# Viral biogeography of gastrointestinal tract and parenchymal organs in two representative species of mammals

Andrey N. Shkoporov*†[1,2], Stephen R. Stockdale*[1], Aonghus Lavelle[1], Ivanela Kondova[3], Cara Heuston[1], Aditya Upadrasta[1], Ekaterina V. Khokhlova[1], Imme van der Kamp[1], Boudewijn Ouwerling[3], Lorraine A. Draper[1], Jan A.M. Langermans[3,4], R Paul Ross[1,2], Colin Hill†[1,2]

[1] APC Microbiome Ireland, University College Cork, Cork, Ireland

[2] School of Microbiology, University College Cork, Cork, Ireland

[3] Biomedical Primate Research Centre, Rijswijk, The Netherlands

[4] Department of Population Health Sciences, Veterinary Faculty, Utrecht University, Utrecht, The Netherlands

* These authors contributed equally to the study

**† Corresponding authors: andrey.shkoporov@ucc.ie, c.hill@ucc.ie**

## Supplemental materials

### Supplemental Methods

### Ethical approval

The study design was developed with consideration to the three Rs for ethical use of animals in science: replacement, reduction, and refinement. The proposed euthanasia only study was reviewed by the Animal Welfare Body (AWB) of University College Cork (Euthanasia Only Authorisation 17-005). With authorisation and under the remit of authorised and experienced personnel, the study was performed succinctly and with minimal distress to the animals involved.

### Animal sampling procedures

*(i). Sus scrofa domesticus* – pigs. Six healthy female Landrace pigs (body mass approximately 30 kg, approximately 3 months of age) were sourced from a local farm in Cork, Ireland. All pigs were raised in a shared environment and on the same diet, although the relatedness of their parentage is unknown. Pigs were transported to the research facility on the morning they were to be euthanised, with two animals sampled back-to-back per day. Before euthanasia, work surfaces and necessary tools were disinfected using Virkon S disinfectant. Following anaesthetic overdose with Pentobarbital (150mg/kg) death was confirmed by an authorised person, and tissue samples were collected.

All biopsies (min. 3 cm × 3 cm) were minimally handled on site. Therefore, samples were not washed or stored in a buffer but placed directly into 50 ml Falcon tubes and stored on dry ice and then at -80$^{o}$C. Initially, external biopsies of the tongue and skin were collected. Once external

biopsies were obtained, pigs were rolled onto their back and a midline incision was performed from below the neckline of the animals to immediately preceding the genitalia. The complete gastrointestinal tract was removed from the abdominal cavity, with the connective tissue severed where required. Surgical thread was used to seal sections of the gastrointestinal tract. Two knots, approximately 2 cm apart, were tied tightly without severing the gastrointestinal tract. Subsequently, sections of the GI tract were separated by cutting between the tied knots that prevented the intestinal contents from leaking. Both the small and large intestines of animals were sealed in three approximately equal length sections to represent the proximal, medial, and distal regions. All sections of the GI tract were treated similarly. Briefly, an opening into the sealed GI tract tube was created and the contents removed before large representative sections of the bowel were cut and stored. Skin biopsy was taken from around the shoulder. Stomach mucosa was from fundic region. Parenchymal organs were removed from the abdominal cavity of animals and once more large biopsies sections stored for later analysis. The processing time per animal was approximately 3 hours.

(ii) *Macaca mulatta* – rhesus macaques. Six healthy Indian-origin, female adult rhesus macaques aged 5-12 years with bodyweight 5.3 to 10.6 kg were used. All animals were born and raised in naturalistic multi-generational breeding groups at the Biomedical Primate Research Centre (BPRC), Rijswijk, The Netherlands, in comparable environments. All enclosures contained environmental enrichment and bedding to stimulate their natural behaviour. They were daily fed monkey chow pellets (Ssniff, Soest, Germany) in the morning, complemented with fruit and vegetables. Over a period of 5 months animals were euthanized using pentobarbital (70 mg/kg) following sedation with ketamin (10 mg/kg). The necropsy and collection of samples were done immediately after euthanasia. The sterility of the necropsy table and the surgical instruments were assured using Virkon S, sterilization procedures and use of disposable scalpels. Macaque tissue samples were retrieved and stored similarly to the procedures outlined for pigs. The parenchymal organs were collected first followed by the sampling of the GIT. All samples were immediately placed on dry ice and stored at -80$^{o}$C. Tissue samples were transported on dry ice to APC Microbiome Ireland for further processing

**Biopsy preparation procedure, VLP enrichment, and nucleic acid sequencing**

GI and parenchymal organ sections of pigs and macaques were processed identically, in the same research facility, by the same team members, but on different days. Tissue samples were thawed on ice until completely defrosted. Excess faecal material on caecal and colon tissue sections were washed with sterile SM buffer (50 mM Tris-HCl; 100 mM NaCl; 8.5 mM $MgSO_4$; pH 7.5). Tissue sections were stretched and pinned to a Styrofoam board using sterile syringes. Defined volume pinch biopsies of mucosal surfaces were collected with an endoscopic biopsy forceps. A “double-bite” of tissue samples at the same site ensured the accurate and complete loading of the forceps’ jaws. Mucosal pinches were removed from the forceps directly into prelabelled Eppendorf tubes, filled with 400 µL of sterile SM buffer for processing.

To enable comparisons of viral load across biopsy samples, 10 µL of $10^7$ plaque forming units per millilitre of lactococcal phage Q33 were added to all samples. Additionally, Q33 in SM buffer or SM buffer-only were processed as negative controls. Fresh 0.5 M dithiothreitol (DTT) was

prepared in 1 mL of SM buffer. A volume of 16 µL of the DTT stock was added to samples to achieve a final concentration of 20 mM, and samples were incubated at 37°C for 30 minutes. DTT was used to gentle solubilize mucin with minimal disruption of phage virions, as this disulfide bond reducing agent was previously demonstrated to release large quantities of non-mucin proteins from small intestine porcine preparations[1]. Host cellular debris and bacterial cells were pelleted by gentle centrifugation at 4000 g for 30 minutes at room temperature. Subsequently, 400 µL of liquid was aspirated and treated with 40 µL of DNase/RNase buffer (50 mM CaCl2; 10 mM MgCl2), 12 µL of DNase (manufacturer), 4 µL of RNase, and incubated at 37°C for 1 hour with intermittent inversion approx. every 15 minutes. Enzymes were inactivated by incubating at 65°C for 10 minutes.

