_3275 1 Caudovirus 34685bp 1kid
OTU_5195 Parasutterella Caudovirus 22344bp 2kids
OTU_4812 Sutterellaceae Caudovirus 24089bp 2kids
OTU_9069 Parasutterella Caudovirus 11085bp 2kids
OTU_2142 Sutterellaceae Caudovirus 40360bp 4kids
OTU_4775 4 Caudovirus 24258bp 4kids
OTU_3734 Sutterellaceae Caudovirus 31927bp 1kid
MN098327 Klebsiella phage Mulock, complete genome.
MK448230 Klebsiella phage ST16-OXA48phi5.2, complete genome.
MK416007 Klebsiella phage ST405-OXA48phi1.2, complete genome.
OTU_1684 Enterobacteriaceae Caudovirus 42533bp 5kids
MN166823 Klebsiella phage ST512-KPC3phi13.5, complete genome.
MK416019 Klebsiella phage ST11-VIM1phi8.3, complete genome.
OTU_6747 Enterobacteriaceae Caudovirus 16567bp 3kids
OTU_1828 Klebsiella Caudovirus 41740bp 1kid
OTU_1019 Klebsiella Caudovirus 51614bp 3kids
OTU_3512 Enterobacteriaceae Caudovirus 33552bp 3kids
OTU_2931 Enterobacteriaceae Caudovirus 36548bp 4kids
OTU_1393 Enterobacteriaceae Caudovirus 44870bp 39kids
OTU_4667 Enterobacteriaceae Caudovirus 24932bp 2kids
OTU_2376 Enterobacteriaceae Caudovirus 39374bp 1kid
OTU_1876 Enterobacteriaceae Caudovirus 41531bp 9kids
OTU_4669 Enterobacteriaceae Caudovirus 24929bp 3kids
OTU_1780 Enterobacteriaceae Caudovirus 41930bp 2kids
OTU_1666 Enterobacteriaceae Caudovirus 42609bp 9kids
OTU_1757 Enterobacteriaceae Caudovirus 42060bp 5kids
OTU_1706 Enterobacteriaceae Caudovirus 42417bp 4kids
OTU_1634 Enterobacteriaceae Caudovirus 42791bp 19kids
OTU_1691 Salmonella Caudovirus 42485bp 1kid
KT630649 Salmonella phage SEN34, complete genome.
OTU_7516 Enterobacteriaceae Caudovirus 14621bp 1kid
OTU_5925 Enterobacteriaceae Caudovirus 19214bp 1kid
OTU_2175 Enterobacteriaceae Caudovirus 40228bp 7kids
OTU_9521 Enterobacteriaceae Caudovirus 10310bp 1kid
OTU_8253 Enterobacteriaceae Caudovirus 12733bp 2kids
EU307292 Burkholderia phage Bups phi1 clone 2 partial sequence.