Viral-enriched samples void of free nucleic acids were lysed using the QIAgen Blood and Tissue Kit following the manufacturers guidelines. However, samples were eluted in only 20 µL of AE elution buffer to increase the final concentration of nucleic acid obtained.

**Virome shotgun library preparation and sequencing**

Reverse transcription (RT) reaction was performed using SuperScript IV First Strand Synthesis System (Invitrogen/ThermoFisher Scientific) with 11 µl of purified VLP nucleic acids sample and random hexamer oligonucleotides according to manufacturer's protocol. Concentration of DNA purified using DNeasy Blood & Tissue kit (QIAGEN) was determined using the Qubit dsDNA HS kit and the Qubit 3 fluorometer (Invitrogen/ThermoFisher Scientific).

Library preparation was carried out using Accel-NGS 1S Plus kit (Swift Biosciences) according to manufacturer's instructions. Briefly, 20 µl of RT product (see above) were taken for sonication after adjusting the volume to 52.5 µl with low-EDTA TE buffer. Shearing of unamplified DNA/cDNA mixture (variable amounts of DNA) was performed on M220 Focused-Ultrasonicator (Covaris) with the following settings: peak power of 50 W, duty factor of 20%, 200 cycles per burst, total duration of 35 s. All following steps were performed in accordance with the manufacturer's protocol. A 0.8 DNA/AMPure beads v/v ratio was used across all purification steps in the Accel-NGS 1S Plus protocol. A single-indexed pooled library was sequenced using 2×150 nt paired-end sequencing run on an Illumina NovaSeq platform at GENEWIZ (Leipzig, Germany).

**Bacterial 16S rRNA amplicon sequencing**

During the biopsy preparation procedure, the porcine and macaque biopsy samples were reduced by DTT followed by centrifugation to reduce host tissue and bacterial cells and enrich the viral-like particles. However, the bacterial-containing pellet was used as the starting material for complementary 16S rRNA analysis of bacterial communities associated with the same biopsy samples analysed with respect viromes. The preparation and sequencing of 16S rRNA gene V3-V4 segment libraries followed the procedure outlined previously[2].

**Analysis of virome shotgun sequencing data**

Raw reads were processed using Cutadapt v2.4 to remove adaptor sequences. Trimmomatic v0.36[3] was used for quality-based trimming and filtration of reads with the following parameters:

‘SLIDINGWINDOW:4:20 MINLEN:60 HEADCROP:10’. Reads aligning to mammalian genomes were identified using Kraken v1.1.1 in combination with *Macaca mulatta* (GCF_000772875.2_Mmul_8.0.1) and *Sus scrofa* (GCF_000003025.6_Sscrofa11.1) reference genome files.

Following removal of mammalian reads, levels of contamination with bacterial genomic reads were assessed by aligning reads to a database of bacterial *cpn60* genes as described before[2]. Reads were then assembled into scaffolds on a per sample basis using SPAdes assembler v3.13.0 in metagenomic mode with standard parameters[4]. Additionally, in attempt to assemble low-abundance genomes, reads were pooled by animal and assembled using MEGAHIT v1.2.1-beta[5]. All scaffolds > 1 kb were then pooled together and an all-vs-all BLASTn search was performed with *e-value* cut-off of ≤ 1E-20. Scaffold redundancy was removed by identifying pairs sharing 90% identity over 90% of the length (of the shorter scaffold in each pair) retaining the longest scaffold in each case.

To extract viral scaffolds from a background of bacterial contamination several selection criteria were used. Firstly, scaffolds aligning using BLASTn v2.10.0+[6] against viral sequences in NCBI RefSeq database (release 203), Gut Virome Database[7], JGI IMG/VR database (v3, release 12-10-2020)[8], our in-house database of crAss-like phage genomes (n=1,576), with at least 50% identity over 85% of contig length (*e-value* cut-off of individual hits ≤ 1E-10) were deemed as viral. Secondly, contigs identified as viral using VirSorter[9] software (all categories) were added. Lastly, ORFs were identified in contigs using Prodigal[10] (*-meta* mode) and translated protein sequences were searched against pVOGs database[11] of phage-specific protein hidden Markov models, using hmmscan (HMMER v3.1b2; *e-value* cut-off of ≤ 1E-5); viral protein sequences from NCBI nr (release 28-11-2020) and RefSeq (release 203) databases and crAss-like phage proteins from an in-house database (n=7,356) using BLASTp v2.10.0+. (*e-value* cut-off of ≤1E-10). Scaffolds coding for at least 3 viral/phage proteins which account for ≥50% of the total number of proteins encoded, or having as few as 2 ORFs, both of which encode viral proteins, were deemed as viral. Viral genomic scaffolds identified by these approaches constituted the final non-redundant viral sequence catalogue (n = 70,615).

Circular genomic scaffolds were identified using LASTZ. Assignment of scaffolds to viral families was accomplished using Demovir script (https://github.com/feargalr/Demovir), as described before[12]. Clustering of viral genomic scaffolds (only for scaffolds with >3 kb in length) into viral clusters (VCs, approximately genus-level operational taxonomic groups) was done using vConTACT2 software[13] with the following optional parameters: *--rel-mode Diamond --db ProkaryoticViralRefSeq85-Merged --pcs-mode MCL --vcs-mode ClusterONE*. Completeness level of viral genomic scaffolds was determined using CheckV[14] with default parameters. Functional annotation of viral genomic scaffolds was performed using Prokka[15] assuming standard genetic code for all scaffolds. Viral HMM database pVOGs[11] and viral RefSeq protein database (release 203), supplemented with protein products encoded by human gut crAss-like phages[16] were used for protein similarity searches. Viral genomic scaffold catalogue was further manually curated to remove coliphage phiX174 genome (commonly used as a spike-in by sequencing facilities), as well as some large scaffolds which on manual examination turned out to be bacterial chromosomal islands containing several prophages in tandem orientation.