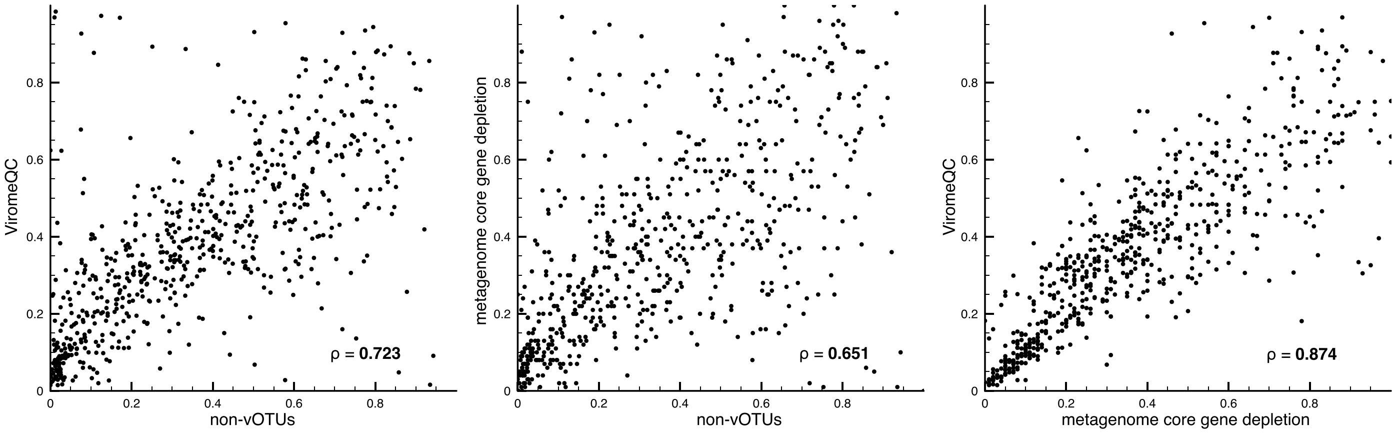
Extended Data Fig. 2 | Clickable gene map of vOTUs belonging to the the *Ingridviridae* family. Available online at http://copsac.com/earlyvir/f1y/families/Ingridviridae.svg along with similar maps for the remaining 247 families, available via http://copsac.com/earlyvir/f1y/fig1.svg. Small vertical gaps between vOTUs denote genus boundaries, while large gaps denote subfamily boundaries. Ordering of the vOTUs follows the order in the APS tree and thus, related vOTUs are next to each other. ORFs are aligned vertically based on strandedness and colored by VOG affiliation. VOG definitions against the PhROGs database[87] can be looked up by clicking on each ORF. ORF gene product (GP) numbers are displayed by mouse-over hovering. GenBank files for each vOTU can be viewed along with virus and host taxonomy by clicking on the OTU name. Caudoviral maps were inverted and zeroed according to TerL gene coordinates, while the GenBank files were not. Reference phages that belong to the same family were also included in the maps and are indicated by GenBank accession numbers.