For prediction of phage hosts data was aggregated from several sources. Firstly, previously predicted hosts for viral species included into IMG/VR database were assigned to viral scaffolds in our catalogue belonging to the same species (≥95% identity over ≥85% of viral genomic scaffold length, in accordance with MIUViG criteria for viral species demarcation in metagenomic sequence data[17]). Secondly, a search against an in-house CRISPR spacer database was performed as described before[12] to assign hosts to viral scaffolds, missing close homologs in the IMG/VR database. Thirdly, BLASTn similarity of viral scaffolds to closely related ≥90% identity over ≥90% of viral scaffold length) prophages in bacterial genomes (RefSeq database of bacterial genomes, release 99; HMP Reference genomes database[18]) was used to assign hosts where neither IMG/VR nor CRISPR approaches were useful. Lastly, tRNA gene hits against NCBI nt database (release 28-11-2020) and bacterial RefSeq database (release 99) were used to predict hosts for cases where all other methods would have failed.

Quantitative analysis of viral metagenomic data was performed essentially as described before[12]. Quality filtered reads were aligned to the curated viral scaffold database on a per sample basis using Bowtie2 v2.3.4.1 in the 'end-to-end' mode. A count table of scaffolds versus samples was subsequently generated using SAMTools v1.7. Sequence coverage was calculated per nucleotide position per scaffold per sample using SAMTools 'mpileup' command. Read counts for scaffolds in samples showing less than a minimum of 1x coverage of 75% of a scaffold length, were set to zero[12].

Absolute viral counts were calculated for viral genomic scaffolds based on comparison of their relative abundance with that of the externally added standard (lactococcal phage Q33). Only viral scaffolds with estimated completeness of >50% were taken into account, on assumption that additional genomic fragments, which together constitute the remaining <50% portion of the complete genome, will not be counted and therefore will not artificially inflate calculated total viral loads.

**Analysis of bacterial 16S rRNA amplicon sequencing data**

Bacterial 16S rRNA amplicon sequencing data was processed using a pipeline based on USEARCH v8.1 (64 bit). Forward and reverse reads of 16S rRNA V3-V4 segment were merged together allowing for an expected error rate of <0.5 per nucleotide position at overlap. Merged sequences were truncated to remove forward (first 17 nt) and reverse (last 21 nt) 16S rRNA primers. Reads were then dereplicated and singletons were removed, followed by clustering into OTUs at 97% sequence identity level. Chimeras were removed using *-uchime_ref* function with *rdp_gold* reference database. Individual reads were then assigned to OTUs generated above at 97% sequence identity cut-off and read count matrix was generated. Finally, taxonomic assignment of OTUs was performed using RDP Classifier v2.12.

**Statistical methods**

All statistical analysis of sequencing data was carried out in R environment v4.1.0. Descriptive statistical visualisations were created using ggplot2 v3.3.3. Network visualisations were created using igraph v1.2.6 (**Fig. 5**; **Fig. S14**, **S21-S23**). Heat maps were produced using gplots

v3.1.1 (**Fig. S8**, **S9c-d**, **S10-S12**). Sankey diagrams were made using networkD3 v0.4 (**Fig. 3-4**; **Supplementary files 1-14**). Virome β-diversity was visualised through ordination of Bray-Curtis distances using NMDS or dbRDA [metaMDS() and capscale() functions in Vegan v2.5-7 with default parameters; **Fig. 2a**; **Fig. S7**]. Permutational multivariate analysis was performed using adonis() function in Vegan with Bray-Curtis distances. Comparison of Bray-Curtis distances between viromes within organs was done using Wilcoxon test with Benjamini-Hochberg corrections (**Fig. S9a**). Differentially abundant VCs between animal species, organs and tissues were identified using Kruskal-Wallis test with Benjamini-Hochberg correction, followed by a *post hoc* Wilcoxon test for pairwise comparisons (**Fig. S10-S12**). Correlations between fractional abundances of VCs and bacterial genera were calculated using Spearman rank correlation method with Benjamini-Hochberg correction (**Fig. S14**). Correlations between fractional abundances of individual viral genomic scaffolds (viral strains) and bacterial OTUs were calculated using Spearman rank correlation method with Bonferroni correction (**Fig. S21-S23**).

## Data and code availability

All data needed to evaluate the conclusions in the paper are present in the paper, supplementary materials, or additional datasets available at https://figshare.com/s/7da5da849d6ae8aee3e3. Raw sequencing data are available from NCBI databases under BioProject PRJNA753514. Further information and requests for data and resources should be directed to and will be fulfilled by Prof. Andrey Shkoporov (andrey.shkoporov@ucc.ie) and Prof. Colin Hill (c.hill@ucc.ie).

## Supplemental Results

### Overview of virome sequencing results

Illumina NovaSeq sequencing of DNA and cDNA prepared from VLP-enriched fractions yielded 1.8B reads, or 6.2±5.8M per sample (median±IQR; **Fig. S1**) after trimming and quality-based filtration of raw data. Aligning reads against mammalian genomes (*Macaca mulatta*, *Sus scrofa domesticus*, *Homo sapiens*, *Mus musculus*) eliminated 3.4±7.3M reads per sample (1.3B or 70.6% in total), originating from host DNA/RNA contamination.

Our simple viral nucleic acids extraction and sequencing protocol was aimed at an accurate and relatively unbiased representation of viral sequences in the mammalian gut. At the same time, it appears to be prone to considerable amounts of contamination from non-viral sequences. **Fig. S1A** shows that luminal samples from the large intestine deliver largest fractions of non-mammalian Illumina reads that can be aligned to viral genomic scaffolds (**Fig. S1C**). At the same time, mucosal samples from both animal species and luminal small intestinal samples from macaques largely contain mammalian sequences, which can be interpreted as extreme scarcity of viral DNA leading to relatively higher representations of contaminating host genomic DNA.