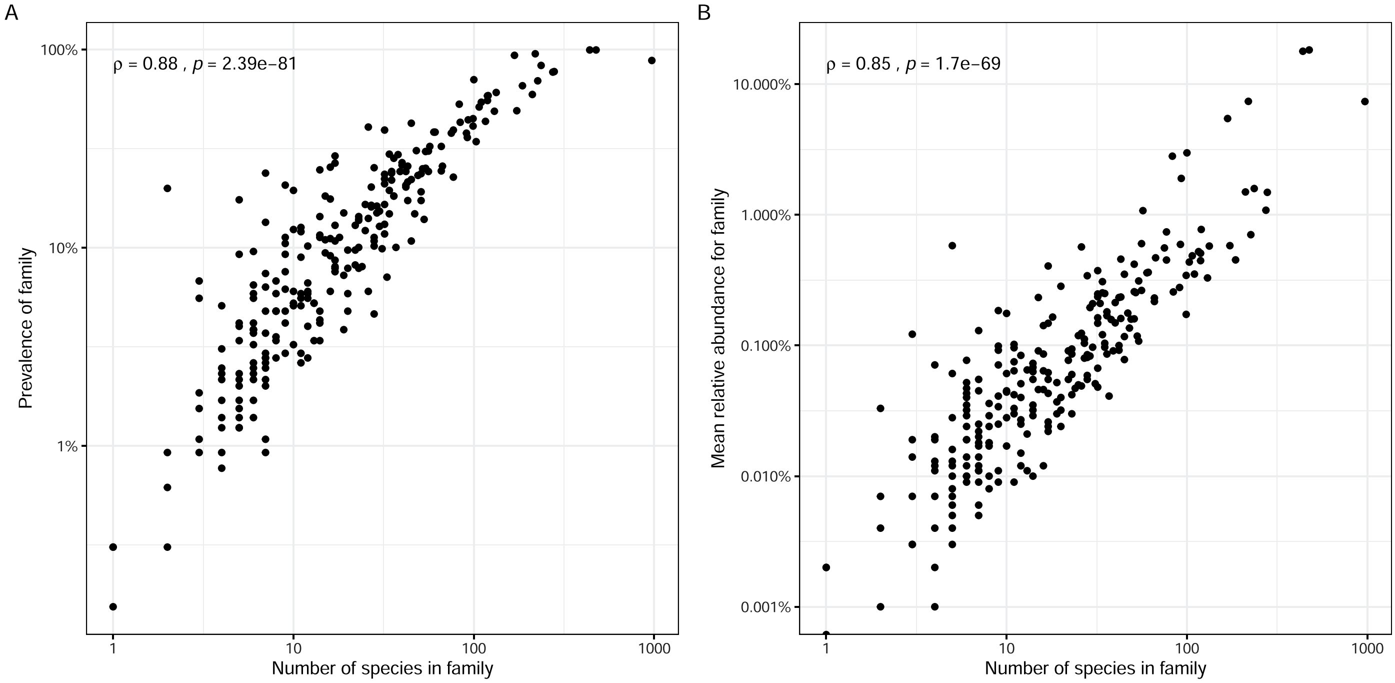
Extended Data Fig. 3 | From assembly to curated vOTUs in numbers. After assembly, species-level deduplication and manual decontamination, most sequence clusters were inferred to be non-viral and had small sizes while viral OTUs were much fewer but longer (A). After mapping, vOTUs accounted for roughly half of the reads (B). 97% of the reads originally comprised “dark matter” but only 7% was left after resolution (C). The 10,021 curated vOTUs fell within five viral classes (caudoviruses [dsDNA], microviruses [ssDNA], anelloviruses [ssDNA], inoviruses [ssDNA] and adenoviruses [dsDNA]). Distributions of the viral classes by: mapped reads (D), MRAs, after normalising read counts for sequencing depth and genome size (E) and species richness, that is number of vOTUs (F) are shown. G) Same as F but at viral order-level, with orders colored as in Fig. 2.