Additionally, **Fig. S1B** shows that skin, tongue, and stomach in both species, as well as small intestine in macaques, contain high proportion of bacterial genomic DNA contamination, which is still present, but is largely displaced by enrichment of viral nucleic acids in other organs. Altogether,

0.78±3.0M reads per sample (or 550M reads in total) survived removal of mammalian sequences and 0.17±0.83M reads per sample (200M in total) could be aligned to the viral scaffold catalogue.

## Catalogue of viral genomic scaffolds from pig and macaque GIT

Non-mammalian trimmed and filtered Illumina reads were assembled into scaffolds using a combined approach, both from individual biological samples (metaSPAdes assembler) and from per-animal read pools (MEGAHIT). The non-redundant set of scaffolds was further decontaminated of non-viral sequences using a combination of approaches (similarity to nucleotide sequences from virus databases, identification of ≥50% of ORFs as encoding viral proteins, or a prediction made by VirSorter tool[9]). The final catalogue includes 70,615 scaffolds (**Datasets S1 and S2**), ranging in size from 1,000 bp to 285,911 bp and representing both complete circular (n=94), nearly complete high-quality genomes (n=1,253), as well as smaller genome fragments (**Fig. S2A,B**).

As shown in Fig. S2A, a single approach for assigning viral contigs may be insufficient, as only 6,635 contigs were recognised as viral by VirSorter. Moreover, many of the apparently complete, circular viral genomes were categorised by VirSorter as prophages (categories 4-5 in **Fig. S2A**). When aligned against the genome databases of cultured and characterised viruses (viral portion of NCBI RefSeq[19]), as well as databases of complete and partial viral genomes extracted from metagenomic data (crAss-like phage genomes; Gut Virome Database, GVD; JGI IMG/VR v3[8]), a total of 36,024 scaffolds had sequence identity of ≥50% with previously reported sequences, over ≥85% of their length. Of them, 308 aligned to NCBI RefSeq entries, 432 aligned to crAss-like phages, 19,329 were similar to GVD entries, and 28,677 were similar with IMG/VR v3 sequences. At the same time, when a recently proposed threshold (≥95% identity over ≥85% of a contig length[17]) for delineation of uncultured viral species is used, the majority of scaffolds aligned to database entries appear to be novel viral species (72% of scaffolds with hits to IMG/VR database).

As shown in **Fig. S2B** and **C** only a small fraction of scaffolds represent complete or nearly complete viral genomes, whereas the majority of genomes are highly fragmented. Nevertheless, high-quality genomes appear to be the most abundant in the virome samples and recruit highest proportion of Illumina reads (**Fig. S2D**).

Taxonomic assignment of viral genomic scaffolds to viral families revealed that, although the vast majority of them (n=57,642, **Fig. S3**) could not be reliably classified, the highest percentage of identifiable scaffolds belong to tailed bacteriophages (*Siphoviridae*, *Podoviridae*, *Myoviridae*[20,21], crAss-like phages[22], together 12,036 scaffolds) and small phages of the family *Microviridae* (n=436). Of interest is presence of three unique *Leviviridae* genomic scaffolds. These ssRNA bacterial viruses were shown to be highly diverse and omnipresent, but are often overlooked in the metagenomic studies of gut viromes[23]. In addition to these highly diverse phage populations, both animal species carry a small core of eukaryotic viruses, *Astroviridae*, *Caliciviridae*, *Circoviridae*, *Cruciviridae*, *Genomoviridae* and *Parvoviridae*, often represented by complete or nearly complete high-quality genomic scaffolds.

## Viral diversity across individual animals and along the GIT longitudinal axis

Aligning Illumina reads back to the scaffolds catalogue resulted in recruitment rates which differed depending on the anatomical location and type of the sample (mucosa vs. luminal content). Samples taken from the upper GIT, parenchymal organs and mucosal tissue invariably produced low fractional counts of reads which could be aligned to the viral catalogue (**Fig. S1C**). This is largely explained by the differences in total viral loads (**Fig. 1**). Samples with higher viral loads (large intestine lumen) also tend to have larger fractions of reads aligned to the viral scaffold catalogue, despite higher α-diversity of viruses in those samples (**Fig. 2B**). Differences in viral α-diversity between anatomical locations were dramatic, from one or a few viral genomic scaffolds dominating the virome of small intestine and parenchymal organs to many thousands of scaffolds in the large intestine lumen and mucosa samples (**Fig. S4**). Despite this high diversity in the lower gut, there was a tendency for a single dominant virus in many of the samples: such as a genomic scaffold classified as a small *Podoviridae* phage with possible links to *Streptococcus* and *Faecalibacterium* (17.0 kb) in colonic mucosa of macaques M3-M5; or a 96.7 kb genomic scaffold of a crAss-like phage with unknown host present in large intestine lumen in pig E4 (**Fig. S4**, **Fig. S5**).

In order to find out whether the high diversity of viruses in the large intestine was a product of artificial inflation due to highly fragmented assemblies, we employed a recently published CheckV[14] tool to identify proportions of Illumina reads aligned to high-quality/complete genomic scaffolds on one hand, and small genome fragments on the other. As shown in **Fig. S6A,B** there was a tendency for large intestine samples to include even higher proportion of high-quality genomic scaffolds, compared to other body sites, therefore ruling out assembly fragmentation as a source of increased α-diversity in the lower gut.

As discussed in the main text, the extent of virus sharing between individual animals within each of the two species was not very high. Unlike the metagenomic analysis of bacteriomes, which is typically conducted at the level of OTUs, with taxonomic classifications being more robust at genus level[24], viral metagenomic is inherently strain-level, given the rapid evolution and diversity of viruses in the microbiomes[7,25,26]. It is known from human studies that strain-level diversity of the gut virome/phageome is very high and that only a small core of viral genomes is typically shared between unrelated individuals[12,27–29]. Strain level diversity of bacterial hosts, host genetics, individual dietary habits and co-habitation are factors, typically used to explain diversity and individual specificity of human viromes[28,30,31]. It was therefore surprising to see that relatively inbred animals, kept in the same facility and fed with a standardised diet, displayed high level of individual virome variation. Only 27-38% viral scaffolds were present in more than one animal in pig and macaque cohorts, with 2-5% being shared by all members of a cohort. Comparisons of sparse, zero-inflated metagenomic count matrices coming from such divergent viromes are unlikely to reveal any common biological signal[29]. To overcome this we used vConTACT2 algorithm[13] to cluster individual viral genomic scaffolds, both high-quality and fragmented into Viral Clusters (VC), provisional taxonomic units identified based off gene sharing patterns. VCs identified by vConTACT2 roughly correspond to the level of genus in current ICTV (International Committee on Taxonomy of Viruses) taxonomy[17,20]. From 3,888 VCs obtained from clustering of 12,633 individual viral scaffolds (singletons were not taken into account), 59-73% were shared between at least two animals in a cohort and 4-10% were shared across all animals (**Fig. S6C,D**).