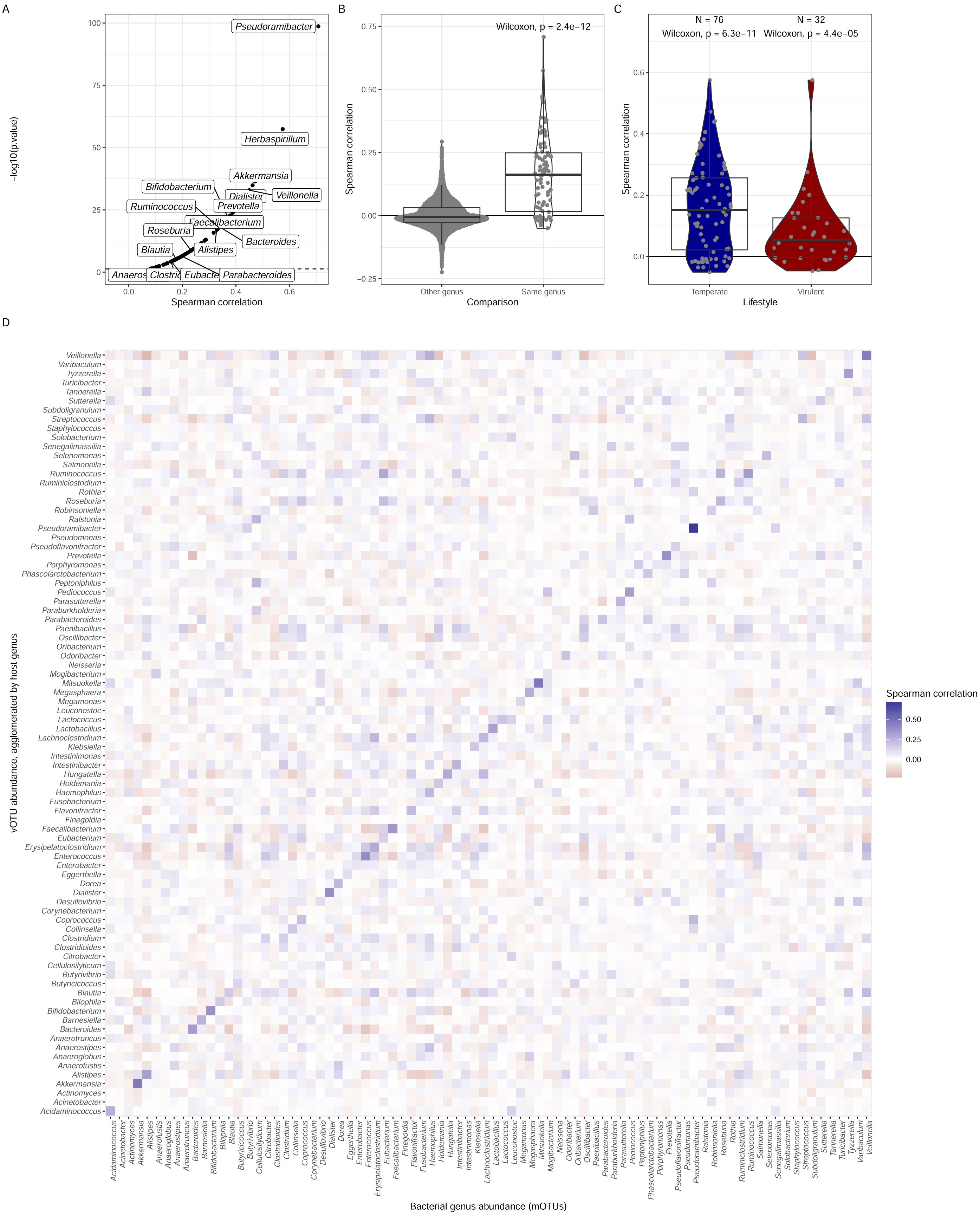
Extended Data Fig. 4 | Features of vOTUs versus non-viral sequence clusters within data. Distribution of size, MRA and sample prevalence for contaminant non-viral sequence clusters and curated vOTUs respectively. The vOTU size distribution shows peaks corresponding to genome lengths for the three major classes of viruses in the dataset, namely anelloviruses, microviruses and caudoviruses (3 kb, 5.5 kb, and 40 kb). The contaminant size distribution peaks at the contig inclusion cutoff (1 kb) continuing with a long uniform tail, consistent with the unspecific origin expected for contaminating DNA. Curated vOTUs were more abundant and prevalent than contaminating species. The majority of the contaminating sequences were sample-specific, in contrast to most curated vOTUs which were found in more than one sample. The latter is consistent with their bacterial chromosomal origin, as unspecific subsampling of the large bacterial genome space is unlikely to yield overlaps between samples.