Results of NMDS ordinations produced with Bray-Curtis dissimilarities calculated from both the individual viral scaffold-based (**Dataset S3**), and the VC-based fractional abundance community tables, were very similar (**Fig. S7**). As outlined in the main results section, permutational analysis of variance using ADONIS revealed that within-species inter-individual virome variation was almost as strong as the variation between different organs within a particular animal (14% vs 19.6% variance explained in ADONIS, $p = 0.001$). Together with the lack of viral scaffold sharing between individual animals this results in a sparse, zero-inflated community table and significant amount of stress (0.167) on NMDS ordination axes. Aggregation of community table to the VC level resulted in greater sharing of viral entities between individual animals, but did not result in an improvement of NMDS ordination (stress value 0.17). We then proceeded to identify most significant covariates, associated with virome variance using constrained analysis of principal coordinates with Bray-Curtis distances. Together with organ-specific variation, total viral load and α-diversity metrics (e.g. Shannon diversity) seem to explain a sizeable amount of total virome variance (18.7%, $p = 0.001$ in a permutational ANOVA of a capscale model). In a capscale ordination both the Shannon diversity and the total viral load were strongly associated with the first constrained axis, which provides the greatest separation between organs (46.5% variance explained). This confirms that certain compositional differences between the stomach, small intestine, large intestine, and other organs occur along the ascending gradient of virome diversity and total viral load (**Fig. 2b**).

**Viral diversity at neighbouring mucosal sites versus distant sites in the GIT**

In a single pig (E6) and macaque (M6) animals we performed additional paired sampling of mucosal lining (1 cm apart) in order to reveal the level of local mucosal virome variance, and to find out whether viromes of two closely located mucosal sites are significantly more alike, compared to more distantly spaced sites within the same alimentary tract organ. Bray-Curtis distances, based on virome composition at the level of individual viral genomic scaffolds, were calculated for all possible within-animal combinations of two anatomical sites, and compared between paired mucosal samples taken from proximal, medial and distal segments of SI and LI, caeca and stomach on one hand (pig E6 and macaque M6), and mucosal and luminal samples taken from different segments of the SI and LI on the other hand (**Fig. S9a**).

For paired mucosal samples there was a tendency for caecal and LI sites to be more compositionally conserved than SI and stomach, and more related to each other than mucosal samples taken from different segments in LI. These differences, however, do not reach statistical significance due to the small number of samples analysed. At the same time, between-segment differences in LI mucosa were shown to be considerably greater than between-segment differences in LI luminal virome ($p = 0.003$ in Wilcoxon test with Benjamini-Hochberg correction), but less pronounced than between-segment variation of the SI mucosal virome ($p = 0.027$ in Wilcoxon test with Benjamini-Hochberg correction). This findings are in line with a previous report on macaque gut bacteriome[32] which observed greater biogeographic variation of mucosal sites, compared to luminal contents.

These results can at least partly be explained by much lower levels of viral load and α-diversity in the SI sites compared to the LI, and potentially higher level of stochasticity of

qualitative and quantitative virome composition revealed by the metagenomic sequencing, due to a low DNA input (**Fig. 2b; Fig. S9b**).

While paired mucosal samples may not always show more significant similarity between them than that with more distant mucosal sites, hierarchical clustering of all samples by β-diversity reveals that many individual pairs were in fact closer to each other than to any other sample from the same animal. This effect was especially evident in LI and caecum (but not SI) mucosa of macaque M6 (**Fig. S9c**), and SI and caecum (but not LI) mucosa of pig E6 (**Fig. S9d**).

**Differentially abundant viral scaffolds**

In order to identify VCs driving the separation between two different mammalian species, their different GIT organs, as well as between luminal and mucosal viromes of the alimentary tract organs, the following battery of statistical tests was applied. We first identified VCs that were differentially abundant between the two animal species (n = 1,272; $p < 0.05$ in Kruskal-Wallis test with Benjamini-Hochberg correction; **Fig. S10**), and VCs that were differentially abundant between organs across the two animal species (n = 905; $p < 0.05$ in Kruskal-Wallis test with Benjamini-Hochberg correction; **Fig. S11**). We then proceeded with a *post hoc* Wilcoxon test with correction to identify VCs that were discriminatory ($p < 0.05$) between pairs of organs in the following order: Tongue-Stomach (n = 5), Stomach-SI (n = 42), SI-Caecum (n = 427), SI-LI (n = 569), Caecum-LI (n = 33). We also applied the same type of *post hoc* test to identify VCs discriminatory between luminal and mucosal sites in each of the individual organs (n = 736 across all organs; **Fig. S12**). Importantly, we failed to detect a single VC that would be significantly overrepresented in the mucosal samples, compared to matched luminal samples. Instead, all of the 736 differentially abundant VCs were depleted in mucosa.

We then tried to put the differences in virome composition between the alimentary tract organs into the context of similar differences displayed by the bacteriome[32]. To do that, we looked for rank correlations of fractional abundance between differentially abundant VCs (n = 905), which approximately correspond to the taxonomic level of genus, and bacterial genera (n = 379; **Fig. S13**). We set a threshold at the level of strong to very strong relationships (Spearman $\rho \geq 0.6$, $p < 0.05$ with Benjamini-Hochberg correction). By doing this we were able to observe 826 correlated VC-bacterial genus pairs in domestic pigs and 436 in rhesus macaques. With very few exceptions, all detected correlations were positive. Only 8 VC-bacteria pairs in pigs and only 16 in macaques demonstrated negative relationships between them. As shown in **Fig. S14**, bacterial genera, involved in positive correlations with organ-discriminatory VCs, often represent taxonomic groups that are hallmarks of the microbiome of a particular segment or organ in the alimentary tract. For example, in macaques bacterial genera such as *Blautia*, *Dorea*, *Faecalibacterium*, *Gemmiger*, *Oscillibacter* and *Treponema*, themselves typical for the LI bacteriomes, were involved in dense networks of positive correlation with VCs, strongly associated with the LI and caecal viromes. Bacterial genera such as Veilonella, Neisseria, Moraxella, and their associated bacteriophages were linked with the upper GIT organs (**Fig. S13, S14a**). Similar, but distinct, patterns of organ-specific viral and bacterial communities can be seen in pigs (**Fig. S14b**).

**Eukaryotic viruses shared between different anatomic locations in pigs and macaques**

In both species of mammals, parenchymal organs, skin and the alimentary tract organs were found to share collections of eukaryotic viruses, belonging to at least five different viral families (*Astroviridae*, *Parvoviridae*, *Circoviridae*, *Caliciviridae*, *Anelloviridae*) both within individual animals, and across animals. While all five families make up the eukaryotic virome in pigs, macaque viromes are mainly composed of *Circoviridae* and *Caliciviridae* (**Fig. S19**).

For the purposes of this study, clusters of closely related individual viral genomic scaffolds were collapsed together when reaching the threshold of 90% nucleotide identity over 90% of a shorter scaffold length. Therefore, true diversity of viruses at strains, and perhaps even species level cannot not be revealed using the approach used here. Nevertheless, up to 31 non-redundant viral genomic scaffolds (up to 13 per viral family; **Fig. S20**) could be identified in a single animal. These included scaffolds with high level of nucleotide similarity (≥95% identity over ≥85% of viral genomic scaffold length) to known viruses (Porcine parvovirus 5, Porcine bocavirus 5/JS677, Adeno-associated virus 2), as well as genomic sequences of potentially novel viral species, falling below that threshold of similarity to existing species.

Patterns of viral colonisation in all animals were individual, with respect to both, the composition of eukaryotic viruses detected, their distribution between organs and total eukaryotic virus loads. Some viral species seem to be conserved across animals (e.g. 7,507 nt *Caliciviridae* genomic scaffold present in all macaques and pigs), whereas others are individual-specific (e.g. a 3,074 bp *Circoviridae* genomic scaffold found only in macaque M1). The majority of viral load in pigs was concentrated in SI, with viral counts exceeding $10^8$ genome copies $g^{-1}$ in some cases. By contrast, in macaques eukaryotic viruses seem to be mainly associated with LI, as well as SI and parenchymal organs (**Fig. S20**).

The majority of detected viral species tend to be broadly distributed in each of the individual animals, rather than be concentrated to a particular body site. Extreme examples of such uniquitous viruses include a 2,224 bp genomic scaffold of *Circoviridae* found in all macaques, and almost equally abundant in GIT sites from tongue to distal LI, as well as in the spleen, lung, liver and on the skin (**Fig. S19**, **Fig. S20**).

**Supplemental Figures and Tables**

**Fig. S1. Total read counts and non-viral sequence contamination. A**, Total Illumina read counts per anatomical location per animal host species after removal of reads aligned to mammalian genomes (*Macaca mulatta, Sus scrofa domesticus, Homo sapiens, Mus musculus*); **B**, % of bacterial DNA in the sample (after removal of reads aligned to mammalian genomes), calculated based on the fractional abundance of reads aligning to bacterial *cpn60* gene database ([2,33]); **C**, Aggregated Illumina read counts per sample per animal including the reads aligned to viral genomic scaffold catalogue, reads eliminated through alignment against mammalian genomes and reads aligning to neither viral nor mammalian genomes (most likely bacterial origin).

**Fig. S2. Catalogue of viral genomic scaffolds assembled from trimmed and filtered Illumina reads of non-mammalian genomic origin**. **A**, Average scaffold read coverage vs. scaffold length, categories of viral genomic scaffolds identified by VirSorter[9] (1-3, complete phages genomes from most confident to least confident; 4-6 "prophages" from most confident to least confident) and circular genomes (or linear genomes containing direct terminal repeats) identified by LASTZ; **B**, Scaffold coverage and length plotted in the same way but colored by MIUViG[17] viral genome quality level as judged by CheckV[14]; **C**, distribution of viral genomic scaffolds by completeness level as predicted by CheckV with high quality draft and complete genomes by MIUViG standard highlighted in blue; **D**, cumulative fractional abundance of genomic scaffolds with different levels of completeness.

**Fig. S3. Taxonomic distribution, size, and completeness of viral genomic scaffolds.** Different viral families are shown in separate panels. Scaffold size is plotted on log10-scaled x-axis. Scaffold completeness is predicted using CheckV.

**Fig. S4. Fractional abundance of top 26,455 viral scaffolds.** Fractional abundance is shown as a fraction of reads aligned to every scaffolds out of the total humber of reads aligned to the viral scaffold database. Only scaffolds with fractional abundance of > 0.01% in any of the samples are shown. Colours are shuffled and randomly assigned. Samples are grouped by individual animals (M1-6 for macaques, E1-6 for pigs)

**Fig. S5. Fractional abundance of viral scaffolds grouped at viral family level.** Fractional abundance is shown as a fraction of reads aligned to every scaffolds out of the total humber of reads aligned to the viral scaffold database. "Other" category includes families *Adenoviridae* (2 scaffolds), *Caulimoviridae* (2 scaffolds), *Cruciviridae* (2 scaffolds), *Flaviviridae* (1 scaffold), *Picobirnaviridae* (8 scaffolds), *Picornaviridae* (3 scaffolds), *Retroviridae* (1 scaffold), *Virgaviridae* (1 scaffold). Samples are grouped by individual animals (M1-6 for macaques, E1-6 for pigs)

**Fig. S6. Fractions of reads aligned to viral scaffolds of different quality levels, and sharing of viral scaffolds and clusters between animals.** A and B, fraction of reads (out of total number of reads aligned to the viral genomic scaffold database) aligned to either high-quality scaffolds or smaller genome fragments (assigned by CheckV[14]) in different anatomical locations in pigs and macaques; C, number of viral genomic scaffolds or viral clusters (VC) shared between individual animals in macaque and pig cohorts; D, percentage of viral scaffolds (out of 70,614) and VCs (out of 3,887) shared between certain number of animals within macaque and pig cohorts.