Extended Data Fig. 5 | Comparison of three approaches for estimating the proportion of bacterial contamination. Each graph has 647 dots, one for each sample. Axes denote the proportion of bacterial contamination as estimated by the indicated method. Each graph is a pairwise comparison of two different methods. A) mappings to non-viral sequence clusters versus ViromeQC B) non-viral sequence cluster mappings versus metagenome core gene depletion C) metagenome core gene depletion versus ViromeQC. Spearman's correlation coefficients (ρ) are given for all three comparisons.

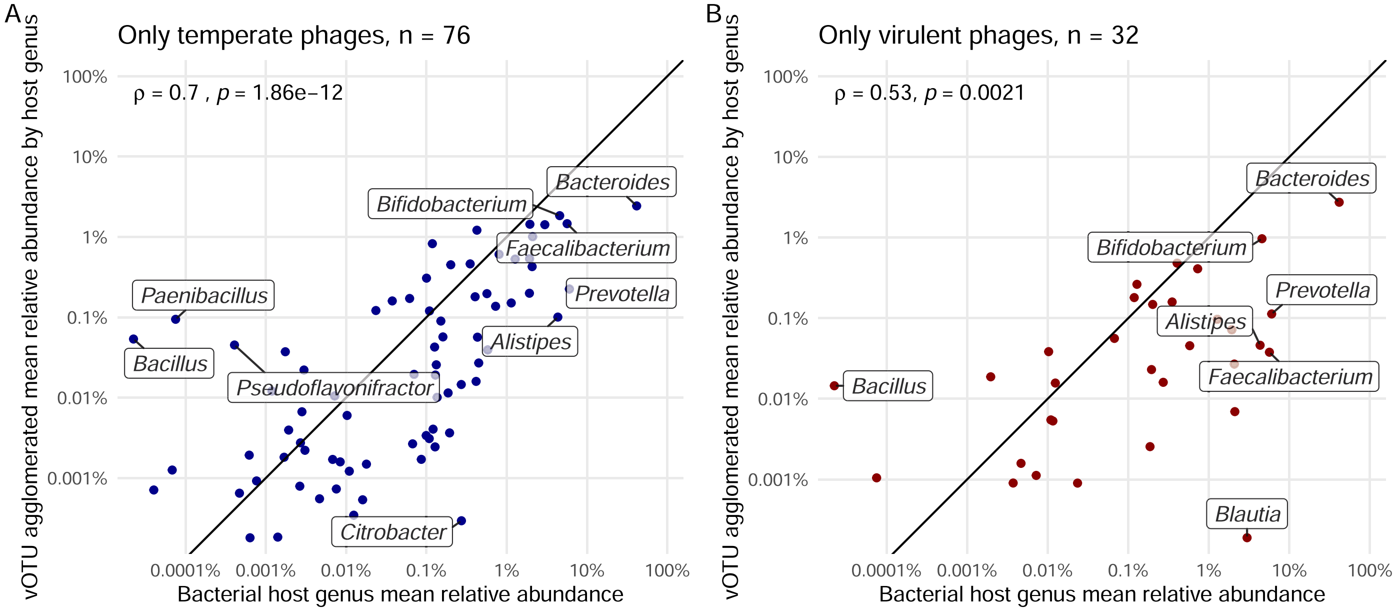


Extended Data Fig. 6 | Viral family species-richness is linked to prevalence and abundance. The species-richness within a family is highly correlated with both its prevalence (A) and the MRA across samples (B), shown here with Spearman's correlation tests (two-sided *P* values). MRAs are correspondingly correlated with prevalence as already shown in Fig. 3. The correlation between all three measures is in line with predictions made by the neutral community model. MRA, mean relative abundance.