**Fig. S7. NMDS ordination of viral communities in various anatomical sites in pigs and macaques.** A, Ordination performed with Bray-Curtis distances calculated from fractional abundance community tables based on individual viral genomic scaffolds (fractions of reads aligned to each of the scaffolds out of the total number of reads aligned to the entire catalogue for a given

anatomical site), NMDS stress 0.167; B, Same ordination done with fractional abundance community table, aggregated to the level of viral clusters (VC), NMDS stress 0.17.

**Fig. S8. Aggregated fractional abundance of viral families across all samples in the study.** Data is log10-transformed.

**Fig. S9. Viral diversity at neighbouring mucosal sites versus distant sites in the GIT.** A, Bray-Curtis distances between mucosal sites separated by 1 cm distance ("1cm_apart", macaque M6 and pig E6) versus same distances between all combinations of mucosal and luminal sites (proximal, medial, distal locations) within an entire organ in each of the 12 animals separately; LI, large intestine; SI, small intestine; Lum, lumen; Muc mucosa; differences did not reach statistical significance in Wilcoxon tests with FDR correction; B, Fractional abundance of viral scaffolds in macaque M6 and pig E6, paired mucosal sites are marked by arrows; fractional abundance is shown as a fraction of reads aligned to every scaffolds out of the total humber of reads aligned to the viral scaffold database; only scaffolds with fractional abundance of > 0.01% in any of the samples are shown; C and D, pairwise Bray-Curtis distances between all mucosal and luminal sites in all macaques (C) and pigs (D); black bars indicate cases where paired mucosal sites were closer to each other than to any other site.

**Fig. S10. VCs (n = 1,272) differentially abundant between pig and macaque animal cohorts across all body sites.** Differentially abundant VCs were selected using Kruskal-Wallis test with FDR correction ($p < 0.05$) by comparing all pig body sites against all macaque body sites; Data is plotted as log10-transformed values.

**Fig. S11. VCs (n = 676) differentially abundant between GIT organs across both animal species.** Differentially abundant VCs were selected using Kruskal-Wallis test with FDR correction ($p < 0.05$) followed by post-hoc test (Wilcoxon with FDR correction, $p < 0.05$) for the following specific anatomic location pairs: Tongue-Stomach, Stomach-SI, SI-Caecum, SI-LI, Caecum-LI; Data is plotted as log10-transformed values.

**Fig. S12. VCs (n = 736) differentially abundant between mucosal and luminal samples at various GIT sites across both animal species.** Differentially abundant VCs were selected using Wilcoxon test with FDR correction ($p < 0.05$), by comparing mucosal and luminal samples on organ-by-organ basis. Data is plotted as log10-transformed values.

**Fig. S13. Fractional abundance of key members of bacterial microbiota at family level (with >20% abundance in any of the samples) determined by 16S rRNA gene amplicon sequencing.** Samples are grouped by individual animals (M1-6 for macaques, E1-6 for pigs); NA denotes a DNA extraction negative control sample ("kitome").

**Fig. S14. Co-abundance networks based on Spearman rank correlation between fractional abundances of organ-discriminatory VCs and bacterial genera.** In macaques, 89 VCs, differentially abundant between GIT organs, had significant ($p < 0.05$) and strong correlations ($|\rho| > 0.6$) with 45 bacterial genera. Similarly, in pigs 149 VCs were correlated with 27 bacterial genera; Network plot was created using iGraph and laid out using Fruchterman-Reingold algorithm; vertices represent organs (orange squares), bacterial taxa (red circles), and VCs discriminatory between different organ pairs; edge thickness represents fractional abundance in a given organ (grey edges), and positive or negative correlation (green or red edges respectively) in pairs of viral and bacterial taxa.

**Fig. S15. Fraction of virome diversity shared between pairs of organs in domestic pigs.** Shared fraction is expressed as percentage of total number of viral genomic scaffolds from one anatomical location ("Organs of origin", aggregate of mucosal and luminal samples) also found in a different location ("Organs of destination"); SI, small intestine; LI, large intestine; Prox/Mid/Dist, proximal, medial and distal portions, respectively.

**Fig. S16. Fraction of virome diversity shared between pairs of organs in rhesus macaques.** Shared fraction is expressed as percentage of total number of viral genomic scaffolds from one anatomical location ("Organs of origin", aggregate of mucosal and luminal samples) also found in a different location ("Organs of destination"); SI, small intestine; LI, large intestine; Prox/Mid/Dist, proximal, medial and distal portions, respectively.

**Fig. S17. Numbers of viral genomic scaffolds shared between pairs of organs in pigs and macaques.** Numbers of shared scaffolds are expressed as aggregate counts of unique scaffolds shared between sites across all animals for each of the two species; SI, small intestine; LI, large intestine; Prox/Mid/Dist, proximal, medial and distal portions, respectively.

**Fig. S18. Absolute counts of some of the most ubiquitous viral genomic scaffolds present in pigs E1-E6.** Only scaffolds shared between 6 or more sites in any of the animals are displayed. Each line corresponds to an individual genomic scaffold (viral strain). Colours are according to viral families. Each panel represent an individual animal.

**Fig. S19. Absolute counts of eukaryotic viral genomic scaffolds in all tested body sites in pigs and macaques.** Each line corresponds to an individual genomic scaffold (viral strain). Colours are according to viral families. Each panel represent an individual animal.

**Fig. S20. Distribution of absolute abundance of individual eukaryotic viral genomic scaffolds across body sites in pigs and macaques.** Each genomic scaffold (horizontal axis) in each

of the animals (individual panels) roughly corresponds to a viral species. Vertical axis shows stacked absolute abundance of individual viruses across body sites (coloured bars).