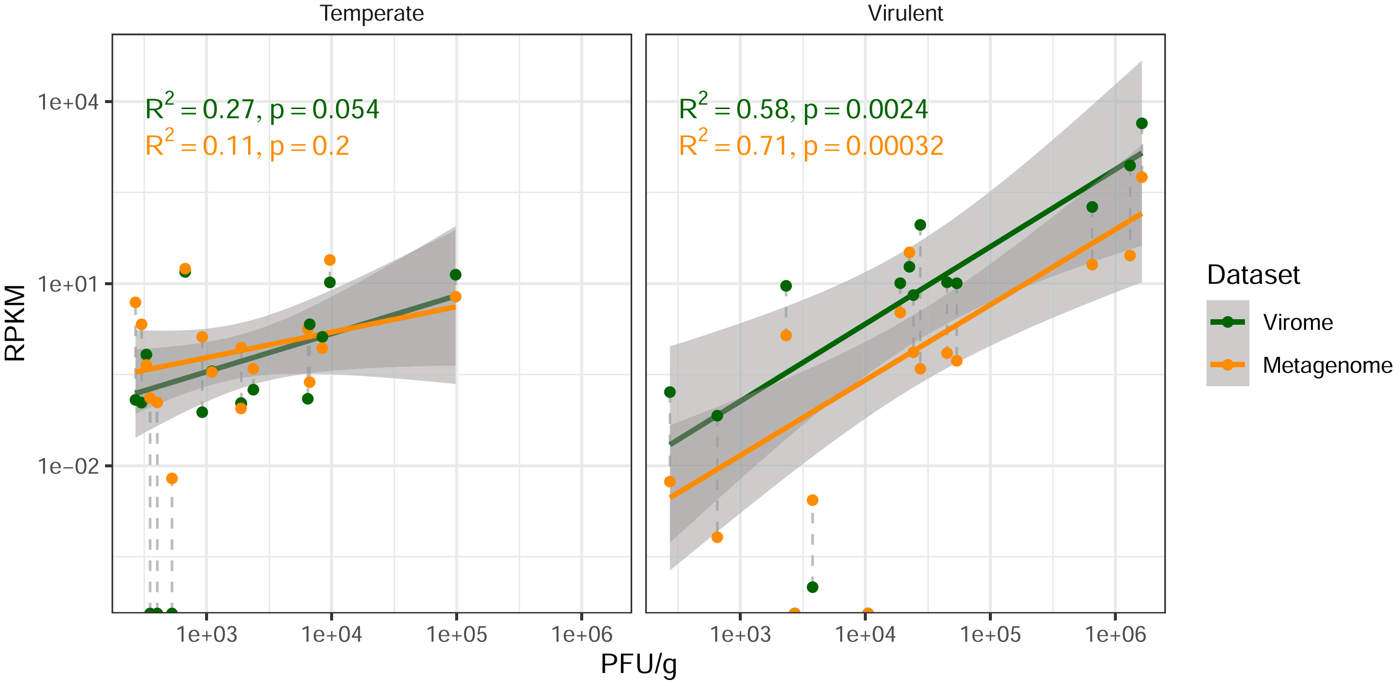


Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Sample-to-sample co-abundance of phages in the virome and host-bacteria in the metagenome. Correlations between host bacterial relative genus abundances in the metagenomes with aggregate relative abundances for phages predicted to infect those host genera in the virome, compared across all children. A) Volcano plot showing how all significant correlations between phage-host pairs were positive ($\rho > 0$; $n = 87$ genera, Spearman's correlation tests, two-sided *P* values). B) The distribution of these correlation values was significantly higher than zero (One-sample Wilcoxon test, two-sided $P = 2.4 \cdot 10^{-12}$, $n = 87$, right side), whereas random non-matched phage-host pairs were centered around zero (left side). C) These correlations were positive regardless of phage lifestyle (one-sample Wilcoxon tests with two-sided *P* values), and D) stood out against the background of all genus combinations tested (same data shown in panel B, diagonal is matched phage-host pairs and off-diagonal are non-matched pairs). Boxplots demonstrate median, middle line; lower and upper quartile, box bounds; and most extreme observations within 1.5 x interquartile range above/below box, whiskers. All individual data points are overlaid on the boxplots.



Extended Data Fig. 8 | Mean co-abundance of phages and hosts regardless of viral lifestyle. Correspondence between host genus abundances in the metagenome with aggregate abundances for all phages infecting those genera in the virome, as stratified by virus lifestyle, namely, temperate phages (A) and virulent phages (B). The MRA of both virulent and temperate phages correlates positively with host MRA. MRA, mean relative abundance. Correlations were tested using Spearman's rank test (two-sided *P* values).



Extended Data Fig. 9 | The sMDA amplified viromes are quantitative for dsDNA phages. The relationship between experimentally determined PFU/g of faeces for 32 coliphages[65], against mapped virome and metagenome reads per kilobase per million (RPKM), from the corresponding 32 samples. The two panels show data for temperate and virulent coliphages respectively. Axes were log-transformed to capture the dynamic range. A linear model was fit following log-transformation. Temperate coliphages show only a tendency of being associated presumably because read-mappings were shared between induced phage DNA and bacterial chromosomal DNA. For the virulent coliphages, however, the relationship was quantitative throughout the range of PFU counts (from 270 to 1.6 M). The sMDA amplified virome is no less quantitative than the unamplified metagenomes for the same samples. sMDA: short multiple-displacement amplification. Paired viromes/metagenomes from the same samples are connected using dashed lines. Regression lines are drawn using linear models, the shaded area represents the 95% confidence band for the regression line. *P* values correspond to Spearman's rank correlation tests, are two-sided and were not adjusted for multiple comparisons.