**Fig. S21. Co-abundance networks based on correlation between fractional abundances of individual viral genomic scaffolds and bacterial OTUs in pigs.** Network plot was created using iGraph and laid out using Fruchterman-Reingold algorithm; vertices represent organs (orange squares), bacterial OTUs (red circles), and viral genomic scaffolds (blue circles); edge thickness represents fractional abundance in a given organ (grey edges), and positive or negative correlation (green or red edges respectively) in pairs of viral and bacterial OTUs.

**Fig. S22. Co-abundance networks based on correlation between fractional abundances of individual viral genomic scaffolds and bacterial OTUs in macaques.** Network plot was created using iGraph and laid out using Fruchterman-Reingold algorithm; vertices represent organs (orange squares), bacterial OTUs (red circles), and viral genomic scaffolds (blue circles); edge thickness represents fractional abundance in a given organ (grey edges), and positive or negative correlation (green or red edges respectively) in pairs of viral and bacterial OTUs.

**Fig. S23. Co-abundance networks of individual viral genomic scaffolds and bacterial OTUs.** Limited to cases where correlation between a virus and a particular host agrees with host prediction (to genus level) from viral sequence analysis.

**Table S1. Sample metadata table. (https://figshare.com/s/7da5da849d6ae8aee3e3)**

**Table S2. Bacterial OTU (16S rRNA gene) read count matrix. (https://figshare.com/s/7da5da849d6ae8aee3e3)**

**Table S3. List of of viral genomic scaffolds and bacterial OTUs correlated by their fractional abundance. Only viral scaffolds also linked to same bacterial genera by CRISPR spacer hits (or other host prediction methods) are included. (https://figshare.com/s/7da5da849d6ae8aee3e3)**

**Dataset S1. Properties of viral genomic scaffolds (n=70,615) recovered in this study. (https://figshare.com/s/7da5da849d6ae8aee3e3)**

**Dataset S2. Sequences of viral genomic scaffolds (n=70,615) recovered in this study. (https://figshare.com/s/7da5da849d6ae8aee3e3)**

**Dataset S2. Viral scaffold read count matrix. (https://figshare.com/s/7da5da849d6ae8aee3e3)**

**Supplementary Files 1 – 14. Interactive graphs showing sharing of viral genomic scaffolds between different anatomical sites in individual pigs (files 1-6) and macaques (files 7-12), as well aggregated result for all pigs (file 13) and all macaques (file 14). (https://figshare.com/s/7da5da849d6ae8aee3e3)**

## References


1. Meldrum, O. W. *et al.* Mucin gel assembly is controlled by a collective action of non-mucin proteins, disulfide bridges, Ca 2+ -mediated links, and hydrogen bonding. *Scientific Reports* **8**, 5802 (2018).
2. Shkoporov, A. N. *et al.* Reproducible protocols for metagenomic analysis of human faecal phageomes. *Microbiome* **6**, 68 (2018).
3. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
4. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
5. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
6. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402 (1997).
7. Gregory, A. C. *et al.* The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut. *Cell Host & Microbe* (2020) doi:10.1016/j.chom.2020.08.003.
8. Roux, S. *et al.* IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Research* **49**, D764–D775 (2021).
9. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985 (2015).
10. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).

11. Grazziotin, A. L., Koonin, E. V. & Kristensen, D. M. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Res* **45**, D491–D498 (2017).
12. Shkoporov, A. N. *et al.* The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. *Cell Host & Microbe* **26**, 527-541.e5 (2019).
13. Jang, H. B. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nature Biotechnology* 1 (2019) doi:10.1038/s41587-019-0100-8.
14. Nayfach, S. *et al.* CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nature Biotechnology* 1–8 (2020) doi:10.1038/s41587-020-00774-7.
15. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
16. 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).
17. Roux, S. *et al.* Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nature Biotechnology* **37**, 29–37 (2019).
18. The Human Microbiome Jumpstart Reference Strains Consortium. A Catalog of Reference Genomes from the Human Microbiome. *Science* **328**, 994–999 (2010).
19. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733–D745 (2016).
20. Adriaenssens, E. M. *et al.* Taxonomy of prokaryotic viruses: 2018-2019 update from the ICTV Bacterial and Archaeal Viruses Subcommittee. *Arch Virol* **165**, 1253–1260 (2020).
21. Koonin, E. V. *et al.* Global Organization and Proposed Megataxonomy of the Virus World. *Microbiol. Mol. Biol. Rev.* **84**, (2020).
22. Koonin, E. V. & Yutin, N. The crAss-like Phage Group: How Metagenomics Reshaped the Human Virome. *Trends in Microbiology* **28**, 349–359 (2020).
23. Callanan, J. *et al.* Expansion of known ssRNA phage genomes: From tens to over a thousand. *Science Advances* **6**, eaay5981 (2020).
24. Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods* **12**, 902–903 (2015).

25. Minot, S. *et al.* Rapid evolution of the human gut virome. *Proc Natl Acad Sci U S A* **110**, 12450–12455 (2013).
26. 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).
27. Manrique, P. *et al.* Healthy human gut phageome. *Proc Natl Acad Sci U S A* **113**, 10400–10405 (2016).
28. Norman, J. M. *et al.* Disease-specific Alterations in the Enteric Virome in Inflammatory Bowel Disease. *Cell* **160**, 447–460 (2015).
29. 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).
30. Minot, S. *et al.* The human gut virome: Inter-individual variation and dynamic response to diet. *Genome Res* **21**, 1616–1625 (2011).
31. Siranosian, B. A., Tamburini, F. B., Sherlock, G. & Bhatt, A. S. Acquisition, transmission and strain diversity of human gut-colonizing crAss-like phages. *Nature Communications* **11**, 280 (2020).
32. Yasuda, K. *et al.* Biogeography of the Intestinal Mucosal and Lumenal Microbiome in the Rhesus Macaque. *Cell Host & Microbe* **17**, 385–391 (2015).
33. 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).