nature portfolio

Corresponding author(s): Dennis S. Nielsen

Last updated by author(s): Feb 8, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	□	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection
Data analysis	fastx-toolkit 0.0.14, cutadapt 2.0, vsearch 2.4.3, SPAdes 3.10.1, blat v35, prodigal 2.6.3, FASTA 36.3.6f, mcl 14-137, MUSCLE v3.8.425, hhsuite v3.0-beta.3, rapidnj 2.3.0.2, bwa 0.7.17-r1188, samtools 1.9, msamtools 0.9.6, Circos v0.69-8, R 4.0.2 with libraries phyloseq, , ggplot2, ggraph, ggforce, ggpubr, ggrepel, ggstance, patchwork and custom code (https://github.com/shiraz-shah/VFCs)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Viral genome sequences, taxonomy and host predictions and VOGs for all viruses are available through the online version of Figure 1 on http://copsac.com/earlyvir/f1y/fig1.svg as well as the FigShare repository https://doi.org/10.6084/m9.figshare.21102805. Benchmarking data including the non-viral sequence clusters is also available through the above as well as via http://copsac.com/earlyvir/f1y/benchmark.tsv. Sequencing FASTQ files can be accessed through the European Nucleotide Archive (ebi.ac.uk) using the project number PRJEB46943. Reference phages were obtained from the INPHARED database on millardlab.org. Reference Bacterial cpn60 sequences were obtained from cpndb.ca.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender

Although information on sex was collected for the participants of the cohort at birth in accordance with the ethics statement below, it was not used in this study. The aim of the current study was to explore the diversity of the human infant gut virome regardless of host sex.

Population characteristics

COPSAC2010 is a polulation-based mother-child cohort recruited in Copenhagen and Næstved, Denmark with the overall aim of studying the mechanisms that lead to chronic disease in childhood. The samples used here were faecal samples from 1-year-old infants.

Recruitment

The COPSAC2010 cohort is a population-based birth cohort of 700 children recruited in pregnancy and has been followed prospectively at the COPSAC research unit. Details on recruitment can be found in Bisgaard, H. et al. Deep phenotyping of the unselected COPSAC2010 birth cohort study. Clin. Exp. Allergy 43, 1384–1394 (2013) (which is also cited in the manuscript, reference 39)

Ethics oversight

The study was conducted in accordance with the guiding principles of the Declaration of Helsinki and was approved by The National Committee on Health Research Ethics (H-B-2008-093) and the Danish Data Protection Agency (2015-41-3696). Both parents gave written informed consent before enrollment.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

The sample size was equal to the total size of the COPSAC 2010 cohort

Data exclusions

Infants for whom the delivered faecal samples were either exhausted after previous metagenomics analyses, or too small to perform a virome extraction were excluded from the study. After virome extraction and sequencing, samples that produced fewer than 50,000 reads were also excluded.

Replication

The criteria for the definition of viral taxa have been recently revised by the ICTV making replication difficult owing to the sparsity of independent studies that have switched over to the new criteria. However, out of the 248 found viral families, eight (Flandersviridae, Gratiaviridae, and alpha to zeta Crassviridae) were found recently in independent studies and thus replicate our family-definition criterion.

Randomization

Not relevant as the descriptive nature of the study means that we did not use any experiment and control groups.

Blinding

Blinding is not applicable to the current study as there was no group allocation. All cohort members were treated as a single group

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology and archaeology
☒	□	Animals and other organisms
☒	□	Clinical data
☒	□	Dual use research of concern

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging