The human intestinal virome

at the mucosal-luminal interface

Austin Yan

A thesis submitted to the University of Ottawa
in partial fulfillment of the requirements for the degree of
Doctorate in Philosophy, Biochemistry

Department of Biochemistry, Microbiology, and Immunology
Faculty of Medicine
University of Ottawa

all the world's a phage

and all the men and women

merely curators

Acknowledgements

"The plan is to fan this spark into a flame" [1] – a plan that has taken nearly a decade of work, a century of CPU time, and a village of supporters to bring this thesis from conception to completion. I dedicate this work to all the old friends who supported me from the beginning, and the new friends I have met along the way.

I acknowledge that my graduate studies were conducted on the unceded traditional territory of the Algonquin people. I am grateful to have had the opportunity to train and work in this land.

I thank my supervisor Dr. Alain Stintzi for the wisdom, direction, and resources you have provided for me to pursue this project as part of my MD/PhD studies. I also thank my thesis advisory committee members, Dr. David Mack, Dr. Mathieu Lavallée-Adam, and Dr. Jim Sun, for their guiding support and for opening up new possibilities throughout my training.

I am thankful for our clinical collaborators at the CHEO Inflammatory Bowel Disease Centre, including Dr. David Mack and Ruth Singleton, for the passion that enables both day-to-day patient care and longitudinal, patient-centred research. We are all also indebted to the patients and their families involved in our studies. You are the foundation and inspiration for our research efforts in the pursuit for better IBD care.

My studies were supported in part by the University of Ottawa, Ontario Graduate Scholarship, and the Canadian Institutes of Health Research Banting and Best Doctoral Research Award. My research was also made possible by the Digital Research Alliance of Canada (formerly Compute Canada), a non-profit organization that provides Canadian researchers with access to advanced research computing. Robust funding of academic institutions, laboratory resources, and graduate students is essential for building and maintaining successful research in Canada.

Fellow members of the Stintzi Laboratory: I have been blessed to train alongside you all, from the highest of highs to the lowest of lows (in no particular order: poop days, Youth Gut Togethers, Biometals 2018, Pandemic Legacy, and 2020 in general). I would not be the person I am today without the guidance and wisdom of Dr. James Butcher, Jennifer Li, Dr. Laeti Schramm, Gillian Tanabe, Juliana Manoogian, Peter Dobranowski, Sasanka Weerasingha, Amine Belaouad, Dr. Kendra Hodgkinson, Dr. Faiha El Abbar, Dr. Annika Flint, and Dr. Whitney Weigel, among others. I also thank Dr. Daniel Figeys and his laboratory members who helped me to refine my hypotheses and presentations over the years through our joint laboratory meetings.

Thank you to my fellow medical students, co-residents, housemates, and friends from Resurrection Church and McMaster University for your unwavering support and kind tolerance of my ability to segue any conversation to the gut microbiome. And I am ever grateful for my parents, my sister Janice, and my dearest, Elaina, for your enduring love and patience through my literal decades in training. You have always been there when I need it most.

Lastly, I give thanks to the microbial world – of viruses, bacteria, fungi, and archaea. Humankind is only the latest page of an epic saga, and I am humbled to continue my training as a medical microbiologist and study this microscopic world. To paraphrase the paleontologist Stephen Jay Gould: this is truly the age of microbes - as it was in the beginning, is now, and ever shall be.

[1] from “My Shot” (Lin-Manuel Miranda, 2016 in *Hamilton*).

Contents

Abstract

The virome is a core but understudied part of the human gut microbiome. New laboratory and bioinformatic technologies have greatly advanced our understanding of our gut viruses, however most studies are based on fecal sampling. The objective of this research was to study the human virome at the intestinal mucosal surface, which enables multisite sampling and the characterization of mucosal viruses at the site of intestinal inflammation.

First, I developed a protocol to extract viral nucleic acids from human mucosal-luminal interface (MLI) samples, which are obtained during endoscopy. I showed that we could reproducibly profile the mucosal virome community. Second, I performed deep multiomic sequencing on MLI samples from three pediatric participants with ulcerative colitis, demonstrating a distinct mucosa-associated viral community in comparison to stool. The combination of metavirome, metagenome, and metatranscriptome sequencing enabled studies of microbial interactions, including bacteriome-virome relationships and prophage analysis. Lastly, we applied our virome methodology in a cohort of over fifty pediatric subjects undergoing investigation for inflammatory bowel disease. We demonstrated that viromes are highly individualized and found that *Crassvirales* were enriched in the proximal colon in participants without IBD. We also performed longitudinal sampling to identify a persistent and highly abundant viral subpopulation.

Altogether, we have contributed methodology and commentary for virome sequencing of an important human sample type, extended the repertoire of virome sequences, and characterized the mucosa-associated virome in the setting of inflammatory bowel disease. These efforts provide practical guidance for future studies on the intestinal virome in human health and disease.

List of abbreviations

CD	Crohn's disease
DC	distal colon
IBD	inflammatory bowel disease
IC	integration-capable
MLI	mucosal-luminal interface
NI	non-integrated
NRA	normalized relative abundance
ORF	open reading frame
PC	proximal colon
STL	stool
TA	transcriptionally-active
UC	ulcerative colitis
VC	viral contig
VLP	virus-like particle

1. Introduction

1.1. The human virome: from dark matter to a new frontier

The human microbiota is the collection of bacteria, viruses, fungi, and archaea that inhabit the human body, predominantly in our gastrointestinal tract. The collective genomes, or metagenome, of our gut microbiota exceeds our human genome by over 100-fold.[1] These microbes contribute to our health, particularly at the intersections of: metabolism, matching the liver in its metabolic potential and influencing host endocrine function; infection, defending against exogenous pathogens and acting as a reservoir for opportunistic infection; and immunity, training our own immune defense systems and contributing to inflammatory processes.[2-4]

Our knowledge of the human microbiome has greatly expanded with advancements in sequencing technology and bioinformatics, yet most microbiome research to date has centered around bacteriology. The original announcement of the "Human Microbiome Project" in 2007 made no mention of viruses, while research on the human virome (i.e. the collection of viruses that inhabit us) continues to lag behind its bacterial counterpart.[5,6] This overlooking of viruses is unfortunate, as viruses inhabit all ecosystems with an estimated 10^{31} viruses on Earth, with 10^{13} viruses inhabiting each human body with potential interactions with our health, either directly through infection or immunity, or indirectly through interactions with bacteria.[7] The lack of virome methodologies and studies had led some researchers to refer to our viromes as microbial "dark matter."[8,9] This last decade (~2013-2023) however has seen a surge in virome research that has been called a "phage biology renaissance."[10]

The following sections outline the foundations and the gaps in our understanding of the human virome that led to my research questions, hypotheses, and contributions in the subsequent

chapters. In sections 1.2 and 1.3, I outline the technological advancements that enabled the discovery and study of the human gut virome. In sections 1.4 and 1.5, I highlight our current knowledge of the human gut virome and its role in health and disease including inflammatory bowel disease. Lastly, in Sections 1.6 and 1.7, I discuss the limitations of fecal sampling and explain our rationale for investigating the mucosal-luminal interface virome and its role in pediatric inflammatory bowel disease.

1.2. Laboratory approaches in human virome research

1.2.1. Modern virology, briefly

Viruses are microscopic entities that contain nucleic acid (either DNA or RNA), a protein coat (the "capsid"), and sometimes a lipid envelope. In order to replicate, viruses rely on infecting living cells: eukaryotes (including human viruses like smallpox, HIV, and SARS-CoV-2) and prokaryotes (including bacteria and archaea). Viruses are: small – impossible to see on a light microscope with the rare exception; diverse – varying greatly in size, structure, and methods of replication; and host-specific – with many potential interactions with their hosts at the cellular, population or organism, and community level.

Virology has come a long way since the first report of a cowpox vaccination in 1796 and the first isolation of a virus – tobacco mosaic virus – in 1892 using ultrafiltration.[11] Our advances in virology is perhaps best highlighted by SARS-CoV-2: within weeks of initial reports of a "pneumonia of unknown etiology" in December 2019, the virus was isolated, sequenced, cultured, visualized, and modelled, aiding the global response to the novel coronavirus and guiding the unprecedented rapid development of effective vaccines and therapeutics.[12-14] Progress in human gut virome research have also been enabled by these technological advances, including viral purification, viral particle visualization, and virome sequencing.

Early efforts to study the human virome were based on established techniques in marine virology. Seawater viruses served as a starting point for virome research, due to the relative uniqueness of viral particles in this sample type (i.e. in size and physical properties) and the absence of free nucleic acids. Thus marine phage researchers pioneered the techniques for the purification

and concentration of viral material including filtration and centrifugation efforts that would later be implemented for human samples.[15]

Microscopy has also played an important role in virome research, with the first electron micrograph of viruses performed in 1938.[11] Both fluorescence techniques and electron microscopy have been used to visualize viral particles in human specimens and confirm proper purification including the removal of bacterial cells.[16,17] Microscopic techniques also allowed for phenotypic classification of intestinal viruses, which are dominated by bacteriophages (or phages, bacteria-infecting viruses).[18] These bacteriophages can be lytic phages – infecting their bacterial cell and leading to cell lysis – or temperate phages, which are able to switch between lytic and lysogenic phases by integrating with host DNA.[19] Most observed bacteriophages were of the former viral order *Caudovirales*, i.e. "tailed viruses" that could be further categorized based on tail length and baseplate.[18,20] Fluorescence microscopy targeted at phages have also been used to estimate viral quantity and study development of the infant virome.[17,21]

Fundamentally though, virome research has been reshaped by advances in sequencing technologies. The first genome ever sequenced was bacteriophage MS2 by Fiers *et al.* (1976) followed by bacteriophage ϕX174 by Sanger *et al.* (1977).[22,23] These early developments allowed for improved virus detection and the detailed study of specific viruses but were low-throughput, resource-intensive, and could not be practically applied at the virome-level.

Sequencing efforts in the early 2000s used linker-amplified shotgun libraries which could amplify hundreds of DNA fragments from viruses and virus-like particle extractions, kickstarting virome (or metavirome) studies.[17] DNA microarrays were also developed which included panels of known viruses.[24-26] Multiplex viral panels are now commonplace in the clinical microbiology laboratory, though most panels target pathogenic human or vertebrate viruses which are not

designed to study the endogenous, diverse human intestinal virome, though custom microarrays for bacteriophages have been developed.[17] Other technologies included randomly primed reverse-transcription PCR for RNA viruses and targeted sequence capture panels utilizing probe-based hybridization such as ViroCap and VirCapSeq-VERT.[24,27,28]

While bacterial communities could be profiled using the 16S rRNA gene, which includes both highly conserved and highly variable regions needed for amplicon-based fingerprinting and identification, there is no analogous viral sequence.[29,30] Some polymerases and capsid proteins may be conserved within viral families and other viral taxons, but cannot be used as a "universal" viral target.[10] The lack of a viral 16S rRNA gene equivalent has also made novel virus discovery more difficult. The majority of virome studies have thus relied on whole metagenome sequencing as a culture-independent and sequence-independent approach in order to capture a given sample's virome.[18,31] These techniques however require increased sequencing depth (compared to 16S rRNA gene amplicon sequencing), further increasing costs and raising the barrier for virome studies.[31] Despite these challenges, targeted viral particle purification approaches, advances in bioinformatic techniques, and the increasing affordability of sequencing have allowed virome studies to proliferate over the past decade, including the work presented here in this thesis.

1.2.2. Viral particle isolation and nucleic acid extraction

Intestinal and stool samples often contain bacteria, fungi, host cells, food particles, and other fecal debris. These DNA and RNA in these components may exceed the viral nucleic acid content by several orders of magnitude. Thus, while shotgun sequencing can be directly applied to clinical samples (i.e. "whole metagenome sequencing), most virome studies involve isolating viruses prior to sequencing for a more targeted and cost-effective approach.[32-44] Whole metagenome sequencing approaches are discussed later in sections 1.2.4 and 1.3.3. General steps

in each virus-like particle isolation protocol include the homogenization of samples in buffer, the removal of fecal debris and cells using physical, chemical, and enzymatic techniques, followed by the subsequent lysis of the purified viral particles.[29,44]

The first proof-of-concept human stool virome study utilized 500 g of fecal material, though more recent studies use 1 g or less of stool.[16] Fecal samples are commonly resuspended in a buffered solution, often saline-magnesium (SM) buffer or phosphate-buffered saline, which facilitates the suspension and dislodging of viruses into the supernatant. This mixture is usually then subjected to low-speed centrifugation to remove cells and debris followed by membrane filtration to further remove bacteria.[41] Various membrane materials may be used including polyethersulfone (PES), polyvinylidene fluoride (PVDF), and cellulose though there is no consensus choice or direct head-to-head comparison for virus-like particle (VLP) purification.[41] Filter pore sizes commonly range from 0.22 μm to 0.8 μm, with many studies utilizing 0.45 μm filters. Filter sizes that are too large may allow passage on smaller bacteria, while filter sizes that are too small are easy to clog, risk reducing viral recovery, and exclude larger viruses including *Prevotella*-infecting Megaphages and giant viruses (targeting amoeba and algae).[41,45] Liang and Bushman (2021) reported that there were "no differences in contamination with bacterial 16S rRNA gene sequences" between 0.22 and 0.45 μm filters, though the supporting data was unpublished.

Due to the resuspension of viruses in solution and possible low initial VLP concentrations, there is often a need for VLP concentration, which may be performed using density gradient centrifugation, precipitation, or membrane elution. Caesium chloride density gradients are thus a commonly used method for viral isolation and can serve to both concentrate and purify VLPs as an alternative or subsequent technique to membrane purification. Caesium chloride centrifugation

is especially useful for high-volume, low virus concentration samples such as seawater.[37] However, this technique is low-throughput, requires specialized centrifugation equipment and high operator expertise, and has been shown to have: significant yield loss compared to membrane purification, high inter-operator variability, and reduced reproducibility.[37,40,41] Given these disadvantages and the relative higher density of VLPs in intestinal samples compared to seawater, caesium chloride gradients are less commonly employed in human virome studies.

Another popular method for VLP concentration is precipitation with polyethylene glycol due to volume exclusion, also referred to as "pegylation".[46] This method, initially used for the purification of specific viruses, was later applied to aquatic viromes and is now commonly used for human intestinal viromes due to its relative affordability, efficiency, and scalability.[40,47] Commercial membrane kits may also be used to concentrate and/or purify viral nucleic acids but may be unsuitable for larger volume preparations. Tangential flow filtration may also be used to assist with viral concentration and desalting, and has similar efficacy to pegylation.[47,48]

Following VLP filtration and concentration, additional purification steps may include lysozyme and/or chloroform treatment, which aim to lyse remaining bacteria or human cells to further reduce non-viral DNA contamination.[32] Chloroform significantly reduces bacterial DNA (over 99% with 10-20% chloroform), however also destabilizes enveloped viruses and impair their recovery.[41] Because most human intestinal viruses are non-enveloped bacteriophages, chloroform treatment is still often employed. Cell lysis efforts are usually paired with endonuclease treatment to digest free DNA from lysed cell, while viral nucleic acids remain protected inside intact virus-like particles.[40,41]

Thus, using these physical, chemical, and enzymatic techniques, complex clinical material including mucosal samples or stool may be processed to obtain highly purified viral concentrates.

A summary of select virome protocols is presented in Figure 1. These steps may be adapted to target or exclude specific elements of the virome or to improve purification at the cost of viral recovery. Additional techniques including flow cytometry and hydroxyapatite chromatography have also been applied to select for viruses of specific size prior to further analysis.[49-51] Many of these purification techniques retain virus viability, and thus these concentrates may be used to study infectivity or be used for transplantation studies. These viral concentrates may also be examined directly using electron microscopy for quantification and phenotypic characterization.[47] For most virome studies though, viral particle purification is followed by nucleic acid extraction and sequencing, described in the following section.

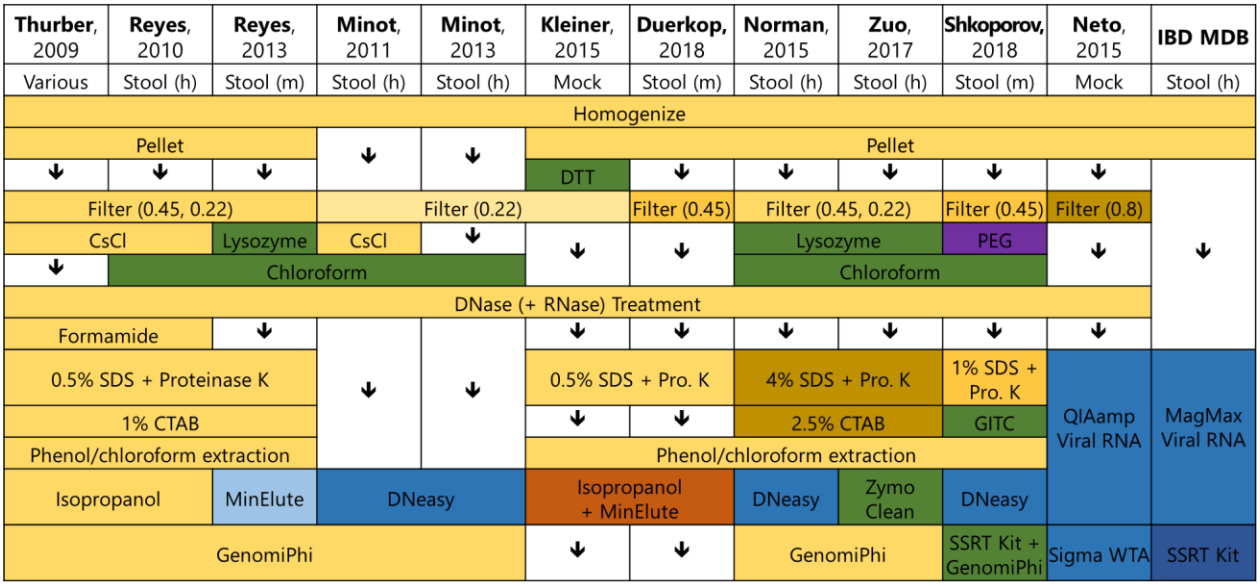


Figure 1: Select virome protocols published over the past decade.

Protocols for purifying virus-like particles and extracting viral nucleic acids were designed for mock communities or stool from humans (h) or mice (m). Filter pore sizes are listed in micrometers. Steps employed by Thurber *et al.* are shown in yellow, with further changes highlighted by various colours. IBD MDB: Inflammatory Bowel Disease Multiomics Database; DTT: dithiothreitol; CsCl: caesium chloride; SDS: sodium dodecyl sulfate; Pro. K: proteinase K; CTAB: cetyl trimethylammonium bromide; GITC: guanidinium isothiocyanate; SSRT: SuperScript IV Reverse Transcriptase (Thermo Fisher); Sigma WTA: Complete Whole Transcriptome Amplification Kit (Sigma); MagMax TNA: MagMAX Total.[32-43]

1.2.3. Nucleic acid extraction and amplification

Once viral particles are isolated, viral nucleic acids can be extracted then purified. Commonly, viral capsids are broken down using a combination of detergents, proteinases, and/or chaotropic salts, such as proteinase K and sodium dodecyl sulfate (SDS).[40,47] Viral nucleic acids may then be precipitated and purified. Frequently used methods include: commercial nucleic acid kits (i.e. spin columns with silica fibres or magnetic beads); cetrimonium bromide (which precipitates and stabilizes nucleic acids), and phenol/chloroform extraction.[32,41,48] Often, a combination of nucleic acid extraction and purification techniques are used (see Figure 1). These techniques may be adjusted to optimize for DNA or RNA extraction.

A key challenge in virome sequencing efforts is low viral nucleic acid quantities, often falling below the minimum quantity of DNA required for most sequencing platforms.[10] An estimated 500 ng of viral DNA may be isolated from 2-5 grams of feces, but this quantity varies greatly depending on the viral particle isolation and nucleic acid extraction techniques.[33] Thus, nucleic acid amplification steps are frequently used to obtain sufficient input DNA. In the absence of a universal viral PCR target, amplification techniques usually aim to amplify all input nucleic acids. The most frequently employed technique is multiple displacement amplification (MDA), which was initially developed by Dean *et al.* (2002) for the Human Genome Project.[29] This isothermal technique uses the Φ29 DNA polymerase and random exonuclease primers by rolling-circle amplification, which reduces amplification bias compared to other PCR-based methods, has a low error rate, and avoids degradation in DNA quality due to repeated high-temperature denaturation.[52] However, MDA approaches also introduce bias toward small circular DNA viruses including those in the family *Microviridae*.[53] Additionally, chimeras may form during the strand

displacement process, creating amplification artifacts.[54] One virome database study, however, found no significant difference in viral recovery between MDA and non-MDA virome datasets.[55]

Nextera is a commercial method developed by Epicentre that only requires a small amount of input DNA that is fragmented by a "transposome"; a transposon sequence is then added to the ends of these segments and used for PCR amplification.[56] This technique avoids the chimeras and selection of circular viruses that are associated with MDA approaches and has become increasingly popular in recent years, but can cause biases based on sequence GC content.[56,57]

Linker-amplified shotgun library construction is an alternate method used by many early virome studies, initially performed with Sanger sequencing but later adapted for next-generation sequencing platforms.[17,50] However, this method requires a higher initial viral load and is less used in comparison to MDA and Nextera approaches. Other methods include sequence-independent single-primer amplification, degenerate oligonucleotide-primed PCR, and multiple-annealing and looping-based amplification cycles, discussed further in Regnault *et al.* (2021).[58]

After viral nucleic acid extraction, viral RNA may be converted to cDNA to facilitate sequencing. This step is often accomplished with commercial kits that incorporate a reverse transcriptase, such as a whole transcriptome amplification kit.[41] Further rRNA depletion steps may be taken to remove bacterial RNA, though these techniques have not been broadly adopted across virome studies.[59] However, many protocols exclude RNA viruses to reduce protocol complexity, especially as most intestinal viruses are DNA viruses, likely leading to an underrepresentation of RNA viruses in virome datasets.

1.2.4. Virome sequencing

Virome research has been enabled by the increasing affordability and ongoing development of DNA sequencing, from Sanger sequencing to massively parallel, short-read "second-generation" or "next-generation" sequencing (i.e. 454 Life Sciences, Illumina, Ion Torrent, etc.) to long-read "third-generation" sequencing (i.e. Pacific Biosciences, Nanopore). According to data from the National Human Genome Research Institute, the cost per megabase (Mb) has decreased from $5,292.39 USD in 2001 to $0.23 USD in 2011 to $0.013 in 2016 (the start of this project) to $0.006 in 2022.[60]

Given the increased sequencing depths required for virome sequencing compared to amplicon-based (i.e. 16S rRNA gene) bacteriome sequencing, these ongoing reductions in cost have been essential for allowing virome research to flourish in recent years. The first human "metavirome" study was published in 2003, when Breitbart *et al.* used a linker-amplified shotgun library and Sanger sequencing to investigate 532 viral sequences from a single human stool sample.[16] The majority of virome studies over the past decade, primarily using short-read sequencing technologies on the Illumina platform, range from 100,000 to over 50 million reads per sample.

Because most bioinformatic tools and databases available today have been built on virome analyses from short-read sequencing technologies, they are limited by the shortcomings of these platforms including challenges in viral genome assembly. Several recent efforts have explored the use of long-read sequencing technologies for virome research, which may be hampered by relatively lower sequence quality.[61] Ongoing improvements in long-read sequencing and implementation of hybrid approaches have been helpful in improving the recovery of viral genomes and their assembly quality.[61-63]

1.2.5. Whole metagenomic sequencing and other "multiomic" approaches

The increasing affordability of DNA sequencing has also opened a new avenue of virome analysis: "whole metagenome sequencing" (also referred to as bulk metagenome sequencing, whole community metagenomes, mixed community metagenomes, etc.). In the absence of viral particle purification, the whole microbiome – including viruses – may be captured in a single sequencing library. These approaches have enabled improved bacteriome analysis when compared to amplicon-based sequencing approaches, but also allow for viral DNA identification. While these techniques may avoid some of the biases introduced by virus particle isolation procedures, variations in sample processing may still skew the efficacy of viral nucleic acid extraction and limit interstudy comparisons. Several bioinformatic pipelines have been designed to "mine" whole metagenome datasets for viral sequences which are discussed in section 1.3.3.

Paired virome and whole metagenome datasets can capture distinct viral populations and can also provide insight into the limitations of each approach.[64] Virome sequencing efforts may be dominated by a subset of lytic viruses and cannot capture temperate phages in the lysogenic cycle, i.e. inactive (or dormant) prophages. On the other hand, whole metagenome analysis can capture viral genomes that are absent from the purified virome sequences. These bulk metagenome-only viruses could represent: integrated prophage sequences (which may be intact or decayed), extrachromosomal sequences, "plasmid prophages", chronic intracellular phages that may or may not produce phage virions but do not kill the host bacterium, or false positive annotations.[65,66] Comparison of the two datasets can help to characterize prophage activation and improve understanding of viral population dynamics.[65,67]

Viromics is increasingly included as part of "multiomic" or systemics biology approaches to human microbiome studies, alongside other analyses including metatranscriptomics (i.e. community-level RNAseq), proteomics, and metabolomics.[67-70] While viromics were left out of

the initial human metagenome project, its second phase – the "Integrative Human Microbiome Project" specifically highlights the importance of multiomic analyses and includes viromics as one of its primary data types.[71] Today, there are limited tools that incorporate multiple layers of data, however this represents a key frontier to further our virome research and our understanding of its transkingdom interactions.

1.3. Bioinformatic approaches to human metaviromics

1.3.1. Expanding viral databases

Most nucleic acid sequencing analysis relies on the availability of sequence databases, which provide a reference for read or assembly annotation. The lack of viral databases, which stalled behind bacterial databases due to the lack of conserved PCR targets and the relative difficulty of viral culture, created significant technical challenges for early virome studies. Brietbart *et al.* noted in 2003 that "the majority of the [viral] sequences have been novel."[16] Most virome analyses before the early 2010s had 60-99% of their reads demonstrate no similarity to known sequences, with a "lack of suitable bioinformatic method[s] to characterize the unknown sequences."[7,24,72] These unidentifiable sequences would sometimes be described as viral "dark matter."[10,66,73,74] This lack of prior sequencing would even be used as a quality indicator for a viral sequence, where the absence of alignment to the large NCBI nucleotide ("nt") database has been used to help assign a contig to be of likely viral origin.[40,75]

Refseq, a public repository for nucleic acid sequences as part of the National Center for Biotechnology Information (NCBI), has had a steadily growing number of viral sequences since the early 2010s.[76] Many key virome papers in the early 2010s used Refseq for viral sequence annotation, however this repository is not targeted towards human intestinal viruses (which are mostly bacteriophages) and includes a diverse range of viruses including mammalian and plant pathogens. Of 11,700 Refseq complete genomes in 2024, about 4200 are bacteriophages.[77]

Several curated databases for the human gut virome have since been developed, including the "Gut Virome Database" and "Gut Phage Database."[55,78] These databases primarily mined whole gut metagenome and metaviromics datasets and used clustering processes to generate a set of non-

redundant viral genomes.[79] These human gut datasets exceed the size of Refseq's total collection by up to a thousand-fold, and instead of relying on manual, uploader-guided sequence annotation, these viral databases often employed bioinformatic tools that could systemically perform quality scoring, taxonomic annotation, and phage host prediction in a systemic process that was more useful for metavirome analysis.[79] Another catalog, IMG/VR (Integrated Microbial Genomes / Virus) also served as a large global virome database which includes both environmental and human-associated viromes. Version 1, published in 2016, became the largest public database with 264,413 viral contigs from over 6000 environmental and human samples;[80] Version 4, published in 2023 includes over 15 million viral genomes and fragments.[81] A list of select viral databases is shown in Table 1.

As a result, the proportion of unaligned viral reads have dropped in recent years, where the majority of viral reads can now be mapped to existing databases, though many previously sequenced viruses still lack basic taxonomic annotation.[79,82] Furthermore, while all viral databases including quality control efforts, viral sequences are subject to artefacts such as annotation, sequencing, and assembly errors. These databases are also prone to reporting bias, such as the exclusion of RNA viruses in most protocols and the underrepresentation of samples from Africa and South America.[55,78,79] Yet, the expansion of viral databases in the past few years is encouraging and will likely continue to improve with ongoing sequencing efforts, new bioinformatic tools, and the inclusion of long-read sequencing technologies which will improve viral genome assembly efforts.

Table 1. Human gut virome databases, partially adapted from Li et al. (2022).

Name	Source	Sample size	Viral contigs	Citation
Gut virome database (GVD)	Human gut	2,697	57,605	Gregory *et al.* (2020)
Cenote human virome database (CHVD)	Human gut, oral, skin, nasal, vaginal	5,996	>180,000	Tisza and Buck (2021)
Metagenomic gut virus (MGV)	Human gut	11,810	189,680	Nayfach *et al.* (2021)
Gut phage database (GPD)	Human gut	28,060	697,817	Camarillo-Guerrero *et al.* (2021)
IMG/VR (v4)	Environmental, human	134,822	15,677,623	Camargo *et al.* (2023)

1.3.2. Metavirome analysis

Raw sequencing reads in virome research are processed like other microbiome studies, requiring quality filtering and the removal of human reads. Human reads may be removed by alignment of reads to the human genome using SOAP, BlastN, SNAP, or bowtie2.[85-88] Bacterial sequence removal may be evaluated by detecting the presence of conserved bacterial sequences including 16S rDNA or the housekeeping *cpn60* gene.[40,57,89] The resulting output should be a set of high-quality viral reads, which can be directly aligned to viral databases and/or assembled.

Read-based analysis is often limited to the completeness of viral databases and will likely involve a significant proportion of unmapped reads, though this proportion has decreased in recent years. As viral databases improve, unmapped reads could represent novel viruses but could also indicate contaminating sequences (including human, bacterial, or other microbial sequences). While some studies choose to omit unmapped reads from analysis, this analysis could result in misleading interpretations, as demonstrated by Clooney *et al.* who re-analyzed a prior study using database-independent approaches and were unable to previously reported findings of viral richness when accounting for these omitted reads.[8]

Affordable virome sequencing and smaller genomes (compared to their bacterial counterparts) have enabled metaviromic assembly that would be unattainable for bacteriomes at similar sequencing depths.[21] Compared to read-based analysis, these longer sequences allow for higher-resolution taxonomic annotation, improved functional gene assignments, and enables the study of whole viral genomes. Researchers often employ *de novo* metagenome assemblers such as MetaSPAdes, MetaviralSPAdes, and MEGAHIT, which are usually used to assemble viral contigs on a sample-by-sample basis, though each has its limitations especially when there is low read coverage or genomic repeats.[90-93] A panel of 16 assemblers were assessed by Sutton *et al.* (2019)

on their ability to metaviromic assembly performance, showing that different algorithms may alter the predicted viral composition, though software choice may depend on sequencing input and computational resources.[93] Bioinformatic tools like Hecatomb that assemble reads from multiple samples and/or datasets to facilitate comparative metagenomics are referred to as cross-assemblers.[94] While virome assemblies, like viral reads, are also limited by viral databases, longer sequences are more useful for virus prediction software while unannotated assemblies can still be readily used in database-independent analyses including viral richness and diversity studies.

After assembly, contigs may be analyzed for quality using programs like CheckV and viralComplete.[92,95] Key metrics include circularization, genome quality (including an assessment for bacterial contamination), and genome completeness. Quality-control is often incorporated into prebuilt or custom bioinformatic pipelines aimed to remove low-quality or contaminating viral sequences.[96] Many viral pipelines have been published over the past few years including VIROME (2012), VirusFinder (2013), Metavir 2 (2014), VIP (2016), VirusSeeker (2017), VIBRANT (2020), VirSorter2 (2021), and Cenote-Taker2 (2021).[83,97-103]. Kraken2 has also incorporated increasing number of viral sequences into its databases.[104] While earlier programs relied on direct viral sequencing matching using tools like BLAST, newer tools have had to utilize machine-learning approaches in order to process larger datasets and databases. New bioinformatic tools are also being built to incorporate long-read sequencing technology, such as viralFlye (2022).[105]

Virome assemblies may be further clustered to facilitate intraindividual analysis or to construct viral databases. Clustering techniques include DNA-sequence based approaches such as CD-HIT and clustergenomes, as well as protein-based clustering approaches including vContact2, MMSeqs2, and HHBlits.[106-110]

While taxonomic groupings have been a useful tool in bacteriome analysis, such as phylum-level observations (i.e. *Bacteroidetes* and *Firmicutes* ratio) or genera-level groupings (i.e. *Enterobacter*, *Prevotella*), viral taxonomy has lacked similar organization. Many bacteriophages had been grouped based on viral morphology (i.e. *Siphoviridae* with long, noncontractile tails) instead of genomic similarity, despite most modern virome studies centered around sequencing approaches.[111] There have been significant efforts to improve viral classification, most notably with the 2022 ICTV taxonomy release, which abolished the viral order *Caudovirales* (the predominant order of intestinal bacteriophages) and its *Siphoviridae*, *Myoviridae*, and *Podoviridae* subfamilies in favour of the class *Caudoviricetes* and many new viral orders.[111,112] There is also an effort to standardize virus naming towards binomial species names.

Because of challenges in viral taxonomic organization, some virome research rely instead on DNA or protein clustering techniques for *de novo* viral phylogeny analysis.[45] Viral phylogeny analysis can also be performed using relatively conserved viral genes, including capsid-encoding genes and the large terminase subunit *terL*, though these studies are usually limited to a specific viral family or other taxon due to high viral genome variability.[113] Efforts to track viral evolutionary phylogeny and classify deep viral taxons are further hampered by the high rate of horizontal gene transfer among viruses.[114]

While these efforts to improve and standardize viral taxonomy holds promise in the coming decade, the heterogeneity in existing databases and bioinformatic programs complicates current efforts for virome analysis. Due to challenges in viral taxonomic assignment, there are also many gaps in the viral taxonomic tree (i.e. unassigned families and genera within *Caudoviricetes*) which also can also make taxon-level analysis difficult.[114]

1.3.3. Whole metagenome sequencing analysis

As discussed in section 1.2.5, whole metagenome sequencing refers to shotgun sequencing approaches that omit the pre-selection of virus-like particles.[79] Often, these techniques have been used for primarily bacteriome analysis, but also may contain viral sequences which could reflect both extracellular and intracellular viral particles depending on the DNA/RNA extraction procedure as well as prophages, which are present in more than half of all bacteria.[66] Viral sequences can thus be identified (or "mined") from whole metagenome datasets, include those that did need initially seek to perform metaviromics analysis. One large benefit of virome mining is data "reuse" – i.e. the ability for new bioinformatic insights without the additional cost of sample extraction and sequencing.[115]

Programs designed for viral sequence identification include MARVEL, Virfinder, VirMiner, and VIBRANT.[98,116-118] Several programs including Cenote-Taker and VirSorter also have adjustable parameters for both metavirome and whole metagenome applications.[83,99] These programs are usually trained on a curated database of viral sequences or protein families, and use unique viral features to identify viral sequences. Assemblies that contain bacteria-specific sequences or predicted proteins may also be identified and excluded; VirSorter refers to this function as "virome decontamination mode."[99] VIBRANT and Cenote-Taker can also identify prophages when both host and viral open reading frames are detected on the same contig; these sequences can then be split and analyzed separately from their associated bacterial genomes.[98] About 2-10% of a whole metagenome dataset may be represented by viral sequences, including an average of 8.6% of reads in a large stool metagenome study of over 10,000 publicly available stool metagenomes.[84]

1.3.4. Inferring virome function and virome-bacteriome interactions

Most virome tools and bioinformatic pipelines utilize the detection and translation of open reading frames to predict viral proteins with programs like Prodigal and MetaGeneMark[119,120] While many virome studies before the mid 2010's used general or microbial protein databases such as KEGG, PFAM, and COG, multiple curated databases of viral proteins (or orthologous groups) are also now available, namely pVOG and VOGDB.[33,121-124] Viral protein prediction can be used to assess viral genome completeness, infer lysogenic potential, and infer viral homology and diversity.

Several bioinformatic tools also include bacteriophage-host prediction. These tools may use: alignment-based methods, which compare viral genomes to phage or phage elements with a known bacterial host; sequence composition similarity, as viruses tend to use similar codons and have similar oligonucleotide frequences as their target host; and machine-learning models inferred from known phage-bacterial pairs.[125-127] CRISPR-Cas9 – the bacterial and archaeal defense mechanism that incorporates short sequences of mobile genetic elements such as bacteriophages – may link a viral target to its putative host.[79] Viral sequences may aligned against a reference database of known CRISPR spacer sequences while paired bacteriome sequencing may be subjected to CRISPR spacer detection programs to create a curated set of possible bacteriophage-host pairs.[128-131]

Prophage studies are also helpful for identifying bacteriophage-host relationships.[66] Whole metagenome studies can demonstrate prophage integration and identify the host bacterium based on flanking sequences, while presence of an integrase gene or genome partitioning systems may be used as markers of a temperate phage.[57] Several prophage related tools including PHASTER and PHASTEST, Prophage Hunter, and BACPHLIP can be used to identify and classify prophages based on expected activity.[132-134] Studies that have paired bacteriome and virome sequencing can

also utilize correlation networks between bacteria and known phages may identify co-occurring phages.[135]

These *in silico* predictions can be further bolstered and confirmed by culture-based methods and viral-tagging based approaches, though these methods are relatively labour intensive.[18] Furthermore, most of these phage-host prediction models assume highly specific (i.e. single) phage-host relationships, when some bacteriophages could have a broader host range.[136] These tools remain an active area of development for virome research.

1.4. Composition of the human gut virome

1.4.1. Gut viromes: individualized bacteriophage communities

The rapid development of virome sequencing tools have enabled the characterization of the human virome at various sites, including: the gastrointestinal tract including the oral cavity and saliva, intestinal contests, and feces;[21,33,36,45,137] the skin virome;[57,138] the respiratory tract;[139,140] and the genitourinary tract.[141] Virome composition varies significantly by site: while the gut virome is predominantly made of bacteriophages, eukaryotic viruses may be more dominant in skin and respiratory viromes.[138,140] The remainder of this chapter will focus on the gut virome – the most abundant and most studied population of human viruses. Most "gut virome" studies have relied on fecal sampling, thus representing colonic luminal contents; the biogeography of the intestinal virome is discussed later in section 1.6.

Based on microscopy, human feces are estimated to have ~10^9 of virus-like particles per gram, with a total virus count in our body within the same order of magnitude as the number of bacteria or human cells.[20,113] The human virome is highly personalized, more than the bacteriome, with large studies suggesting that most viruses are only sporadically detected (i.e. present in <0.5% of all populations) with no single virus present in over half of all individuals, though this is limited by current virome sequencing efforts including studies with low sequencing depth and varying VLP extraction methods.[45,55] The concept of a "core virome" remains debated;[8,45,142-144] regardless, the lack of shared viruses complicates virome comparisons between individuals.[145] Gut viromes are also influenced by geography, diet (which can be influenced by culture and ethnicity), industrialization, and urbanization.[146,147]

The large majority of human intestinal viruses are prokaryotic (bacteria and archaea infecting) viruses. In a survey by Gregory *et al.* (2020) of nearly 2,000 metaviromes, an average of 97.7% of gut viruses were bacteriophages, followed by 2.1% eukaryotic viruses and 0.1% archaeal viruses.[55] Most other studies demonstrate similarly low levels of eukaryotic viruses, though this fraction is elevated in the setting of viral gastroenteritis.[45,148] The following sections describe prokaryotic gut viruses with special attention to the viral order *Crassvirales*, followed by a brief discussion on eukaryotic viruses and the assembly of viruses during infancy.

1.4.2. Prokaryotic gut viruses

Prokaryotic viruses are often grouped by their genome form (i.e DNA or RNA, double or singled stranded, linear or circular), morphology, and replication style, which can be virulent or temperate. Virulent or lytic phages replicate only by the lytic cycle, while temperate viruses may also replicate through a lysogenic cycle which involves bacteriophage integration into the bacterial genome (i.e. a prophage).

Most intestinal bacteriophages are tailed, double-stranded DNA (dsDNA) phages that belong to the class *Caudoviricetes* (replacing the defunct viral order *Caudovirales*).[112] Historically, these phages were grouped into *Myoviridae*, *Siphoviridae*, and *Podoviridae* based on their morphology, though this taxonomy was recently abolished with new families *Straboviridae*, *Autographiviridae*, and *Drexlerviridae*, respectively (with modifications to classification based on genomic characteristics).[111,149] *Caudoviricetes* also includes the new viral order *Crassvirales*, which are among the most abundant intestinal phages (described in the following section). *Caudoviricetes* have linear dsDNA genomes that range from 16-140 kb in length, can be virulent or temperate, and share a distinctive major capsid protein (MCP) and terminase gene (large subunit: TerL) that may be used for phylogenetic studies.[150]

The other main taxonomic group of gut bacteriophages are *Microviridae*, which are icosahedral (symmetrical, twenty-faced polyhedral structures), non-tailed, single-stranded DNA (ssDNA) viruses. *Microviridae* have circular genomes of 3-7 kb, are usually lytic viruses, and may be overrepresented when samples are subjected to multiple displacement amplification (MDA) prior to virome sequencing.[53] Other ssDNA bacteriophages seen in gut virome studies include *Inoviridae*.[45]

In small cohort, longitudinal studies involving healthy patients, human gut viromes appear to be relatively stable over time, with the majority of viral contigs representing highly abundant bacteriophages that persist for several years.[35,45] Many of these bacteriophages – *Caudoviricetes* and *Microviridae* – are predicted to infect abundant and persistent gut bacteria including *Bacteroides* and *Faecalibacterium* species.[45] Some studies suggest that persistent viruses are more likely to have genes related to lysogeny (i.e. temperate phages).[33,36,151] Temperate phages may also enter a "pseudo-lysogenic" state including persistence through low copy number plasmids while stable low-level production of phage particles has also been observed.[151,153] Shkoporov *et al.* (2019) noted that many persistent viruses were not temperate, suggesting the presence of dormant lytic phages.[152]

Notably, prokaryotic viruses have substitution rates as high as 4% in some *Microviridae*, which may contribute to virome microheterogeneity, i.e. the presence of multiple subpopulations with various single nucleotide polymorphisms (SNPs).[36,45] The high rate of viral substitution may also explain the individualization of viromes that limit identification of a "core virome."

1.4.3. Crassvirales

Viruses in the order *Crassvirales* (also referred to crAssphages/crassphages, crassviruses, and crass-like phages) have become the posterchild of the "phage renaissance." CrAssphages were only first reported in 2014, named for the "cross-assembly" bioinformatic tools that led to its identification; prior to this, these sequences were hidden among viral "dark matter".[74,94,111,154] Since its discovery, phages of the order *Crassvirales* have been found to be prevalent and ubiquitous: accounting for 1.68% of all fecal whole metagenome sequencing; representing up to 90% of an individual's gut virome; and present in 77% of human fecal samples worldwide.[154] CrAssphages have been found in other primates but rarely in other animals, suggesting that crAssphage and humans may have co-evolved.[155-157] Given their high prevalence and association with humans, *Crassvirales* have been proposed as markers of human fecal contamination in environmental samples and have been used to normalize SARS-CoV-2 wastewater surveillance studies, outperforming other markers including pepper mild mottle virus and human 18S rRNA.[157-160]

CrAssphages have received significant attention from the International Committee on Taxonomy of Viruses (with the creation of a *Crassvirales* Study Group). In the 2022 taxonomy update, these viruses were given its own viral order, *Crassvirales*, which contains four novel viral families (*Steigviridae*, *Intestiviridae*, *Crevaviridae*, and *Suoliviridae*), 42 genera, and 73 species.[111] *Crassvirales* are within the class *Caudoviricetes*, have linear dsDNA genomes ranging from 83-106 kb, have a short-tailed podovirus-like morphology, and infect hosts within the phylum *Bacteroidetes*.[111,161] Interestingly, crAssphages are not lysogenic but also do not have typical virus-host dynamics of lytic viruses, instead demonstrating long-term, phage-host co-existence.[162,163]

Several crAssphage databases have been created, including a collection of 249 genomes by Guerin *et al.* (2018) and 694 genomes Yutin *et al.* (2021).[164,165] These databases have enabled

genome analyses and have been used as a curated database to identify other *Crassvirales* within metagenomes. Only a few CrAssphages have been isolated in viral culture to date. ΦcrAss001 (family *Steigviridae*, species *Kehishuvirus primarius*) has underwent extensive study including genome analysis, protein structure, cryo-electron microscopy reconstruction, and replication study in its host *Bacteroides intestinalis*.[162,166] ΦCrAss002 (family *Intestiviridae*, species *Jahgtovirus secundus*) has also been isolated and found to infect *Bacteroides xylanisolvens*.[167]

1.4.4. Eukaryotic gut viruses

Eukaryotic viruses, including human and plant viruses, are also commonly detected in the human stool samples, though usually at much lower abundances than bacteriophages.[24] Many eukaryotic viruses are enveloped and/or RNA viruses, which may contribute to their underreporting in most virome studies. Human pathogens including rotavirus, enterovirus, adenovirus, and norovirus are shed in the setting of viral gastroenteritis, and may also be found in asymptomatic patients.[168,169] Outside the setting of enteric infection, the most frequently sequenced eukaryotic viruses are *Anelloviridae* (a family of non-enveloped, ssDNA virus with small circular genomes) and *Geminiviridae*.[170,171] Plant viruses such as *Virgaviridae* may also be found in stool, presumably from dietary sources.[20] Like the "phageome", eukaryotic viromes also appear to be personalized; its stability over time is not as well studied.[172]

1.4.5. Virome development through life

Many studies have aimed to characterize the assembly and development of the virome in early life. Meconium – the first feces passed by newborn – contains very low numbers of viruses. Using fluorescent microscopy, Breitbart *et al.* (2008) did not detect any viral particles in meconium

(n=1) while Liang *et al.* (2020) were able to detect VLPs in 3 of 20 samples that were labelled as "meconium / early stools."

The infant virome is established quickly, reaching $10^8 – 10^9$ VLPs/gram of stool (similar to children and adults) in the first few weeks of life.[17,173] This early virome is much less diverse than an adult virome (especially in terms of viral richness), but is rapidly evolving: half of the viruses detected at one week of age were absent at two weeks.[17,174] Viruses are likely acquired from a combination of oral intake and prophage induction from the bacteriome.[17,175] Some authors propose a "stepwise" development of the infant virome, starting with pioneer bacteria and their respective bacteriophages, then evolving depending on dietary factors, immune development, environmental exposures, and bacteriome changes.[170,172,173] When compared to formula, breastfeeding appears to reduce the acquisition of eukaryotic viruses while increasing the abundance of *Bifidobacterium*, *Lactobacillus*, and *Lactococcus* phages.[82,173] Vaginal delivery was also associated with *Lactobacillus* phages (*Lactobacillus* being a common commensal bacteria of the female genital tract), but may not make a significant impact on maternal-infant virome similarity.[176,177] Virome composition remains highly personalized, even in infancy.[172]

Over the following months, initially dominant phages are replaced by other enteric-bacteria infecting phages including an increasing proportion of *Microviridae*, constituting a virome that more closely resembles the maternal virome.[82] Human-infecting viruses are more abundant in infancy compared to adulthood, including *Anelloviridae* and several pathogenic viruses (parechoviruses, enteroviruses, caliciviruses, and rotaviruses), suggesting colonization even in the absence of active symptoms.[82,148,170] *Crassvirales* have also been identified in infants, though at lower prevalence and abundance when compared to adults.[75,155,170,174]

By age one, the virome is dominated by *Caudoviricetes*, *Microviridae*, and *Anneloviridae.*[75] *Anelloviridae* abundance remains high during infancy but decreases after fifteen months and through early childhood.[55,170] One study demonstrated changes to the viral community and reduced Shannon diversity around this age in infants born by Caesarean-section compared to spontaneous vaginal delivery, in the absence of corresponding bacteriome alterations in the same dataset, but the longer term impact are unknown.[75]

Few studies have investigated the virome in older childhood and adolescent years, but there is interval development of increased species richness towards adulthood, including the acquisition of *Crassvirales*.[178] *Microviridae*, abundant in early childhood, decreases during later childhood and adolescent years, before returning as a predominant member of the adult virome.[178] Most healthy adult populations demonstrate stable populations of primarily temperate phages.[45,170,179] Viral richness in the elderly may be decreased, though in a healthy population of older adults including centenarians, increased viral diversity and increased *Crassvirales* abundance were noted.[55,180]

These longitudinal studies provide significant insight into the acquisition and evolution of the gut virome over time while highlighting early dynamic changes during infancy that progress towards a stable and rich virome of adulthood. Many of these studies however are limited in size and usually reflect single institution studies, making these virome studies an active area of ongoing research.

1.5. The human gut virome and health

1.5.1. Virome-host interactions

Relationships between the virome, bacteriome, and human health have been referred to as trans-kingdom, tri-kingdom, or cross-kingdom interactions, reflecting our own complex ecology and biology. In this section, I will first discuss the virome-human relationship followed by the virome-bacteriome relationship, as modulation of our bacterial community is a prominent feature of our bacteriophage-dominated virome. Next, I will outline virome alterations in human diseases with a focus on inflammatory bowel disease, before discussing potential clinical applications of modulating the human gut virome.

Eukaryotic viruses colonize the intestine from early infancy, appearing most abundant during childhood but persisting throughout adulthood.[82,170] Some of these viruses may cause human infection (particularly RNA viruses in setting of foodborne diarrheal illnesses) but are often present even in asymptomatic individuals, while other viruses like *Anelloviridae* frequently colonize the intestinal virome but have no clear pathogenic potential.[168,181]

Eukaryotic viruses also appear to assist with maintaining gut homeostasis and host immune development.[170,182] Mouse models have demonstrated that colonization with murine norovirus and various herpesviruses offer protection against bacterial pathogens like *Yersinia pestis* and *Listeria monocytogenes*, possibly by maintaining low-level activation of macrophages and interferon gamma release.[168,183] In human studies, the relative abundance of *Anelloviridae*, appear to be directly dependent on host immune status, with enrichment of *Anelloviridae* in the settings of inflammatory bowel disease and immunosuppression.[171,181] The eukaryotic virome could thus be considered a reflection of the bacteriome, where many of its members have a commensal or

mutualistic relationship with the human body, with some viruses being pathobionts that could cause infection in select settings.[184]

The direct relationship between humans and prokaryotic viruses is less clear. Bacteriophages, which do not infect human cells, do interact with the mammalian immune system in both anti-inflammatory and pro-inflammatory processes.[24,182] Like *Anelloviridae*, phages may help to regulate cytokine release in order to maintain immune homeostasis, provide protection against bacterial infection, or modulate response to bacterial antigens.[182,184] Intestinal phages are also able to translocate into the bloodstream and other tissues, and have been proposed to downregulate immune cell activity which could be helpful in reducing graft rejection in the setting of organ transplantation.[184,185] Conversely, in rat and germ-free mouse models, treatment with bacteriophages can induce a phage-specific immune response and increase gut permeability.[186,187]

1.5.2. Virome-bacteriome interactions

Bacteriophages are directly involved in the homeostasis of the gut bacteriome. Phages often co-occur with their host bacteria, i.e. *Staphylococcus* and *Propionibacterium* phages are more abundant in our skin, while most intestinal phages target enteric bacteria.[57,138] Given bacteriophages' ability to infect and lyse bacteria, and their high abundance at sites of bacterial colonization including mucosal barriers, our virome may be considered a non-host component of our own immune system.[188] Despite this potential for predation, the virome appears to help stabilize the gut microbiome with a symbiotic virome-bacteriome relationship in contrast with other environmental microbiomes.[189]

In marine environments where bacterial density is low, viruses are classically modelled using a lytic, "predatory-prey", "kill-the-winner" relationship with bacteria, where rapidly

expanding bacterial strains are likely to be infected, lysed, and have their population tightly controlled by the virome.[189-191] In the intestinal lumen and fecal microbiome where bacterial density is high, viruses have been proposed to follow a lysogenic "piggyback-the-winner" approach, supported by observations of: longitudinal virome stability; positive bacteriome-virome correlations in diversity and richness; high relative abundance of temperate phages; and lower mutation rates in temperate phages in comparison to lytic phages.[178,190,191] The intestinal mucosa, which contains lower numbers of bacteria and higher levels of bacteriophages due to phage adherence to mucin, may favour more lytic "kill-the-winner" relationships.[191] These two models however are likely too simplistic, and many additional models of bacteriome-virome relationships have been proposed to account for the complex biogeography of the gastrointestinal tract, the dynamic changes of early life, the persistence of bacterial strains, and the lysogenic-like behaviour of some lytic phages.[170,190,192,193]

Prophage activation or induction involve the switching of temperate phages from their lysogenic phase to a lytic phase. This process is thought to be triggered by environmental factors such as nitric oxide, antibiotics, and host inflammation, as well as the lower phage-to-bacteria ratio environments like the intestinal mucosa.[20,191,194]

While the virome overall appears to be beneficial for our immune system, bacteriophages may also be involved in the pathogenicity of bacterial infections. Some bacteria contain phage-encoded virulence factors such as the cholera toxin of *Vibrio cholerae* and the Shiga toxin-producing isolates of *Escherichia coli*.[168,188] Phages may also be involved in the promotion of biofilm formation and transfer of antibiotic resistance genes, the latter increasing in the setting of antibiotic treatment.[24,168,195]

1.5.3. Virome alterations in human disease

Like the bacteriome, the virome is altered in a wide range of human illnesses, including inflammatory bowel disease (IBD), diabetes, obesity, colorectal cancer, liver disease, acute diarrhea, and SARS-CoV-2 infection.[182,185,189,190,196-198] Key attributes of the virome include viral diversity and the relative abundance of *Caudoviricetes*, *Crassvirales*, and select eukaryotic viruses. A summary of key findings to date can be found in Table 1 of a review by Zhang and Wang (2023), with increased relative abundance of *Caudoviricetes* seen in many human diseases including Crohn's disease and ulcerative colitis, lupus, liver cirrhosis, and type 2 diabetes. *Crassvirales* have also been associated with Crohn's disease and lupus, though in a cohort of 1,135 individuals there were no significant correlations between crAssphage and any health or disease parameters.[155,199,200] High interpersonal heterogeneity however confounds many of these studies, which may be further biased by virus purification techniques, nucleic acid amplification, and viral database limitations.[8,145]

Eukaryotic viruses can be detected in the setting of acute diarrheal diseases (with viral multiplex assays often used in diagnostic settings), while potentially carcinogenic viruses including Epstein-Barr virus (EBV), human papillomavirus (HPV), and cytomegalovirus (CMV) are associated with viral infections, gastric cancer, and colorectal cancer.[168,182,201] Several phages including *Inovirus* species as well as *Bacteroides* and *E. coli* phages have been also associated with colorectal tumors, with induction of bacterial biofilm development and chronic inflammation thought to be possible mechanisms of action.[168,182,198] The presence of rotaviruses and enteroviruses in children also appear to be an independent risk factor for type 1 diabetes, interestingly predating the development of autoimmunity.[197]

Given the ability of the virome to "stabilize" or modulate the bacteriome, viruses may also help to maintain an altered bacteriome and contribute to disease severity and duration.[197] This

hypothesis was tested in an *in vitro* setting where phages from children with severe malabsorption maintained a dysbiotic (*Proteobacteria*-dominant) fecal microbial culture in cross-infection experiments.[135] Virome transplantation experiments also provide evidence of the virome's role in our health, with several mouse models showing symptoms of obesity and diabetes on VLP transplantation from obese, high-fat diet fed mice.[182,202]

Virome association studies, like their bacteriome counterparts, are faced with the same "chicken or the egg" dilemma: whether microbiome alterations are drivers for disease pathogenesis or are downstream consequences of a separate disease process. Virome transplantation efforts in both mouse models and possible future human studies provide evidence that viromes can influence host health processes, yet there is an ongoing need for method standardization and mechanistic studies to delineate the virome's role in human health.

1.5.4. Virome alterations in inflammatory bowel disease

One of the main disease settings for virome research is inflammatory bowel disease (IBD). IBD is a chronic, relapsing disorder with no known cure that affects over 320,000 Canadians (as of 2023), and is increasing in incidence both nationally and globally.[203,204] IBD encompasses both Crohn's disease (CD) and ulcerative colitis (UC) which have distinct clinical presentations, histopathological features, and treatment options: CD causes transmural ulceration that may be patchy throughout the entire GI tract though most commonly affected the terminal ileum and cecum, while in UC the ulceration is typically limited to the colonic mucosa starting at the rectum and progressing proximally.[205] The pathogenesis of IBD involves an interplay of host genetics, the gut microbiome, and environmental factors.[206-209] Alterations in both the virome and bacteriome have been observed in IBD, and are hypothesized to be involved in disease pathogenesis and severity.[210]

Early studies of the IBD virome used microscopy and sequence-independent single-primer amplification methods, demonstrating an increase in tailed phages (*Caudovirales*, now *Caudoviricetes*) in Crohn's disease based on ileal and colonic sampling.[211,212] The first major IBD virome study was published in 2014 in a cohort of 130 subjects including healthy controls and subjects with CD and UC at varying points of their illness, i.e. remission and flares.[32] Viral richness (overall and among *Caudovirales*) were increased in both CD and UC with disease-specific patterns, though this data analysis excluded unaligned reads – up to 85% of reads in a given sample.[32] Repeated database-independent analysis by Clooney *et al.* (2019) redemonstrated increased *Caudovirales* diversity in CD but not in UC, and not in the overall virome.[8] Intestinal inflammation is thought to prevent maintenance of a stable, virulent core virome, while potentially inducing prophages to enter their lytic cycle resulting in the dominance of select phages.[8,45]

In subjects with very early-onset IBD (IBD onset before age six, one of the fastest growing demographics of IBD), *Caudovirales* and *Anelloviridae* were enriched with no significant change in overall viral richness.[171] In a small cohort of pediatric IBD subjects (CD, n = 7; UC, n = 5) and non-IBD patients (n = 12), *Caudovirales* were enriched in CD compared to UC, while *Microviridae* richness was reduced in CD patients compared to controls.[137] A study of rectal mucosa samples in UC subjects (n = 91) compared to healthy controls (n = 76) showed decreased virome diversity and richness in UC patients with increased abundance of *Caudoviricetes* and a loss of bacteriome-virome correlations, suggesting that inflammation may disrupt the virome-bacteriome equilibrium.[213] More recently, Stockdale *et al.* (2023) performed an unamplified fecal virome study (40 controls, 19 CD, 20 UC) and concluded that disease associations may be obscured by the highly personalized nature of the fecal virome and vast fluctuations in abundance.[145] No clear differences

in virome diversity, evenness, and richness were noted, especially in comparison with paired 16S rDNA analysis of the bacteriome.[145]

Few studies have explored the virome during IBD treatment. Norman *et al.* (2015) included 17 subjects with longitudinally collected samples but did not identify any changes to viral diversity or taxonomy over time in relation to disease activity. Jansen *et al.* (2023) studied 181 IBD subjects who underwent biologic therapy and were able to stratify patients into two "viral community types" which could predict clinical outcomes: patients who entered endoscopic remission were more likely to have low viral diversity and high relative abundance of non-*Crassvirales Caudoviricetes* while the relapsing group had higher viral diversity and increased *Crassvirales* and *Malgrandaviricetes* (which includes *Microviridae*).[214]

Altogether, a few trends emerge: CD and UC have different altered viromes, with CD usually demonstrating increased virome diversity particularly within *Caudoviricetes*; UC may be associated with decreased virome diversity; and decreases in *Caudoviricetes* abundance are often paired with increased abundance of the main other taxon, *Malgrandaviricetes* / *Microviridae*. Given the significant heterogeneity in patient population and virome sequencing methodologies, these studies are not directly comparable with each other and may be subject to regional or methodological biases.[93,145,215] Ongoing studies on the IBD virome are required, with particular need for epidemiological studies that incorporate larger sample sizes, diverse geographic ranges, unamplified viromes, multiomic approaches, and alternatives to fecal sampling. Furthermore, mechanistic studies including virome transplantation may provide more insight into the virome's role in IBD.

1.5.5. Clinical applications of phages and the virome

Part of the "phage renaissance" has been a growing interest in using phages for clinical therapy, with applications ranging from directed treatment of a bacterial infection with a single phage to whole virome transplantation.

Phage therapy has been proposed to help treat antibiotic-resistant bacterial infections, with many ongoing trials worldwide.[216-218] Phages are selected to target a specific drug-resistant isolate and administered to the patient orally, intravenously, or topically.[217] Intestinal viromes are a potential source for bacteriophages that may be screened against pathogens of interest.[219] Additionally, phage-host relationships that are identified in virome studies may guide bacteriophage selection.[220] Alternatively, bacteriophages could be used for precision editing of the microbiome, such as the use of *E. coli* targeting phages to reduce *E. coli* populations.[221] Phages may also be engineered as vehicles for drug delivery and vaccinations.[143]

One of the most interesting applications of the gut virome is its role in fecal microbiome transplant (FMT). FMT has been an exciting therapeutic option for recurrent *C. difficile* infection and is also being actively explored for treatment for IBD and necrotizing enterocolitis.[222,223] Several studies suggest that bacteriophages may be contributing to the efficacy of FMT therapy in *C. difficile* infection. First, both lytic and temperate bacteriophages are transferred during transplantation and are able to demonstrate colonization for up to twelve months.[142,224] Second, sterile fecal filtrates containing bacteriophages but not bacteria provide therapeutic benefit in reducing *C. difficile* recurrence.[225] Third, higher levels of donor bacteriophages were associated with FMT treatment response.[39] While the exact mechanisms of FMT efficacy remain unclear, the virome is thought to prevent *C. difficile* infection by modulating the overall bacteriome population. Outside of *C. difficile* studies, virome transplants have been shown to influence bacterial diversity and specific populations, including increasing the relative abundance of the naturally occurring

probiotic *A. muciniphila*.[226] A small pilot study has also safely implemented a fecal virome transplantation in the setting of metabolic symdrome.[227] Lastly, mouse models demonstrate the ability for fecal virome transplantation to impact weight gain, glucose tolerance, and fertility.[202,226]

Bacteriophage cocktails are thus a potential therapy for modulating the gut microbiome. Practically, many phage-based therapies have been designated "Generally Recognized As Safe" by the U.S. Food and Drug Administration, based on: no documented significant adverse reactions, their theoretical inability to infect human cells, and the lower likelihood of inducing a severe inflammatory response.[221,228,229] Sterile fecal filtrates may also be an alternative to FMT that would reduce the risk of bacterial infection and transmission of multidrug resistant organisms, though would still require screening against pathogenic viruses.[224,230]

In all, virome transplantation research is an exciting field for studying the human virome that offers both clinical potential and improved mechanistic study on how the virome interacts with the bacteriome and our health.

1.6. The mucosal-luminal interface

1.6.1. Phages at the mucosal-luminal interface

While most virome studies use stool sampling, the gastrointestinal tract is a complex environment with distinct ecologies along its longitudinal axis (i.e. from oral cavity to rectum) as well as its cross-sectional axis (i.e. from lumen to mucosa to epithelium), alongside microanatomical features such as intestinal crypts.[231] The mucosa is a particularly interesting site for intestinal viruses, as many bacteriophages include immunoglobulin-like domains that facilitate binding to mucin and other glycoproteins and promote bacterial encounters.[232,233] A diverse study of mucosal surfaces across anemones, fish, mouse intestine, and human gingiva showed an average 4.4-fold increase in phage to bacteria ratios when compared to adjacent environments.[232] In the mucosa, phages are thought to support host immune defenses by preventing bacterial infection, and may be under different ecological pressures due to increased virus-microbe ratios that promote bacterial lysis.[232,234,235] Interestingly, Lourenço *et al.* (2020) present an alternate hypothesis based on murine models, suggesting that the deep mucosa (particularly at the ileum) may be phage-inaccessible and serve as a refuge for bacteria. These data highlight the importance in capturing mucosa-associated viruses in human virome studies.

1.6.2. Limitations of fecal and intestinal biopsy sampling

Most gut virome studies rely on fecal sampling, which is non-invasive and relatively easy to collect. However, bacteriome studies of mucosal tissues as early as 2002 described distinct mucosal communities in comparison to stool sampling, suggesting that mucosal sites represented distinct microanatomic niches instead of a homogenous community or gradient across the proximal-distal axis of the colon.[236-241] Metagenomic, metabolomic, and metaproteomic analyses

have demonstrated biogeographic differences between different regions of the colon.[242] Differences between mucosa-associated and stool bacteria are also dependent on local inflammation.[237] Therefore, fecal samples primarily represent the colonic luminal microbiome, likely underrepresenting the diversity and richness of the mucosa-associated microbiome and potentially missing mucosa-adherent microbiota.[236,239,242] Given the intimate bacteriome-virome relationship, these data suggest that the mucosa-associated virome is distinct from the stool virome and further supports sampling of the intestinal mucosa.

The minority of microbiome studies that sample the intestinal mucosa usually involve endoscopic biopsies.[213,243,244] These biopsies are limited by low microbial biomass (with high host DNA content). While bacteriome studies may use 16S rRNA gene amplicon studies, virome studies are limited by the impractical use of shotgun metagenomic sequencing. Nucleic acid amplification strategies may be used but may be prone to bias especially with low input DNA.[245] Other mucosal microbiome studies utilize resected bowel, however this sample type is not practical for larger studies.[246]

1.6.3. Sampling the mucosal-luminal interface

Our laboratory has previously demonstrated that the mucosal-luminal interface (MLI) is an ideal sample for studying the microbiome.[247] These samples are obtained during endoscopy by flushing the loose mucus layer from the intestinal wall with sterile water followed by aspiration.[248] This mucosal layer also serves as a barrier layer that is in close proximity with the host, which may better capture the microbial community involved in transkingdom interactions.[248] These samples have been previously suggested as a "critical locale" for studying intestinal pathologies and can be considered a proxy for intestinal biopsies.[248,249] Paired MLI and biopsy samples show similar species diversity and evenness that are distinct than stool.[247]

Compared to stool, MLI sampling enables multisite sampling and may capture distinct viruses that are not present in the fecal virome (Figure 2). This benefit is particularly relevant in inflammatory bowel disease, where inflammation may be localized to specific regions of the gastrointestinal tract. The ability to sample both inflamed and non-inflamed sites may be key to understanding the virome's role in disease pathogenesis, especially given significant interpersonal variation of the human gut virome.[237]

Compared to tissue biopsies, MLI sampling result in increased biomass that both reduces the risk of contamination and enables the shotgun sequencing approaches required for virome sequencing.[249,250] MLI sampling may also be used for metagenomic, metaproteomic, metatranscriptomic, and metabolic analysis, which permits multiomic analysis.[242,251-255] Bacteriophage proteins have been detected in proteomic analysis of MLI samples, though proteomic viral databases are currently very limited.[253] Furthermore, MLI samples may reduce the risk of bleeding during sample collection, especially in patients with underlying inflammation.[256]

While MLI sampling does not replace fecal sampling (which remains most accessible) or tissue biopsy (required for histopathological diagnosis of various intestinal diseases), the MLI is a practical sample for studying the IBD virome, especially as these patients undergo endoscopy as part of standard of care.

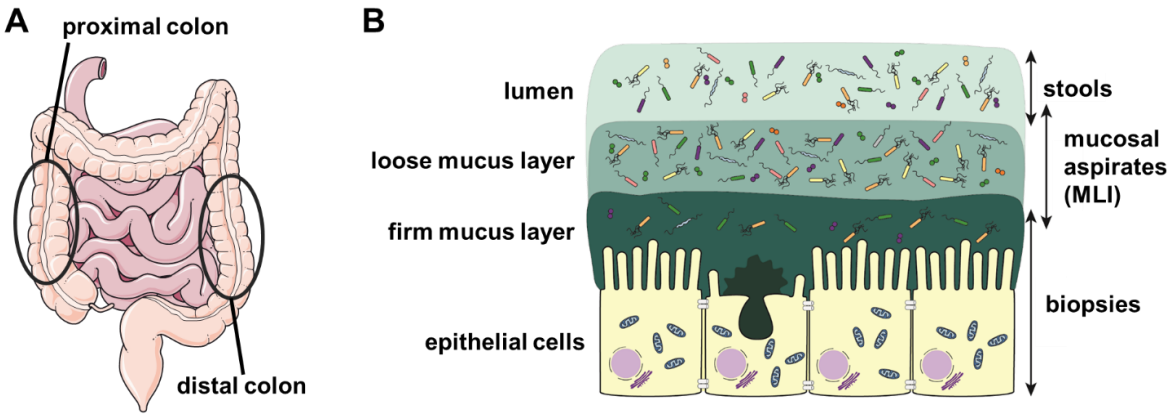


Figure 2: Sampling the mucosal-luminal interface enables site-specific study of the intestinal microbiome.

MLI aspirates are collected during endoscopy, thus allowing for site-specific sampling of the gastrointestinal tract. In the subsequent chapters, we compare MLI samples taken from the proximal/ascending and distal/descending colon (A).[32-43] These samples are collected by performing washing the mucosal wall during endoscopy with subsequent aspiration, which primarily captures the loose mucus layer (B).

1.7. Rationale and hypotheses

The mucosal-luminal interface thus serves as an interesting sample type for studying the intestinal virome to help elucidate host-microbe interactions in health and disease. Our laboratory's expertise in microbiome research in collaboration with the IBD Centre at the Children's Hospital of Eastern Ontario (CHEO) has enabled our group to investigate the colonic virome in pediatric inflammatory bowel disease. I proposed that with advances in virome techniques and bioinformatics, we could investigate the MLI virome to provide new insights on the role of the virome in IBD. This research would guide future therapeutic considerations for IBD, particularly regarding microbiome modulating therapies, while providing a set of techniques that could be applied in virome research in similar sample types for a wide range of clinical interests.

In **Chapter 2**, published in Yan *et al.* (2020), I demonstrated that we could reproducibly profile the virome at the mucosal-luminal interface in clinical samples. In **Chapter 3**, published in Yan *et al.* (2023), I applied these techniques to perform deep metaviromic sequencing on MLI samples with paired metagenomic and metatranscriptomic datasets to characterize the MLI virome, with added context of gene expression and prophage integration. We also reported differences in the viral communities at the MLI compared to stool, further supporting efforts to sequencing the MLI virome. In **Chapter 4**, we applied these methods to study a cohort of 51 pediatric subjects undergoing investigation for inflammatory bowel disease. This cohort included children who were diagnosed with Crohn's disease, ulcerative colitis, or non-IBD, to highlight differences in the treatment-naïve, IBD virome at the mucosal-luminal interface. In **Chapter 5**, I will provide a brief discussion on the results of these three articles and highlight the ongoing development in human gut virome research.

2. Article: Virome sequencing of the human intestinal mucosal-luminal interface

2.1. Preface

This chapter has been previously published as the following research article:

> **Yan A**, Butcher J, Mack D, Stintzi A. Virome Sequencing of the Human Intestinal Mucosal-Luminal Interface. Front Cell Infect Microbiol. 2020 Oct 22;10:582187. doi: 10.3389/fcimb.2020.582187. PMID: 33194818; PMCID: PMC7642909.

Specific author contributions are as follows:

Austin Yan: performed all experiments, sequencing, data analysis, and wrote the initial manuscript.

James Butcher: provided input on the bioinformatic analysis and manuscript.

David Mack: recruited patients, performed endoscopy on patients in collaboration with the CHEO IBD Centre, provided clinical and demographic information, and revised the manuscript prior to submission for peer review.

Alain Stintzi: provided reagent and research materials, obtained funding, supervised research, conceptualized project, and revised the manuscript prior to submission for peer review.

AY performed all experiments, sequencing, data analysis, and wrote the initial manuscript. DM recruited and performed endoscopy on patients and provided access to relevant clinical metadata. AY and AS designed the experiments. JB. All authors reviewed and provided comments on the final manuscript.

2.1.1. Conflicts of Interest

AS and DM are co-founders of MedBiome, a clinical microbiomics company. The other authors have no competing interests to declare.

2.1.2. Funding

AY is supported by the Frederick Banting and Charles Best Canada Graduate Scholarships Doctoral Award from the Canadian Institutes of Health Research. This work was funded by the Government of Canada through Genome Canada and the Ontario Genomics Institute (OGI-149), the Canadian Institutes of Health Research (ECD-144627), and the Ontario Ministry of Economic Development and Innovation (Project 13440). The funders had no role in study design, data collection and analysis, or preparation of the manuscript.

2.1.3. Acknowledgments

The authors would like to acknowledge the patients and their families for their participation in our study. We acknowledge Ruth Singleton for her help in enrolling patients and assistance in collecting intestinal aspirate samples. We also thank Dr. Christine M. Szymanski and Clay S. Crippen for providing the NCTC 12673 phage used for spike-in experiments. The whole metagenome annotations presented here was enabled in part by WestGrid (www.westgrid.ca) and Compute Canada (www.computecanada.ca).

2.1.4. Data Availability Statement

Host-removed, sequencing data are available under BioProject PRJNA645218.

2.2. Abstract

While the human gut virome has been increasingly explored in recent years, nearly all studies have been limited to fecal sampling. The mucosal-luminal interface has been established as a viable sample type for profiling the microbial biogeography of the gastrointestinal tract. We have developed a protocol to extract nucleic acids from viruses at the mucosal-luminal interface of the proximal and distal colon. Colonic viromes from pediatric patients with Crohn's disease demonstrated high interpatient diversity and low but significant intrapatient variation between sites. Whole metagenomics was also performed to explore virome-bacteriome interactions and to compare the viral communities observed in virome and whole metagenome sequencing. Site-specific study of the human gut virome is a necessary step to advance our understanding of virome-bacteriome-host interactions in human diseases.

2.3. Introduction

The human microbiome represents a complex ecosystem of microbes, including bacteria, viruses, fungi, protozoa, and archaea. These microbes mostly reside in the gastrointestinal tract and are implicated in human health and disease, ranging from immune system development [259] to nutrient and drug metabolism [260] to involvement in conditions including obesity, inflammatory bowel disease, and cancer [2,208,261]. Despite major developments in microbiome research, most existing knowledge is focused on the bacteriome. The virome, consisting mostly of bacteriophages, lacks conserved marker genes and requires viral purification for cost-effective shotgun metagenomic sequencing [143]. Existing databases are also limited, hindering the interpretation of virome sequencing data, most of which cannot be aligned to any known viral genome [73]. The NCBI

Genome Resource as of August 2020 contains over 250,000 bacterial genomes but less than 40,000 viral genomes, of which eukaryotic viruses are overrepresented [262]. Yet, bacteriophages can modulate the bacteriome [8,263,264] and have a potential role in microbiome transplantation or manipulation therapies [39,202,224,265]; examining the virome is essential for any comprehensive model of the host-microbiome relationship.

Recent advances in the study of the human virome include virus-like particle purification protocols optimized for stool [40,266] and improved bioinformatic tools and databases [55,96,116]. These tools have enabled the study of the gut virome in inflammatory bowel disease [8,267], its temporal stability in healthy adults [45], and its development during infancy [75,173] and into senescence [55]. However, nearly all studies profiled stool samples, which overlook the complex biogeography of the gastrointestinal tract [268]. Moreover, the potential for mucin-bacteriophage interactions [232] and varied microbial concentrations across the intestinal mucosa provide unique ecological pressures and niches that cannot be captured by simply sampling stool [38,190,191,269].

Site-specific differences in the intestinal virome have been explored in mice [270] and in rhesus macaques, where viromes from the terminal ileum were distinct from the colon and rectum [271]. Early research on the human intestinal virome characterized ileal and cecal viral particles, but these studies lacked modern high-throughput sequencing approaches [211,212,272]. Two recent investigations used biopsies to study the ileal eukaryotic virome [273] and rectal virome [213] in inflammatory bowel disease, the latter finding an altered virome with intestinal inflammation. Yet, there has not been a focused effort to characterize the virome along the gastrointestinal tract. This information would inform our understanding of virome-bacteriome-host interactions, especially for site-specific conditions like Crohn's disease.

In this study, a protocol was developed to characterize the virome at the human intestinal mucosal-luminal interface. Aspirates obtained during endoscopy were subjected to virus-like particle enrichment through filtering and polyethylene glycol precipitation, followed by the removal of remaining bacteria and their nucleic acids. Viral DNA was then extracted using Proteinase K and phenol-chloroform, then subjected to multiple displacement amplification (MDA) prior to sequencing. Mucosal-luminal interface aspirates can be collected at various sites along the gastrointestinal tract while providing more microbial DNA than intestinal biopsies, enabling whole metagenomic sequencing [247]. Hence, samples were processed for both virome and metagenome sequencing to validate the virome sequencing protocol, measure diversity within and between subjects, and explore virome-bacteriome relationships. We also compare our virome sequencing efforts with the viral signals identified in whole metagenome sequencing. Combining both methods has been hypothesized to "improve *de novo* viral recovery" [55].

Our samples were obtained from pediatric subjects with Crohn's disease including two individuals with active colonic inflammation, demonstrating the applicability of this protocol to investigate clinically informative samples. Thus, we provide a sample type and methodology that can be used by clinicians and researchers to study the human virome along the gastrointestinal tract.

2.4. Materials and Methods

2.4.1. Ethics Approval and Patient Recruitment

Sample collection from pediatric subjects was approved by the Research Ethics Board of the Children's Hospital of Eastern Ontario (CHEO) in Ottawa, Canada with informed

consent/assent obtained from parents and/or subjects. Samples from five patients were used in this study, which were obtained during routine endoscopy in the diagnosis and care of Crohn's disease. Subjects with infectious gastroenteritis in the past two months or antibiotic treatment in the past four weeks were excluded from this study.

2.4.2. Sample Collection and Phage Spike In

The collection of mucosal-luminal interface (MLI) aspirates has been described previously [247]. In brief, sterile water was used to wash the bowel wall during colonoscopy to remove the loosely adherent mucous layer. The wash was then aspirated into a sterile container and stored at -80°C. Samples were obtained from three distinct sites: the terminal ileum (TI), proximal colon (PC), and distal colon (DC). Aliquots of 10 ml were used for virus-like particle purification and viral DNA extraction; 2 ml was used for whole metagenomic DNA extraction. To estimate viral load, an exogenous phage, NCTC 12673 [274], was added to two samples at final concentrations of 10^5, 10^6, and 10^7 plaque-forming units (pfu) ml^{-1} for virus-like particle purification and sequencing. NCTC 12673 phage was similarly added to three samples prior to whole metagenome sequencing at concentrations of 10^7 pfu ml^{-1} with one of those samples also spiked at 5 x 10^7 pfu ml^{-1}.

2.4.3. Virus-like Particle Purification and Nucleic Acid Extraction

A protocol to purify virus-like particles from mucosal aspirates (summarized in Supplementary Figure 1) was developed by adapting existing methods for stool [32,40]. Mucosal aspirates were first subjected to centrifugation twice (4,696 *g*, 10 min, 4°C) to remove debris. Samples were then sequentially filtered through 5 µm and two 0.45 µm PVDF filters to remove host and bacterial cells. Virus-like particles were precipitated by overnight incubation with 10% w/v PEG-8000 and 0.5 M NaCl at 4°C and subjected to centrifugation (4,696 *g*, 20 min, 4°C) the

following day. The pellet was suspended in 400 µl saline-magnesium buffer. Remaining bacterial cells were lysed by treatment with 1 mg/ml lysozyme (Sigma) for 30 min at 37°C followed by 0.2 volumes of chloroform (10 min, room temperature). After centrifugation (5 min, 2,500 *g*), the aqueous mixture was treated with DNase (TURBO™ DNase, Thermo Scientific) and RNaseI (Thermo Scientific) in a buffer of 1 mM $CaCl_2$ and 5 mM $MgCl_2$ for 1 hour at 37°C to degrade remaining bacterial nucleic acids. Enzymes were inactivated at 70°C for 10 minutes. Virus-like particles were lysed with Proteinase K (3.2 µg/ml) in 3.2% SDS for 20 minutes at 55°C, then treated with 2.5% cetyltrimethylammonium bromide and 0.5 M NaCl for 10 minutes at 65°C. Viral DNA was then extracted by adding 1 volume of phenol-chloroform-isoamyl alcohol (25:24:1, pH 6.7) to each mixture, which was vortexed and subjected to centrifugation (10 min, 8,000 *g*); this step was repeated with chloroform to remove trace phenol. Nucleic acids were purified from the aqueous layer using the DNeasy Blood and Tissue Kit (QIAGEN) and eluted in 50 µl of water. DNA was concentrated using an Eppendorf™ Vacufuge™ Concentrator to 3 µl to maximize the input DNA load for the GenomiPhi™ V2 DNA Amplification polymerase kit. Multiple displacement amplification (MDA) reactions using 1 µl of input DNA were run in triplicate, then pooled and purified with the DNeasy Blood and Tissue Kit. DNA was quantified fluorescently using the Qubit dsDNA HS Assay Kit (Thermo Fisher). This protocol was also tested on a sample of sterile water as a negative control.

2.4.4. Whole Metagenome Extraction, Library Preparation, and DNA Sequencing

Whole metagenomic DNA was extracted using the FastDNA Spin Kit for DNA Isolation (MP Biomedicals), eluted in water, and quantified using the Qubit High Sensitivity dsDNA Assay Kit as previously described [251].

Shotgun metagenomic sequencing libraries for both virome and metagenome DNA were prepared and barcoded using the Ion Xpress Fragment Library Kit and Ion Xpress Barcode Adapters (Thermo Fisher), with sonication performed on the Covaris S220 Ultra Sonicator following the manufacturer's instructions. Libraries were visualized with the High Sensitivity DNA Kit (Agilent) on the 2100 Bioanalyzer. Samples were templated and loaded on two Ion PI Chips (virome and metagenome samples on separate chips) by an Ion Chef using the Hi-Q Chef Kit and sequenced on an Ion Proton with the Hi-Q Sequencing 200 Kit following manufacturer's instructions.

2.4.5. DNA Pre-processing, Host DNA Removal, and Bacteriome Annotation

Our bioinformatic pipeline is summarized in Supplementary Figure 2. High quality sequencing reads from the Ion Proton were trimmed using seqtk 1.2-r94 [275] at the default error rate threshold of 0.05; reads <50 bp were also removed. Remaining reads were mapped to various databases to examine host, bacterial, and viral content using bowtie2 version 2.3.4.1 [87] with the default settings unless otherwise specified. Host and spike-in reads were detected by mapping reads to the human genome (GRCh38 with bowtie2's ultra-sensitive mode) [276] and the NCTC 12673 genome [274]; these reads were counted and removed from further analysis using samtools 1.7 [277]. Samples sequenced multiple times to assess phage-spike in loads were also analyzed to evaluate reproducibility; these reads were then merged for subsequent analyses. Bacterial contamination and viral content were assessed by aligning reads to the cpn60 database (ultra-sensitive mode) [89], the Gut Virome Database [55], known crAssphages [164], and all viral genomes available on the NCBI Viral Genomes Resource [77] as of May 11, 2020 (12,194 genomes). Host-removed, whole metagenome sequencing reads were aligned to a database of all non-redundant sequences using DIAMOND 0.9.2 [278] and annotated using MEGAN 6.18.3 [279]. When performing bacteriome

analysis, non-bacterial taxa were excluded from the whole metagenome results (an average of 0.32% of reads/sample mapped to viruses; 1.2% to eukaryotes).

2.4.6. Identification of Viral Contigs in Virome and Whole Metagenome Sequencing Data

Host and spike-in decontaminated reads from each virome sample were assembled using MEGAHIT; contigs longer than 1,000 bp from all samples were clustered using ClusterGenomes [108] at 90% identity and a minimum length of 90% of the shorter contig. Open reading frames were predicted using Prodigal v2.6.3 [119] in metagenomic mode. Clustered contigs were then subjected to the following viral selection criteria: positive (category 1 or 2) or circular identification by Virsorter v1.0.5 [116] in virome decontamination mode (db = Viromedb); alignment to the NCBI Viral Genomes Resource or the crAssphage database (e-value < 10^{-10}) using nucleotide-nucleotide BLAST 2.9.0+ [86]; or identification of ≥ 3 open reading frames aligning to a 2016 database of prokaryotic viral orthologous groups [121] with an e-value < 10^{-5} with at least two hits / kb of contig length, assessed by hmmscan in HMMER 3.1b2 [280]. Contigs ≥ 3 kb with no blastn alignments (e-value < 10^{-10}) to the nt database (November 2019) were also retained. Contigs were removed if they had three or more open reading frames aligning to ribosomal proteins in the Clusters of Orthologous Groups of proteins database [281] using blastp (e-value < 10^{-10}) in BLAST 2.9.0+. Lastly, virome sequencing reads (host and spike-in reads removed) were remapped to the putative viral contigs; any contig that did not have a minimum horizontal coverage of 75% in at least one sample was likely misassembled and thus removed. These breadth of coverage statistics were calculated using samtools, idxstats, and mpileup [277].

The same pipeline was used to identify viral contigs in the whole metagenome data, with Virsorter run in its default mode instead of virome decontamination mode. This set of contigs is referred to as the metagenome viral contigs (mVCs). Both virome and metagenome sequencing

reads from each sample were mapped to the VCs and mVCs with bowtie2 and indexed with samtools for further analysis.

2.4.7. Viral Contig Clustering, Taxonomic Annotation, and Bacterial Host Prediction

Viral contigs were clustered using vConTACT2 [282], which uses ClusterONE to detect and interpret protein-level relationships between contigs. Open-reading frames were first generated using Prodigal in metagenomic mode, while vConTACT2 0.9.19 was run with pc-inflation and vc-inflation set to 1.5, pcs-mode set to MCL, and vcs-mode set to ClusterONE, as has been previously used in other virome studies [8]. These clusters were viewed using Cytoscape 3.7.2 with the default Perfuse Force Directed Layout using vConTACT2-derived edge-weights.

Viral annotations were performed using Demovir [283] which uses amino acid homology searches against viral references to assign viral order and family. We used the pre-built database of non-redundant viruses from TrEMBL available at figshare.com/articles/NR_Viral_TrEMBL/5822166. Bacterial hosts were predicted using WIsH 1.0 [126], which identified the most likely bacterial host for each viral contig among a set of all reference and representative RefSeq bacteria genomes (n = 9,523) that were available in August 2020. The NCBI Viral Genome Resource (12,194 genomes) were used to generate null parameters for each host genome. If no bacterial genome was matched with a p-value ≤ 0.05, the VC was not assigned a putative host.

2.4.8. Statistical Analysis

Our bioinformatic pipeline generated three main datasets: virome sequencing reads mapped to VCs, whole metagenome sequencing reads mapped to mVCs, and whole metagenome sequencing annotated using the non-redundant protein database which primarily characterized

bacteria (B). We also mapped the virome reads to mVCs and metagenome reads to VCs when examining differences between the two viral populations.

For alignment comparisons and alpha-diversity analysis, read counts were used, with the latter subsetted to the sample with the lowest number of mapped reads (Virome / VC: 1,877,966; Virome mVC: 528,240; metagenome / VC: 5,358; metagenome / mVC: 5,867; bacteriome: 38,711). When stated, we applied a 75% horizontal coverage filter for each sample's viral contigs [45]; counts for contigs below this threshold were set to zero. For beta-diversity (Bray-Curtis) analysis, viral read counts were normalized to reads per kilobase per million mapped reads (RPKM) while bacteriome counts were normalized by relative abundance; additionally, viral hits or bacteria taxa that never exceeded a minimum 0.01% relative abundance in any sample were filtered out to remove very low abundance hits and potential false positives.

Analysis and plotting was performed in R 3.6.0 using phyloseq 1.30.0 [284], reshape2 1.4.4 [285], ggplot2 3.3.0 [286], ggthemes 4.2.0 [287], ggpubr [288], ggnewscale 0.4.1 [289], HMisc 4.4.0 [290], and corrplot 0.84 [291].

2.5. Results

2.5.1. Sample Descriptions and Sequencing Statistics

Samples from five pediatric subjects (11.3–16.6 years old) with Crohn's disease were obtained between June and September 2018 at the Children's Hospital of Eastern Ontario in Ottawa, Canada (Table 1). Subjects A, B, and C had known Crohn's disease and underwent colonoscopy that was required for their ongoing medical care. Subjects D and E were treatment-

naïve subjects undergoing colonoscopy for confirmation of their clinically suspected Crohn's disease.

Twelve samples were processed for virome extraction: MLI aspirates from the proximal colon (PC) and distal colon (DC) were collected from all patients; MLI aspirates from the terminal ileum of subjects A and B were also available for analyses. All samples were processed for virome and whole metagenome sequencing (Supplementary Figure 1). Both terminal ileum samples, and the distal colon sample from patient B did not yield sufficient viral nucleic acids for sequencing library construction (< 50 ng). The remaining nine samples were subjected to virome sequencing. Shotgun sequencing of the whole metagenomes (i.e. not subject to virus-like particle purification) of seven of these nine samples was also performed. A negative control of sterile water subjected to the virome protocol yielded no detectable quantities of nucleic acids.

For virome sequencing, a total of 98.0 million reads were obtained and subjected to quality filtering and trimming, resulting in 95.7 million high-quality reads (mean length = 198.2 bp). An average of 10.6 million reads (2,002,855–25,566,766) was obtained per sample (Table 1). For metagenomic sequencing, a total of 102.3 million reads were similarly processed, resulting in 100.6 million high quality reads (mean length = 188.8 bp), or an average of 14.4 million reads/sample (158,638–33,429,010). Corresponding virome and metagenomic reads were matched for analysis.

2.5.2. Virus-Like Particle Purification Removes Host and Bacterial Content

The alignment of virome and whole metagenome sequencing reads to human, bacterial, and viral databases is shown in Figure 1. While host DNA is usually low in stool (< 10% and often much lower) [292,293], an average of 39.0% of metagenome reads from the mucosal-luminal interface samples aligned to the human genome, though varying from 0.362–90.0%. In contrast, an average

of 0.05% (0.0028–0.17%) of virome sequencing reads aligned to the human genome, representing a mean 3500-fold decrease in host content. This effect was most evident when host content was high in the original sample (from 5.7-fold decrease in A-PC to 20,800-fold decrease in D-PC), suggesting that host DNA is efficiently removed during the virus-like particle purification.

The removal of bacterial DNA was assessed by aligning reads to a database of chaperonins, which are found in nearly all bacteria and have thus been used to estimate bacterial contamination [40,89]. The average proportion of host-removed, whole metagenome reads aligning to the cpn60 database was 0.00545%; this decreased to 0.0000401% in virome reads in the same matched samples, including two samples with no matching reads. Overall, the 136-fold decrease across the entire dataset suggests that level of bacterial contamination remaining after virus-like particle purification is low.

Aligning virome sequencing reads to viral databases further corroborated the removal of host and bacterial DNA. An average of 76.6% of virome reads aligned to 10,673 of 33,242 viral sequences in the Gut Virome Database (66.2–87.5%) [55], though decreasing to 41.7% (11.0–86.5%) after applying a 75% breadth of coverage filter (372 viral sequences). Thus, a significant portion of virome sequencing reads in each sample were previously identified in other virome studies, providing some evidence of a common gut viral community. These same samples had far fewer alignments to a database of all NCBI RefSeq Viruses (0.00152–12.7%), underscoring the current lack of well-annotated human gut bacteriophage genomes. Lastly, reads were mapped to a set of 249 crAss-like phage contigs, representing the most abundant human gut phage [164]; four samples had 5% or more reads matching known crAssphages. We observed a nearly forty-fold increase in crAssphage reads in A-DC compared to A-PC, suggesting site-specific differences in the human gut virome of this subject.

2.5.3. Estimation of Viral Load at the Proximal Colon Mucosal-Luminal Interface

The addition of an exogenous phage, at concentrations of 10^5, 10^6, and 10^7 pfu ml^{-1}, was used to estimate viral load in A-PC and B-PC samples. We used the phage NCTC 12673, a *Campylobacter jejuni* bacteriophage that was first isolated from poultry [274]. As these patients are prescreened to ensure that they do not have infectious colitis (i.e. an active *Campylobacter* infection), this phage should be naturally absent from patient viromes. Indeed, no reads aligning to NCTC 12673 were detected in 4 patient viromes (A-DC, C-PC, D-DC, and E-DC) and 2 samples (D-PC, E-DC) contained a single matching read each (< 0.0001%). The remaining sample (C-DC) contained a very low abundance of reads matching NCTC 12673 (0.00015%, a 300-fold decrease from the lowest tested phage titre). Thus, NCTC 12673 is a suitable exogenous phage that could be added at 10^5–10^6 pfu ml^{-1} as a virome standard in patients without an active *Campylobacter* infection.

In these six phage-spiked samples, 0.044–23.3% of reads aligned to NCTC 12673, increasing linearly with phage titres ($R^2 > 0.99$) (Supplementary Figure 3). Assuming an average phage genome size of 40 kb [294], total viral loads were estimated as $6.21 \pm 0.13 \times 10^8$ ml^{-1} viral particles in A-PC and $1.80 \pm 0.31 \times 10^8$ ml^{-1} viral particles in B-PC.

NCTC 12673 was also added to three mucosal-luminal interface aspirates (A-PC, A-DC, and B-PC) and processed for whole metagenome sequencing. No reads in the 5×10^7 pfu ml^{-1} spiked sample (A-DC) and no more than one read in any 1×10^7 pfu ml^{-1} spiked samples (< 0.000012%) was mapped to NCTC 12673. These results suggest that exogenous, extracellular phage particles like NCTC 12673 are essentially undetectable at these spike-in concentrations using standard DNA extraction kits like the FastDNA Kit used in this study.

2.5.4. Assembling and Annotating Putative Viral Contigs at the Mucosal-Luminal Interface

Virome sequencing reads from each sample were assembled to identify putative viral contigs (VCs). A total of 12160 contigs across nine samples were pooled, clustered, and filtered for viral features, resulting in 2,511 VCs (Figure 2A). The mean contig length was 8,413 bp, ranging from 1,001–120,543 bp. Mapped read counts were adjusted for contig length for downstream analyses, adjusting for the 120-fold difference in VC size. Between 86.8–96.6% of reads from each virome sequencing sample could be mapped to these contigs. There was a correlation between unmapped virome sequencing reads to the proportion of reads aligning to the cpn60 database, suggesting that these unmapped reads could represent low-abundance, bacterial reads (Pearson correlation = 0.795, p = 0.0105).

At a minimum of 75% contig coverage, each sample contained an average of 648 VCs (169–1,066) with an average of 892 VCs (208-1,066) per subject. Only three VCs were present in all subjects and only one VC was present in all samples; 44 VCs were present in at least three of five subjects. Figure 2B shows that at this level of coverage, most VCs (n = 2,229) were only seen in one subject. 435 VCs were also site-specific (excluding the 871 VCs only observed in B-PC). These site-specific VCs represented up to 44.6% of a sample's total observed VCs (C-DC). Breadth of coverage filtering can be confounded by sequencing depth, thus we analyzed a rarefied dataset and found similar results (Supplementary Figure 4).

Whole metagenome sequencing was also used for viral sequence mining following a similar bioinformatic pipeline used to assemble and filter VCs. Overall, the ratio of viral contigs to sequencing reads was 3.56e-5 using a metagenome-mining approach compared to 2.69e-5 in the VLP-enriched approach, representing a 32.2% increase (Table 2). A total of 3,122 metagenome-derived viral contigs (mVCs) with an average length of 6,233 bp were identified and compared

with the virome-derived VCs. Viral families were predicted using Demovir, which annotated 62.4% of VCs and 27.9% of mVCs, as shown in Table 2. *Caudovirales*, the predominant viral order observed in the gut, represented the vast majority of annotated contigs (97.6% of VCs, 96.9% of mVCs), with *Siphoviridae*, *Myoviridae*, and *Podoviridae* observed in decreasing frequency. *Anelloviridae* and *Microviridae* were observed in the virome dataset (14 and 24 VCs, respectively) while only 2 *Microviridae* were seen among mVCs. *Herpesviridae*, *Iridoviridae*, *Mimiviridae*, *Poxviridae*, and *Phycodnaviridae* were only seen among mVCs, except for one *Phycodnaviridae* among VCs (compared to 17 mVCs).

2.5.5. The Mucosal-Luminal Interface Virome is Subject Specific and Distinct from the Viral Community Observed in Whole Metagenome Sequencing

Viral communities identified in virome and whole metagenome sequencing are compared in Figure 3. Panel A shows the mapping of virome and metagenome reads against VCs and mVCs. On average, 93.7% of virome sequencing reads aligned to VCs while 33.6% of reads could align to mVCs. In comparison, 6.24% of whole metagenome sequencing reads aligned to mVCs, while 4.35% of reads aligned to VCs.

Using vConTACT2, the 5,633 viral contigs were organized into 4,473 clusters (including singletons and outliers). 634 clusters contained multiple contigs, ranging from two to fifteen contigs. Of these clusters, 251 contained at least one VC and one mVC; 186 clusters contained only VCs, and 197 clusters containing only mVCs (Figure 3B). Viral contig network maps are shown in Supplementary Figure 5. The viral clusters were then used to merge virome and metagenome sequencing reads (each aligned against their respective viral contigs, then aggregated by viral cluster). Within the same subject, there was no significant difference in Bray-Curtis distance measured among the virus-enriched virome or the viral portion of the whole metagenome. While

intersubject beta-diversity was higher overall, these distances were significantly lower in the viral portion of the whole metagenome than the virus-enriched virome ($p = 6.2e-10$). The bacteriome is also subject-specific (Supplementary Figure 6) though it is more conserved between subjects compared to VCs or mVCs.

2.5.6. Technical Replicates Demonstrate Protocol Reproducibility and Virome Variation between Locations

Beta-diversity was also used to assess reproducibility of the virome protocol by comparing replicates of A-PC and B-PC (Figure 4), which were processed and sequenced in triplicate. While merged for previous analyses, aligning each replicate separately to the assembled viral contigs revealed a significant difference ($p = 0.0087$) in Bray-Curtis distances between proximal and distal colon viromes (0.383 ± 0.159) compared to replicates from the same site (0.081 ± 0.063). These results demonstrate the reproducibility of the protocol while emphasizing site-specific differences in the human intestinal virome.

2.5.7. Virome-Bacteriome Relationships at the Mucosal-Luminal Interface

The Chao1 index were used to measure the alpha diversity in each sample's virome and bacteriome (Supplementary Figure 7). Spearman correlations were performed between these datasets (Figure 5A). These results show a trend towards positive correlations between the alpha diversities of all viral communities and inverse correlations between the virome and the bacteriome.

WIsH was used to assign each VC with its most likely host among a reference set of 9,523 bacterial genomes. 2,056 of 2,511 VCs (81.9%) was assigned a putative host. Of these annotated VCs, Firmicutes (73.8%), Bacteroidetes (16.2%), and Proteobacteria (7.25%) were the most predominant bacterial host phyla (Figure 5B), corresponding to the bacterial species observed

through whole metagenome sequencing (Figure 5C). 52 of 593 unique predicted hosts were detected in the annotated bacteriome; these strains were the putative targets of 426 VCs. Spearman correlations between the observed VC and its putative host were calculated across the seven paired samples (411 VCs, 52 hosts) that were subjected to both virome and whole metagenome sequencing (Figure 5D). Overall, Spearman correlations between VC-host pairs were significantly higher than correlations between VCs and the other 51 non-paired strains ($p = 2.67e-5$).

2.6. Discussion

2.6.1. The Mucosal-Luminal Interface Enables Site-Specific Study of the Human Gut Virome

Research in the human gut virome is still in its early years, lagging well behind its bacterial counterpart. Like its predecessor, early virome studies have worked to solve similar challenges, namely the need for reproducible, standard protocols for both laboratory and bioinformatic techniques. These developments have led to the necessary exploratory phase, where the virome is being characterized in various states of human health and disease, from infancy to advanced age, and in its relation to diet, drugs, and disease [143,190]. Our efforts here expand on these new techniques and aim to extend our knowledge of the "gut" virome beyond fecal samples and into the gastrointestinal tract.

We demonstrate that the mucosal-luminal interface is a promising sample type for studying the virome in specific regions accessible to colonoscopic sampling. We have optimized a reproducible virome protocol for this sample type which removes contaminant host and bacterial nucleic acids and has low background noise (no amplification from a negative control). An exogenous phage added in at known titres could be reliably quantified, while most virome

sequencing reads could be aligned to previously sequenced viral populations. The assembly of filtered, viral contigs represented ~94% of all virome sequences in this dataset and could be used for the analysis of viral communities with greater discerning capability than existing databases. Furthermore, this protocol was optimized in subjects with Crohn's disease at various stages in their disease history, demonstrating applicability to clinically relevant conditions.

2.6.2. Characterizing the Human Colonic Virome

Through our virome sequencing efforts at the mucosal-luminal interface, we were able to estimate viral load, quantify virome diversity, and compare viromes between sites and subjects. We estimated total viral loads at the proximal colon of Subjects A and B to be $6.21 \pm 0.13 \times 10^8$ ml^{-1} and $1.80 \pm 0.31 \times 10^8$ ml^{-1}, respectively. This represents an order of magnitude fewer viral particles than in stool using similar purification, MDA amplification, and sequencing techniques [40]. The decreased viral load could reflect the sampling methodology and/or the decreased microbial load in the proximal colon compared to stool [295]. Further characterization of the viral load across the gastrointestinal tract in healthy and disease states using quantitative virome sequencing approaches would provide important context for future virome and metagenome studies.

We identified 2511 viral contigs across nine virome samples. Over sixty percent of VCs could be annotated by Demovir, with *Caudovirales* representing nearly all classified VCs. 38 viral contigs were classified as *Anelloviridae* and *Microviridae*, which represent families of ssDNA viruses that are preferentially amplified by multiple-displacement amplification [296]. These were enriched compared to two *Microviridae* identified among metagenome-derived viral contigs.

We observed an average of 648 VCs per sample that were present at ≥75% horizontal coverage; most of these were subject-specific while 435 VCs across four subjects were also site-

specific. Beta-diversity analysis showed that intrasubject viromes were similar but not identical, demonstrating a higher Bray-Curtis distance than same-sample replicates and highlighting virome diversity across the gastrointestinal tract. Intersubject viral diversity was also higher than bacterial diversity. We did not identify a core virome, which has been previously reported in healthy adult cohorts [45,144]; this result could reflect higher viral diversity in inflammatory bowel disease and/or a pediatric population. However, the specific impact of Crohn's disease and other clinical metadata on VC diversity or composition were not examined due to the low number of subjects and could be the focus of future studies.

2.6.3. Interpreting Viral Sequences in Whole Metagenome Data

The mucosal-luminal interface samples yields sufficient microbial content for whole metagenome sequencing, which provides additional context for virome studies that is missed when only using 16S rRNA gene sequencing. While host sequencing proportions averaged 40% per sample and was as high as 90% in one sample, the mucosal-luminal interface still offers a significant improvement from intestinal biopsies, which yields > 95% host DNA [297]. Our data suggests that this variation could be due to disease activity: host DNA was higher in treatment-naïve patients at diagnosis (Subjects D, E: 25.2–90.0%) than during remission (A, B: 0.362–2.63%), in line with what other studies have reported in stool [298].

We were able to use the shotgun metagenomic sequencing reads to perform viral contig assembly from the whole metagenome, compare viral populations, and perform virome-bacteriome analysis at a species level. We assembled 3,122 metagenome-derived viral contigs, though only 28% of contigs could be annotated. The reduced classification compared to the virome could be due to the shorter contig length or potential misassembles including bacterial reads that were not filtered out, suggesting the need for further decontamination tools. Like the VCs, most mVCs were

Caudovirales, though small numbers of *Phycodnaviridae*, *Mimiviridae*, *Iridoviridae*, and *Herpesviridae* were enriched compared to the VCs. These eukaryotic viruses could represent viruses that are selected against in the VLP extraction protocol (such as filters that exclude larger *Megavirales*) or possible false annotations with Demovir [299].

An average of 6.24% and a high of 10.9% of host-removed reads in each metagenome sample could align to the assembled mVCs, decreasing to a mean of 4.35% and maximum of 9.94% of reads aligning to VCs. The significance of viral contig alignments in whole metagenome data is not well understood, limited by the lack of viral databases and the predominant use of 16S rRNA gene sequencing to characterize the bacteriome, even in virome-metagenome studies. Several bioinformatic approaches have been developed to mine existing whole metagenome data for viruses, including VirSorter [96], VirFinder [116], virMine [300], PhagePhisher [301], and others [302]. A reanalysis of a human microbiome gene catalog that originally attributed 0.1% of its contents to both eukaryotes and viruses had 1.31–38.43% of contigs predicted as viral; the authors attributed this discrepancy to prophages [300]. Efforts to infer phage attributes include alignments to well-characterized phages and prophages, searches for integrases and transposes, and tests for circularity; yet it remains difficult to interpret the presence of phage assemblies and viral alignments from short-read metagenomic studies. Given the dynamic potential for bacteriophages to be present as integrated prophages, extrachromosomal elements, intracellular phages, extracellular free or membrane-bound phages, and other forms [19], sequencing approaches will need to be paired with other fields of study to contextualize these results.

We were unable to meaningfully detect our spike-in phage through whole metagenome sequencing. This result can be attributed to the fact that our microbiome extraction methods do not aim to retain or lyse viral particles. Alternatively, or in concert with kit limitations, the quantity of

free phage DNA at physiological concentrations may be outcompeted by microbial DNA to a degree that renders deep shotgun sequencing ineffective for detection. These findings should be further tested with a diverse range of phages and additional samples with various DNA extraction methods; regardless, viral contig alignments in our whole metagenome sequences are unlikely to represent free phages. These reads are thus more likely to represent prophages or other forms of intracellular phages, while the virome sequencing reads would exclude prophages, and could explain the alignment gaps between VCs and mVCs. Both our clustering efforts and beta-diversity analysis demonstrate that the VCs and mVCs represent distinct viral communities with some overlap represented by 251 clusters containing 837 contigs. Bray-Curtis distances suggest that the viral portion of the whole metagenome may be more conserved than the VLP-enriched virome, which could include a community of prophages that are only induced under specific conditions.

Importantly, these results indicate that virome studies that employ metagenome-mining techniques should be interpreted differently than VLP-enriched viromes, echoed by recent analysis by Gregory *et al.* (2020). We also similarly report an increased number of viral contigs identified in the whole metagenome with a decreased average contig length when comparing paired samples. Differences between these two datasets can only be investigated when both virome-focused and whole metagenome sequencing are performed in parallel, an analysis that has been rarely performed in virome studies [55]. These comparisons are made easier by mucosal-luminal interface sampling, though these analyses will need to consider biases in each methodology, including the choice of extraction kit and use of amplification techniques (e.g. MDA).

2.6.4. Virome-bacteriome interactions

We observed inverse alpha diversities between viral and bacterial communities, which was previously observed in the infant virome [170]. Lim *et al*. attributed this relationship to dynamic

changes in the developing infant gut microbiome; whether this effect is also demonstrated in these patients will require a control cohort of subjects without Crohn's disease.

We also identified specific VC–host species pairs based on WIsH predictions against a large set of reference bacterial genomes. We focused on VC-host pairs involving bacterial strains also identified in our bacteriome. For 411 VC-host pairs involving 52 unique bacterial species, Spearman correlations across seven samples were calculated. Positive correlations reflect stable host-phage interactions [33,170], suggestive of "piggyback-the-winner" relationships that have been proposed for the mucosa-associated virome [191]; negative correlations suggest predatory-prey relationships. We were able to visualize these potential interactions with species-level resolution. Compared to unpaired VC-host interactions (i.e. a null comparison), we observed significantly higher Spearman correlations in our VC-host pairs ($p = 4.78e\text{-}6$). This result suggests that VC-host pairs in the mucosal-luminal interface microbiome tend to be positively correlated. Interpretation of these data is limited by sample size and complicated by host inflammation; further studies are required to explore these interactions and investigate how virome-bacteriome dynamics are affected by additional stresses like host disease. Additionally, incorporation of CRISPR spacers and other annotation tools could strengthen these analyses.

2.6.5. Protocol Limitations

While we were able to characterize the colonic mucosal-luminal interface virome, we were unable to recover viruses from the terminal ileum. This limitation may be due to a lower viral load, though a distal colon sample also lacked sufficient DNA; further optimization of techniques or a higher biomass may be required.

Our protocol employed multiple displacement amplification to increase DNA input for virome sequencing. While frequently used, MDA approaches can skew the observed viral community, selecting for small, circular ssDNA viruses [296]. Newer approaches including alternative linker amplification or tagmentation may be implemented within our protocol to reduce this bias [303]. Like other virome sequencing efforts, we also did not attempt to characterize enveloped or RNA viruses. RNA viruses tend to represent transient, plant pathogens that are less likely to be involved in human health [304]. Additionally, enveloped RNA viruses encompass many human viruses including SARS-CoV-2; these human pathogens tend to be comparatively well characterized and/or have existing tools to enable their direct study. Moreover, outside of viral infections, eukaryotic viruses are typically found in low-abundance in the human gut [143]; their role in microbiome-implicated human diseases like inflammatory bowel disease has also been suggested to be limited [305], though potential eukaryotic virome signatures have been reported [273].

2.6.6. Future Directions

We demonstrate that our protocol enables the site-specific study of the human gut virome in the context of the whole metagenome at the mucosal-luminal interface. Future research can apply these techniques to further investigate many of the hypotheses discussed here, while also examining the virome in site-specific pathologies like Crohn's disease.

Many key questions about the human gut virome remain, including whether a core healthy human virome exists and whether phage treatments could be a viable approach in microbiome modulation therapies. These questions cannot be fully answered with stool alone. A focused effort to characterize the virome along the length and cross-section of the gastrointestinal tract is required to provide a higher-resolution understanding of virome-bacteriome-host interactions. Studying the

mucosal-luminal interface virome is one step forward in these efforts to develop a more comprehensive model of the human microbiome.

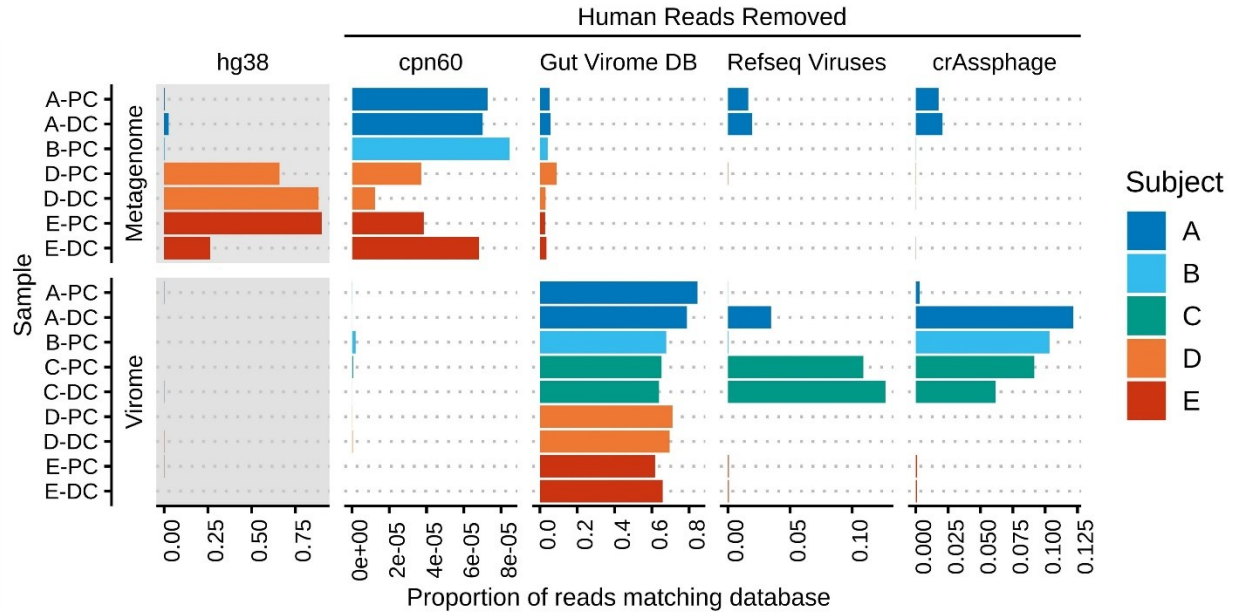


Figure 2-1: Mapping of metagenome and virome sequencing reads to human, bacterial, and viral databases.

Metagenome and virome sequencing reads were mapped to the human genome (hg38), a bacterial chaperonin database (cpn60), the Gut Virome Database, all NCBI RefSeq Viruses (May 2020), and 249 crAssphage-like contigs. Bars indicate the proportion of reads aligning to each database; human read counts were only included in the gray panels while subsequent panels were rescaled after human and phage-spike in reads were removed. PC: proximal colon; DC: distal colon.

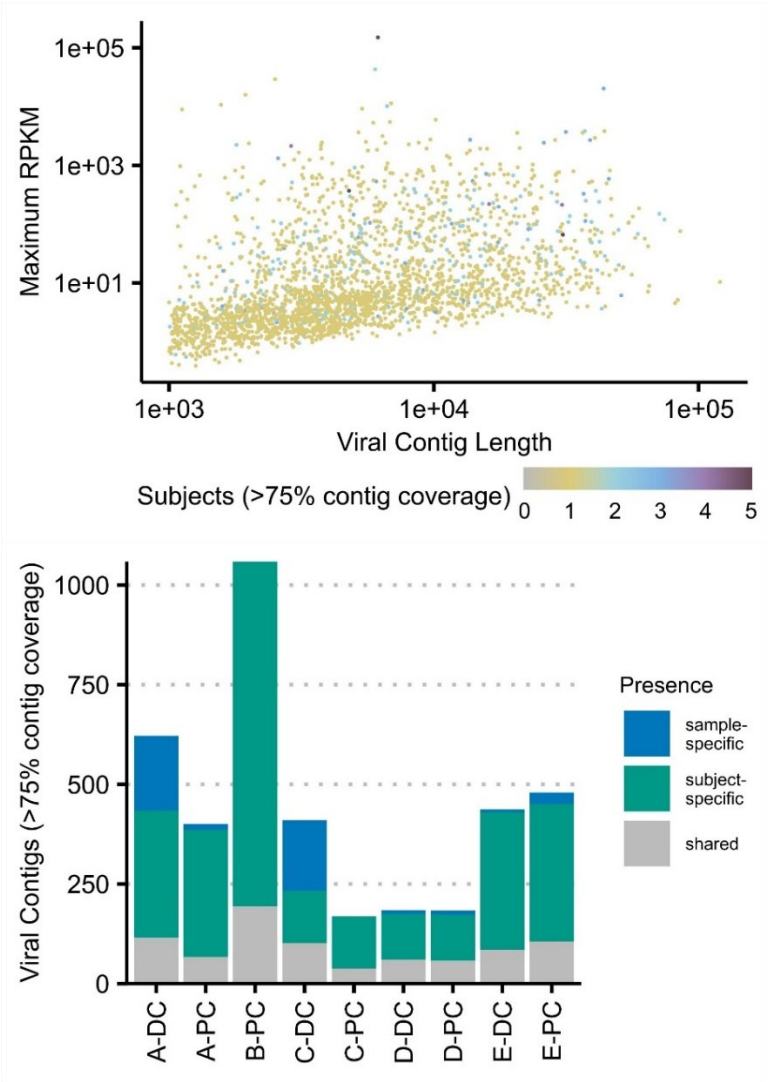


Figure 2-2: Viral contigs at the colonic mucosal-luminal interface.

(A) All viral contigs (VC) were plotted by their maximum observed abundance (RPKM-adjusted) vs length. VCs were coloured by the number of subjects where the contig was observed at ≥75% horizontal contig coverage. (B) The number of contigs present at ≥75% horizontal coverage was plotted for each sample, shaded by whether the contig was only observed in that sample, that subject, or in two or more subjects ("shared"). The same plots with each sample subsetted to two million reads is shown in Supplementary Figure 4. PC: proximal colon; DC: distal colon.

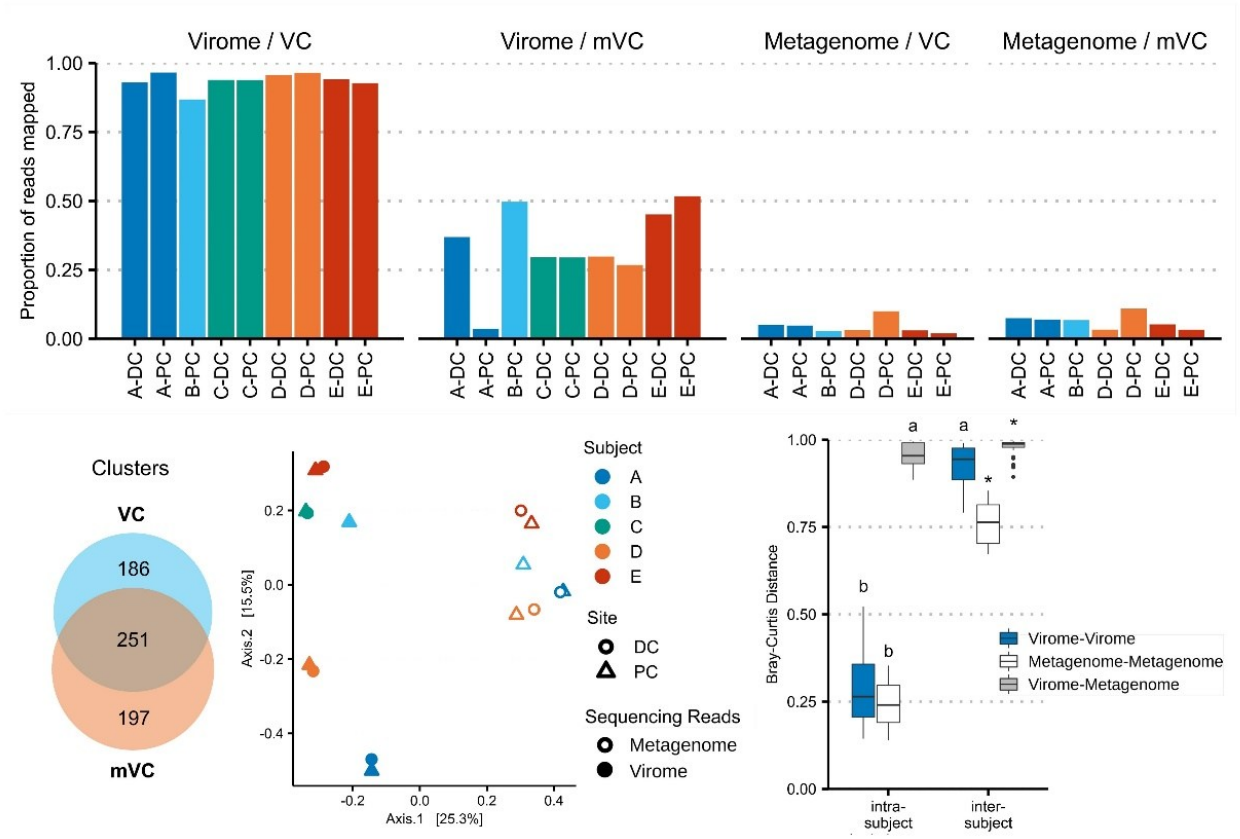


Figure 2-3: Viral contigs derived from virome and whole microbiome sequencing represent different viral populations.

(A) Virome and whole microbiome sequencing reads were each mapped to the 2511 virome-derived viral clusters (VCs) and 3122 metagenome-derived viral clusters (mVCs). The proportion of reads mapped is shown for each sample. **(B)** VCs and mVCs were clustered using vConTACT2, forming 634 clusters of two or more contigs. The Venn diagram illustrates the number of clusters that include only VCs, only mVCs, or both. Network maps are shown in Supplementary Figure 6. **(C)** Virome sequencing read counts (mapped to VCs) and metagenome sequencing read counts (mapped to mVCs) were merged using all vConTACT2-generated viral clusters. Bray-Curtis

distances were calculated and plotted using principal coordinate analysis. **(D)** The Bray-Curtis distances were plotted by analysis type and compared within and between patients. * indicates an FDR corrected p-value of <0.05 against every other subgroup; *a* indicates significance against all other comparisons except *a*; *b* indicates significance against all other comparisons except *b*.

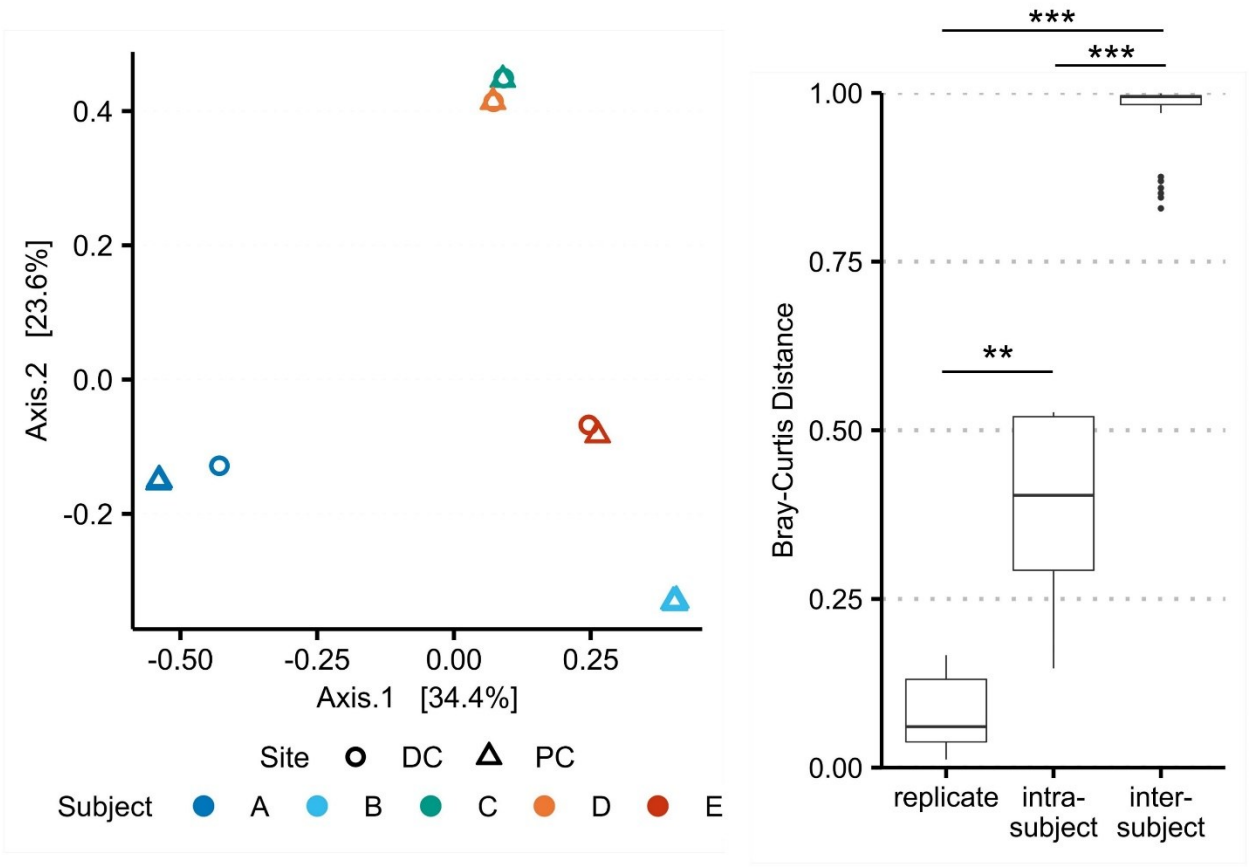


Figure 2-4: Beta-diversity between technical replicates demonstrate protocol reproducibility and intrasubject variation.

Bray-Curtis dissimilarities for viral populations in the virome (as assessed using assembled viral contigs) were recalculated with A-PC and B-PC triplicates. **(A)** Principal coordinate analysis shows significant overlap of the technical replicates. **(B)** Bray-Curtis distances were compared between replicates, intrasubject samples, and intersubject samples. ** marks an FDR-corrected p-value of < 0.01; *** < 0.001.

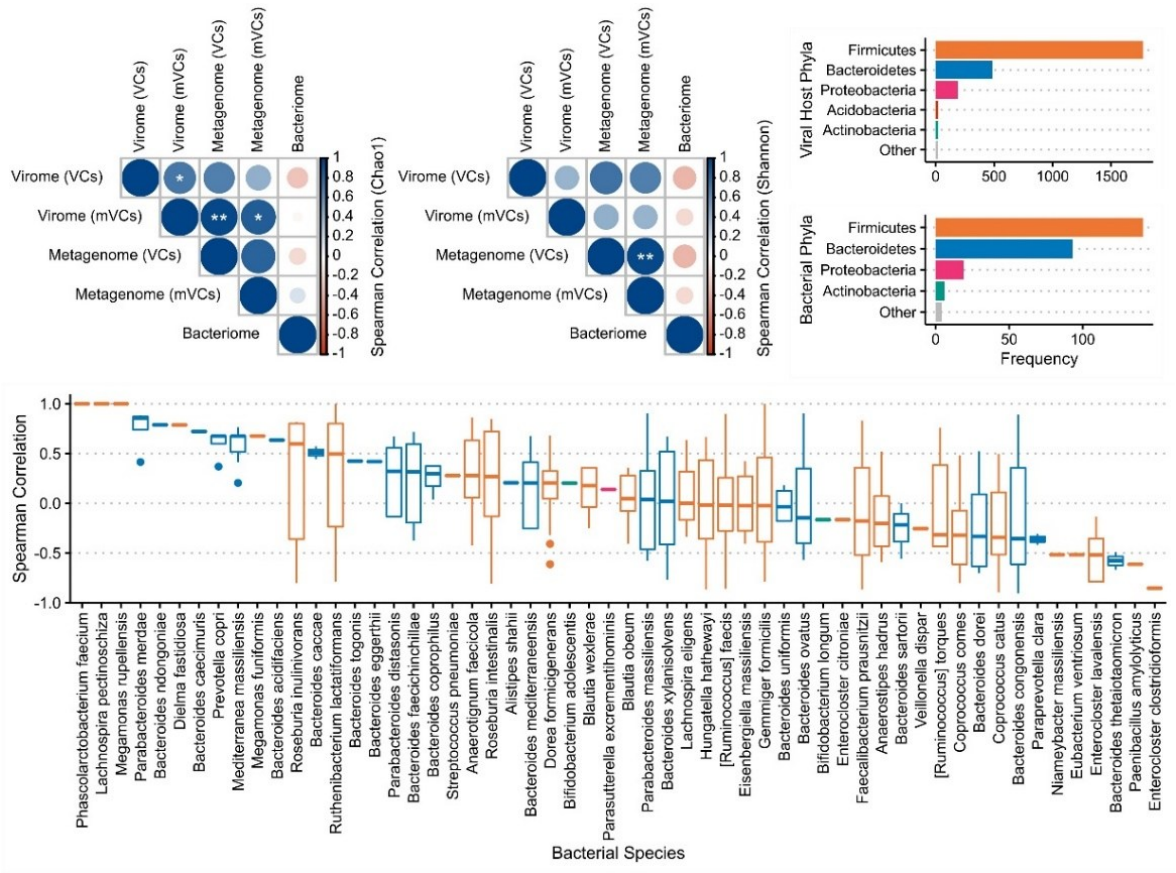


Figure 2-5: Virome-bacteriome relationships at the mucosal-luminal interface.

(A) Chao1 indices were used to calculate the alpha-diversity of each sample's bacterial and viral communities, the latter measured with both virome-derived viral contigs (VC) and metagenome-derived viral contigs (mVC). Spearman correlations between these alpha diversities were calculated and shown here. * marks an FDR-corrected p-value of < 0.05; ** < 0.01. Alpha diversities for each sample were plotted in Supplementary Figure 7. **(B)** Putative bacterial hosts, mainly representing Firmicutes, Bacteroides, and Proteobacteria, were assigned to 2056 of the 2511 VCs and plotted in comparison to **(C)** the number of bacterial species in each phylum identified in the bacteriome. **(D)** 52 bacterial species that were present in the metagenome were identified as the most likely host for 411 VCs (excluding Subject C). The Spearman correlation between each VC

and associated host species across all samples was plotted. Each boxplot is coloured by the bacterial host phyla as presented in Panels B and C.

Table 2-1: Subject and sample descriptions.

TI: terminal ileum; PC: proximal colon; DC: distal colon. Read counts reflect quality-filtered reads.

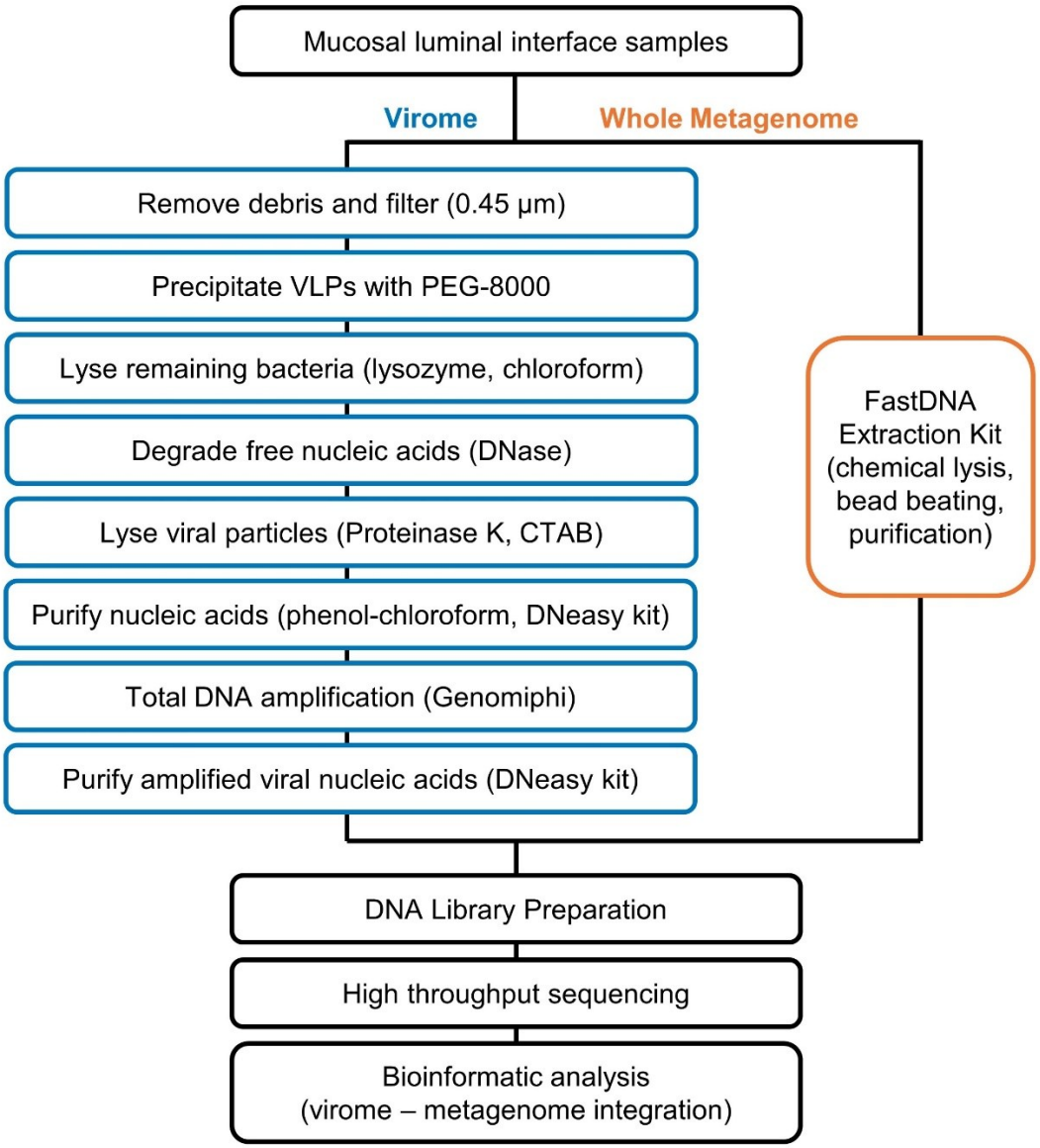
* indicates sequencing efforts that were evaluated with the addition of an exogenous phage.

Subject	Sex	Age (years)	Disease Phase	Site	Site Mucosal Inflammation	Virome Reads	Metagenome Reads
A	Male	16.6	Remission	TI	No	Insufficient DNA	Not performed
				PC	No	15,429,836*	22,190,375*
				DC	No	9,960,257	33,429,010*
B	Female	11.3	Remission	TI	No	Insufficient DNA	Not performed
				PC	No	25,566,766*	21,652,739*
				DC	No	Insufficient DNA	Not performed
C	Female	13.5	Flare	PC	Yes	2,002,855	Not performed
				DC	Yes	18,134,356	Not performed
D	Female	13.7	Diagnosis	PC	No	5,660,849	158,638
				DC	No	6,620,231	4,165,957
E	Male	14.1	Diagnosis	PC	Yes	6,008,391	4,932,195
				DC	Yes	6,334,986	14,043,455

Table 2-2: Virome and whole metagenome assembled viral contigs.

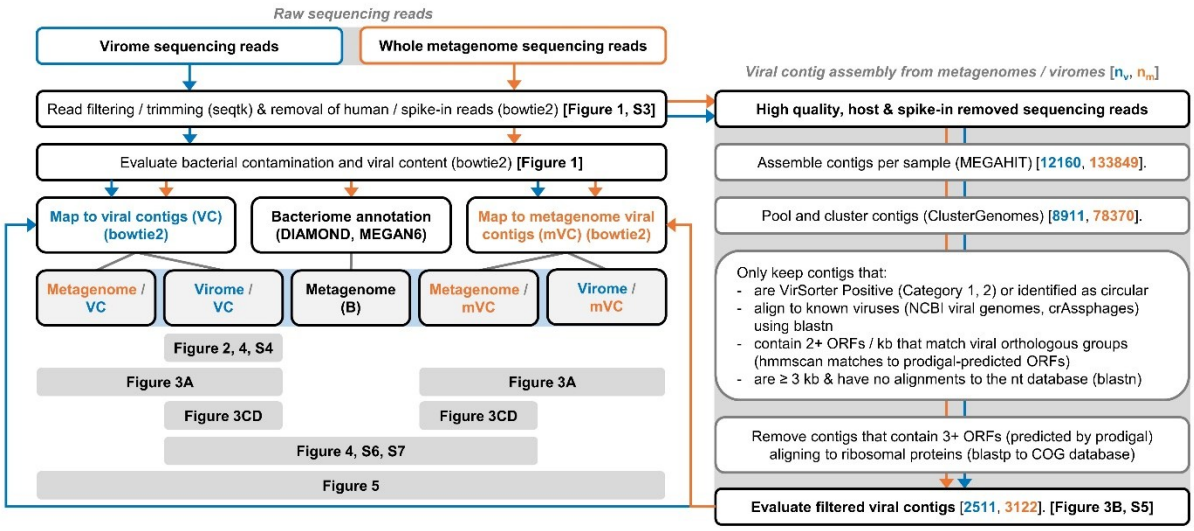
Virome and whole metagenome sequencing reads were assembled and filtered for viral properties, resulting in 2511 viral contigs (VCs) and 3122 metagenome-derived viral contigs (mVCs). The source read counts represent host and spike-in removed reads that were subjected to quality filtering and trimming. Contigs were annotated using Demovir.

	Viral Contigs (VCs)	Metagenome-derived Viral Contigs (mVCs)
Source	Virome Sequencing (93.2 million reads)	Whole Metagenome Sequencing (87.6 million reads)
Samples	9 / 9 mucosal luminal interface samples	7 / 9 mucosal luminal interface samples
Total Viral Contigs	2511	3122
Contigs / Source Reads	2.70e-5	3.56e-5
Mean Contig Length	8413 bp	6233 bp
Viral Family		
Caudovirales		
Myoviridae	275	204
Podoviridae	90	41
Siphoviridae	904	430
Unassigned *Caudovirales*	285	169
Other viruses		
Anelloviridae	14	0
Herpesviridae	0	2
Iridoviridae	0	2
Microviridae	24	2
Mimiviridae	0	2
Phycodnaviridae	1	17
Unclassified	918	2251



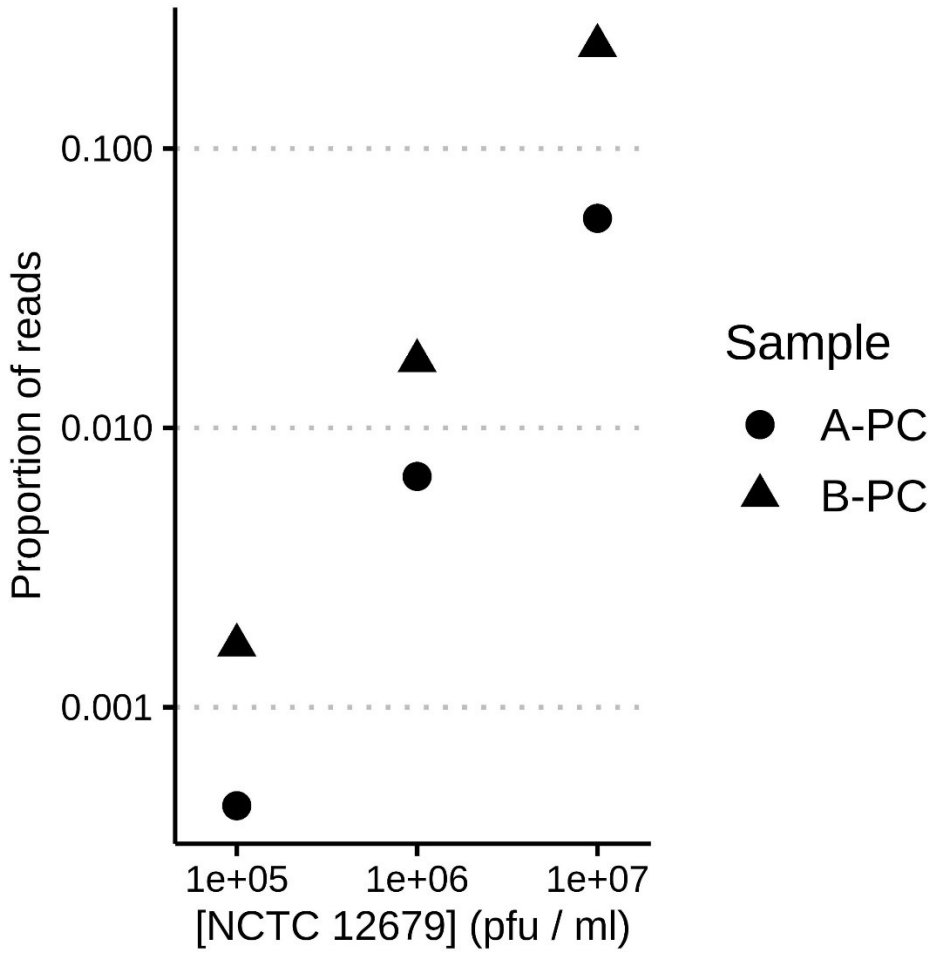
Supplementary Figure 2-1: Summary of virome and whole metagenome DNA extraction and sequencing protocols.

Viromes and whole metagenomes were extracted from mucosal-luminal interface samples. Virus-like particle purification and the removal of remaining bacterial cells was required for efficient virome sequencing. The full protocol is described in Materials and Methods. VLPs: virus-like particles; PEG: polyethylene glycol; CTAB: cetyltrimethylammonium bromide.



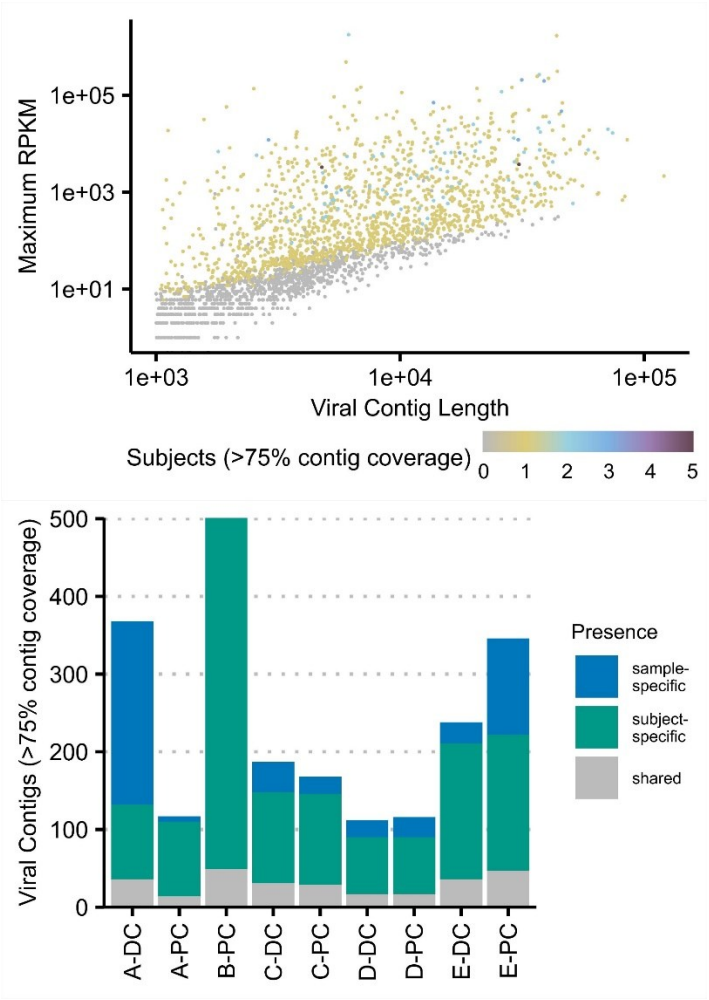
Supplementary Figure 2-2: Summary of bioinformatic pipeline and subsequent analysis.

Sequencing reads were first quality filtered and subjected to host-read removal. Both virome and metagenome sequencing reads were then assembled into contigs which were subjected to a viral contig identification pipeline. Sequencing reads could then be mapped to these contigs for further analysis. Bacteriome annotation of the whole metagenome sequencing reads was also performed. Full details and programs are described in Materials and Methods. VC: virome-derived viral contigs; mVC: metagenome-derived viral contigs; B: bacteriome; ORFs: open reading frames; COG: Clusters of Orthologous Groups of proteins.



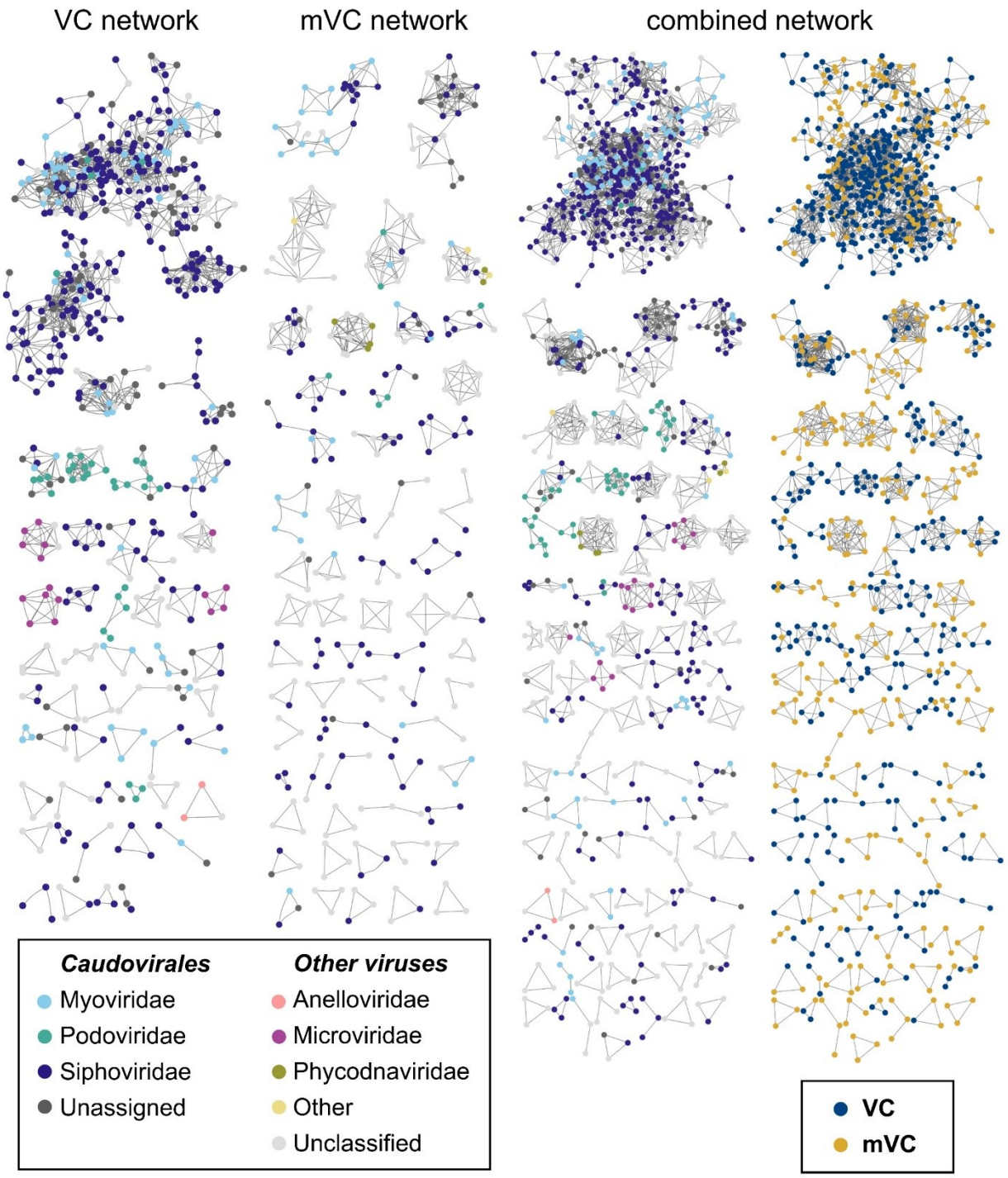
Supplementary Figure 2-3: Virome sequencing reads matching exogenous phage are linearly correlated with spike-in phage titres.

An exogenous phage, NCTC 12673, was added to mucosal-luminal interface aspirates from the proximal colon (PC) of subjects A and B at concentrations of 10^5, 10^6, and 10^7 pfu ml^{-1}. Virome sequencing reads were mapped to the phage genome. The proportion of reads aligning to NCTC 12673 were plotted against the phage titres, showing a linear relationship ($R^2 > 0.99$).



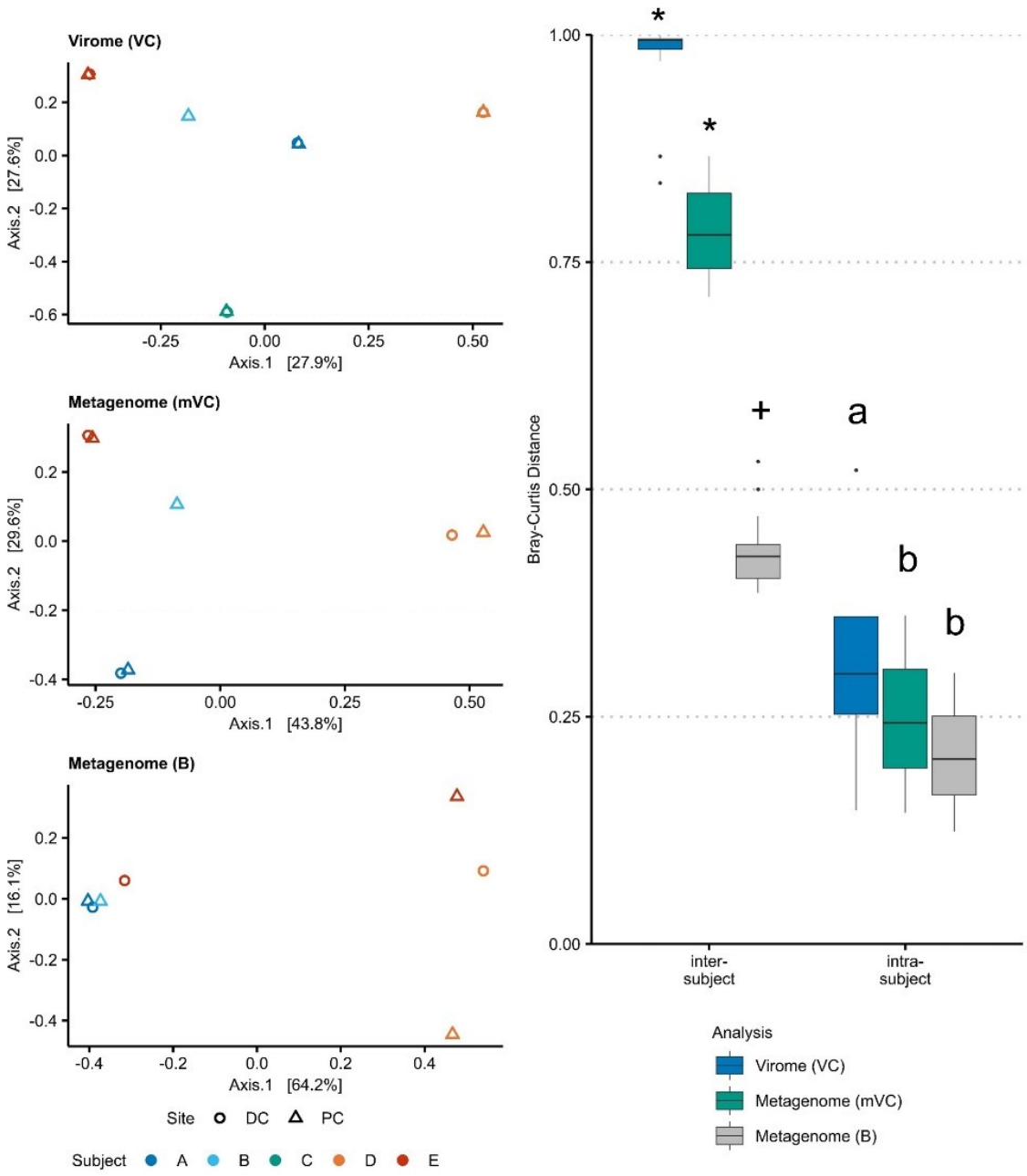
Supplementary Figure 2-4: Viral contigs at the colonic mucosal-luminal interface (subsetted dataset).

Prior to viral contig mapping, virome sequencing reads for each sample were randomly subsetted to two million reads to represent an even sequencing depth. **(A)** All viral contigs (VC) were plotted by their maximum observed abundance (RPKM-adjusted) vs length. VCs were coloured by the number of subjects where the contig was observed at ≥75% horizontal contig coverage. **(B)** The number of contigs present at ≥75% horizontal coverage was plotted for each sample, shaded by whether the contig was only observed in that sample, that subject, or in two or more subjects ("shared"). PC: proximal colon; DC: distal colon.



Supplementary Figure 2-5: Clustering of virome-derived and metagenome-derived viral contigs.

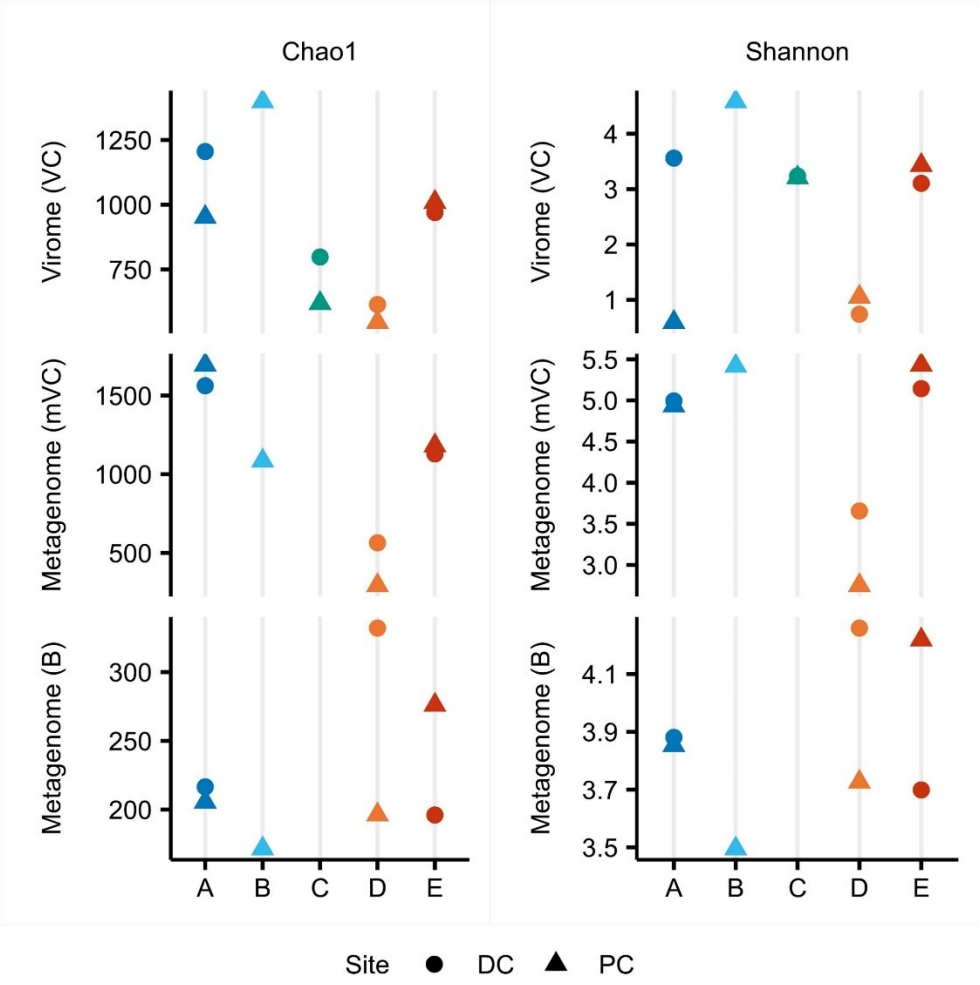
Viral contigs from virome (VC) and metagenome (mVC) datasets were pooled and clustered using vConTACT2. Contig networks were plotted using Cytoscape 3.7.2 using its default Perfuse Force Directed Layout using vConTACT2-derived edge weights. Outliers, singletons, and doubletons were excluded from the network visualizations. Contigs are coloured by viral annotations from Demovir for each network, while the combined network is also shown coloured by VCs and mVCs in the right panel.



Supplementary Figure 2-6: Beta-diversity of the mucosal-luminal interface virome and bacteriome.

Bray-Curtis distances between samples were calculated using the assembled viral contigs (VC) in the virome, metagenome-derived viral contigs (mVC) in the whole metagenome, and bacterial taxa (B) in the whole metagenome. **(A)** Principal coordinate analyses were plotted for each dataset. **(B)**

Bray-Curtis distances were compared between subjects and within subjects for each dataset. * indicates an FDR-corrected p-value of < 0.05 against every other subgroup; + indicates significance against all other comparisons except *a*; *a* indicates significance against all other inter-subject comparisons except +; *b* indicates significance against inter-subject comparisons only.



Supplementary Figure 2-7: Alpha-diversity of the mucosal-luminal interface virome and bacteriome.

Using the Chao1 index, alpha diversity was measured using the relative abundance of viral contigs (VCs) and metagenome-derived viral contigs (mVCs) in the virome, VCs, and mVCs in the whole metagenome, and bacterial taxa in the metagenome (B). For each dataset, read counts were first subsetted to the sample with the lowest number of mapped reads (Virome / VC: 1,877,966; Virome mVC: 528,240; metagenome / VC: 5,358; metagenome / mVC: 5,867; bacteriome: 38,711).

3. Multiomic spatial analysis reveals a distinct mucosa-associated virome

3.1. Preface

This chapter has been previously published as the following research article:

> **Yan A**, Butcher J, Schramm L, Mack DR, Stintzi A. Multiomic spatial analysis reveals a distinct mucosa-associated virome. Gut Microbes. 2023 Jan-Dec;15(1):2177488. doi: 10.1080/19490976.2023.2177488. PMID: 36823020; PMCID: PMC9980608.

Specific author contributions are as follows:

Austin Yan: performed all virome experiments, sequencing, multiomic data analysis, and wrote the initial analysis.

James Butcher: provided input on the bioinformatic analysis and manuscript, assisted with analysis of metatranscriptomic data.

Laetitia Schramm: performed all RNA-seq sample processing.

David Mack: recruited patients, provided clinical and demographic information, and revised the manuscript prior to submission for peer review.

Alain Stintzi: provided reagent and research materials, obtained funding, supervised research, conceptualized project, and revised the manuscript prior to submission for peer review.

AY and AS designed the experiments. AY performed all virome extractions, data analysis, and wrote the initial manuscript. Bulk metagenome DNA extractions were performed by AY, while LS performed the RNA extractions. DM was responsible for sample collection and clinical data

generation. JB assisted in bioinformatic analysis and data interpretation. All authors reviewed and provided comments on the final manuscript.

3.1.1. Disclosure Statement

AS and DM are co-founders of MedBiome, a clinical microbiomics company. The other authors have no competing interests to declare.

3.2. Abstract

The human gut virome has been increasingly explored in recent years. However, nearly all virome sequencing efforts rely solely on fecal samples and few studies leverage multiomic approaches to investigate phage-host relationships. Here, we combine metagenomics, metaviromics, and metatranscriptomics to study virome-bacteriome interactions at the colonic mucosal-luminal interface in a cohort of three individuals with inflammatory bowel disease; non-IBD controls were not included in this study. We show that the mucosal viral population is distinct from the stool virome and houses abundant crAss-like phages that are undetectable by fecal sampling. Through viral protein prediction and metatranscriptomic analysis, we explore viral gene transcription, prophage activation, and the relationship between the presence of integrase and temperate phages in IBD subjects. We also show the impact of deep sequencing on virus recovery and offer guidelines for selecting optimal sequencing depths in future metaviromic studies. Systems biology approaches such as those presented in this report will enhance our understanding of the human virome and its interactions with our microbiome and our health.

3.3. Plain Language Summary

The human gut hosts many bacteria-infecting viruses, also known as phages. These phages interact with other gut microbes and affect human health, yet they are understudied. We advance the study of human gut phages on three frontiers by studying gut viruses in three patients with bowel diseases. First, nearly all phage studies use fecal sampling, however these phages may be transient and are less relevant to our health. By obtaining samples during colonoscopy, we studied phages from the colonic wall and show that they are different than those found in our stool. Second, we use RNA sequencing to explore how phages interact with bacteria within a complex microbial ecosystem. By applying these new techniques, we advance our study from "what phages are there?" to "what are they doing?" Lastly, this study reports one of the most extensive per-sample sequencing efforts to date, allowing us to make practical recommendations for other researchers aiming to design phage studies. Further studies will expand our understanding of our gut phages and how they interact with human health.

3.4. Introduction

The human gut virome, which encompasses the vast and diverse collection of viruses in our gastrointestinal tract, has been linked to several human diseases including inflammatory bowel disease, cancer, and diabetes, and may impact the efficacy of microbiome-modulating therapies.[20,83,190] Our gut viromes are highly individualized, temporally stable, and consist mainly of double-stranded DNA bacteriophages from the *Caudovirales* order.[45]

Virome research has lagged behind its bacterial counterpart due to difficulties in sample processing, unannotated viral "dark matter", and high viral heterogeneity. Yet recent advances have led to the exploration of virome assembly and development in neonates, a significant expansion of viral databases, and the characterization and isolation of several crAss-like phages which were only first described in 2014.[55,84,163,164,173] Most studies rely either on the sequencing of virus-like particle (VLP) enriched metagenomes, herein referred to as metaviromes, or the identification of viral sequences within whole metagenomes (also referred to as bulk or whole-community metagenomes). Each approach captures a distinct subset of the virome community and, when used together, can increase the recovery of viral species and provide additional context about viral populations.[55,257,306] Paired metagenomic and metaviromic sampling also enables the study of interactions between viral species and their bacterial hosts.[306] Efforts to integrate other experimental technologies, including metatranscriptomics and metaproteomics, are also evolving, bringing a systems biology approach to viromics and its transkingdom interactions.[83,143]

To date, metatranscriptomic approaches have been rarely applied in the context of human metaviromics. This may be due to the additional technical challenges of obtaining high quality RNA, the lack of procedures to enrich for viral transcripts, and limited viral sequence databases.

Oral viromes have been investigated using RNA-seq; one study was able to map 30% of the reads to viral populations, but lacked the sequencing depth for further downstream analysis such as assembly.[69] Data mining efforts in the past few years have also identified ssRNA phages in activated sludge and aquatic environments, significantly expanding the number of known ssRNA phage genomes.[307] Due to the challenge of studying viromes at the community level, transcriptomic analysis has thus been primarily used for more focused phage-host studies such as phage profiling of *Salmonella* and *Yersinia*,[308,309] and more recently ΦcrAss001 and its host *Bacteroides intestinalis*.[162]

Furthermore, most of our understanding of the "gut" virome is based on fecal viromes, which, like the bacteriome, likely do not represent the complex biogeography of our gastrointestinal tract.[247] A recent profiling of gastrointestinal tract viromes in the domestic pig and rhesus macaque show that virome composition was specific to anatomical region, with differences in viral load and diversity between luminal and mucosal samples.[310] The mucosal virome is hypothesized to have different ecological pressures than the primarily luminally-derived stool virome and may have greater influence over the host-associated bacteriome.[38,190,191,269] Rectal samples have been used to study viromes in patients with ulcerative colitis while a recent study utilized colon resections and ileostomy fluid.[213,246] These human studies, however, did not compare the virome across multiple sites or compare colonic sampling to fecal viromes.

We recently demonstrated the use of mucosal-luminal interface samples to study the gut virome at specific sites within the gastrointestinal tract.[257] This sampling technique also provides sufficient microbial content for matched whole metagenomic and metatranscriptomic shotgun sequencing. Here, we leverage these advantages to demonstrate a multiomic spatial characterization of the virome in three treatment-naïve, pediatric patients with ulcerative colitis, an

inflammatory bowel disease. We also highlight the power of matched metatranscriptomic sequencing to reveal viral gene transcription to better understand virome-bacteriome interactions. We did not seek to compare the IBD or ulcerative colitis virome to non-IBD controls in this study.

3.5. Patients and Methods

3.5.1. Resource Availability

Lead contact: Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Alain Stintzi (astintzi@uottawa.ca).

Materials availability: This study did not generate new unique reagents.

Date availability: The datasets generated during this study are available at NCBI BioProject PRJNA818303.

3.5.2. Experimental Model and Participant Details

Sample collection from pediatric patients was approved by the Research Ethics Board of the Children's Hospital of Eastern Ontario (CHEO) in Ottawa, Canada with informed consent/assent obtained from parents and/or participants. Mucosal luminal interface samples from three treatment-naïve patients were obtained during diagnostic endoscopy following a standard colonoscopy preparation protocol.[311] Demographic and clinical information is shown in Table 1. Stool samples were collected either before or after endoscopy and stored at -80°C.

3.5.3. Sample Collection

The collection of mucosal-luminal interface (MLI) aspirates has been described previously.[247] In brief, sterile water was used to wash the bowel wall at the proximal and distal colon during colonoscopy to remove the loosely adherent mucous layer. The wash was then

aspirated into a sterile container and stored at -80°C. Aliquots of 10 ml were used for virus-like particle purification and viral DNA extraction; 20 ml were used for RNA extraction, and 2 ml was used for whole metagenomic DNA extraction.

3.5.4. Virus-like Particle Purification and Nucleic Acid Extraction

We have previously described a protocol to purify virus-like particles from mucosal aspirates.[257] In summary, 10 ml aliquots of mucosal aspirates or 0.5 g of stool homogenized in 10 ml saline-magnesium buffer were subjected to centrifugation and sequential filtration at 5.0 and 0.45 µm filters (Sigma-Aldrich, SLSV025LS and SLHV033RB) to remove debris and cellular content. Virus-like particles were precipitated by overnight incubation at 10% w/v PEG-8000 (Fisher Scientific, BP233) and resuspended in saline-magnesium buffer. Remaining bacterial cells were lysed by treatment with 1 mg/ml lysozyme (Sigma, L4919) for 30 min at 37°C followed by 0.2 volumes of chloroform (10 min, room temperature). After centrifugation (5 min, 2500 *g*), the aqueous mixture was treated with TURBO DNase (Thermo Scientific, AM2238) and RNaseI (Life Technologies, EN0602) in a buffer of 1 mM $CaCl_2$ and 5 mM $MgCl_2$ for 1 hour at 37°C to degrade remaining bacterial nucleic acids. Enzymes were inactivated at 70°C for 10 minutes. Virus-like particles were lysed with Proteinase K (3.2 µg/ml; Fisher Scientific, BP1700) in 3.2% SDS for 20 minutes at 55°C, then treated by 2.5% cetyltrimethylammonium bromide (Fisher Scientific, O3042) with 0.5 M NaCl for 10 minutes at 65°C. Viral DNA was then extracted by adding 1 volume of phenol-chloroform-isoamyl alcohol (25:24:1, pH 6.7) to each mixture, which was vortexed and subjected to centrifugation (10 min, 8000 *g*); this step was repeated with chloroform to remove trace phenol. Nucleic acids were purified from the aqueous layer using the Dneasy Blood and Tissue Kit (QIAGEN, 69506) and eluted in 50 µl of water. DNA was concentrated using an Eppendorf™ Vacufuge™ Concentrator to 3 µl to maximize the input DNA for the GenomiPhi™

V2 DNA Amplification polymerase kit (GE Life Science, 25660032). Reactions using 1 µl of input DNA were run in triplicate, then pooled and purified with the Dneasy Blood and Tissue Kit. DNA was quantified fluorescently using the Qubit dsDNA HS Assay Kit (Thermo Fisher, Q32854).

3.5.5. Whole Metagenomic and Metatranscriptomic Extraction

Whole metagenomic DNA was extracted using the FastDNA Spin Kit for DNA Isolation (MP Biomedicals, 116540600), eluted in water, and quantified using the Qubit High Sensitivity dsDNA Assay Kit as previously described.[251]

Total RNA was freshly extracted from pellets obtained from MLI aspirates, using a modified phenol/chloroform extraction protocol adapted for samples with high mucin content.[312] Briefly, fresh 20 ml MLI aliquots were subjected to two sequential centrifugations at 4°C (700 *g* for 5 min; 13000 rpm for 20 min) to remove cellular debris and pellet bacterial cells. The pellets were resuspended in denaturing buffer (4 M guanidine thiocyanate, 25 mM sodium citrate, 0.5% N-lauroylsarcosine, 1 M 2-mercaptoethanol, 1% N-acetyl cysteine) and 0.1 volumes of 1 M sodium acetate pH 5.2. The samples were incubated at 65°C for 2 min to lyse the cells, preheated buffer-saturated phenol pH 4.3 was added at a 1:1 ratio and incubated at 65°C for 10 min with frequent inverting. The samples were then chilled by placing them directly on ice for 15 min, chloroform was added (0.5 volumes) and the tubes inverted 10 times. The aqueous phase was separated by centrifugation at 13000 rpm for 25 min at 4°C and the RNA precipitated at -80°C by adding 2.5 volumes of ethanol, 0.1 volumes 3 M sodium acetate pH 5.2 and EDTA added to a final concentration of 1 mM. Precipitated RNA extractions were further cleaned by washing the pellets 5 times with chilled 80% ethanol with 5 min 13000 rpm centrifugations at 4°C. The RNA was then resuspended in Rnase free water (Ambion, AM9937) and stored at -80°C. Contaminating DNA was removed using sequential Dnase I (Thermo Scientific, EN0521) treatments and further purified

and concentrated using the RNA Clean & Concentrator-25 (Zymo Research, R1017). The absence of human and bacterial DNA was then assessed using PCR with primers targeting human actin (hACTBf2: AGTCCTACGGAAAACGGCAG, hACTBr2: CACCCTGAAGTACCCCATCG) and the bacterial V4-16S rRNA hypervariable region (V4_bc45: CCATCTCATCCCTGCGTGTCTCCGACTCAGTAACGCGTTCGAYTGGGYD TAAAGNG, V4_revs: CCTCTCTATGGGCAGTCGGTGATTA CNVGGGTATCTAATCC).[251] PCR products were analyzed on a 1% agarose gel and the Dnase I treatment repeated if a band was present. RNA quantity and quality were evaluated using the Qubit RNA HS Assay Kit (Thermo Fisher, EN0602) and the RNA 6000 Nano Kit (Agilent Technologies, 5067-1511) on the 2100 Bioanalyzer. Samples with an RNA Integrity Number (RIN) ≥ 7 were deemed suitable for sequencing.

3.5.6. DNA sequencing, quality-filtering, and host-read removal

DNA and RNA sequencing were performed at the Génome Québec CES using the NEB Ultra II and NEB rRNA-depleted stranded library preparation kits, respectively, with subsequent sequencing on the Illumina NovaSeq 6000 platform. Cutadapt 2.10 was used to trim Illumina's universal adapters (AGATCGGAAGAG; AGATCGGAAGAG) while Trimmomatic 0.36 was used in paired-end mode to retain high quality sequences with the settings SLIDINGWINDOW:4:20, MINLEN:60, and HEADCROP:10. Reads mapping to the human genome (GRCh38 with bowtie2's ultra-sensitive mode)[276] were removed from further analysis using samtools 1.7.[277] Finally, low complexity reads were removed using Komplexity.[313] Bacterial contamination was assessed by aligning reads to the *cpn60* database with bowtie2's ultra-sensitive mode.[89]

An average of 230 million paired end reads (173-280 million) were obtained for each metavirome sample, totaling 34 Gbps, while an average of 154 million paired end reads (137-177

million) were obtained for each metagenome sample, totaling 23 Gbps. Quality filtering resulted in the removal of 6.16% of reads (4.91-9.10%) across all 18 samples, with no significant difference between metavirome and metagenome sequencing. Host read filtering removed 0.020 – 0.319% of high-quality metavirome reads and 0.0849 – 94.3% of high-quality metagenomic reads ($p = 0.0005$ between the two groups). Low complexity sequence filtering, which removes human microsatellite DNA and aids downstream assembly and analysis, was performed using Komplexity, which removed another 0.235% - 5.79% high-quality reads.[313] Overall, this resulted in a mean of 211 (157 – 254) million paired-end metavirome reads and 75.4 (6.83 – 162) million paired-end metagenomic reads per sample.

3.5.7. Metatranscriptomic analysis

Metatranscriptome sequencing reads were subject to adapter trimming and quality filtering using cutadapt and Trimmomatic as described above. TopHat2 was used to identity reads aligning to human transcripts and these were removed using samtools 1.9.[277,314] SortMeRNA and Komplexity was then used to remove ribosomal and low complexity reads, resulting in remaining high-quality, non-human, and non-ribosomal reads.[315] For RdRp and ssRNA marker identification, MEGAHIT v1.2.7 was used to assemble these reads at default settings (minimum contig length of 200 bp). ORFs were predicted using Prodigal 2.6.3 in its metagenomic mode,[119] and searched against databases of RNA-dependent RNA polymerases and capsids used by Cenote-Taker 2.1.3 and ssRNA marker genes using hmmsearch with a minimum E-value of 1E-05.[83,280,307]

3.5.8. Metagenomic assembly, viral contig identification, and viral gene annotation

The bioinformatic pipeline for VC identification is summarized in Supplementary Figure 2. Host-decontaminated, quality-filtered reads from each metagenome and metavirome sample

were assembled using MEGAHIT v1.2.7 with a minimum contig length of 3000 bp.[91] Metagenome and metavirome assemblies were each pooled and clustered using ClusterGenomes at 95% identity across 90% of the length of the smaller contig.[108] Viral contigs were identified using Cenote-Taker2 using default settings for whole metagenomic assembly (including pruning of flanking host regions) and virus-like particle preparation assembly, respectively. To examine the relationship between sequencing depth, this process was repeated with quality-filtered, metaviromic sequencing reads subsetted at increments from 10,000 to 100 million paired end reads per sample to simulate various sequencing depths.

All putative viral contigs were further pooled and clustered, resulting in 2,171 VCs. During this step, all metavirome-derived VCs were also clustered against the set of metagenomic non-viral contigs: non-viral contigs clustering within a viral contig were removed from the set of non-viral contigs, while metavirome-derived viral contigs clustering within non-viral contigs were considered prophages (along with host DNA flanked prophages identified in metagenome). Non-viral contigs that contained one or more prophages had these regions masked. Whole metagenomic assemblies that were not identified as viral contigs (n = 39,735) were annotated using the Contig Annotation Tool.[316]

VCs were assessed using CheckV,[95] annotated using Demovir,[283] then clustered with vContact2.[282] Members of a viral cluster were annotated by majority vote. As Demovir has not yet been updated to match the latest viral taxonomy established by ICTV, we manually removed reference to the families *Siphoviridae*, *Myoviridae* and *Podoviridae*, but opted to retain references to the order *Caudovirales* (now class *Caudoviricetes*) to allow for straightforward comparisons to previous published datasets. Open reading frame (ORF) prediction was performed using Prokka 1.14.5 with –meta enabled and –kingdom Bacteria or –Viruses against the set of non-viral and viral

contigs, respectively.[317] All open reading frames were clustered using Mmseqs2 13.45111,[109] and annotated using hmmsearch (minimum E-value of 1E-05), a part of HMMER 3.3,[280] against the following databases: VOG (release 206),[122] pVOG,[121] PFAM (34.0),[124] KEGG (98.0),[123] and TIGRFAM (15.0).[318] Genes were preferentially annotated in the order of the databases listed above. crAss-like phages were additionally annotated using a curated set of crAss-like phage protein profiles.[165] SNP analysis was performed on un-clustered crAss-like phages using Snippy 4.6.0.[319] A viral contig was considered to contain an integrase if it contained one or more of the 603 genes annotated with the term "integrase"; the majority of these (591/603) were annotated by VOG00035.

All metagenomic, metatranscriptome, and metaviromic reads were mapped to the pooled set of viral and non-viral contigs using bowtie2,[87,320] with counts and breadth of coverage calculated using samtools depth.[277] A breadth of coverage threshold of >=75% of the contig should have a minimum depth of >=1 read per bp was applied to metagenome and metavirome count tables; counts under this threshold was set to 0. The normalized relative abundance (NRA) of each VC was calculated as $\frac{x_i}{\sum x_i}$ where x_i is the number of reads mapping to a contig divided by the contig length. Gene counts were calculated using subread featureCounts.[321]

3.5.9. Viral gene transcription and host prediction

A viral contig was considered to be transcriptionally active if its NRA in the metatranscriptome was greater than 0.0001%. Gene features within transcriptionally active VCs with >1 counts were considered to be transcribed. Gene transcription was approximated using the ratio of NRA in the metatranscriptome to NRA in the metagenome.

WiSH 1.1 was run using the set of non-viral contigs against VCs.[126] Only virus-host scores with an adjusted p-value < 10^{-5} were selected for downstream analysis. Correlation with their viral hosts was calculated using a Spearman correlation.

3.5.10. Statistical analysis

Alpha-diversity and beta-diversity analysis, host-phage correlations, statistical analysis, and plotting were performed in R 4.1.3 using phyloseq 1.36.0.[284] Figures were created with the following packages: ggplot2 3.3.0, ggthemes 4.2.0, ggpubr 0.4.0, gggenes 0.4.1, ggalluvial 0.12.3, eulerr 6.1.0, and UpSetR 1.4.0. Alpha diversity analysis was performed using read counts rarefied to 1,104,709 reads (the lowest number of viral reads identified in a sample across the dataset). The Chao1 index was calculated as $S_0 + \frac{a_1^2}{2a_2}$ where S_0 was the total observed species, a_1 the number of species seen once and a_2 the number of species seen twice; and the Shannon index was calculated as $-\sum_i^S P_i \ln P_i$ where S was the total number of species and P_i the proportion of the total population composed of species i. The likelihood ratio was calculated as Sensitivity / (1 – Specificity).

3.6. Results and Discussion

3.6.1. Multiomic sequencing of the mucosal-luminal interface microbiome

Samples were obtained from three pediatric patients at the Children's Hospital of Eastern Ontario in Ottawa, Canada (Table 1). These participants (hereafter referred to as Participants F, G, and H) were treatment-naïve patients undergoing colonoscopy for confirmation of their clinically-suspected ulcerative colitis. Washes of the mucosal-luminal interface (MLI) from the proximal colon (PC) and distal colon (DC) were collected, along with a stool (STL) sample that was obtained between two days prior to endoscopy and twenty days after endoscopy. All samples were processed for metavirome and whole metagenome sequencing as previously described.[257] MLI samples were also subjected to metatranscriptome sequencing; H-DC had insufficient quality after library preparation and was not sequenced. After removal of low-quality, low-complexity, and host reads, a mean of 211 (ranging from 157 – 254) million paired-end metavirome reads were obtained per sample, providing one of the deepest metavirome sequencing efforts to date. Whole metagenome sequencing yielded an average of 75.4 (6.83 – 162) million paired-end reads per sample while metatranscriptomic sequencing yielded an average of 119 (107 – 139) million reads per sample.

As noted previously, whole metagenome sequencing at the mucosal-luminal interface is prone to high host contamination at inflamed sites (Supplementary Figure 1A).[257] Host contamination has been suggested as a possible biomarker for inflammation, reflecting epithelial and blood cells that are more likely shed in intestinal diseases.[322] In this study, all aspirates taken from sites of Mayo Score 2, which is indicative of moderate disease, had >69.8% host contamination.[323] In contrast, aspirates from non-inflamed sites (Mayo 0) or mildly inflamed sites (Mayo 1) had no more than 3.7% host contamination. Similarly, host content in the stool

metagenomes was markedly increased if Mayo Grade 2 inflammation in the proximal or distal colon was observed. These results suggest that moderate inflammation (Mayo 2) which includes erosions, complete loss of vascular pattern, and significant erythema but not mild inflammation (Mayo 1) is correlated with increased host content. Furthermore, both MLI and stool sampling are affected by mucosal inflammation.

Bacterial contamination in the metavirome sequencing data was assessed by aligning reads against a *cpn60* housekeeping gene database.[89] Metavirome samples had a mean 172-fold decrease of mapped *cpn60* reads compared to their whole metagenomic counterparts (Supplementary Figure 1B), demonstrating efficient viral purification in this dataset consistent with prior metavirome studies.[40,257]

3.6.2. Viral contig identification across multiomic datasets

High-quality, host-removed sequencing reads were assembled per sample and clustered (Supplementary Figure 2). Using Cenote-Taker2, we identified 1,880 putative viral contigs (VCs) from the metavirome, and an additional 785 VCs from the whole metagenome, including 242 pruned sequences with flanking host regions. These putative VCs were pooled and further clustered, resulting in a set of 2,171 VCs, of which 330 were present in both the metagenome and metavirome datasets. Most VCs (1,499) were only assembled from the metavirome, while 342 were only found in the metagenome. By integrating metagenomics and metaviromics, we could infer bacteriophage lysogeny by the presence of bacterial flanking regions. VCs were defined as integration-capable (IC-VCs) if their cluster included a metagenome-derived VC with pruned flanking regions,[84] or a virome-derived VC that could be clustered within a non-viral contig (296 of 2,171 VCs). These 296 IC-VCs were linked to 309 non-viral contigs representing their host genomes. Of these IC-VCs, 105 were only identified from the metagenome and were thus likely

inactive (i.e. prophages), while the remaining 191 were also present in the metavirome, suggesting simultaneous prophage activation as temperate phages. The other 1,875 VCs are likely primarily composed of free phages but could also include pseudolysogenic phages; here, we define these phages as non-integrated VCs (NI-VCs). While NI-VCs may be capable of phage integration, they were not detected as prophages in our study.

Viral contigs were assessed for quality using CheckV and annotated using Demovir and vContact2.[95,282,283] 94.5% of VCs (2,051/2,171) could be annotated at the viral order level (Table 2) and primarily represented the order *Caudovirales* (n = 1,892). The remaining VCs were mainly *Microviridae* and *Circoviridae*, which were predominantly identified from metavirome sequencing. Non-viral contigs in the metagenome (n = 39,735) were taxonomically annotated and subsequently used for phage-host analysis. High-quality sequencing reads across all three datasets were mapped to the set of viral and non-viral contigs; counts were normalized for contig length, resulting in a normalized relative abundance (NRA). In all, 86.4-98.3% of metagenome reads, 67.3-97.6% of transcriptome reads, and 88.9-98.2% of metavirome reads mapped to a contig; 4.74-11.8%, 0.96-2.52%, and 52.9-91.8% of reads mapped to viral contigs, respectively. Metagenome and metaviromic datasets were further filtered for a minimum breadth of coverage of 75% with >=1 read per bp for subsequent analysis.

VCs are plotted in Figure 1A by viral order, quality, contig length, maximum observed NRA in the metavirome, and dataset identified (i.e. metavirome, metagenome, or both). VCs identified in both the metavirome and metagenome were more likely to be longer (p = 6.52e-11), abundant (p < 2e-16), and less likely to be low-quality (p < 2e-16). Thus, co-presence in both datasets could be used as a marker for high-quality VCs. *Caudovirales* contigs were significantly longer than those belonging to other viral orders (mean length of 24.8 kb vs. 5.3 kb, p < 2e-16 by

Wilcoxon rank sum test) which reflects the smaller genome sizes of *Microviridae* and *Circoviridae* species; this difference was even greater when only including complete and high-quality VCs (48.9 kb vs. 4.5 kb, $p < 2e\text{-}16$). Lower-quality viral contigs were also significantly more likely to be shorter. As viral databases continue to grow, we recommend including the type of source dataset (i.e. VLP-enriched vs. whole metagenome) as key annotations alongside taxonomy and quality to help evaluate the influence of sequencing methodologies on human virome profiles.

3.6.3. Viral contigs across metagenomic, metaviromic, and metatranscriptomic datasets

Viral contigs represented 75.1% of the metavirome (by NRA), 4.85% of the metagenome, and 1.90% of the metatranscriptome (Figure 1B). The 2.5-fold decrease from the metagenome to the metatranscriptome suggests that the virome may be less transcriptionally active than the bacteriome, consistent with hypotheses of a predominantly temperate phage-host dynamic in the intestinal microbiome.[34,190]

While only 15.8% (342/2,171) of VCs were identified from both the metagenome and metavirome, these VCs were highly abundant, representing an average of 36.4%, 35.6%, and 73.1% of the observed viral community in the metagenomic, metatranscriptomic, and metaviromic datasets by NRA, respectively. The metagenome also contained metavirome-derived VCs that were not identified from the whole metagenomic assemblies. This finding may be due to improved assembly of viral contigs in VLP-enriched samples (with greater sequencing depth per VC), and due to different thresholds employed by Cenote-Taker2 when identifying viral assemblies in whole metagenomes as compared to VLP-enriched metaviromes, the former being more stringent to reduce false positives. It remains pertinent to consider the impact of bioinformatic tools and approaches in the identification of viral sequences,[8,324] where multiomic approaches could be utilized to corroborate and contextualize viral contig identification.

Viral contigs of the order *Caudovirales* represented the most abundant viruses of the metagenome and metatranscriptome while other viral taxa, primarily *Microviridae*, were enriched in the metavirome. Random displacement amplification, a common step employed in virome sequencing efforts, including this study, has been reported to introduce positive bias for small, circular ssDNA viruses such as *Microviridae*.[53,325] However, the patterns of ssDNA virus enrichment in our dataset appear to be participant-specific: all virome samples from participant G were *Caudovirales*-dominant while all samples from participant F and H were *Microviridae*-dominant. We have also previously shown that this virome sequencing protocol does not interfere with relative quantification of a spike-in phage in mucosal-luminal interface samples.[257] Thus, the expansion of *Microviridae* in participants F and H is likely suggestive of a true relative enrichment of these viruses in comparison to participant G.

We also observed that metaviromic reads were more likely to map to complete or high-quality viral contigs compared to metagenomic or metatranscriptomic reads. This supports the use of CheckV annotations to infer viral genome completeness. Lower quality genomes may include incomplete prophages, chimeric assemblies, or low abundance phages that lacked sufficient coverage for full genome assembly. We thus chose to use these CheckV quality thresholds when examining viral gene transcription and prophage induction as described in subsequent sections.

Next, we assessed the virome's alpha-diversity as observed across multiomic datasets; read counts to VCs were rarefied for these calculations. The Chao1 index, which evaluates species richness, was significantly lower in the metatranscriptome ($p = 0.0015$, Wilcoxon rank sum test), suggesting that only a subset of viruses are transcriptionally active (Figure 1C). The similar Chao1 indices between metagenome and metavirome suggest that both methods capture a similar number of VCs per viral sequencing read, as has been previously reported.[55] However, a metagenome

sample would require a greater sequencing depth of 1-2 orders of magnitude to be comparable to metaviromic sequencing. The Shannon index and population variance show that the viral community as seen in the metavirome is less evenly distributed than in the metagenome ($p = 0.00012$) or metatranscriptome ($p = 0.0015$); the Shannon index also shows that the metatranscriptome population is less evenly distributed than the metagenome ($p = 0.0015$). The reduced evenness in these datasets may be due to prophage activation, virulent phage replication, or virulence factors, that enable select viruses to become highly abundant in the metatranscriptome and metavirome, including the *Microviridae*-dominance seen in select virome samples.

In all, we demonstrate that each sequencing method provides a distinct snapshot of the human gut virome. With the majority of virome studies utilizing either metagenomic or metaviromic approaches, it is important to recognize each technology's biases on the observed taxonomy, quality, and community dynamics of identified viral populations, and to utilize multiomic approaches where possible to capture the true complexity of these viral communities.

3.6.4. The virome at the colonic mucosal-luminal interface is distinct from the stool virome

Several studies have investigated differences between the luminal and mucosal bacteriome along the length of the gastrointestinal tract.[247,326,327] Different ecological pressures would therefore be expected to impact the intestinal viral community along both radial and longitudinal dimensions, though there have been very few efforts for multi-site, or spatial, characterization of the human virome. Here we provide a direct comparison between the proximal colon MLI, distal colon MLI, and stool (Figure 2A). Among the three participants, 24.6% to 40.4% of VCs were identified at all three sites, representing a shared viral community throughout an individual's colonic mucosa and lumen. With 32.7 to 61.0% of VCs unique to MLI samples and 9.6 to 23.6% of VCs unique to stool, multiple site sampling substantially expands an individual's known intestinal virome. Bray-

Curtis dissimilarities between these samples show the MLI samples of participants G and H clustering apart from the stool, while participant F, the only patient to have pancolitis, has all three samples clustering together (Figure 2B). Multi-site sampling allows for an increased spatial resolution to understand viral-host interactions, especially in conditions like inflammatory bowel disease.

Consistent with other studies demonstrating personalized microbiomes and viromes, inter-participant metagenome and metavirome diversity was significantly greater than between intra-participant sites (Figure 2C). Bray-Curtis dissimilarities of the non-viral metagenomic community demonstrated reduced distances as compared to the virome ($p \leq 0.009$) indicating that the virome is more individualized than the metagenome. Within the same patient, Bray-Curtis distances between the proximal and distal colon samples were smaller than those between the colonic mucosa and stool ($p = 0.037$), suggesting that despite heterogeneity in local inflammation (such as a non-inflamed proximal colon and an inflamed distal colon), the colonic mucosa provides a common niche that is distinct from the luminal, stool virome. We hypothesize that bacteriophage adherence to mucin, facilitated by interactions between mucin glycoproteins and phage capsid proteins, supports the development and persistence of a mucosa-associated virome.[232,234] Our data also demonstrated a trend for greater species richness (i.e. Chao1 index) at both MLI sites in comparison to stool in each participant. Larger studies are required to further investigate these observations and to assess whether these findings extend beyond pediatric subjects with inflammatory bowel disease. Characterizing the mucosa-associated virome will expand the known human virome while enabling the study of microbial communities that are more likely to interact with human health and intestinal diseases.

3.6.5. Viral contig transcription and prophage activation in the metatranscriptome

A total of 59,217 open reading frames (ORFs) were predicted across the set of 2,171 VCs, which were further clustered into 28,112 viral genes and annotated using a combination of five protein family databases: two viral orthologous group databases VOGDB and pVOG,[121,122] and three general-purpose databases PFAM, KEGG, and TIGRFAM.[123,124,318] This approach annotated 9,646 (34.3%) viral genes, with PFAM annotating the most viral genes (Supplementary Figure 3A). However, VOG and pVOG annotations were more effective at annotating abundant viral genes detected in the metavirome, supporting the use of more tailored viral databases (Supplementary Figure 3B). We thus assigned viral functions in the order of the databases listed above, with VOGDB being the primary database used in this study.

To assess viral gene transcription, we focused on the subset of complete and high-quality VCs (580/2,171). We classified these VCs as "present" in a sample's metagenome and/or metavirome if detected with a minimum breadth of coverage of 75%, or as "transcriptionally active" (TA) if they had a minimum NRA >=0.0001% in the sample's metatranscriptome. These TA-VCs, highlighted in Figure 3A, were more likely ($p = 0.016$) to be present in both metagenome and metavirome datasets (64.8%) than non-transcriptionally active VCs (17.2%). We hypothesize that TA-VCs which are present in all three multi-omic datasets reflect the host-dependent production process of free viral particles. TA-VCs were less likely to be derived solely from the metavirome, which could represent viral particles without available hosts including transient, luminal VLPs.

Most TA-VCs were *Caudovirales* (96.0%, compared to 64.2% of non-transcriptionally active VCs, $p = 0.018$) with only a small minority of *Microviridae* detected in the metatranscriptome. This observation could suggest that the non-*Caudovirales* viruses are more likely to be transient while also reflecting the sequencing bias of small circular genomes in

commonly used metavirome sequencing efforts, though future multiomic studies with greater sample sizes are required to further explore these findings TA-VCs were also more likely to be observed with flanking host regions (47.0% vs. 7.5%, p = 0.0095) and contain an integrase (62.4% vs. 27.8%, p = 0.016); however, the latter observation was no longer significant when excluding non-*Caudovirales* VCs, which rarely contain an integrase.

Using metagenomic, metaviromic, and metatranscriptomic sequencing data, we plotted virome transcription profiles of each sample (Figure 3B). The NRA of viral contigs in the metatranscriptome was most correlated with their NRA in the metagenome (Pearson correlation = 0.395, p = 3.9e-12). The NRA of viral contigs in the metavirome was also positively correlated with their NRA in the metatranscriptome (Pearson correlation = 0.307, p = 1.1e-7), suggesting that observed increases in viral transcription may translate to increased VLP production. The NRA of VCs in the metavirome was also correlated with their NRA in the metagenome (Pearson correlation = 0.136, p = 0.021); the relatively weaker correlation between the metagenome and metavirome once again demonstrates the differences between whole metagenome and VLP-enriched sequencing approaches. The similarity between colonic sites is further visualized in Supplementary Figure 3C, reflecting the inter-participant variability and intra-participant stability of the virome as described in the previous section. More samples are required to investigate differences along the gastrointestinal tract and to compare viromes between subjects with and without IBD, while longitudinal studies will provide further insight into temporal variations in the virome's transcriptomic activity.

Within the subset of complete and high-quality TA-VCs, we also examined the metatranscriptome dataset at the feature level. The most frequently occurring genes are shown in Figure 3C, led by repressor protein cI (n = 227), integrase (207), and packaging protein 1 (148).

Gene transcription varied from 39% to 95%, with several genome processing genes (DNA helicase uvsW, 89%; integrase, 84%; reverse transcriptase 82%) showing higher transcription than structural proteins (major capsid proteins, 49-73%; baseplate proteins, 45-57%). While inferring population-level trends remains difficult given significant viral heterogeneity, we hypothesize that this pattern reflects the increased presence of early-stage bacteriophage genes, many of which are supportive of phage lysogeny, as compared to late-stage genes. The greater transcription of repressor protein cI (81%) compared to the anti-repressor protein ant (49%) may also provide further evidence of an overall temperate viral community.

3.6.6. Two abundant crAss-like phages identified at the MLI

Two circular *Caudovirales* VCs, V014264 (97.91 kb) and V016904 (99.54 kb), had high similarity to existing delta crAss-like phages: ERR844016_ms_1 and ERR844030_ms respectively.[164] The comparative phage genome tool VIRIDIC showed that each VC had 95.3% and 89.9% identity with their aligned crAss-like phage, 15.3% with each other, and less than 2% with the first isolated crAssphage.[154,328] These were the only two VCs in our dataset that contained all three markers (portal, TerL, major capsid protein) utilized in a recent profiling of crAss-like phages.[329] While crAss-like phages have been observed in up to 90% of metavirome samples,[154] the rarity of these phages in our IBD-only dataset could reflect a recent study which showed a depletion of crAss-like viruses in patients with inflammatory bowel disease.[329] The prevalence and diversity of crAss-like phages also appear to increase with age;[330] here, we profiled adolescents, a rarely studied demographic within virome studies. A larger dataset is required to better capture the prevalence of crAss-like phages in the human colonic mucosa, including in the context of human disease.

Both crAss-like phages were highly abundant in the proximal and distal colon metaviromes of participant G (Supplementary Figure 4), with V014264 being the third most abundant VC (NRA of 4.82%) in the proximal colon. We did not detect any single-nucleotide polymorphisms in either VC between the two sites.[319] These crAss-like phages were detected in the metagenome and were among the subset of complete and high-quality, transcriptionally-active VCs (highlighted in Supplementary Figure 3C). Both crAss-like phages were also present at very low abundance (< 2 x 10^{-4} %) in the distal colon metavirome of patient H.

While abundant in MLI metaviromes, both crAss-like phages were nearly undetectable in the stool metavirome of participant G (2.6 x 10^{-6} % for V014264; 1.4 x 10^{-7} % for V016904). These reads were below our filtering thresholds and exceeded a million-fold decrease from the MLI samples. At a more conventional sequencing depth, these crAss-like phages would be effectively undetectable by stool sampling. Both VCs were among the 61.0% of MLI-specific VCs detected in participant G (Figure 2A), demonstrating the importance of distinct viral niches across our complex biogeography.

Multiomic coverage maps of V014264 and V016904 are shown in Figure 4 and Supplementary Figure 5. Functional annotation of the crAss-like phages was performed as described above using Prokka and several viral and general-purpose databases, yet this only annotated 11.9% and 2.4% of ORFs on V014264 and V016904, respectively (Supplementary Figure 6A). Utilizing a curated set of crAss-like phage protein families described in Yutin *et al.* (2021), we were able to increase the proportion of annotated reads to 39.6% and 32.1%, albeit including limited annotations such as “uncharacterized protein of delta crassfamily delta group phages.” Metatranscriptomic sequencing reads, when mapped to forward and reverse strands, corroborated the orientation of the predicted crAss-like phage genes and enabled gene-level

analysis (Figure 4). For V014624, 55.4% of the contig was transcribed in the proximal colon while 70.7% was transcribed in the distal colon (i.e. with a minimum sequencing depth of 1). Between the two sites, metavirome map depths were highly correlated (Spearman correlation = 0.983), compared to 0.585 in the metagenome and 0.504 in the metatranscriptome, suggesting that the transcription profiles of these crAss-like phages were similar in both the non-inflamed proximal colon and inflamed distal colon.

Considering all transcribed genes, V016904 had a higher mean gene transcription ratio (transcriptome / metagenome) than V014264 (0.62 vs. 0.23, p = 4.3e-15), while VC16904 was significantly more transcribed in the proximal colon than the distal colon (0.80 vs. 0.44, p = 0.00078) (Supplementary Figure 6B). The most transcribed gene clusters are highlighted in Supplementary Figure 6C, with a GIY-YIG endonuclease and major capsid protein among the highest transcribed crAss-like phage genes in this dataset. We also note the presence of highly transcribed regions without annotated ORFs (i.e. a 500 bp region at ~44.5 kb) on V014264, that could be involved in some form of regulatory role to support the stable relationship between crAss-like phages and their hosts.[162,331] Despite significant improvement with targeted databases, virome studies continue to be limited by poor viral genome annotation. As these databases continue to grow with our understanding of bacteriophage function, metatranscriptomics has great potential in capturing the *in vivo* functional capabilities of our human viromes.

3.6.7. Host-phage prediction and transcription

While there are multiple methods to predict bacterial host-phage relationships, we utilized two approaches that leveraged our paired metavirome and whole metagenomic sequencing (Figure 5). First, we used a probabilistic modelling approach (WiSH) to assign VCs to their most likely host within our set of whole metagenome-assembled, non-viral contigs. After applying a false-

discovery rate adjusted *p*-value threshold of 10^{-5}, 855 (39.4%) VCs were assigned a host contig. Nearly all VCs assigned a host were *Caudovirales* (Figure 5A). Few *Circoviridae* (2/86) and *Microviridae* (1/53) were assigned hosts. While ssDNA viruses have shorter genomes (*Microviridae* and *Circoviridae* have a mean length of 5.2 kb and 4 kb in our dataset, respectively) and there is a correlation between VC length and host assignment, short genomes do not preclude host assignment: 25.8% of all our host-assigned VCs are less than 6 kb. Moreover, ssDNA viruses have been reported to be less adaptive to their host genomes and have a higher mutation rate,[36,45,66] which could limit our ability to assign their hosts. Additionally, ssDNA virus enrichment due to random displacement amplification could result in a lack of host-assignment if its corresponding bacterial genome was below our sequencing or contig assembly detection threshold. Lastly, three *Phycodnaviridae* VCs were assigned a bacterial host, which could represent a false prediction or an incorrect annotation, as *Phycodnaviridae* are known to infect algae rather than bacteria.

VCs identified in the metagenome were also more likely to be assigned a bacterial host than those identified solely in the metavirome (Figure 5A). We hypothesize that whole metagenome sequencing enables the identification of prophages and intracellular phage particles within their host bacteria that may be missed by metavirome sequencing, while VLP-enriched methods may capture transient VLPs that are unrelated to the resident bacteriome. Figure 5C shows the breakdown of host-viral pairs by the predicted VC family and bacterial host contig order, with most VCs assigned to bacteria in the order Clostridiales.

Our second approach for studying virome-bacteriome interactions involved our subset of IC-VCs with flanking host regions. These 296 IC-VCs were associated with 309 unique host contigs, with 320 unique VC-host pairs (23 VCs were present in 2 or 3 host contigs; 11 host contigs contained 2 VCs). Like WiSH-assigned VCs, IC-VCs were nearly all *Caudovirales* and more likely

to be derived from the metagenome or both the metagenome and metavirome (Figure 5B). IC-VC hosts were also most commonly of the bacterial order Clostridiales, though compared to WIsH assignments, Bacteroidales hosts were proportionally increased from 7.3% to 18.1% (Figure 5D).

Spearman correlations were calculated between the NRA of each VC in the metavirome and the NRA of its predicted host in the metagenome across the nine paired datasets (Figure 5E, 5F). The NRA of both WiSH-assigned VC-host pairs and IC-VC-host pairs were significantly ($p < 2.2e-16$) more likely to be positively correlated in comparison to the NRAs of unmatched pairs, with median Spearman correlations of 0.325 and 0.713, respectively. Viral contig transcriptional enrichment (i.e. NRA of contig in the metatranscriptome/metagenome) also positively correlated with host contig transcriptional enrichment (Figure 5G, 5H). These results are indicative of phage-host coexistence in the mammalian gut mucosa, and suggests that the transkingdom equilibrium as demonstrated between ΦcrAss001 and its host *B. intestinalis* may also be reflected at the community level.[162] The gut mucosa is thought to enable a heterogenous distribution of phages and their bacterial hosts that supports their co-existance.[193] In our small IBD cohort, this apparent co-existence is maintained despite the presence of local host inflammation, though the impact of inflammation on virome-bacteriome interactions at the MLI requires further study.

3.6.8. The presence of an integrase is a weak predictor for observed viral lysogeny

Phage integrases are viral enzymes that facilitate site-specific recombination, allowing for the integration of a viral genome into its host. Many model bacteriophages, such as *Escherichia virus Lambda*, require an integrase for lysogeny.[332,333] Additionally, integrases tend to be relatively prevalent in virome datasets including this study, where integrases were the most common viral gene annotation overall (second-most common in complete and high-quality VCs). Integrases have

therefore been commonly employed as prophage markers and as indicators of a temperate lifestyle.[36,45,57,132]

Among our 2,171 VCs, we identified 603 integrase genes (591 VOG00035 "Integrase", 10 pVOG0275 "integrase", 1 VOG11667 "DDE-type integrase/transposase/recombinase", and 1 PF13495 "Phage integrase, N-terminal SAM-like domain") on 497 VCs. All identified PFAM-annotated integrases (PF00589), commonly used for integrase identification, were encompassed within the VOG00035 integrases. These integrases were present on 25.8% of *Caudovirales* VCs (489/1892) and absent in all other viral orders. Longer and higher-quality VCs were more likely to contain an integrase gene. We thus focused on the subset of complete and high-quality *Caudovirales* VCs, of which 54.9% (219/399) contained an integrase.

IC-VCs were more likely to contain an integrase (69.6%, 71/102) than non-integrated VCs (49.8%, 148/297); this finding was independent of contig length. While this result is expected, the majority of integrase-containing VCs were only detected as free phages, and thus suggests that many NI-VCs may be temperate phages observed only in their lytic cycle. We therefore investigated the transcription of all common gene families including integrases (present in 10% or more of high-quality or complete *Caudovirales* VCs) for their association with observed phage integration (Figure 6).

Repressor protein cI (VOG06177) and integrase (VOG00035), both known to be involved in phage lysogeny, were by far the most common gene annotations. However, both were similarly present in IC-VCs and NI-VCs when normalized by contig length and had modest integration likelihood ratios of 1.11 and 1.31, respectively (Figure 6B). In comparison, the highest observed integration likelihood ratios were for ParB-like nuclease domain protein (2.77), packaging protein 1 (2.50), and insertion sequence IS21 putative ATP-binding protein (2.29). We thus suggest that

the presence of an integrase gene alone is a weak predictor of a VC lysogeny, as it is neither a sensitive nor specific indicator of phage integration. Existing databases may limit our capacity to annotate the diverse array of enzymes that enable phage recombination, which may include various integrases, transposases, and recombinases.[333] Our own annotation sensitivity would have been further reduced if we had relied only on the commonly used PFAM integrase family PF00389, which missed 15.9% of integrase genes identified with VOG00035. Successful identification of flanking host regions is also dependent on accurate metagenomic contig assembly and subsequent annotation, and thus limits the utility of using these regions as a gold standard for lysogeny.

Given that the presence of integrase and repressor protein merely infers lysogenic potential but may not clearly differentiate phages in their lytic or lysogenic cycle, we hypothesized that transcription would be a better indicator of phage lysogeny. We thus examined the transcription of these common viral genes (Figure 6C). Integrases and repressor proteins were both highly transcribed, which supports the overall lysogenic nature of the intestinal virome community. Both integrases and repressor protein cI were slightly more transcribed in VCs with flanking host regions (Figure 6C), increasing from 81.4 to 83.9% and 81.8% to 87.3%, respectively, which supports the hypothesis that some temperate prophages are being detected during a lytic cycle. Packaging protein 1, putative nuclease p44, and two uncharacterized proteins were most likely to be transcribed in IC-VCs (+14.3-29.3%), while several terminases and major capsid protein genes were more transcribed in NI-VCs, consistent with these phages being in their lytic phase. Improved viral annotation and further phage studies will assist with the prediction of phage lifestyle, especially since phage-host relationships outside of model organisms remain poorly understood. Notably, crAss-like phages display features of temperate phages including high-level long-term persistence without plaque formation in its host, yet both ΦcrAss001 and ΦcrAss002 do not contain

lysogeny modules.[162,167] We also did not detect integrases within crAss-like phages V014264 and V016904, though integrases were noted within a larger subset of crAss-like phages.[165] Ultimately, phage integration is likely among other lysogenic mechanisms including episomes, pseudolysogeny, phage carrier states, and other uncharacterized means, that are able to sustain stable viral populations.[151]

We thus demonstrate the potential for integrating metagenomics, metaviromics, and metatranscriptomics to evaluate viral lysogeny in a cohort of pediatric, IBD subjects. Through metatranscriptomics, we revealed that lysogenic proteins are highly transcribed at the community level, while the presence and transcription of these proteins could be used to predict phage lysogeny at the genome level. We recommend against inferring phage lifestyle based on the presence of an integrase alone, and instead include additional factors such as integrase transcription, alternative integration-associated genes, and the presence of flanking host DNA. To help overcome limitations in viral gene databases, machine-learning classification tools and alignment-free *K*-mer frequency models are also being developed to assist in phage lifestyle prediction.[334,335]

3.6.9. Metatranscriptomic sequencing does not reveal a significant population of RNA phages

We also searched our metatranscriptomic data for RNA phages, which may be missed in DNA-based approaches for metagenomics and metaviromics. Predicted ORFs on metatransciptomic assemblies were searched against two protein profile databases: a set of RNA-dependent RNA polymerases (RdRps) and capsid genes curated for Cenote-Taker2, and a set of ssRNA marker genes that were used in a recent identification of 15,611 non-redundant environmental ssRNA phages.[83,307] Both MLI samples in participant F and G contained an RdRp originally identified in a dsRNA human picobirnavirus.[336] There were no other RdRps or ssRNA phage markers detected in our metatranscriptomic dataset. Thus, like prior studies, we report a

scarcity of RNA phages in the human virome, though these observations remain hampered by limited RNA phage databases and a small, IBD cohort with no control subjects.[235] Additionally, our metatranscriptome protocol, which like our metagenome protocol involves pelleting bacterial cells prior to DNA/RNA extraction, may exclude viral VLPs that remain in the supernatant.[257] Alternative methods to capture RNA-based VLPs would be required to assess their true contribution to the mucosa-associated virome.

3.6.10. The impact of sequencing depth on virus discovery

Virome sequencing has substantially improved over the past decade, from read counts in the 10,000s to over ten million paired-end (PE) reads. To investigate the impact of sequencing depth on virome discovery, we subsampled our metaviromic sequencing reads at increments spanning 10 k and 100 million PE reads to examine how sequencing depth affects VC identification (Figure 7). Empirically, the yield for new contigs per million bases dropped from 38.6 VCs at 1 million reads to 9.4 VCs at 10 million reads to 2.3 VCs at 100 million reads (Figure 7A). Mean contig length peaked at 20 million PE reads (18.1 kb) but was overall stable between 17.1 – 18.1 kb above 10 million reads (Figure 7B). Maximum contig length appeared to stabilize at around 10 million reads but did continue to increase as the sequencing depth reached 100 million reads. At all sequencing depths above 100,000 reads, VC quality remained remarkedly stable, with 8.1 to 10.3% of all VCs annotated as "complete" by CheckV (Figure 7C). Taxonomic annotations of VCs also remained relatively stable at sequencing depths greater than 2 million PE reads (Figure 7D). This study is one of the deepest sequenced datasets to date and allowed us to explore the impact of sequencing depth on virome identification. We show that after about 10-20 million PE reads, there were no further gains in mean contig length, viral contig quality, or contig annotation. Thus, given the existing bioinformatic tools currently available, it is unlikely that extensive sequencing depth

will ultimately result in a set of "complete genomes". As such, it may be more practical for virome researchers to design experiments to maximize the information obtained from a particular sample, as the "holy grail" of assembling all phages that may be present within a complex virome community is likely unfeasible. We also found that increasing the read depth leads to a decrease in the percent contigs which could have an annotation assigned. While shorter DNA lengths may contribute to this, it is equally likely that rarer, low-abundant viruses are less likely to be characterized and are thus more difficult to classify.

3.7. Conclusions

In summary, we present an in-depth, multiomic spatial analysis of the human virome in a small cohort of pediatric patients with ulcerative colitis. We demonstrated that the mucosal-luminal interface virome was distinct from stool and were able to isolate two highly abundant, mucosa-associated crAss-like phages, highlighting the importance of studying the virome across our complex host biogeography. We then developed practical approaches to leverage metagenomics, metaviromics, and metatranscriptomics to study the virome at the community and contig level. Because each technique yields a different viral subset, multiomic approaches help to corroborate and contextualize observations while revealing biases that are present in metagenome or metavirome only studies. Viral contigs present across all three datasets were more likely to represent higher-quality *Caudovirales* genomes. Many of our observations were consistent with an overall temperate virome, including reduced virome representation in the metatranscriptome, co-existence of predicted phage-host pairs, and high transcription of lysogeny-associated genes. We also assessed integrases and the presence of flanking host DNA regions and show that metatranscriptomic analysis improves our ability to infer phage lysogeny.

While these findings were described in a pediatric IBD cohort and may not extend to the general population, we describe techniques that advance our ability to characterize the human virome. Providing further biogeographic resolution of our virome and employing multiomic approaches are two key frontiers in understanding how the virome interacts with the bacteriome in the context of human health and disease.

3.8. Acknowledgments

The authors would like to acknowledge the patients and their families for their participation in our study. We also thank Ruth Singleton for her help in enrolling patients and assistance in collecting intestinal aspirate samples and Dr. Kendra Hodgkinson for her proofreading and review of this manuscript. The multiomic analyses presented herein were enabled in part by Compute Ontario, and the Digital Research Alliance of Canada (https://alliancecan.ca/en).

AY is supported by the Frederick Banting and Charles Best Canada Graduate Scholarships Doctoral Award from the Canadian Institutes of Health Research. DRM is supported in part through a Distinguished Clinical Chair in Pediatric IBD through the Faculty of Medicine, University of Ottawa. This work was supported by the Government of Canada through Genome Canada and the Ontario Genomics Institute under Grant OGI-149; the Canadian Institutes of Health Research under Grant ECD-144627; and the Ontario Ministry of Economic Development and Innovation under Project 13440. The funders had no role in study design, data collection and analysis, or preparation of the manuscript.

3.9. Supplemental Online Material

See separate attachment for supplemental figures, titles, and legends.

3.10. Figures and Tables

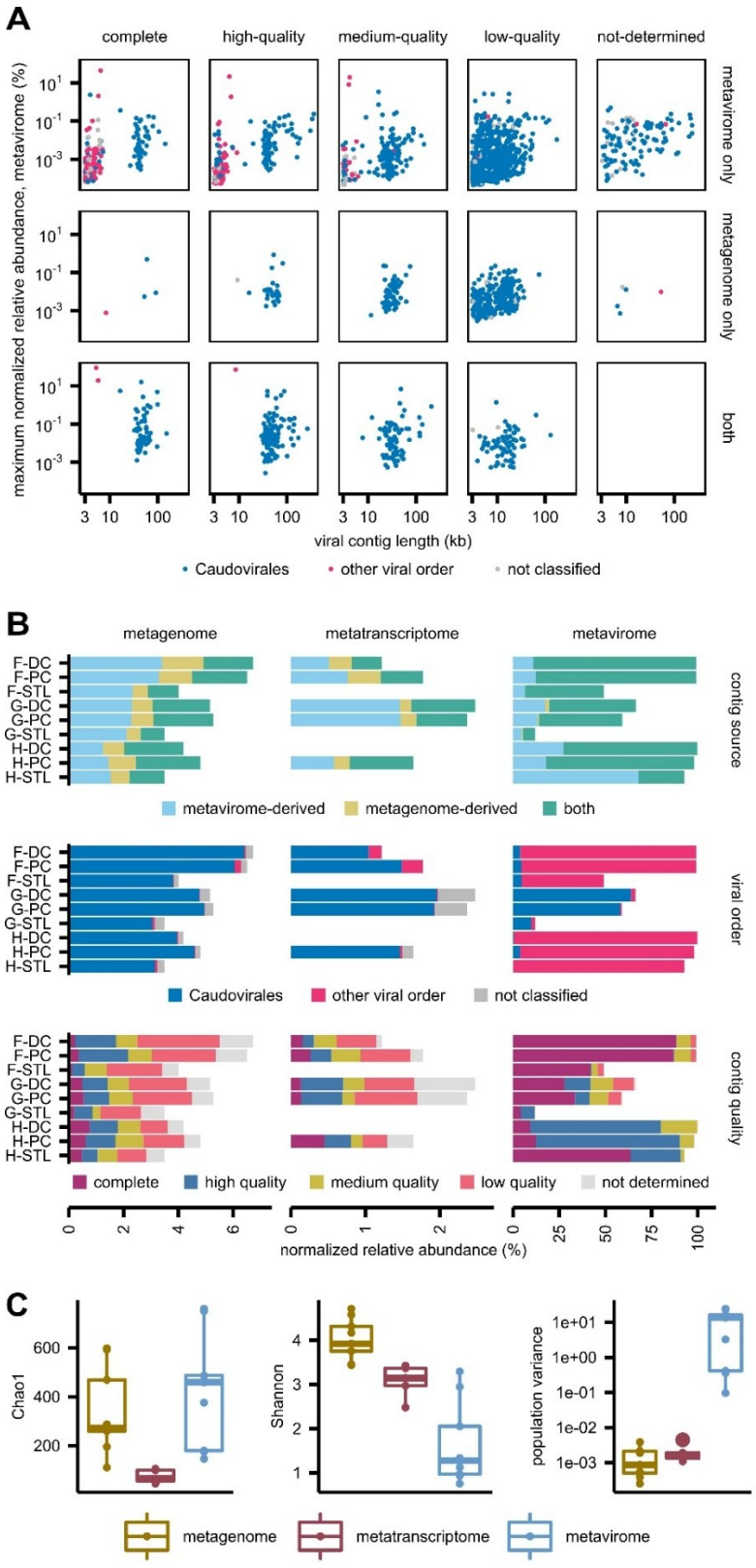


Figure 3-1. Viral contig annotation and alpha-diversity across metagenomic, metatranscriptome, and metaviromic datasets.

(A) 2,171 viral contigs (VCs) were identified across metaviromic and whole metagenomic assemblies and are plotted by: dataset of origin, viral contig quality, maximum normalized relative abundance (NRA) in the metavirome in any given sample, contig length, and viral order. (B) NRA of viral contigs across multiomic datasets for each participant's samples, annotated by source, taxonomy, and quality. (C) Boxplots of Chao1, Shannon diversity and population variance by multiomic dataset. * p-value < 0.05; ** p < 0.001. DC: distal colon, PC: proximal colon, STL: stool.

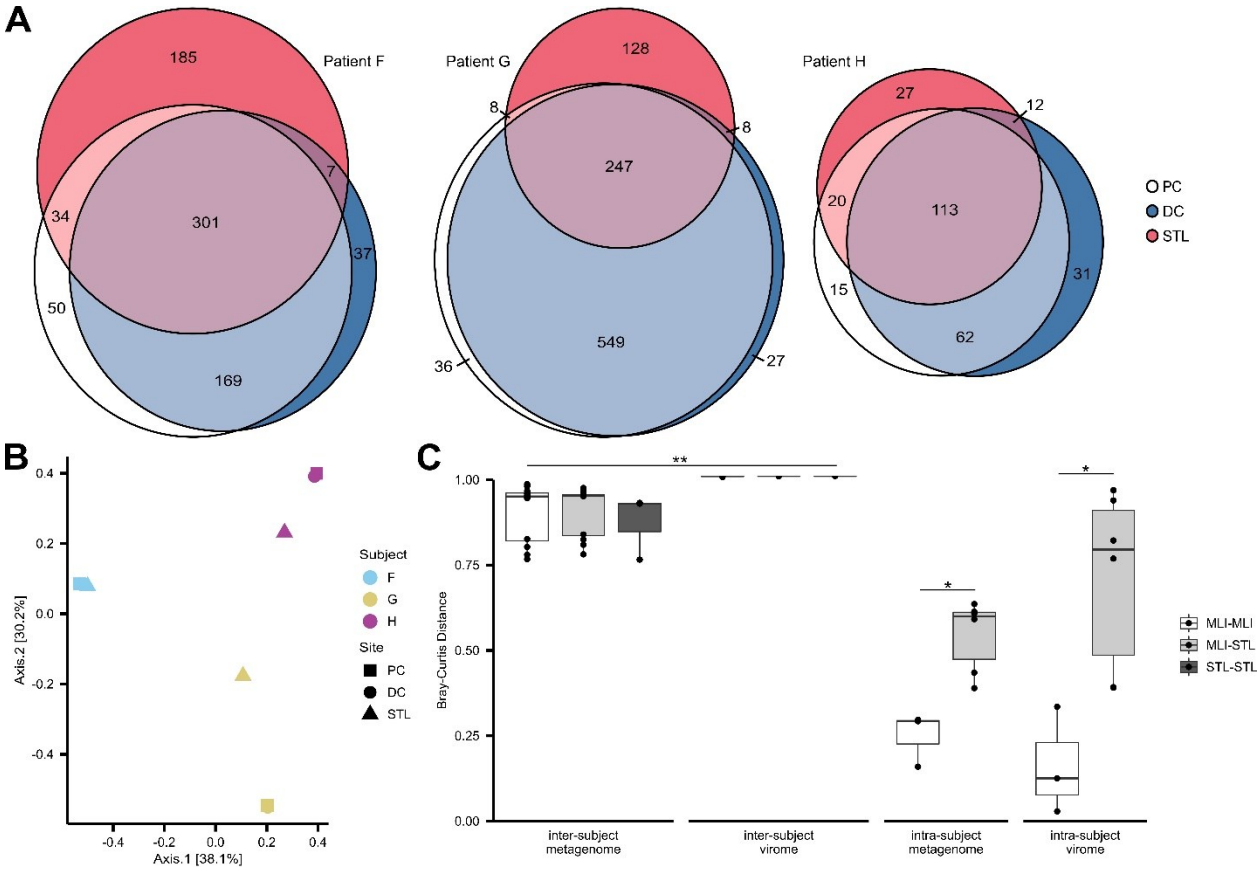


Figure 3-2. Beta diversity of colonic and stool metavirome communities.

(A) Euler diagrams showing the number of shared viral contigs (with a minimum of 75% of breadth of coverage) between sampling sites. (B) Principal coordinate analysis showing Bray-Curtis dissimilarities of metavirome communities. (C) Bray-Curtis dissimilarities of viral and non-viral metagenomic communities between MLI and STL samples, plotted by intra-individual and inter-individual comparisons. * p-value < 0.05. ** all inter-individual metagenome communities had a significantly lower Bray-Curtis distance than inter-individual metavirome communities ($p < 0.01$). DC: distal colon, PC: proximal colon, STL: stool, MLI: mucosal luminal interface.

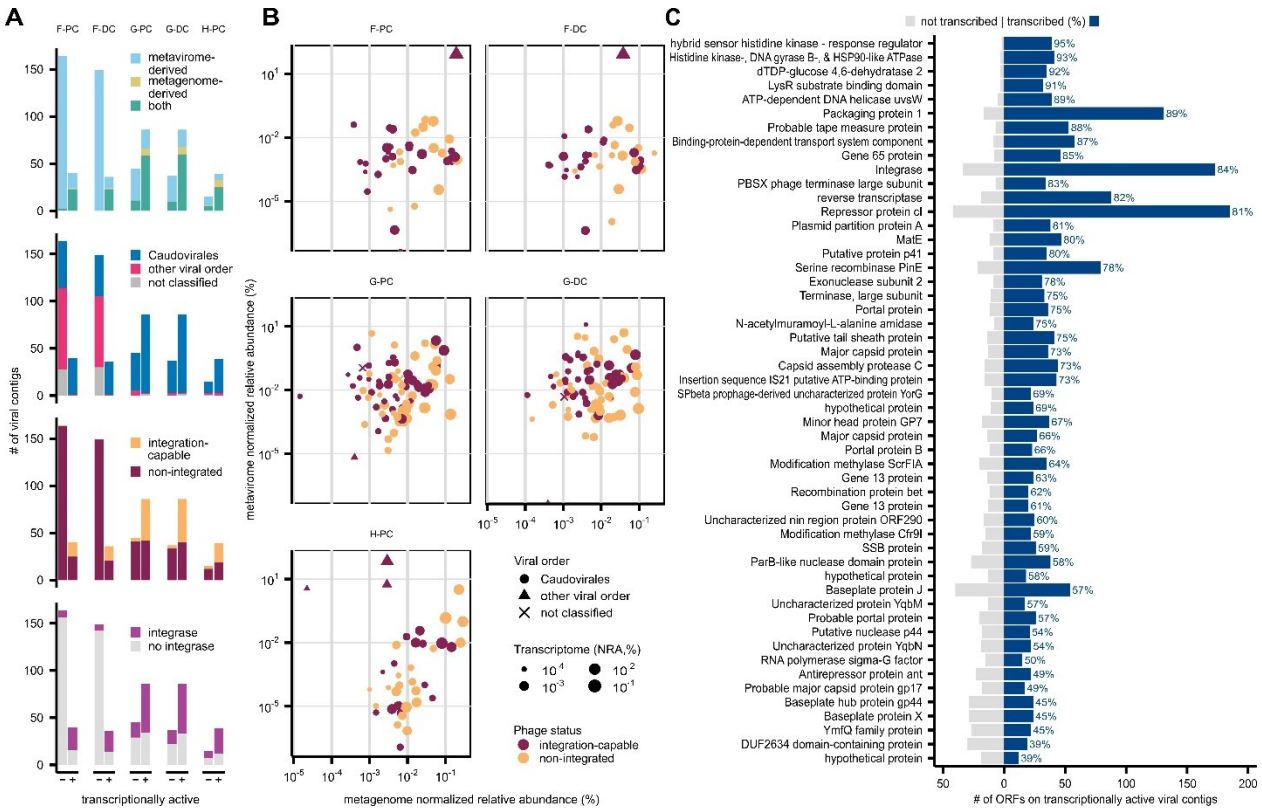


Figure 3-3. Transcriptionally active viromes of the colonic mucosal-luminal interface.

(A) Bar plot showing all complete and high-quality VCs by metatranscriptome activity (minimum abundance of 0.0001%), further annotated by viral contig source, taxonomy, phage integration, and integrase presence. (B) Scatter plots highlighting transcriptionally active VCs by their NRA in the metagenome, metavirome, and metatranscriptome. (C) Bar plot showing annotated ORFs of transcriptionally active VCs present in ≥30 instances (i.e. counted for every VC and every sample). Transcribed ORFs (i.e. present in that sample's transcriptome) are counted on the right side of the plot.

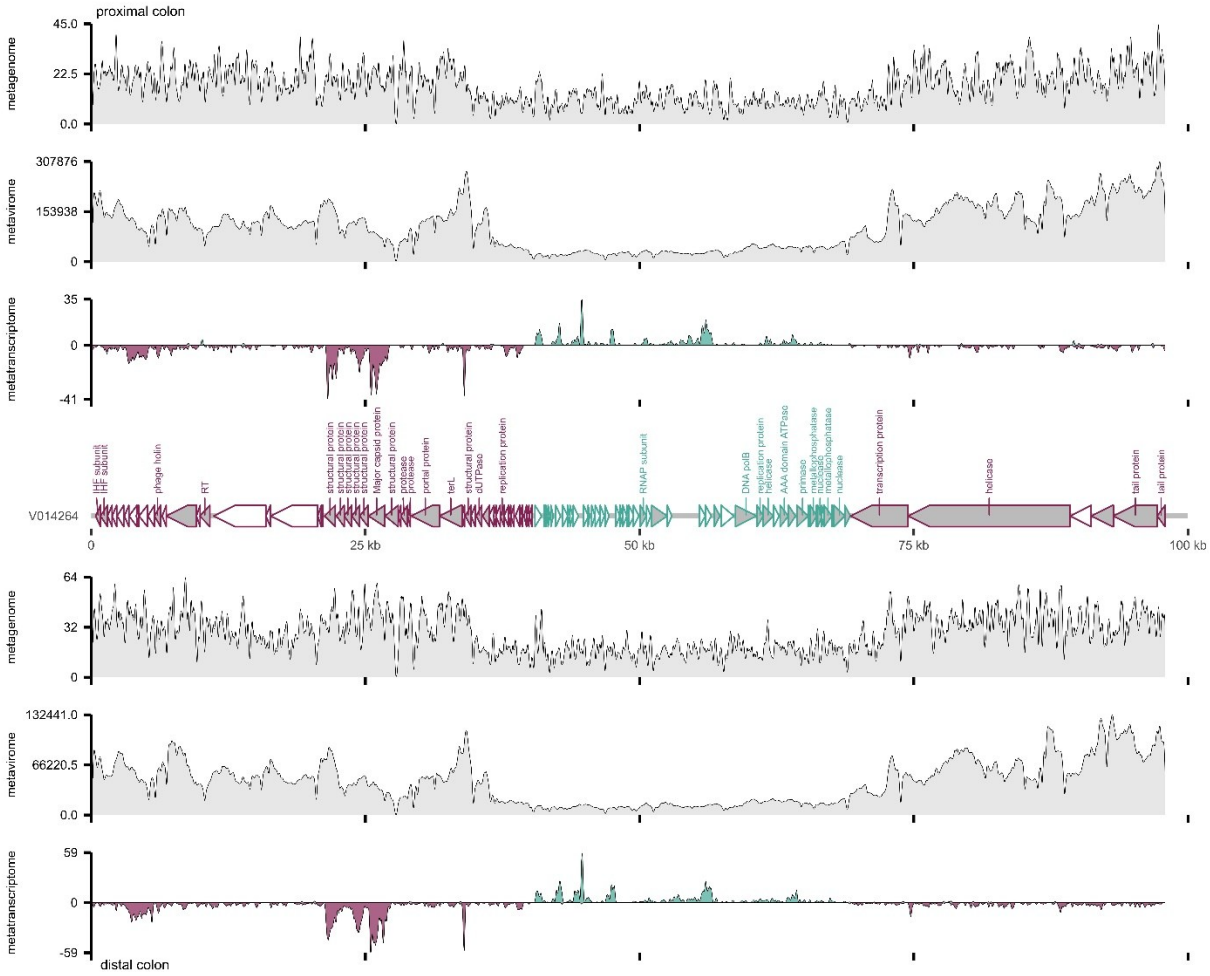


Figure 3-4. Multiomic sequencing map of V014624, a crAss-like phage identified in the colonic mucosal-luminal interface.

Open reading frames (ORFs) were predicted over the 97.9 kb genome and coloured by forward (green) or reverse (purple) orientation. Annotated ORFs are shaded in gray and labelled (excluding 'uncharacterized' or 'hypothetical' proteins). Metagenome, metavirome, and metatranscriptome sequencing depths at the proximal colon (top) and distal colon (below) are plotted using a 151 bp sliding window; metatranscriptome reads are also mapped by strand. Another crAss-like phage, V016904, is shown in Supplementary Figure 5.

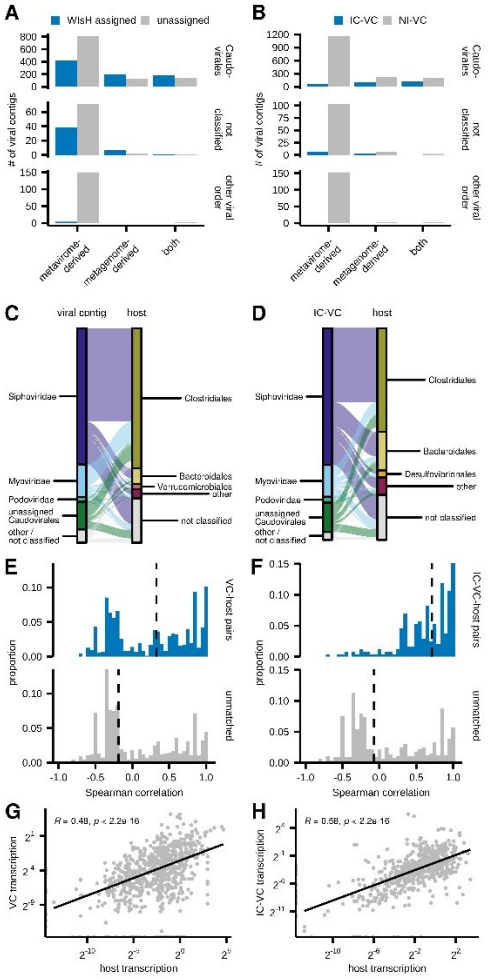


Figure 3-5. Phage-host prediction inferred from metavirome and whole metagenome assemblies.

Phage host relationships as calculated using WiSH (A,C,E,G) or inferred by the presence of non-viral flanking regions of VCs (B,D,F,H). (A-B) Bar plots of all VCs by source, viral order, and their host assignment status. (C-D) VC-host or IC-VC--host combinations by viral and bacterial order. (E-F) Spearman correlations of the NRAs of matched VC-host or IC-VC host pairs, as compared to unmatched pairs. The dashed line indicates the median. (G-H) Scatter plots of transcription (i.e. metatranscriptome/metagenome NRA), with the line and formula representing the linear model and Pearson correlation between VC and host transcription.

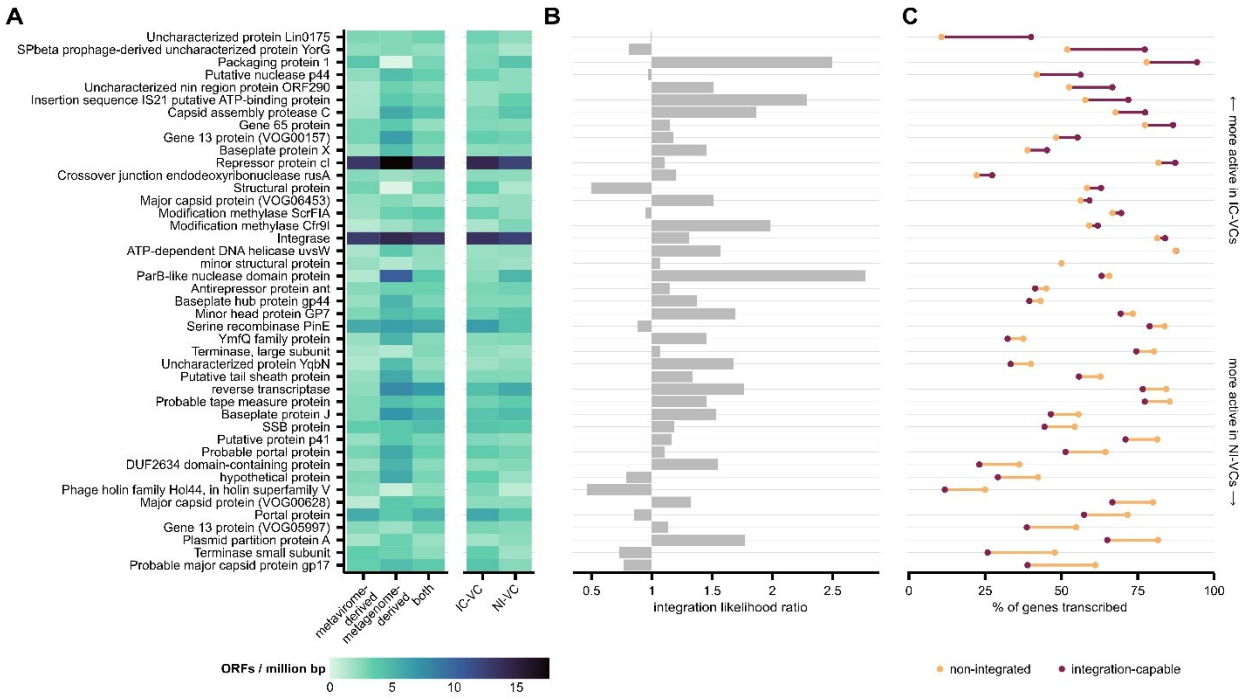


Figure 3-6. Presence and transcription of common viral ORFs including lysogenic proteins on *Caudovirales* VCs.

(A) The frequency of common viral genes (≥10%) present on complete and high-quality *Caudovirales* VCs based on contig source and between IC-VCs and NI-VCs. ORF counts were normalized by contig length. (B) Integration likelihood ratios of based on presence of each gene family. (C) Proportion of genes transcribed in IC-VCs and NI-VCs, sorted by difference.

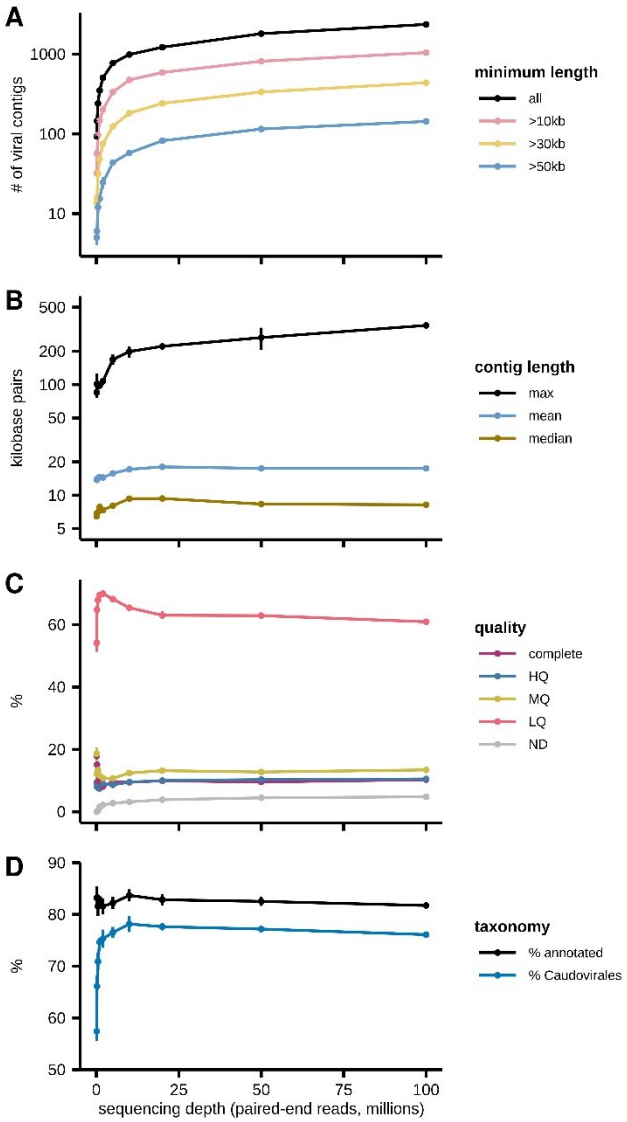


Figure 3-7. Effects of sequencing depth on viral contig identification.

The impact of sequencing depth on (A) number of VCs identified, (B) VC length, (C) contig quality, and (D) annotation. Error bars show standard deviation. HQ: high-quality; MD: medium-quality; LQ: low-quality; ND: not-determined.

Table 3-1: Participant and sample descriptions.

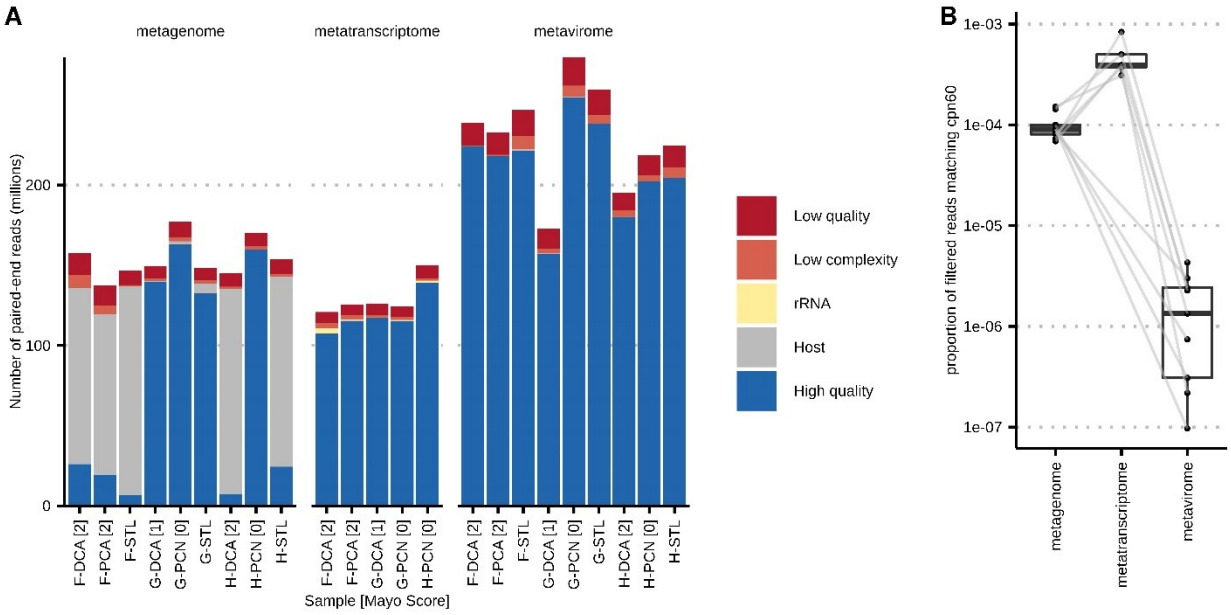
PC : proximal colon; DC : distal colon; STL: stool.

Participant	Sex	Age (years)	Paris Score	Site	Site Mucosal Inflammation / Stool description
F	Female	17.3	E4 S1	PC	Mayo UC – Grade 2
				DC	Mayo UC – Grade 2
				STL	Pre-scope (two days)
G	Male	14.5	E2 S0	PC	Mayo UC – Grade 0
				DC	Mayo UC – Grade 1
				STL	Post-scope (twenty days)
H	Male	16.1	E3 S0	PC	Mayo UC – Grade 0
				DC	Mayo UC – Grade 2
				STL	Pre-scope (one day)

Table 3-2: Viral genomes by viral family.

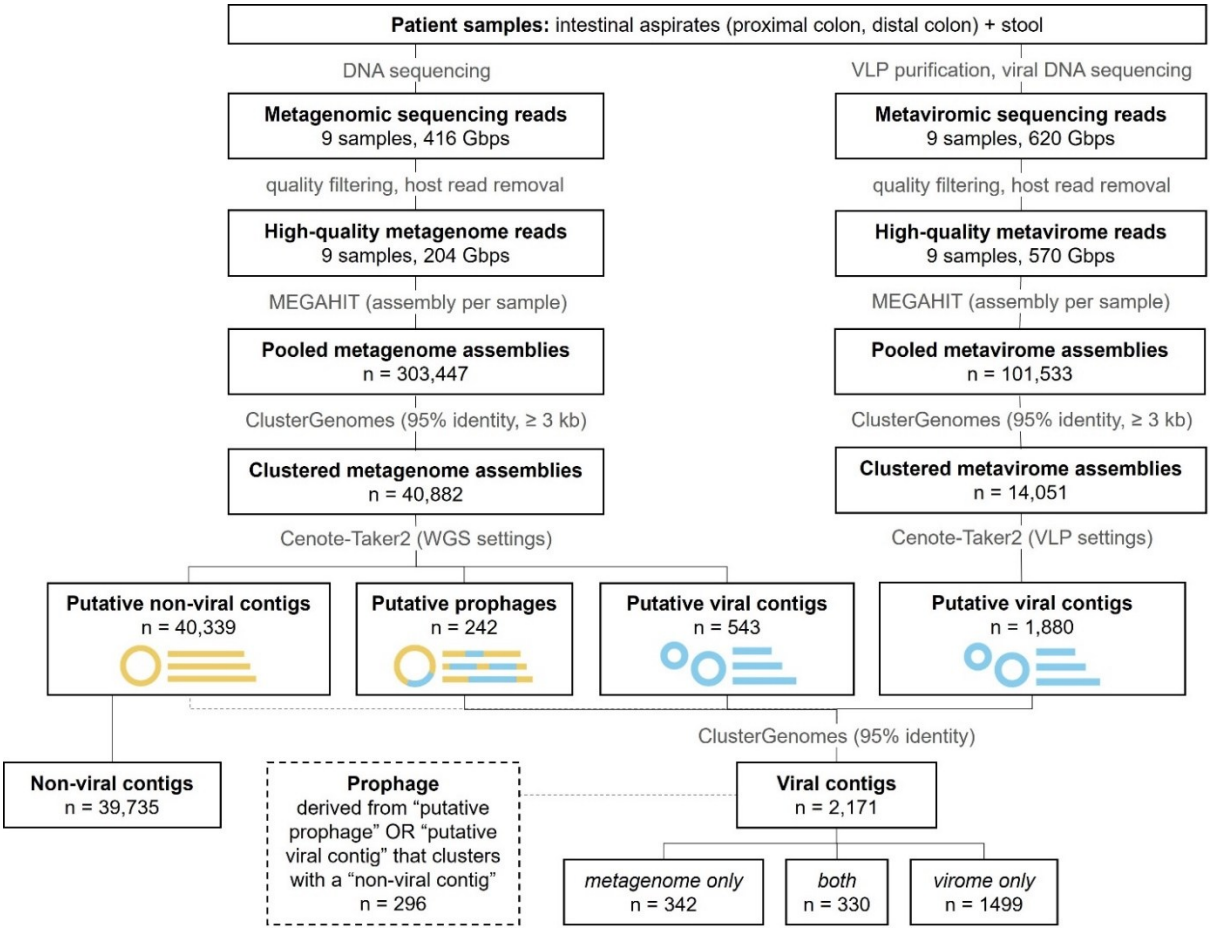
A total of 2,171 viral genomes were identified from metavirome and metagenome assemblies. The table below shows their annotations by source.

	Metavirome only	Metagenome only	Both
Caudovirales families			
Myoviridae	148	65	48
Podoviridae	37	5	12
Siphoviridae	910	214	242
Unassigned *Caudovirales*	138	45	23
Other viral families			
Anelloviridae	6	0	0
Circoviridae	86	0	0
Cruciviridae	5	0	0
Geminiviridae	2	0	0
Genomoviridae	1	0	0
Inoviridae	1	1	0
Microviridae	50	0	3
Mimiviridae	0	1	0
Nanoviridae	1	0	0
Phycodnaviridae	3	1	0
Unassigned viruses	109	9	2
TOTAL	1499	330	342



Supplementary Figure 1: Processing of raw metagenomic, metatranscriptomic, and metaviromic sequencing reads.

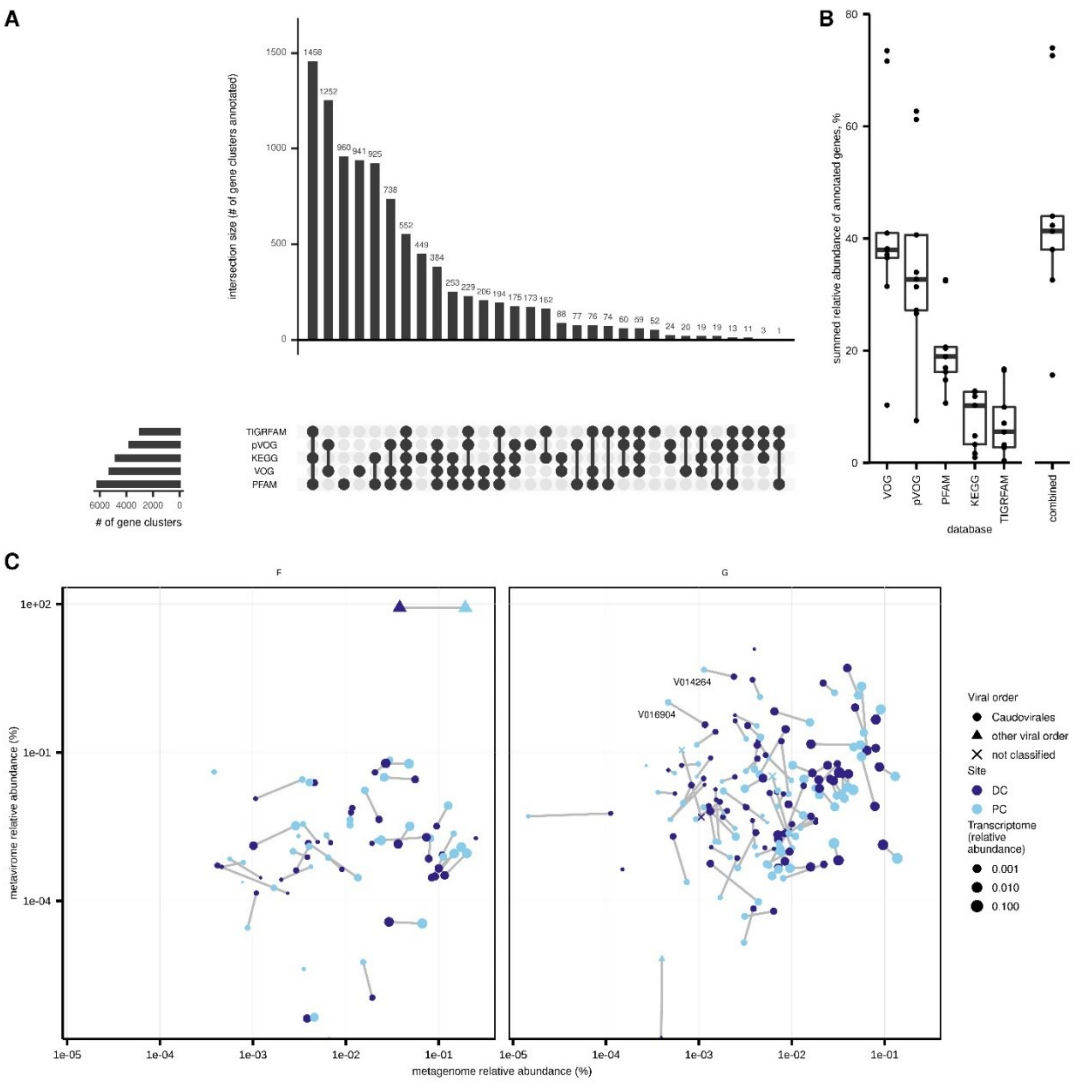
(A) Bar plot showing the filtering of raw sequencing reads based on quality, host contamination, low-complexity, and rRNA (metatranscriptome only) for each sample, which includes the participant, site (PC: proximal colon; DC: distal colon; STL: stool), and inflammation status (A: affected; N: non-inflamed). For aspirate samples, the Mayo endoscopic score of the collection site is noted in square brackets. (B) Boxplot showing presence of *cpn60* genes in high-quality multiomic sequencing datasets to assess bacterial contamination in the VLP-filtered metavirome as compared to the metagenome and metatranscriptome with matched samples connected with lines.



Supplementary Figure 2. Metagenomic and metaviromic sequencing pipelines for assembly and viral contig identification.

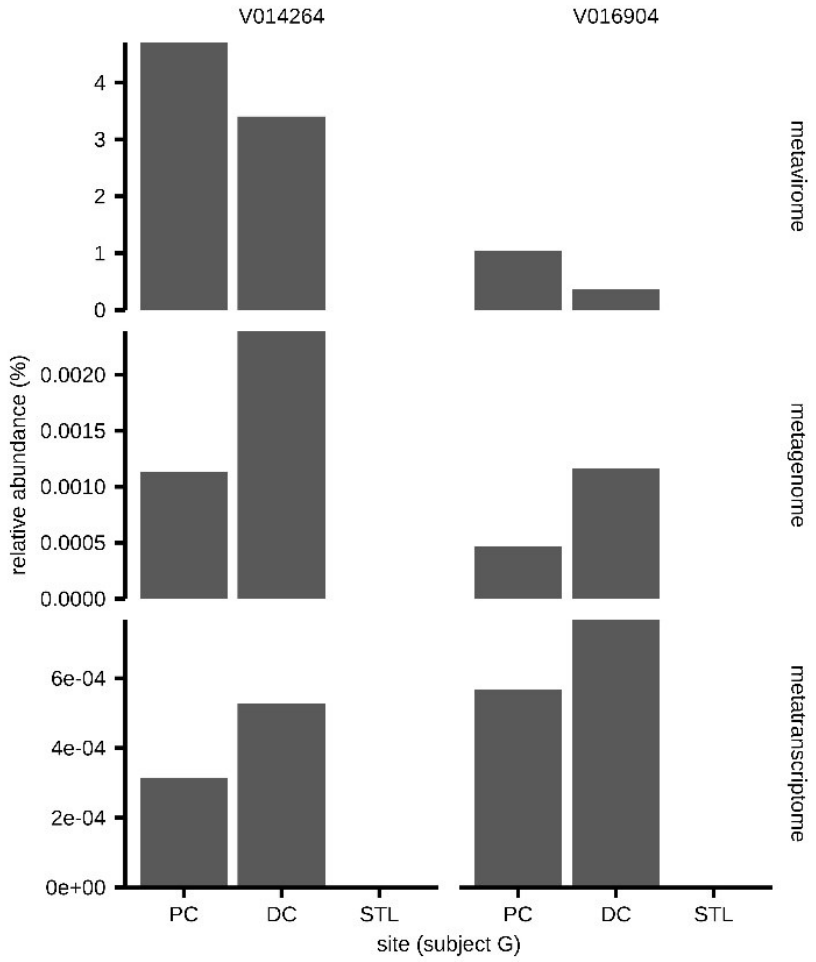
High-quality sequencing reads were assembled per sample using MEGAHIT, then pooled and clustered with ClusterGenomes. These clusters were then subjected to viral sequence identification with Cenote-Taker2, utilizing WGS and VLP oriented settings for the metagenome and metavirome clusters, respectively; metagenome assemblies that were not identified as viral were classified as putative non-viral contigs. Flanking regions of putative prophages were also considered separately as non-viral contigs. Putative viral contigs and prophages from metagenome and metavirome sequencing were further clustered resulting in 2,171 viral contigs; the numbers of

contigs isolated from the metagenome, metavirome, or both are specified. Additionally, viral contigs that were either derived from a putative prophage (i.e. with flanking host regions) or derived from a viral contig clustering with a non-viral contig (thus inferring flanking host regions) were identified as prophages, representing 296 of the 2,171 VCs.



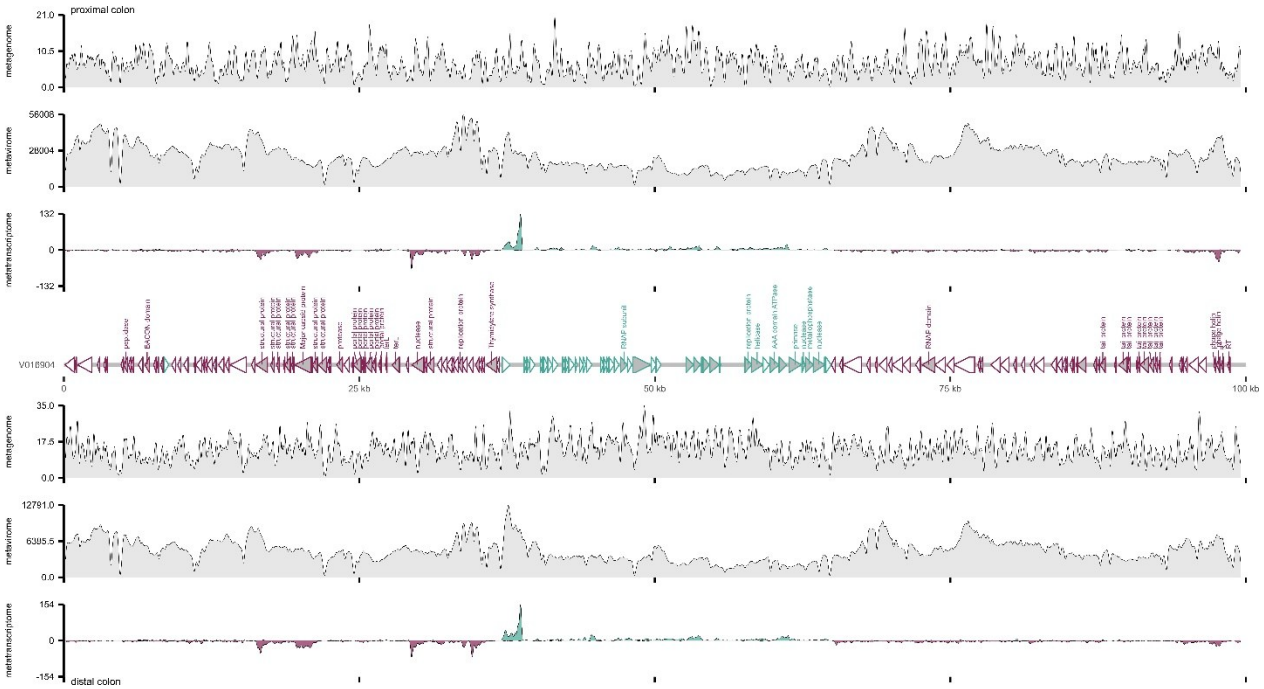
Supplementary Figure 3: Annotating viral gene clusters.

(A) Plot showing the sets of the 9,647 viral gene clusters that were annotated using viral protein family databases (VOG, pVOG) and general purpose databases (PFAM, TIGRFAM, KEGG); the remaining 18,465 gene clusters were annotated. (B) Boxplot showing the combined relative abundance of annotated genes in each metaviromic sample, by individual databases and all combined databases. (C) Scatter plots highlighting transcriptionally active VCs between proximal and distal colon samples in participants F and G. Lines indicate VCs present at both sites. The two mucosa-associated crAss-like phages identified in this study are labelled.



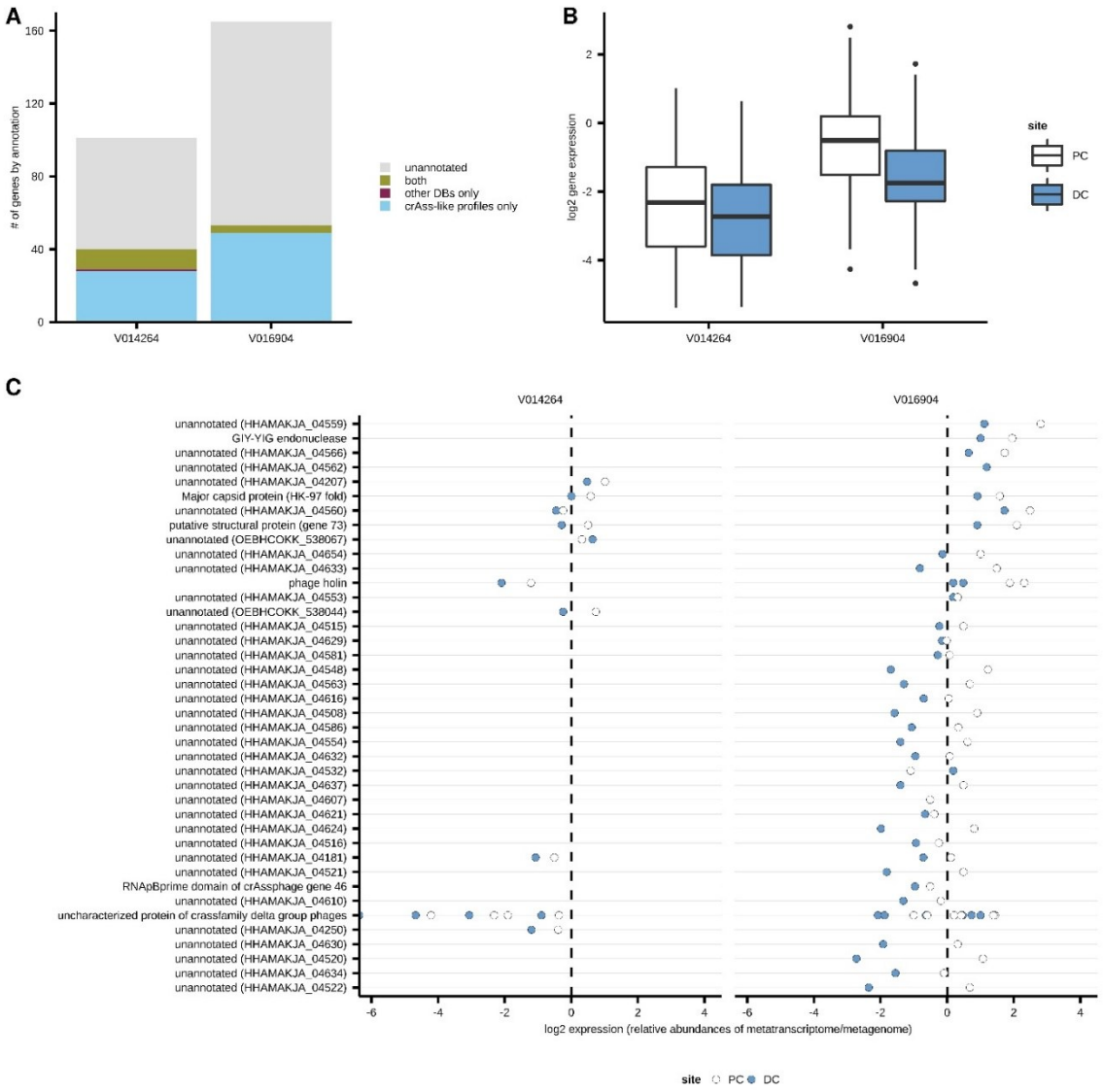
Supplementary Figure 4: Two crAss-like phages highly abundant in the colonic mucosal-luminal interface.

Relative abundance of crAss-like phages V014264 and V016904 across the metavirome, metagenome, and metatranscriptome of participant G. No metatranscriptome samples were obtained from stool. PC: proximal colon; DC: distal colon; STL: stool.



Supplementary Figure 5: Multiomic sequencing map of crAss-like phage V016904.

Open reading frames (ORFs) were predicted over the 99.5 kb genome coloured and oriented by forward (green) and reverse (purple) orientation. Annotated ORFs are shaded in gray and labelled (excluding annotations of 'uncharacterized' or 'hypothetical' proteins). Metagenome, metavirome, and metatranscriptome sequencing depths of the proximal colon (top) and distal colon (below) are plotted with a 151 bp sliding window; metatranscriptome reads were mapped by strand.



Supplementary Figure 6: Gene annotation and expression of two crAss-like phages.

(A) The number of genes in each crAss-like phage that could be annotated using a curated set of crAss-like phage profiles as compared to other viral and generic databases. (B) The expression (ratio of metatranscriptome / metagenome counts) of expressed genes for V014264 and V016904, by sample site of participant G. (C) The 30 most expressed crAss-like phage gene families identified in the proximal and distal colon of participant G. The dashed line indicates an metatranscriptome / metagenome ratio of 1.

Supplementary Table 1: Viral contig annotations

2,171 viral contigs identified in the metagenome and metavirome, with annotations regarding source, taxonomy, presence of integrase or flanking regions, and maximum abundance in any given metaviromics sample.

This table can be found in the supplementary data provided at the link below:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9980608/bin/KGMI_A_2177488_SM8988.zip

4. Article: The Colonic Mucosal Virome in Inflammatory Bowel Disease Reveals Crassvirales Depletion and Disease-Specific Virome Features

4.1. Preface

This chapter has been submitted to *Gut Microbes* as the following research article:

> **Yan A**, Butcher J, Mack D. R., Stintzi A. The Colonic Mucosal Virome in Inflammatory Bowel Disease Reveals Crassvirales Depletion and Disease-Specific Virome Features.

Specific author contributions are as follows:

Austin Yan: performed all virome experiments, led data analysis, and wrote the initial manuscript.

James Butcher: provided input on the bioinformatic analysis and manuscript.

David R. Mack: recruited patients, performed endoscopy on patients in collaboration with the CHEO IBD Centre, provided clinical and demographic information, and revised the manuscript prior to submission for peer review.

Alain Stintzi: provided reagent and research materials, obtained funding, supervised research, and revised the manuscript prior to submission for peer review.

4.1.1. Conflicts of Interest

AS and DM are co-founders of MedBiome, a clinical microbiomics company. The other authors have no competing interests to declare.

4.1.2. Funding

AY is supported by the Frederick Banting and Charles Best Canada Graduate Scholarships Doctoral Award from the Canadian Institutes of Health Research. This work was funded by the Government of Canada through Genome Canada and the Ontario Genomics Institute (OGI-149), the Canadian Institutes of Health Research (ECD-144627), and the Ontario Ministry of Economic Development and Innovation (Project 13440). The funders had no role in study design, data collection and analysis, or preparation of the manuscript.

4.1.3. Acknowledgments

The authors would like to acknowledge the patients and their families for their participation in our study. We acknowledge Ruth Singleton for her help in enrolling patients and assistance in collecting intestinal aspirate samples. This research presented here was enabled in part by WestGrid (www.westgrid.ca) and the Digital Research Alliance of Canada (alliancecan.ca).

4.1.4. Data Availability Statement

Host-removed, high-quality sequencing reads are available under BioProject PRJNA1004560.

4.2. Abstract

The mucosal virome is increasingly recognized for its potential role in shaping intestinal health and disease. Building on previous findings, we analyzed the mucosal virome from 51 individuals, including newly diagnosed treatment naive participants with ulcerative colitis (UC), Crohn's disease (CD), and non-inflammatory bowel disease (non-IBD) controls, incorporating longitudinal sampling for a subset of the participants. Viromes were highly individualized, with no shared or core components across participants. Unlike fecal virome studies, we observed no significant associations between virome diversity and mucosal inflammation, disease subtype, or sampling site. However, a positive correlation between virome and bacteriome diversity, particularly in CD, suggesting the presence of dynamic interactions that influence microbial community structure. We found *Crassvirales* to be abundant in the mucosa layer, comprising up to 82.4% of an individual's virome. Consistent with prior studies, *Crassvirales* abundance was reduced in IBD, irrespective of inflammation status or IBD subtype. These findings highlight their potential as biomarkers of virome health. Longitudinal sampling revealed a persistent subset of viruses, potentially shaping disease progression and remission dynamics. Our study underscores the importance of distinguishing microbial community dynamics across IBD subtypes and highlights *Crassvirales* as key players in mucosal immunity.

4.3. Introduction

Inflammatory bowel disease (IBD) is a chronic, relapsing illness at the interplay of host genetics, environmental factors, and the gut microbiota – the collection of bacteria, viruses, fungi, and archaea in our gastrointestinal tract.[206,337,338] While most microbiome studies have focused on bacteria, gut viruses are increasingly being studied due to the rapid development of new virome-targeted bioinformatic tools and databases.[33,182,190] Our gut virome is individualized and primarily composed of bacteriophages; recent studies have explored the temporal development of our virome,[75,170] identified new highly prevalent intestinal viruses,[154,331] studied the ecological stability of gut viruses,[329] and investigated the virome's role in human health.[83,339,340]

Changes to the gut virome have been reported in the IBD subtypes Crohn's disease (CD) and ulcerative colitis (UC). These observations include the enrichment of temperate phages and reductions in the diversity of both DNA and RNA viromes in CD.[32,213,341] *Crassvirales*, a viral order within the class *Caudoviricetes* (i.e. tailed bacteriophages) that are highly abundant in humans throughout life, are depleted in individuals with IBD.[82,329,342] Viral communities with low diversity and a high abundance of non-*Crassvirales Caudoviricetes* have been linked to lower likelihood of achieving endoscopic remission, as well as an altered bacteriome.[214] Other researchers, however, have suggested that high interpersonal variability in the virome may limit our ability to discern differences between the IBD and non-IBD virome.[145]

As with other microbiome analyses, a critical question remains: do observed correlations indicate causation? This question is particularly relevant in the context of intestinal mucosal inflammation as this is often considered a hallmark of IBD. Interestingly, clinical studies in participants with chronic-relapsing *C. difficile* infection have shown potential benefits when

receiving virome-containing fecal filtrates (and thus devoid of other microbiome components) and virome alterations appear to be associated with fecal microbiome transplant responses.[39,225] Furthermore, a recent study using a mouse model of colitis reported exacerbated inflammation in response to treatment with fecal virus-like particles derived from human participants with UC.[343] Conversely, fecal virome transplantation from healthy human donors ameliorated colitis symptoms in a mouse model, whereas virome transplantation from IBD donors exacerbated inflammation.[344] Together, these findings highlight the potential functional and causal role of the virome in modulating intestinal inflammation and underscore the need for further investigation into its contribution to IBD pathogenesis.

Virome analysis on rectal mucosa biopsies have demonstrated a decrease in *Caudovirales* diversity, richness, and evenness in UC participants.[213] Studies that incorporate sampling throughout the colon are scarce as most IBD virome studies have limited their sampling to stool, but the few studies available support the hypothesis that the intestinal mucosa-associated virome is different from their luminal or fecal counterparts.[246] Similarly, we have previously demonstrated that the colonic mucosal-luminal interface (MLI) fosters a unique viral population that included highly abundant *Crassvirales* phages in the proximal and distal colon that were not detectable in stool.[258]

Here, we expand our study of the MLI microbiome to evaluate the IBD colonic mucosal virome by using MLI samples collected from both the proximal and distal colon during diagnostic colonoscopy. We report significant alterations in the virome associated with IBD subtypes and different stages of disease activity. This approach provided a spatially resolved picture of the IBD virome that cannot be captured by fecal sampling alone.

We also assessed bacteriome-virome interactions in matched samples, identified a reduction in *Crassvirales* in IBD participants, and conducted a preliminary analysis of the MLI virome's longitudinal stability. Our findings significantly enhance our understanding of the colonic mucosal virome in IBD, highlighting the utility of using MLI samples over stool samples to gain a more precise and comprehensive insight into the virome's role in IBD pathology.

4.4. Materials and Methods

4.4.1. Participants

A cross-sectional cohort of 51 pediatric participants were recruited between April 2018 and December 2019. This cohort comprised 35 children newly diagnosed with IBD [CD: n=21 (60%); UC: n=14 (40%)]. Additionally, 16 participants who underwent diagnostic colonoscopy for suspected IBD, based on presenting signs and symptoms, were included. These participants exhibited normal mucosal visual appearance during colonoscopy and showed no histological evidence of inflammation in mucosal biopsy specimens, thus meeting the criteria as a non-IBD control group. All participants were IBD treatment naïve at the time of their initial diagnostic colonoscopy and had mucosal luminal interface (MLI) aspirate sampling performed at both the proximal and distal colon during their procedure. Exclusion criteria for study recruitment and collection of MLI aspirates included (a) presence of diabetes mellitus, (b) presence of documented or suspected infectious diarrhea within the previous two months, and (c) use of antibiotics or probiotics within the past 4 weeks.

Diagnosis of CD or UC was made following thorough clinical, endoscopic, histologic, and radiologic evaluations according to standardized criteria to reduce observer bias of disease sub-type, location and severity.[345] Disease location was described using the Montreal modification to

the Paris classification system.[346] For CD, the clinical severity at initial colonoscopy was reported using the Pediatric Crohn's Disease Activity Index, and categorized as: mild (PCDAI=11-30), moderate (PCDAI=30-37.5), or severe (PCDAI≥ 40).[347] For UC, the Pediatric Ulcerative Colitis Activity Index was used to categorize disease severity as remission (PUCAI<10), mild (PUCAI=11-34), moderate (PUCAI=35-64) or severe (PUCAI≥65).[348] Active colonic IBD (i.e., inflamed mucosa) was described by the visual appearance of colonic mucosa during colonoscopy (e.g. the loss of normal blood vessel appearance, mucosal erythema with mucosal ulcers) and was scored using the Simplified Endoscopic Score for Crohn's disease (SES-CD) or the Modified Mayo Endoscopic Subscore for UC.[349,350] Several participants (n=8) required follow-up colonoscopy as part of their medical care, and regional MLI aspirates samples were also obtained during these serial colonoscopes. Follow-up CD clinical disease activity was based on use of the weighted PCDAI (wPCDAI).[351] Clinical disease severity was categorized as remission (wPCDAI<12.5), mild (wPCDAI=12.5-39.5), moderate (wPCDAI=40-57.5), or severe (wPCDAI>57.5). Clinical data was managed using REDCap, hosted at the CHEO Research Institute. REDCap is a secure, web-based application designed to support data capture for research studies.[352] Choice of therapy for patients was based on published guidelines to minimize variation in care with the final therapeutic choices and was determined collaboratively by the treating physician, patient, and family.[353,354] Three participants (UC-F, UC-G, and UC-H) were previously described in Yan *et al.* (2023).

4.4.2. Ethics Approval

Participants were prospectively enrolled, and samples were collected, as part of a larger Children's Hospital of Eastern Ontario Review Ethics Board (REB) approved study (#09/37X). Informed written consent/assent were obtained from parents and/or participants, as appropriate.

4.4.3. Sample processing and VLP extraction

Mucosal-luminal interface (MLI) aspirates were obtained during colonoscopy as previously described.[247,257,258] In brief, mucosal-luminal interface (MLI) aspirates (40-80 mL) were collected from participants during their diagnostic endoscopy with sampling either being performed or supervised by a single physician to reduce sampling bias. Colonoscopy was performed following a one-day standard colon clean-out preparation,[311] first aspirating and discarding any existing fluid and luminal debris. Sterile water was then flushed through the colonoscope onto the mucosa of the site of interest (i.e. PC or DC) and aspirated through the colonoscope into a sterile container. These regional MLI samples were collected from both the PC and DC of each participant, immediately placed on ice in the endoscopy room, transported to the laboratory, and aliquoted within 30 minutes for storage at -80°C until further processing.

Aliquots of 10 mL were used for virus-like particle (VLP) purification, which involved the following steps: centrifugation and sequential filtration with 5.0 and 0.45 μm filters (Sigma-Aldrich, SLSV025LS and SLHV033RB) to remove cells and debris; virus-like particle precipitation with 10% w/v PEG-8000 (Fisher Scientific, BP233); resuspension in saline-magnesium buffer and bacterial cell lysis with 1 mg/ml lysozyme (Sigma, L4919), then chloroform treatment; centrifugation and DNase and RNaseI (Thermo Scientific, AM2238, Life Technologies, EN0602) treatment of the supernatant to degrade remaining bacterial nucleic acids; and VLP lysis with Proteinase K and cetyltrimethylammonium bromide (Fisher Scientific, BP1700 and O3042) and viral DNA extraction with phenol-chloroform-isoamyl alcohol (25:24:1, pH 6.7).

Viral nucleic acids were purified from the aqueous layer using the Dneasy Blood and Tissue Kit (QIAGEN, 69506), eluted in 50 μL of water, and subjected to centrifugal vacuum concentration to maximize input DNA for the multi-displacement amplification (GenomiPhi V2: GE Life

Science, 25660032). Reactions using 1 µL of input DNA were performed in triplicate, pooled, and purified using the Dneasy Blood and Tissue Kit. DNA quantification was performed using the Qubit dsDNA HS Assay Kit (Thermo Fisher, Q32854).

4.4.4. Virome sequencing and host-read removal

Shotgun DNA sequencing was performed at the Génome Québec CES using the NEB Ultra II library preparation kit and Illumina NovaSeq 6000 platform as previously described.[258] In summary, raw sequencing reads were trimmed and filtered using Cutadapt 2.10 (Illumina's universal adapters: AGATCGGAAGAG; AGATCGGAAGAG) and Trimmomatic 0.36 (SLIDINGWINDOW:4:20, MINLEN:60, and HEADCROP:10). Reads mapping to the human genome with bowtie2 (GRCh38 using ultra-sensitive mode) were removed, along with low complexity reads using komplexity.[276,313] Bacterial contamination was assessed by aligning reads to the *cpn60* database with bowtie2's ultra-sensitive mode.[89]

4.4.5. Metaviromic analysis

Host-decontaminated, high-quality reads from each virome sample were assembled using MEGAHIT v1.2.7 with a minimum contig length of 3 kb.[91] These reads were then mapped to the assemblies using bowtie2. The normalized relative abundance (NRA) of each viral contig was calculated as $\frac{x_i}{\sum x_i}$ where x_i is the number of reads mapping to a contig divided by the contig length.

We used a similar approach to Yutin *et al.* (2021) to identify *Crassvirales* contigs, specifically by selecting contigs harboring all three universally conserved *Crassvirales* genes: portal protein, large terminase subunit (TerL), and major capsid protein (MCP). Open reading frames were predicted using Prodigal v2.6.3 in metagenome mode, with subsequent clustering using vContact2.[119,282] Further functional annotation of viral open reading frames (ORFs) was

performed with Prokka, with mapping to predicted ORFs to VOG (version 211), pVOG, PFAM, TIGRFAM, Kegg, and crAssphage databases.[165,317] Viral contigs were also identified using geNomad 1.5.0 with score-calibration enabled and evaluated using CheckV.[95,355]

4.4.6. 16S rRNA amplicon sequencing and analysis

DNA extraction, V6-16S rRNA gene library construction and sequencing were performed as previously described.[247,251] In short, 2 mL of MLI sample was pelleted and metagenomic DNA extracted using a Fast-DNA Spin Kit. Extracted DNA was normalized to 5 ng/uL and stored at -20°C until library construction. V6-16S library construction was carried out using two successive rounds of PCR, amplicons purified using a MilliPore™ MultiScreen™ PCR96 plate and samples pooled at equimolar ratios. Pooled libraries were gel purified using a QIAquick PCR Purification Kit and then sequenced on an Illumina HiSeq 2500 with 100 bp paired-end reads at The Center for Applied Genomics (Toronto, Canada). Paired-end sequencing reads were demultiplexed and PCR primers removed using cutadapt with paired-reads that contained an ambiguous base-call or were <50 bp in length discarded. Subsequent analysis was performed in R using dada2. Paired-reads were quality filtered to remove pairs with >1 expected error and those aligning to the phiX genome. Forward and reverse reads were denoised separately for each sequencing run. Denoised reads were merged, the merged reads from each run combined together and chimeras identified using pooling on the entire dataset to generate amplicon sequence variants (ASVs). ASVs were subsequently analyzed using phyloseq.

4.4.7. Statistical analysis

Alpha-diversity and beta-diversity analysis, statistical analysis, and plotting were performed in R 4.1.3 using phyloseq 1.36.0,[284] reshape2 1.4.4,[356] ggplot2 3.3.0,[357] ggthemes

4.2.0,[358] and ggpubr.[359] Alpha diversity analysis was performed using read counts rarefied to 11,390,062 reads per sample for metavirome analysis (the lowest number of viral reads identified in a sample across the dataset) and 13,915 sequences for ASV analysis. DESeq2 was used to identify differentially abundant ASVs between the non-IBD controls and IBD-subtypes with biological sex (male/female) and sampling site (PC/DC) used as covariates in the analysis (Supplemental Table 1), fold changes ≥2 with adjusted *p* values <0.05 were considered significant. To maximize our potential to identify low abundant ASVs with differential abundance, we kept ASVs present at >2 reads in at least 4% of the samples and rarefied the dataset to 100,000 reads, removing samples below this threshold. Viral beta-diversity analysis was performed at the viral cluster level (using vContact2 gene-sharing networks) and at the ASV level for the bacteriome. Differential abundance analysis of viral taxa annotations using geNomad was performed with the Kruskal-Wallis test with false-discovery rate correction.

4.5. Results

4.5.1. Study population

There were no statistical differences in age at diagnostic colonoscopy (i.e., diagnosis) between the three groups of participants (CD, UC, and non-IBD) or between the number of males and females in each group (Table 1). Individuals without IBD (non-IBD) were participants undergoing colonoscopy to assess for IBD based on signs and symptoms but ultimately found to have normal mucosal appearance and normal mucosal biopsy histology. There was a similar number of participants with an inflamed proximal colon (PC) in either CD or UC (39% vs. 31%) and as expected there was more UC participants with an inflamed distal colon (DC) than in CD participants (90% vs. 47%). There was one participant with UC that did not have DC inflammation

and was ultimately diagnosed with proctitis. This participant thus had their DC sampling conducted proximal to the site of inflammation. Longitudinal regional colonic MLI samples were available for some participants (n=8, Table 2) as these individuals required a repeat colonoscopy as part of their clinical care.

4.5.2. Isolation of colonic viromes

Virome sequencing was performed on 88 MLI samples from the PC and/or DC from 51 individuals at time of initial diagnostic colonoscopy. We identified 87,658 viral contigs (VCs), which were subsequently annotated and assessed for quality. Among these, a subset of 4,793 VCs (5.5% of all VCs) were classified as "complete" or "high-quality" (both hereafter referred to as "high-quality"). These high-quality VCs accounted for the majority of mapped reads (mean of 69.5% for new-onset samples), consistent with previous findings.[258]

Viral contigs were annotated using geNomad, which reflects recent changes to viral taxonomy nomenclature.[111,355] The majority of assigned VCs belonged to the class *Caudoviricetes*, comprising 31,798 VCs, of which 3,346 were high-quality VCs (Figure 1A). This class previously included the order *Caudovirales* and its morphological subfamilies *Siphoviridae*, *Myoviridae*, and *Podoviridae* (these taxa are now abolished), and now includes the new order *Crassvirales*.[111] The second most common viral class was *Malgrandaviricetes,* with 894 VCs, of which 784 were high-quality VCs (Figure 1A). Nearly all of these were identified as *Microviridae* (892 of 894 VCs). The remaining VCs were predominantly ssDNA viruses from the viral families *Circoviridae, Inoviridae*, and *Genomoviridae*. Of the high-quality genomes, most ssDNA viruses, including *Microviridae,* were 3-5 kb in size, whereas most *Caudoviricetes* ranged from 10-100 kb (Figure 1B).

The 87,658 VCs were further categorized into 56,515 viral clusters (VCLs) using a gene-network approach to facilitate interparticipant comparison.[282] Most viral clusters (n = 48,674, 86.1%) were only present in one participant (Figure 1C). Very few VCLs were present in >10% of all participants (n = 786, 1.39%), and only five VCLs (0.01%) were present in >50% of all participants.

4.5.3. Alterations in the treatment-naïve colonic virome in inflammatory bowel disease

Within the two main viral classes, *Caudoviricetes* and *Malgrandaviricetes*, we observed a trend for increased *Caudoviricetes* among participants with IBD (Figures 1A and 1D), though this was only statistically significant between CD and non-IBD participants at the DC (Figure 1D). Conversely, *Malgrandaviricetes* were significantly enriched in DC samples from non-IBD participants compared to participants with CD (Figure 1E). Analysis of geNomad viral class and viral order annotations revealed no significant differences in viral taxa between UC, CD, and non-IBD participants (Supplemental Table 2).

We also evaluated the viral and bacterial community diversity using the Chao1 and Shannon diversity indices, which reflect species richness and evenness, respectively. No significant differences were observed in Chao1 or Shannon diversity among viral or bacterial communities based on diagnosis or inflamed mucosa (Figure 2A, Figure 2B, Supplementary Figure 1). Although there was a trend towards decreased viral richness and diversity in samples obtained from inflamed mucosa and in participants with CD with a higher clinical activity based on PCDAI score, these results were not statistically significant (Supplementary Figure 1). Additionally, no statistically significant differences were identified when viral communities were partitioned into *Caudoviricetes* and non-*Crassvirales Caudoviricetes* groups. There was a positive correlation between the Shannon diversity of the virome and bacteriome in both the PC (Spearman correlation

0.37, $p = 0.004$) and DC (Spearman correlation 0.32, $p = 0.013$). When subsetted by IBD subtype, this correlation was higher and statistically significant in participants with CD but was reduced in participants with UC, even trending towards an inverse relationship at the PC (Figure 2C).

Beta-diversity analysis did not reveal distinct viromes based on the colonic site of MLI collection, mucosal inflammation, or IBD subtype (Figure 3A, Supplemental Table 3). Instead, our data indicates that mucosal viromes are highly individualized, with significantly reduced intra-individual Bray-Curtis dissimilarity (i.e. between one's PC and DC) compared to inter-individual comparisons ($p < 2E-16$, Figure 3B). There was no significant difference for intra-individual comparisons based on local inflammation (Figure 3C) or IBD subtype (Figure 3D). When analyzing the bacteriome data using the same methods, we found that there is a greater trend for participants with IBD to cluster away from non-IBD controls (Figure 4A, Supplemental Table 3); although this explained a minimal amount of the overall variation (R^2=0.04). We also found that participant bacteriomes were personalized (Figure 4B), although not to the same extent as seen for the virome (median inter-individual Bray Curtis dissimilarity of 0.75 versus almost 1.0 for the virome). There was also no difference between the bacteriome intra-individual comparisons based on local inflammation (Figure 4C) or IBD subtype (Figure 4D) as also seen in the viromic dataset.

4.5.4. Reduced abundance of Crassvirales in the colonic virome

By using a similar approach to Yutin et al. (2021), we identified 109 *Crassvirales* VCs in samples collected during the initial diagnostic colonoscopy (Figure 5A). Most of these were high-quality VCs (68/109) and range in size from 90 to105 kb (64/109). Taxonomic annotations by geNomad identified a greater number of *Crassvirales* VCs (n = 350).[355] However, these primarily included shorter and lower quality contigs: only 45/350 were complete or high-quality and only 43/350 were greater than 90 kb (Figure 5B). Combining the two methods yielded a set of 386

Crassvirales VCs (shown in Figure 5C) which were used to evaluate the relative abundance of *Crassvirales*.

Crassvirales were identified in 40/51 study participants, including 15/21 participants with CD, 11/14 participants with UC, and 14/16 participants without IBD. Interestingly, the normalized relative abundance of *Crassvirales* was significantly higher in the PC of participants without IBD than in participants with IBD, with the DC showing a similar, but not statistically significant, trend (Figure 5D). There was no significant difference in *Crassvirales* levels between participants with UC and CD or between the PC and DC in this dataset.

4.5.5. Alterations in the colonic bacteriome

As we observed potential alterations in *Crassvirales*, we were interested in whether there were corresponding alterations in the colonic bacteriome of these specific participants that followed the same trends. We identified 108, 96 and 99 amplicon sequence variants (ASVs) differentially abundant between CD vs non-IBD, UC vs non-IBD and UC vs CD after controlling for sex and colonic sampling region (Supplemental Table 1). There were more ASVs belonging to the phylum Bacteroidetes, and in particular those belonging to the genera *Bacteroides/Phocaeicola* that were enriched in participants with CD as compared to either UC or non-IBD (Figure 6A-C). In contrast, participants with UC had more Bacteroidetes ASVs depleted as compared to either non-IBD or CD. Interestingly, the dominant Bacteroidetes ASV matched with 100% identity across its entire length to *Phocaeicola* (formally *Bacteroides) vulgatus* and was enriched in both CD and UC subtypes as compared to individuals without IBD. This ASV was also enriched in participants with CD as compared to those with UC. Moreover, we noted striking sex differences in the abundance of this ASV regardless of IBD sub-type (Figure 6D), with males having higher levels as compared to their female counterparts (although this only reached significance at the PC). In contrast, there

was no difference seen between the sexes in those without IBD. We also identified an ASV which matched perfectly to *Phocaeicola* (formally *Bacteroides) dorei* in our dataset (ASV0002), but this ASV was not found to be differentially abundant. We also investigated potential correlations between ASVs and genera and *Crassvirales* abundances across different participant groups and sampling sites. However, no statistically significant correlations were detected in any group or site.

4.5.6. Exploratory analysis assessing the temporal stability of the MLI virome

Longitudinal samples were available from eight IBD participants (6 CD, 2 UC), collected during repeat colonoscopies (Table 2). These samples were collected at a mean of 20.8 weeks after their initial diagnostic colonoscopy (ranging from 14.4 – 32.7 weeks). Participant UC-G had two additional samples obtained at 37 and 63 weeks post-diagnosis. The initial samples represented treatment-naïve viromes, while subsequent samples were influenced by varying illness severities and the introduction of different IBD treatments (i.e., corticosteroids, mesalamine, immunomodulators, and anti-TNFα monoclonal antibodies; Table 2). Thus, gene-sharing based analysis was used to cluster viral contigs for longitudinal comparisons.

This exploratory analysis revealed that intra-individual serial variations in viral communities exhibited significantly greater similarity compared to inter-individual variations (Figures 7A and 7B, $p < 2e\text{-}16$). Within individuals, samples collected at different time points showed greater variation in viral community composition than samples collected simultaneously from different anatomical sites (e.g., proximal colon [PC] and distal colon [DC]; Figure 7B, $p < 0.001$). Furthermore, the overall viral diversity remained relatively stable over time within individuals when accounting for clinical disease severity (Figure 7C).

Between 2.4 and 9.3% of VCs identified at the initial diagnostic colonoscopy were detected at subsequent time-points (Figure 7D). These persistent VCs were more likely to be highly abundant, representing over 25% of the normalized relative abundance of VCs in most participants (Figure 7E). Two participants with serial MLI sampling showed initial VCs persisting at similar levels over the subsequent year.

Crassvirales VCs were detected in the colonic virome of 7/8 individuals with repeat samples available (Figure 7F). While most *Crassvirales* were identified at only one timepoint, several participants had *Crassvirales* VCs that were conserved longitudinally.

4.6. Discussion

4.6.1. Virome diversity at the mucosal-luminal interface

We previously demonstrated that the MLI virome is distinct from the stool virome.[258] In this study, we build upon our previous work by expanding our cohort from three participants with UC to fifty-one individuals, including participants with UC and CD, and individuals without IBD. Furthermore, we performed longitudinal sampling for eight of these participants. The viromes within our cohort were highly personalized, with no shared or core virome identified across individuals.

A fecal virome study reported reduced overall viral alpha-diversity and increased diversity of *Caudovirales* (now classified under *Caudoviricetes*) in CD,[32], although a re-analysis of the same dataset found no significant changes in overall viral alpha-diversity.[8] In contrast, our dataset did not identify statistically significant associations between viral diversity and the presence of mucosal inflammation, CD or UC, or sampling site (i.e. PC vs. DC), even with targeted analysis of *Caudoviricetes* populations. These results are consistent with recent analysis of unamplified fecal

viromes by Stockdale *et al.* (2023), which suggested that high interpersonal variability may substantially confound alpha and beta-diversity comparisons in IBD virome studies, to a much greater extent than for bacteriome analyses.

While Norman *et al.* (2015) initially reported an inverse relationship between bacterial and viral alpha-diversity, more recent fecal studies have shown a positive correlation between gut virome diversity and the bacteriome, an effect that is driven by bacteriophages rather than eukaryotic viruses.[360] Our data also supports a positive correlation between virome and bacteriome alpha diversity, which appears to be primarily driven by our cohort of individuals with CD. This positive correlation suggests that a more diverse bacteriome may provide a broader ecological niche for bacteriophages to proliferate, potentially facilitating dynamic interactions between phages and their bacterial hosts. These interactions could promote bacterial turnover through lysis, thereby shaping microbial community composition and maintaining balance within the gut ecosystem. Alternatively, the increased diversity of bacteriophages may help stabilize bacterial communities by preventing the overgrowth of specific taxa via predator-prey dynamics. Conversely, our results in UC participants are consistent with those of Zuo *et al.* (2019), who reported a weakening of bacterial-viral correlations in mucosal biopsies in individuals with UC compared to those without IBD. These disease-specific findings highlight the importance of recognizing the presence of distinct microbial pathogenesis and community structures in CD and UC. This disruption of bacterial-viral interactions may reflect alterations in the stability and functionality of the mucosal microbiota, potentially contributing to the disrupted microbial ecosystem. Altogether, these findings underscore the importance of distinguishing microbial community dynamics and pathogenesis across different IBD subtypes.

4.6.2. Crassvirales at the mucosal-luminal interface

We used two distinct approaches to identify *Crassvirales* VCs. The VCs identified by geNomad represent short partial *Crassvirales* genomes that do not contain any of the three marker genes used by the first approach.[165] Despite these differences, these two approaches showed strong concordance. Among the 109 VCs containing the three universal *Crassvirales* genes, all were classified by geNomad within the class *Caudoviricetes,* with 73 further assigned to the order *Crassvirales*, while the remaining 36 remained unclassified at this taxonomic level.

In our previous study, we identified two abundant *Crassvirales* species at the MLI that were nearly undetectable in matched stool samples,[258] suggesting that the MLI may serve as a reservoir for *Crassvirales*. In this study, we extended these finding by detecting *Crassvirales* in 40/51 participants, with *Crassvirales* representing up to 82.4% of an individual's virome. While we were unable to identify statistically significant correlations between bacterial taxa and *Crassvirales*, suggesting that these may be either weak, context-dependent, or obscured by other factors influencing microbial community dynamics; our data does reveal that, in a pediatric (primarily adolescent) population, IBD is associated with a reduced relative abundance of *Crassvirales*. However, we did not observe a significant correlation between *Crassvirales* abundance and local inflammation or IBD subtype (Figure 5D). This observation is consistent with findings from studies in adult populations that reported decreased *Crassvirales* in individuals with CD as well as individuals with UC who did not respond to treatment.[8,214] Collectively, our findings along with those from other studies suggest that *Crassvirales* may be markers of a healthy virome, and could potentially serve as biomarkers or modulators of virome health, independent of the local inflammatory status. The presence of *Crassvirales* in the mucus layer, along with the exclusive presence of specific *Crassvirales* phage species,[258] may constitute an antimicrobial barrier that limits mucosal bacteria and protects the underlying epithelium; this model is summarized Figure

8. Supporting this hypothesis, bacteriophages adhering to the mucus have been shown to provide a non-host-derived immunity that actively protects the mucosal surface.[232] Consistent with the essential role of the mucosal virome in host protection, mice humanized with non-IBD colon mucosal viromes were shown to be protected from intestinal inflammation.[246]

Based on these studies, a decrease in mucosal *Crassvirales* abundance could weaken the antimicrobial barrier, allowing the invasion of the mucus layer by *Crassvirales* bacterial hosts. *Crassvirales* are associated with bacteria from the phylum Bacteroidetes and, based on limited isolated virions, they appear to specifically infect *Bacteroides/Phocaeicola* genera.[361] Interestingly, several *Bacteroides/Phocaeicola* species have been shown to induce colitis in mouse models susceptible to IBD.[362] Furthermore, a recent study demonstrated that *P. vulgatus* and *P. dorei* can disrupt colonic epithelial integrity, leading to colitis, a phenotype associated with their proteases.[363] Our observed increase in *P. vulgatus* in both IBD subtypes would support this hypothesis; although we note that several *Bacteroides/Phocaeicola* genera, such as *P. dorei,* were not increased in CD/UC in this study and we were unable to identify specific *Crassvirales*/bacterial correlations. This could be due to differences in the specific species that we were able to identify using our 16S amplicon approach and/or due to sub-strain variation as both *P. vulgatus* and *P. dorei* have been reported to have extensive pangenomes (especially *P. vulgatus*) that could impact the infectivity of specific *Crassvirales* phages.[364] Indeed, *P. vulgatus* genomes typically carry fewer CRISPR-Cas anti-phage systems and have a higher proportion of bacteriophages/insertions sequences as compared to *P. dorei*.[364] It is thus tempting to speculate that the expansion of *P. vulgatus,* but not *P. dorei,* in IBD may be due to decreased selective pressure from endogenous *Crassvirales* phages. In addition, we found evidence for sex-specific differences in *P. vulgatus* abundances only in those diagnosed with CD or UC. This suggests that there are potentially sex-specific differences in the

Crassvirales phage repertoires in pediatric IBD, as has been reported in studies on healthy individuals.[365] Although we did not identify statistically significant differences in our dataset, these potential sex-associated variations could contribute to the high degree of virome personalization observed in this study. These findings highlight the potential significance of *Crassvirales* in protecting their host from colitis and underscore the need for further research to elucidate their role in maintaining virome health and mucosal immunity.

4.6.3. Exploratory analysis of the intestinal virome's longitudinal stability in IBD

While the human fecal virome has been reported to be individualized and longitudinally stable,[45,82,179] those studies primarily involved healthy volunteers. In contrast, the gut virome is likely to experience greater temporal variation in the context of human illnesses, as suggested in fecal virome studies in individuals with IBD as well as those infected with COVID-19.[145,366] Our study demonstrates that the mucosal virome remains individualized, even when comparing treatment-naïve microbiomes to samples taken several months later after initiation of IBD therapies. While most VCs were only present at diagnosis, the 2.4-9.3% of VCs which persisted were more likely to be abundant, often representing a majority of the viral community at repeat colonoscopy. Furthermore, the proportion of shared viruses appear to stabilize over the course of a year, suggesting the presence of both a subset of low-abundance, transient viruses and a persistent virome that resides in the colonic mucosa and remains stable over time, even as individuals transition from active disease with an inflamed mucosa to clinical remission with normal mucosa. The transient virome identified in this study may primarily result from the impact of disease treatment and reduced intestinal inflammation. However, it is unclear whether the long-term persistent virome identified in individuals with IBD has beneficial or detrimental consequences. This persistent virome could either represent the baseline virome expected in healthy individuals

or a reservoir of phages that could promote future disease flares. Further research is necessary to elucidate the role of this persistent virome in IBD pathogenesis and its potential implications for disease management.

4.6.4. Limitations and next steps

Our protocol is designed to identify DNA bacteriophages, which represents the majority of intestinal viruses. Bacteriophages are the primary drivers of virome diversity patterns in the gut, as opposed to eukaryotic viruses, which play a more limited role in shaping these dynamics.[360] We thus did not identify enveloped viruses (which includes many eukaryotic viruses) or RNA viruses. It is important to note that our previous metatranscriptomics study demonstrated the absence of RNA-associated viral signatures in our samples.[258] However, alterations in the RNA virome have been reported in IBD, particularly in individuals with Crohn's disease experiencing clinical flares.[341]. Additionally, specific eukaryotic viruses, such as *Orthohepadnavirus*, have been associated with UC.[367] Alternate sequencing or sampling approaches are required to comprehensively profile RNA viruses. Future studies employing these approaches will be critical to unraveling the contributions of RNA viruses and eukaryotic viruses to intestinal disease processes.

Our study also utilizes multiple displacement amplification (MDA) which introduces biases into both virome composition and diversity analyses. To address this issue, we have previously demonstrated the reliable quantification of a control spike-in phage using MDA.[258] This control allows us to calibrate our amplification process and assess the accuracy of our virome measurements despite the inherent biases introduced by MDA. Nonetheless, the limitations of MDA remain a concern, particularly when analyzing samples with low microbial loads, such as those obtained from the gastrointestinal mucosa including sites proximal to the colon.[145,325] In

addition, the virome at the MLI exhibited greater inter-individual variability than the bacteriome, indicating that larger cohort sizes may be necessary to detect statistically significant correlations as compared to typical bacteriome studies. The collection and use of MLI samples are a strength of this study. The mucosal layer of the gastrointestinal tract represents a critical niche for understanding the gut virome, particularly in the context of IBD where the mucosa is a site of active inflammation and altered microbial composition. Unlike the luminal environment, the mucosa provides a unique microenvironment where bacteriophages constitute an antimicrobial barrier,[232] and microbial communities, including viruses, engage directly with the epithelial barrier and immune system. Previous studies have revealed that several viral members of the gut microbiome, including *Crassvirales,* possess mucus-interacting capabilities, enabling them to attach to mucin.[368] Despite these observations, the mechanisms underlying these interactions and their potential effects on host physiology and disease processes remain largely unexplored. Sampling the mucosal layer offers the unique opportunity to investigate these critical virome-microbes dynamics. However, sampling from the bowel following the clean-out required for safe colonoscopy cannot fully represent native conditions.[369] Nevertheless, it allowed us to investigate the inner mucus composition regionally and assess the tightly adherent MLI communities.[247] The higher-resolution sampling of MLI samples adds a further layer of heterogeneity with further potential impact of varying local disease severity and different spatial patterns of inflammation. However, due to the invasive nature of endoscopic sampling, we lack a traditional "healthy cohort" commonly used in fecal virome studies. Instead, our non-IBD controls may include individuals with underlying conditions that could introduce potential confounding factors, even in the absence of mucosal inflammation.

This study provides valuable insights that can guide future studies involving larger, standardized clinical cohorts that include matched proximal and distal colon, as well as fecal samples. Such an approach will be instrumental in advancing microbiome studies beyond solely utilizing fecal specimens, addressing a critical gap in viromics research. In particular, understanding the exact role of the virome at the site of IBD disease remains a key challenge, as specific phages may exert either beneficial or pathogenic effects. Moreover, the extent to which the virome shapes, or is influenced by, disease states such as active inflammation, remission, and flares is still poorly understood. Future work using large cohorts of participants with IBD with repeated samplings over time will likely be required to answer these questions and overcome the high interpersonal variability seen within the colonic virome.

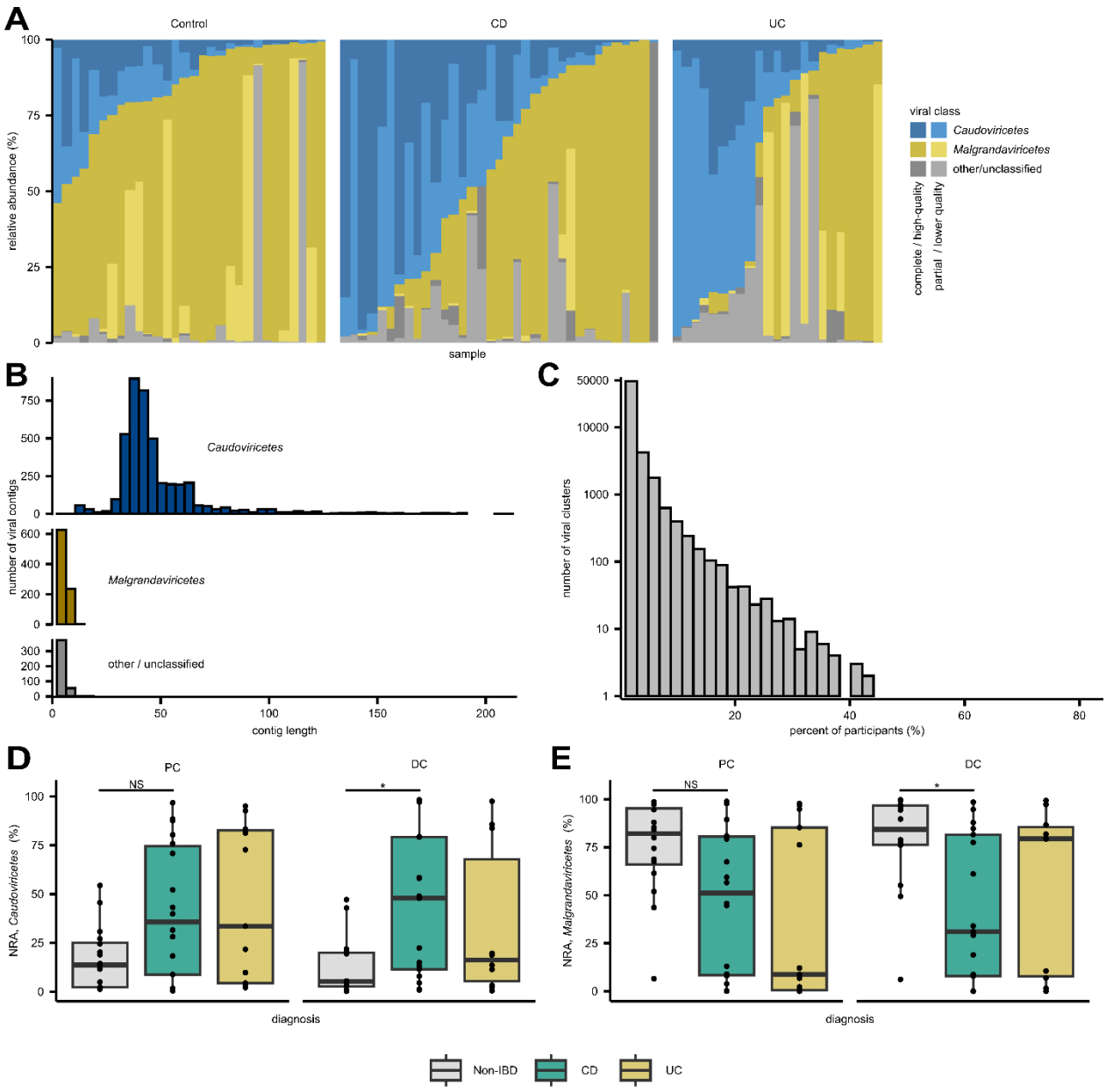


Figure 4-1: Taxonomic annotation and clustering of viral contigs.

Virome sequencing of study participants undergoing their initial diagnostic colonoscopy revealed 87,658 viral contigs which were subject to taxonomic annotation and gene-network clustering. (A) Normalized relative abundance of viral contigs per sample, grouped by diagnosis. (B) Histogram

showing contig length of high-quality VCs based on viral class. (C) Histogram showing the percent of participants that share each viral cluster. (D) Normalized relative abundance of *Caudoviricetes* and *Malgrandaviricetes* by site and diagnosis. CD: Crohn's disease; UC: ulcerative colitis; PC: proximal colon; DC: distal colon; *: $p < 0.05$; NS: not significant.

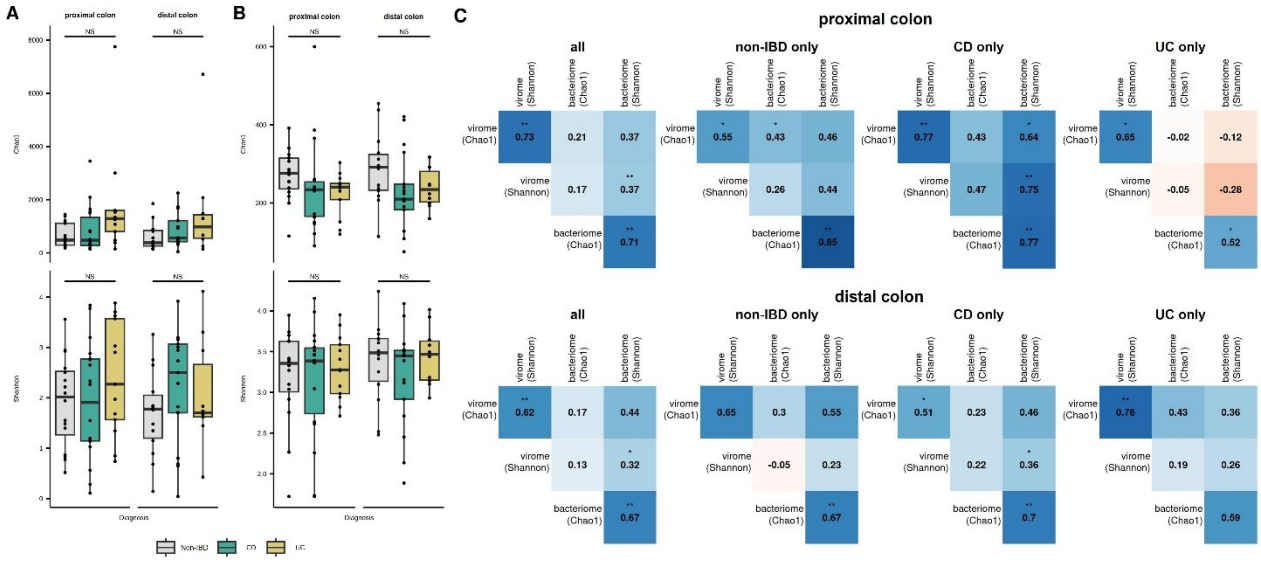


Figure 4-2: Alpha-diversity of the mucosal-luminal interface virome and bacteriome in inflammatory bowel disease.

Chao1 and Shannon diversity were calculated based on rarefied counts to (A) viral contigs and (B) 16S rRNA amplicon sequence variants. (C) Correlations between bacterial vs viral alpha diversity at the proximal and distal colon, including all samples and separated by disease status. **, $P < 0.01$; *, $P < 0.05$; NS: not significant; CD: Crohn's disease; UC: ulcerative colitis.

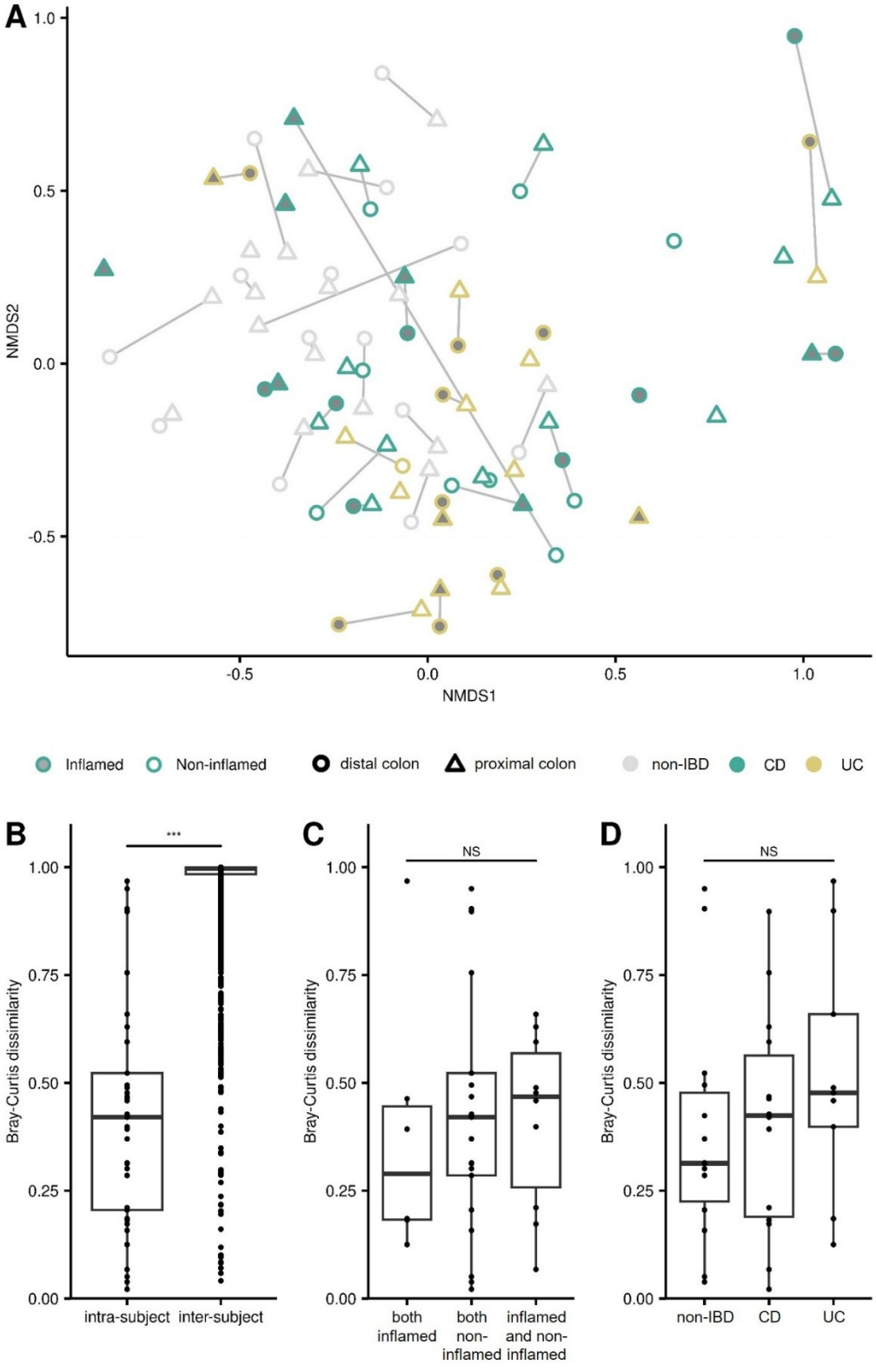

A
NMDS2
NMDS1
1.0
0.5
0.0
-0.5
Inflamed
Non-inflamed
distal colon
proximal colon
non-IBD
CD
UC
B
***
Bray-Curtis dissimilarity
intra-subject
inter-subject
C
NS
both inflamed
both non-inflamed
inflamed and non-inflamed
D
NS
non-IBD
CD
UC

Figure 4-3: Virome beta-diversity at the mucosal-luminal interface in inflammatory bowel disease.

(A) Virome community structure visualized with NMDS plot (stress = 0.251, k = 2) with annotations for site, diagnosis, and local inflammation. (B) Bray-Curtis distances were calculated and compared between and within individuals. Samples from the same patient are connected. Distances between proximal and distal colon sites from the same participant were compared based on (C) local inflammation and (D) diagnosis.***, $P < 0.001$; NS: not significant; CD: Crohn's disease; UC: ulcerative colitis.

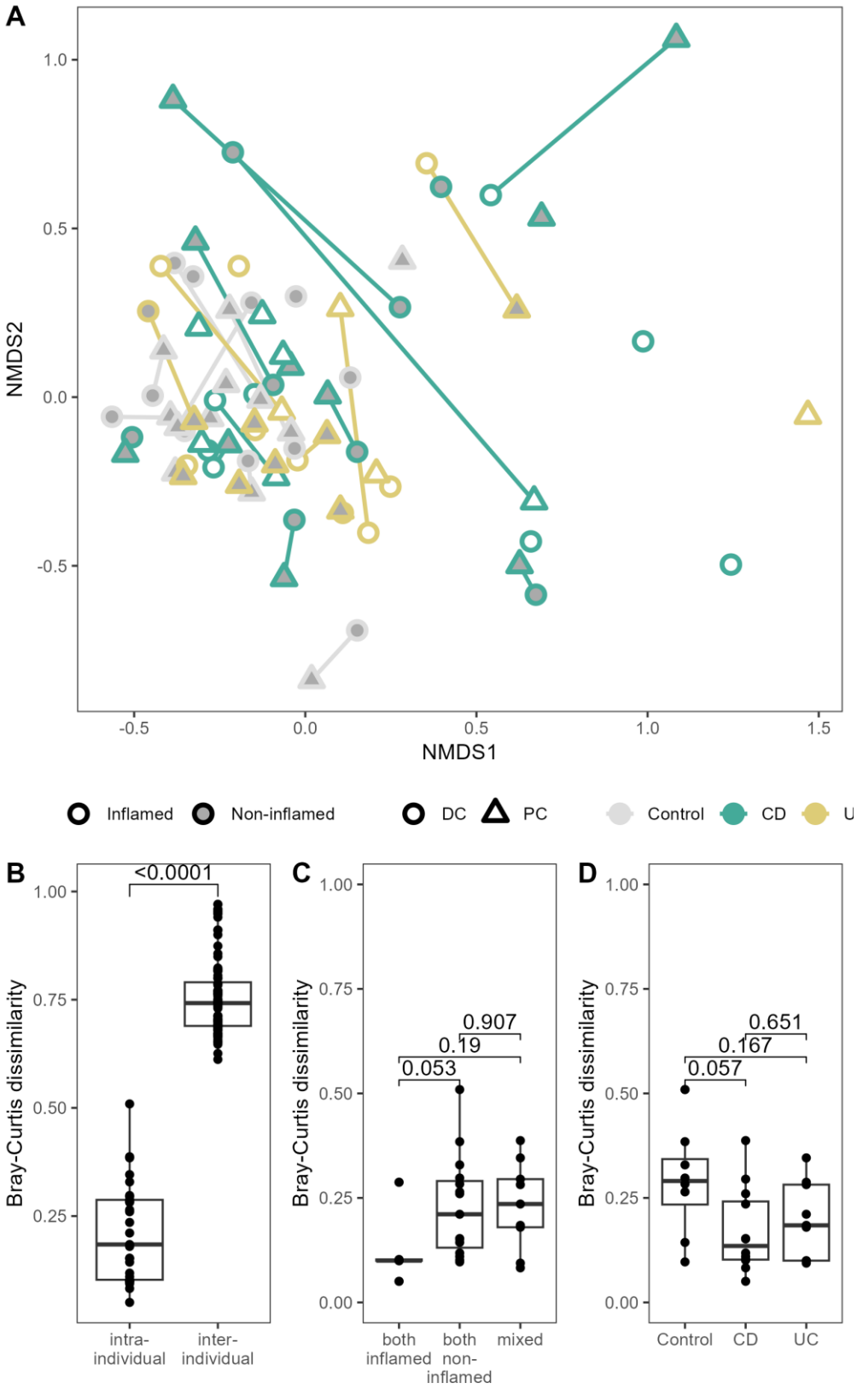

A
NMDS2
NMDS1
1.0
0.5
0.0
-0.5
-0.5
0.0
0.5
1.0
1.5
Inflamed
Non-inflamed
DC
PC
Control
CD
B
Bray-Curtis dissimilarity
<0.0001
intra-individual
inter-individual
C
0.907
0.19
0.053
both inflamed
both non-inflamed
mixed
D
0.651
0.167
0.057
Control
CD
UC

Figure 4-4: Bacteriome beta-diversity at the mucosal-luminal interface in inflammatory bowel disease.

(A) Bacteriome community structure visualized with NMDS plot (stress = 0.193, k = 2) with annotations for site, diagnosis, and local inflammation. (B) Bray-Curtis distances were calculated and compared between and within individuals. Distances between proximal and distal colon sites from the same participant were compared based on (C) local inflammation and (D) diagnosis. ***, $P < 0.001$; NS: not significant; CD: Crohn's disease; UC: ulcerative colitis.

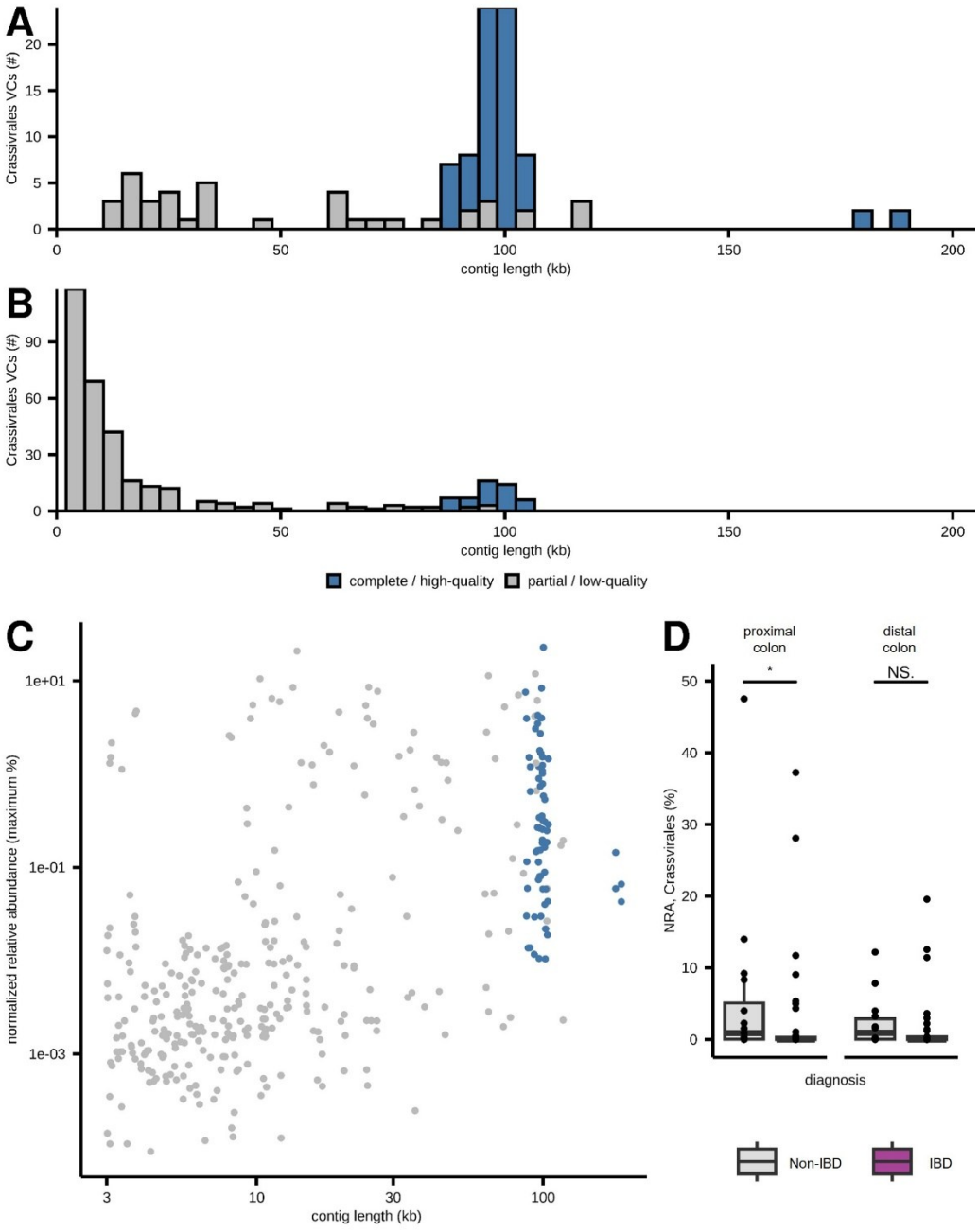


Figure 4-5: *Crassvirales* at the human mucosal-luminal interface.

(A) Quality and length of *Crassvirales* VCs identified using presence of portal protein, TerL, and MCP. (B) *Crassvirales* VCs annotated using geNomad. (C) Relative abundance vs. length of *Crassvirales* VCs at the human MLI. (D) Normalized relative abundance of *Crassvirales* at the proximal and distal colon MLI in participants with and without IBD.

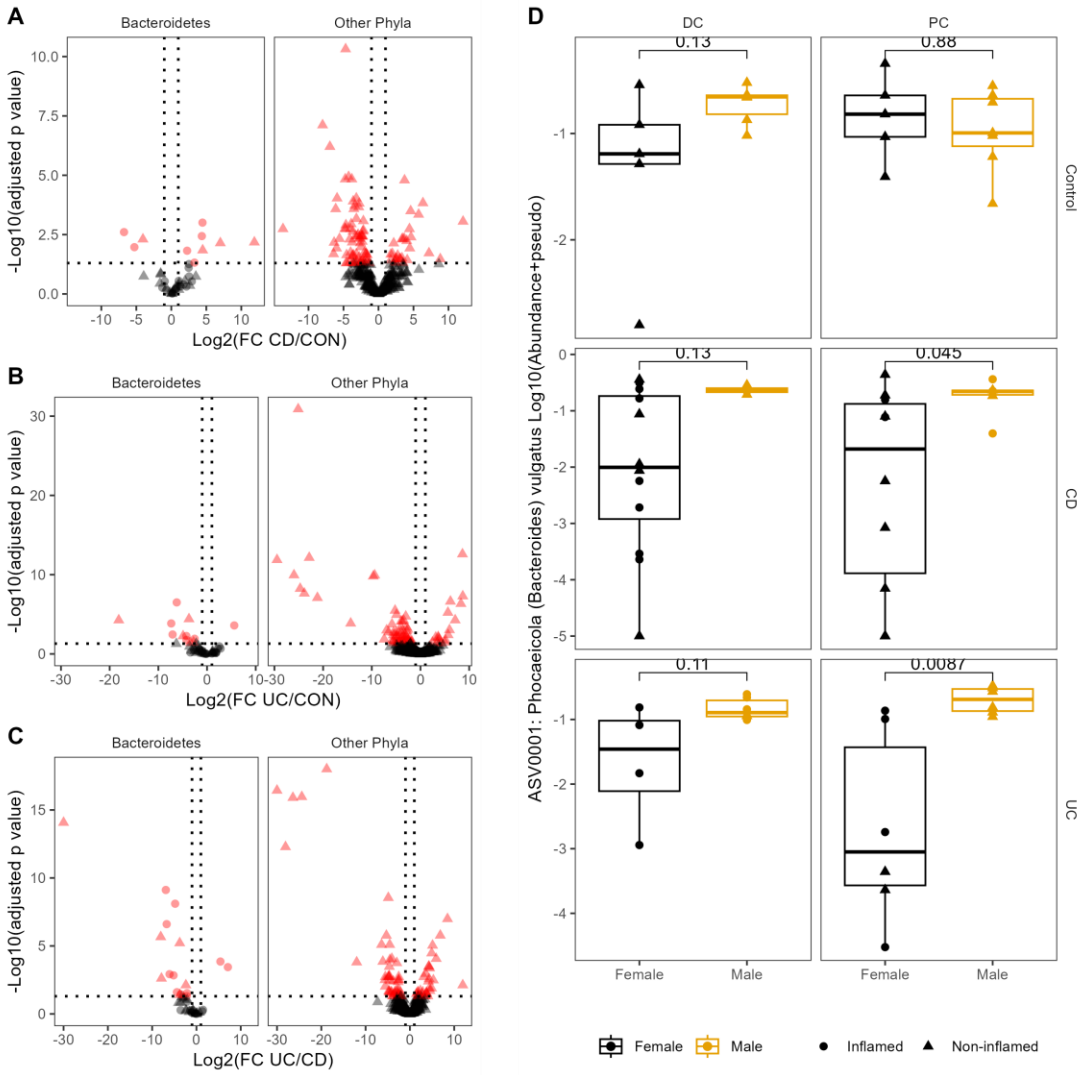


Figure 4-6: Differentially Abundant ASVs Between Non-IBD Controls and IBD Subtypes.

ASVs identified by DESeq2 as differentially abundant (absolute log2 fold change ≥ 1, adjusted p-value ≤ 0.05) for comparisons between CD and non-IBD controls (A), UC and non-IBD controls (B), and UC and CD (C). ASVs belonging to Bacteroidetes are displayed separately from those of other phyla. Relative abundances of ASV0001 (putative *Phocaeicola vulgatus*) are shown for male and female participants in non-IBD, CD, and UC groups across sampled sites. Statistical differences between biological sexes were assessed using the Mann-Whitney U test * $p < 0.05$; ** $p<0.01$; NS: not significant.

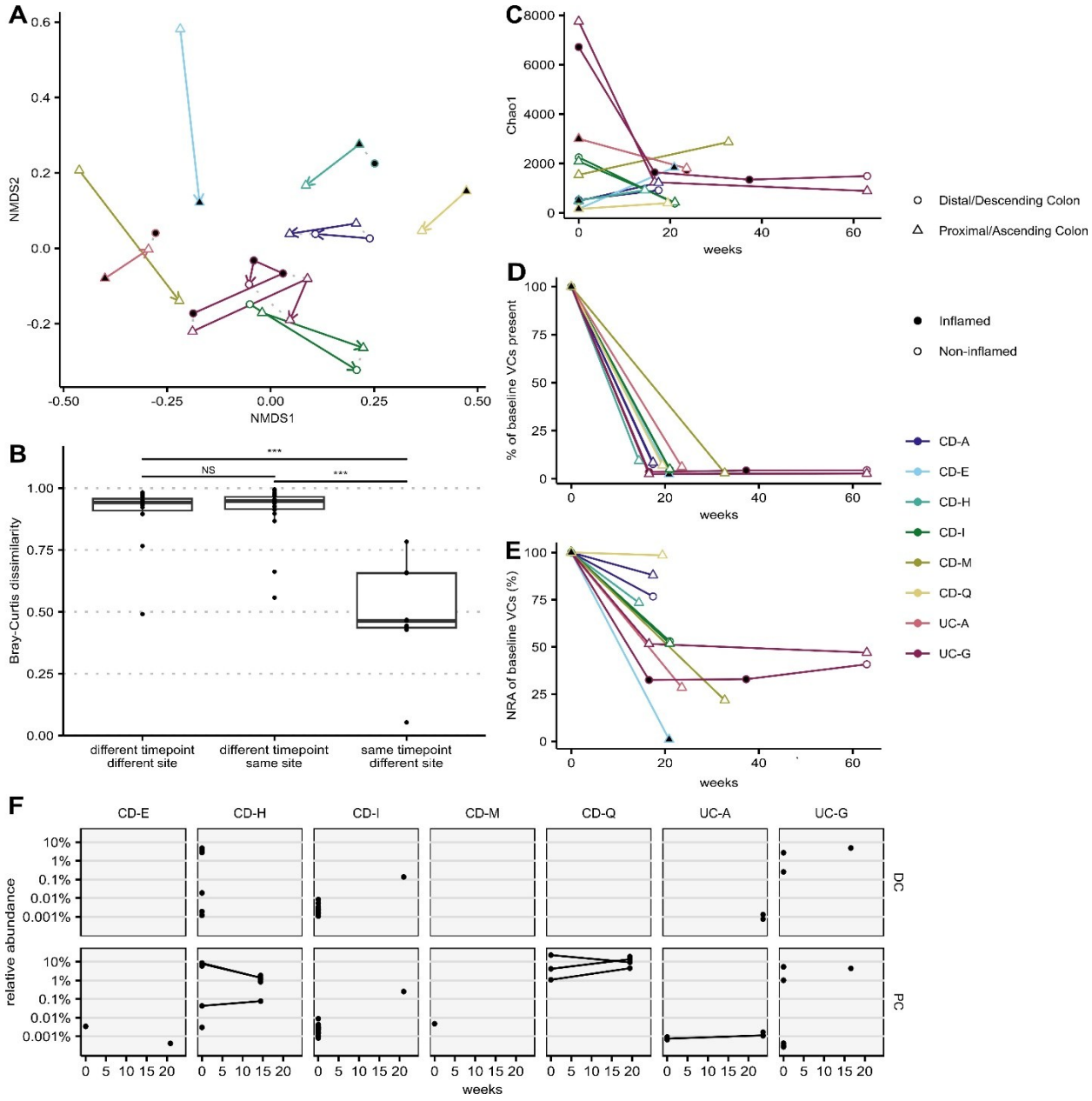

A
NMDS2
NMDS1
B
Bray-Curtis dissimilarity
NS
***
***
different timepoint different site
different timepoint same site
same timepoint different site
C
Chao1
weeks
D
% of baseline VCs present
weeks
E
NRA of baseline VCs (%)
weeks
Distal/Descending Colon
Proximal/Ascending Colon
Inflamed
Non-inflamed
CD-A
CD-E
CD-H
CD-I
CD-M
CD-Q
UC-A
UC-G
F
CD-E
CD-H
CD-I
CD-M
CD-Q
UC-A
UC-G
relative abundance
DC
PC
weeks

Figure 4-7: Longitudinal virome sampling at the MLI in subjects with inflammatory bowel disease.

(A) Longitudinal virome community structure visualized by NMDS plot (stress = 0.175, k = 2) with annotations for participant, site, and local inflammation. Solid coloured lines represent samples from the same participant and site with arrows originating from the first sample; samples taken at the same timepoint are connected with a gray dotted line (B) Intraparticipant Bray-Curtis distances compared between sites and time. NS: not significant; ***: $p < 0.001$. (C) Shannon diversity of longitudinal virome samples. (D) Percent of a sample's viral clusters shared with the participant's initial baseline sample at time of diagnostic endoscopy. (E) Total normalized relative abundance of baseline viral contigs present. (F) Normalized relative abundance of *Crassvirales* clusters identified by longitudinal virome sampling; only participant sites with multiple timepoints are plotted. Clusters representing same *Crassvirales* contigs are connected with lines. $p < 0.001$; PC: proximal colon, CD: Crohn's disease; UC: ulcerative colitis.

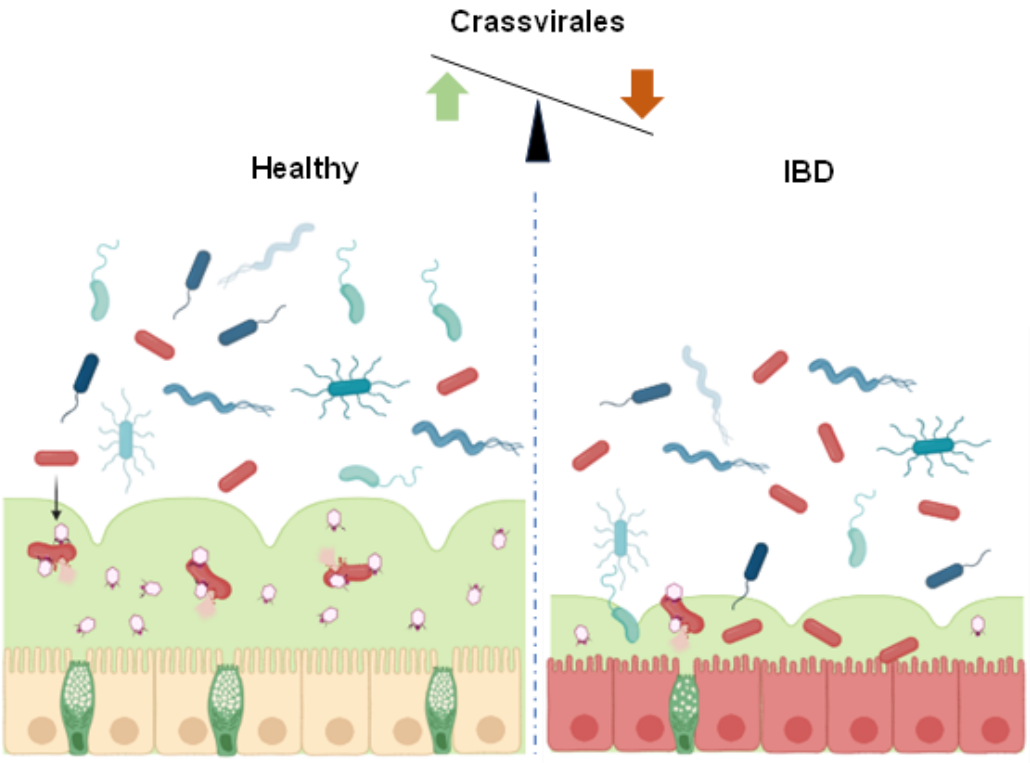


Figure 4-8: Longitudinal virome sampling at the MLI in subjects with inflammatory bowel disease.

In a healthy colonic mucosal-luminal interface, a thick mucus layer embeds *Crassvirales*, enabling their interaction with potentially pathogenic and proteolytic bacteria. This interaction allows *Crassvirales* to initiate their lytic replication cycle, limiting the proximity of these bacteria to the mucosal surface and providing a protective barrier. In contrast, the thinner mucus layer observed in IBD results in reduced *Crassvirales* abundance within the mucosal layer. This reduction permits the expansion of pathobionts, such as *Phocaeicola vulgatus*, which can directly interact with mucosal cells. The proteolytic enzymes produced by these pathobionts would contribute to increased epithelial damage, reduced intestinal barrier integrity, and enhanced bacterial translocation across the epithelial layer, thereby contributing to the inflammation characteristic of IBD.

Table 4-1: Participant and sample summary at time of initial diagnostic colonoscopy.

sd: standard deviation; PC: proximal colon; DC: distal colon; CD: Crohn's disease; UC: ulcerative colitis; L1: distal 1/3 ileal +/- limited cecal disease; L2: colonic; L3: ileocolonic; E1: ulcerative proctitis; E2: left-sided UC; E3: extensive UC; E4: pancolitis.

	Non-IBD	**CD**	**UC**	**Significance**
Participants (n)	16	21	14	NA
Age (s.d.)	13.4 (4.2)	14.0 (2.6)	12.9 (3.5)	p = 0.74
Sex (% female)	7 (44%)	14 (66%)	6 (43%)	p = 0.27
Disease activity (based on PCDAI or PUCAI)	NA	Mild: 2 Moderate: 9 Severe: 3	Mild: 2 Moderate: 9 Severe: 3	NA
Inflammation location	NA	L1: 9 L2: 5 L3: 7	E1: 1 E2: 3 E3: 7 E4: 3	
Number of samples per site:				
PC (n, % inflamed)	16 (0, 0%)	18 (7, 39%)	13 (4, 31%)	p = 0.02 (for inflamed between CD & UC)
DC (n, % inflamed)	14 (0, 0%)	17 (8, 47%)	10 (9, 90%)	p = 0.00006 (for inflamed between CD & UC)

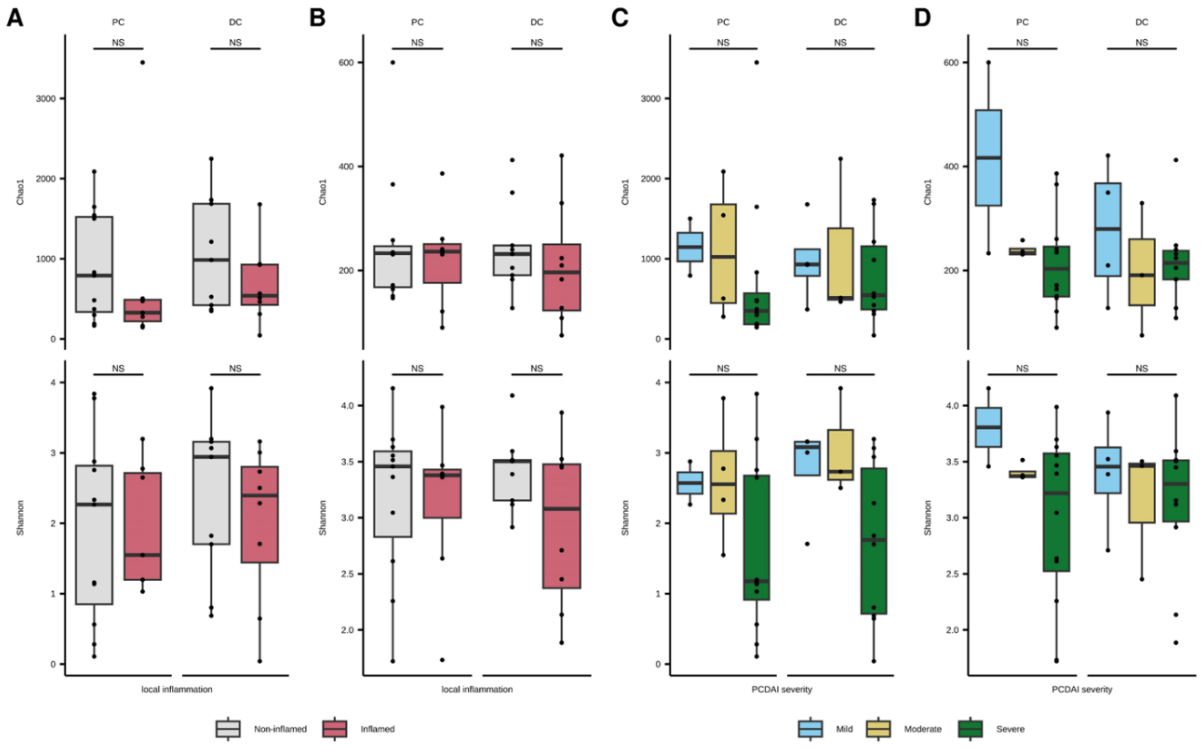
Table 4-2 – Longitudinal sample collection of the mucosal-luminal interface virome.

Collection time reflects the number of weeks from initial diagnostic endoscopy. 5-ASA: mesalamine, 6-MP: mercaptopurine, AZA: azathioprine, MTX: methotrexate, PRED: prednisone, IFX: infliximab.

Diagnosis and subject	Collection time (weeks)	Clinical score (PCDAI or PUCAI)	Sample	Local score (SES-CD or MAYO-UC)	Treatment (induction / maintenance)
CD-A	0	PCDAI 57.5, wPCDAI 72.5 (severe)	PC	0	-
			DC	0	
	17	PCDAI 5, wPCDAI 10 (remission)	PC	0	PRED / AZA + 6-MP
			DC	0	
CD-E	0	PCDAI 45, wPCDAI 65 (severe)	PC	0	-
	21	PCDAI 42.5, wPCDAI 57.5 (severe / moderate)	PC	1	PRED / AZA + 6-MP
CD-H	0	PCDAI 30, wPCDAI 50 (moderate)	PC	3	-
			DC	6	
	14	PCDAI 5, wPCDAI 0 (remission)	PC	0	PRED / AZA + 6-MP
CD-I	0	PCDAI 35, wPCDAI 50 (moderate)	PC	0	-
			DC	0	
	21	PCDAI 15, wPCDAI 20 (moderate)	PC	0	PRED / MTX
			DC	0	
CD-M	0	PCDAI 32.5, wPCDAI 47.5 (moderate)	PC	0	-
	33	PCDAI 30, wPCDAI 42.5 (moderate)	PC	0	PRED / MTX
CD-Q	0	PCDAI 60, wPCDAI 77.5 (severe)	PC	5	-
	19	PCDAI 22.5, wPCDAI 35 (mild)	PC	0	IFX / IFX
UC-A	0	PUCAI: 35 (Moderate)	PC	Grade 1	-
	24	PUCAI: 15 (Mild)	PC	Grade 0	
			DC	Grade 2	
UC-G	0	PUCAI: 30 (Mild)	PC	Grade 0	-
			DC	Grade 1	

	17	PUCAI: 40 (Moderate)	PC	Grade 0	PRED / 5-ASA
			DC	Grade 2	
	37	PUCAI: 40 (Moderate)	DC	Grade 2	PRED / 5-ASA
	63	PUCAI: 0 (Remission)	PC	Grade 0	IFX / IFX
			DC	Grade 0	

4.7. Supplementary Material



Supplementary Figure 4-1: Alpha-diversity of the colonic virome and bacteriome by local inflammation.

Alpha-diversity of the colonic virome and bacteriome by local inflammation. Comparisons between inflamed IBD and not-inflamed IBD MLI samples of the virome (A) and bacteriome (B). Comparisons between virome (C) and bacteriome (D) alpha diversity by the Pediatric Crohn's Disease Activity Index (PCDAI) in participants with CD.

The additional tables are available for download among the supplemental data with our published article:

Supplemental Table 1: Significantly differentially abundant ASVs between participant groups after controlling for sex and sampling location

Supplemental Table 2: PERMANOVA results using virome (viral clusters) (a) and bacteriome (b) Bray-Curtis dissimilarities with participants as strata.

Supplemental Table 3: Differential abundance analysis of viral taxa (viral class and viral order) using geNomad annotations, comparing non-IBD, CD, and UC participants

5. Discussion

5.1. Summary

This thesis provides a narrative of my efforts to characterize the mucosa-associated virome in the context of pediatric inflammatory bowel disease. Our laboratory was the first group to publish virome methods on human mucosal-luminal interface samples, which required adapting existing VLP purification protocols for this distinct sample type. Furthermore, these articles offer new insights through multiomic virome analysis including metatranscriptomics and are among the few studies that offer direct comparison between mucosa-associated and stool viromes.

All of our studies, performed with the support of the patients, families, and care teams at the Children's Hospital of Eastern Ontario, used clinically-relevant samples in an effort to better understand how our microbiome affects our human health. These studies were the first published efforts by our laboratory to explore the intestinal virome, and highlight our ongoing multiomic efforts to improve our understanding of the microbiome's role in inflammatory bowel disease by studying important yet underexplored part of our microbiome's ecology.

In Chapter 2, we surveyed the virome processing tools and protocols discussed in the introduction, to develop and optimize a protocol for virome sequencing of mucosal-luminal interface samples. Using a spike-in phage and technical replicates, we showed that these methods were reproducible. Moreover, we were able to process samples obtained from both the proximal and distal colon, demonstrating that MLI sampling enables multi-site study of the intestinal virome in contrast to stool. Using bioinformatic tools that were available in 2018-2019, we designed a custom bioinformatic pipeline to characterize the virome and were able to perform virome-

bacteriome analysis using paired metagenomic sequencing. This paper thus served as a foundation for further virome analyses using this sample type, including the following two chapters.

Next, we subjected the mucosal-luminal interface samples to extremely deep sequencing (>100,000,000 reads per sample) and multiomic approaches (metaviromics, whole metagenomics, and metatranscriptomics) on samples obtained from a cohort of pediatric patients with ulcerative colitis. This work, highlighted in Chapter 3, allowed our laboratory to contribute towards further virome studies, providing practical suggestions including:

- recommending a virome sequencing depth of 10-20 million reads per sample
- establishing multiomic data analysis techniques, including mapping of metatranscriptome reads to metavirome-derived contigs
- supporting the use of flanking regions to predict phage integration, combining whole metagenomics and metaviromics

Furthermore, by studying the MLI virome in comparison to stool, we further validated the utility of MLI sampling, showing that the mucosa-associated virome is more diverse between individuals than the stool virome. We were also able to identify two *Crassvirales* that were abundant in the mucosal virome but unidentifiable from paired stool viromes and subsequently applied our metatranscriptomic data to study and visualize transcription of the *Crassvirales* genome. We were also able to assess the transcriptional activity of viral genes and show that the presence of an integrase as a predictor of temperate phages, though also examine other viral genes as potential markers of integrated phages and transcriptional activity.

Lastly, we applied these our virome sequencing methods to a cohort of over fifty pediatric patients undergoing investigation for inflammatory bowel disease. We were able to show that

mucosa-associated viromes are highly individualized with no core virome identified and demonstrated that *Crassvirales* were decreased in IBD patients compared to non-IBD controls. In this study, we also performed longitudinal MLI virome sequencing which demonstrated the persistence of highly abundant viruses in the mucosal-associated virome that were stable for up to one year. This data sheds light on the mucosa-associated virome in inflammatory bowel diseases and its importance in understanding our overall microbiome-host relationship.

Altogether, this research advances both our ability to characterize the mucosa-associated virome and our understanding of this viral community in human disease. By publishing this work, we hope to share our methods and findings with other microbiome-focused laboratories from both our wet and dry lab experiences. By uploading our virome sequencing data to public data repositories, we have also directly contributed to global efforts to catalogue the human virome. Below, I outline some of the next steps in human virome research.

5.2. Future work

Given the high variability of human viromes, most studies including the work presented in this thesis are relatively underpowered for clinical studies. By using the MLI sequencing, which provides a higher resolution tool that stool sequencing, studies could employ more samples with relevant clinical metadata (i.e. macroscopic inflammation, pathological findings, clinical epidemiology) that would provide further insight into our virome in intestinal diseases. MLI sampling is scalable to incorporate additional sites (rectum, transverse colon, terminal ileum, and upper GI tract) which could provide further biogeographical resolution of the human virome. By using non-inflamed sites as a control for inflamed sites, MLI sampling may improve our ability to

propose mechanisms of host-virome interactions that are otherwise obscured by high personal variability.

The main challenge for proximal GI tract sampling is reduced viral and microbial load. In data described in Chapter 2 and unpublished data, terminal ileum aspirates often yielded insufficient nucleic acid for sequencing even when subjected to amplification techniques. Novel sequencing techniques that allow for lower (or "ultra-low") input DNA workflows could enable MLI sequencing of more proximal sites in the future. The terminal ileum, one of the commonly affected sites in Crohn's disease, would be of high interest for human virome studies in IBD.

The increasing use of long-read sequencing technology also has significant promise in metaviromics by improving virome genome assembly. These efforts are further supported by new bioinformatic tools and assembly algorithms that can merge short and long read datasets. Improvements in virome genomes will enhance virome databases and annotation, especially with multiomic analysis such as metatranscriptomics.

Efforts to use amplification-free sequencing, to reduce the bias introduced by techniques such as multiple displacement amplification, are also a key frontier in virome studies which will enable more interstudy comparisons. While many of these approaches were not readily available at the start of this thesis, these techniques would likely be easily adopted for MLI samples. Bioinformatic tools, including those for viral contig/genome assembly, viral contig/genome annotation, viral gene annotation, and host-phage prediction, continue to improve and could both improve future studies and be retroactively used to reanalyze existing datasets such as those presented in this thesis.

In all – the work presented in this thesis, from conceptualization to protocol development to clinical study, all took place in the setting of a rapidly developing field. Several of the bioinformatic tools and databases used in Chapter 4 were not yet available when I was working on Chapter 2, highlighting the expansion of virome studies. I am excited – and will be humbled – to see what the next ten years will bring to human virome research.

6. References

1 Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59-65, doi:10.1038/nature08821 (2010).

2 Franzosa, E. A. *et al.* Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nature Microbiology* **4**, doi:10.1038/s41564-018-0306-4 (2019).

3 de Vos, W. M., Tilg, H., Van Hul, M. & Cani, P. D. Gut microbiome and health: mechanistic insights. *Gut* **71**, 1020, doi:10.1136/gutjnl-2021-326789 (2022).

4 Rastelli, M., Cani, P. D. & Knauf, C. The Gut Microbiome Influences Host Endocrine Functions. *Endocrine Reviews* **40**, 1271-1284, doi:10.1210/er.2018-00280 (2019).

5 Turnbaugh, P. J. *et al.* The Human Microbiome Project. *Nature* **449**, 804-810, doi:10.1038/nature06244 (2007).

6 Zou, S., Caler, L., Colombini-Hatch, S., Glynn, S. & Srinivas, P. Research on the human virome: where are we and what is next. *Microbiome* **4**, 32, doi:10.1186/s40168-016-0177-y (2016).

7 Mokili, J. L., Rohwer, F. & Dutilh, B. E. Metagenomics and future perspectives in virus discovery. *Current Opinion in Virology* **2**, 63-77, doi:10.1016/j.coviro.2011.12.004 (2012).

8 Clooney, A. G. *et al.* Whole-Virome Analysis Sheds Light on Viral Dark Matter in Inflammatory Bowel Disease. *Cell Host Microbe* **26**, 764-778 e765, doi:10.1016/j.chom.2019.10.009 (2019).

9 Santiago-Rodriguez, T. M. & Hollister, E. B. Unraveling the viral dark matter through viral metagenomics. *Front Immunol* **13**, 1005107, doi:10.3389/fimmu.2022.1005107 (2022).

10 Reyes, A., Semenkovich, N. P., Whiteson, K., Rohwer, F. & Gordon, J. I. Going viral: next-generation sequencing applied to phage populations in the human gut. *Nature Reviews Microbiology* **10**, 607-617, doi:10.1038/nrmicro2853 (2012).

11 Burrell, C. J., Howard, C. R. & Murphy, F. A. History and Impact of Virology. *Fenner and White's Medical Virology*, 3-14 (2017).

12 Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265-269, doi:10.1038/s41586-020-2008-3 (2020).

13 Colson, P. *et al.* Ultrarapid diagnosis, microscope imaging, genome sequencing, and culture isolation of SARS-CoV-2. *Eur J Clin Microbiol Infect Dis* **39**, 1601-1603, doi:10.1007/s10096-020-03869-w (2020).

14 Wrapp, D. *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260-1263, doi:doi:10.1126/science.abb2507 (2020).

15 Anderson, N. G., Gerin, J. L. & Anderson, N. L. Global Screening for Human Viral Pathogens. *Emerging Infectious Diseases* **9**, 768-773, doi:doi:10.3201/eid0907.030004. (2003).

16 Breitbart, M. *et al.* Metagenomic Analyses of an Uncultured Viral Community from Human Feces. *Journal of Bacteriology* **185**, 6220-6223, doi:doi:10.1128/jb.185.20.6220-6223.2003 (2003).

17 Breitbart, M. *et al.* Viral diversity and dynamics in an infant gut. *Research in Microbiology* **159**, 367-373, doi:https://doi.org/10.1016/j.resmic.2008.04.006 (2008).

18 Mirzaei, M. K. *et al.* Challenges of Studying the Human Virome – Relevant Emerging Technologies. *Trends in Microbiology* **29**, 171-181, doi:https://doi.org/10.1016/j.tim.2020.05.021 (2021).

19 Hobbs, Z. & Abedon, S. T. Diversity of phage infection types and associated terminology: the problem with 'Lytic or lysogenic'. *FEMS Microbiology Letters* **363**, doi:10.1093/femsle/fnw047 (2016).

20 Liang, G. & Bushman, F. D. The human virome: assembly, composition and host interactions. *Nature Reviews Microbiology*, doi:10.1038/s41579-021-00536-5 (2021).

21 Pride, D. T. *et al.* Evidence of a robust resident bacteriophage population revealed through analysis of the human salivary virome. *The ISME Journal* **6**, 915-926, doi:10.1038/ismej.2011.169 (2012).

22 Fiers, W. *et al.* Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260**, 500-507, doi:10.1038/260500a0 (1976).

23 Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463-5467, doi:10.1073/pnas.74.12.5463 (1977).

24 Haynes, M. & Rohwer, F. The Human Virome. *Metagenomics of the Human Body*, 63-77 (2010).

25 Palacios, G. *et al.* Panmicrobial oligonucleotide array for diagnosis of infectious diseases. *Emerg Infect Dis* **13**, 73-81, doi:10.3201/eid1301.060837 (2007).

26 Wang, D. *et al.* Microarray-based detection and genotyping of viral pathogens. *Proc Natl Acad Sci U S A* **99**, 15687-15692, doi:10.1073/pnas.242579699 (2002).

27 Wylie, T. N., Wylie, K. M., Herter, B. N. & Storch, G. A. Enhanced virome sequencing using targeted sequence capture. *Genome Res* **25**, 1910-1920, doi:10.1101/gr.191049.115 (2015).

28 Briese, T. *et al.* Virome Capture Sequencing Enables Sensitive Viral Diagnosis and Comprehensive Virome Analysis. *mBio* **6**, e01491-01415, doi:10.1128/mBio.01491-15 (2015).

29 Goodrich, Julia K. *et al.* Conducting a Microbiome Study. *Cell* **158**, 250-262, doi:10.1016/j.cell.2014.06.037 (2014).

30 Rohwer, F. & Edwards, R. The Phage Proteomic Tree: a genome-based taxonomy for phage. *J Bacteriol* **184**, 4529-4535, doi:10.1128/jb.184.16.4529-4535.2002 (2002).

31 Wang, D. 5 challenges in understanding the role of the virome in health and disease. *PLoS Pathog* **16**, e1008318, doi:10.1371/journal.ppat.1008318 (2020).

32 Norman, J. M. *et al.* Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* **160**, 447-460, doi:10.1016/j.cell.2015.01.002 (2015).

33 Reyes, A. *et al.* Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466**, 334-338, doi:10.1038/nature09199 (2010).

34 Reyes, A., Wu, M., McNulty, N. P., Rohwer, F. L. & Gordon, J. I. Gnotobiotic mouse model of phage–bacterial host dynamics in the human gut. *Proceedings of the National Academy of Sciences* **110**, 20236-20241, doi:10.1073/pnas.1319470110 (2013).

35 Minot, S. *et al.* The human gut virome: Inter-individual variation and dynamic response to diet. *Genome Research* **21**, 1616-1625, doi:10.1101/gr.122705.111 (2011).

36 Minot, S. & Bryson, A. Rapid evolution of the human gut virome. *Proceedings of the ...* **110**, 12450-12455, doi:10.1073/pnas.1300833110/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1300833110 (2013).

37 Kleiner, M., Hooper, L. V. & Duerkop, B. A. Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes. *BMC Genomics* **16**, 7-7, doi:10.1186/s12864-014-1207-4 (2015).

38 Duerkop, B. A. Bacteriophages shift the focus of the mammalian microbiota. *PLOS Pathogens* **14**, e1007310, doi:10.1371/journal.ppat.1007310 (2018).

39 Zuo, T. *et al.* Bacteriophage transfer during faecal microbiota transplantation in *Clostridium difficile* infection is associated with treatment outcome. *Gut*, gutjnl-2017-313952, doi:10.1136/gutjnl-2017-313952 (2017).

40 Shkoporov, A. N. *et al.* Reproducible protocols for metagenomic analysis of human faecal phageomes. *Microbiome*, 1-17 (2018).

41 Conceição-Neto, N. *et al.* Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. *Scientific reports* **5**, 16532-16532, doi:10.1038/srep16532 (2015).

42 IBDMDB Team. Viromics HMP2 Protocol. (2018).

43 Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655-662, doi:10.1038/s41586-019-1237-9 (2019).

44 Kohl, C. *et al.* Protocol for metagenomic virus detection in clinical specimens. *Emerg Infect Dis* **21**, 48-57, doi:10.3201/eid2101.140766 (2015).

45 Shkoporov, A. N. *et al.* The Human Gut Virome Is Highly Diverse , Stable , and Individual Specific. *Cell Host and Microbe* **26**, 527-541.e525, doi:10.1016/j.chom.2019.09.009 (2019).

46 Colombet, J. *et al.* Virioplankton 'pegylation': Use of PEG (polyethylene glycol) to concentrate and purify viruses in pelagic ecosystems. *Journal of Microbiological Methods* **71**, 212-219, doi:https://doi.org/10.1016/j.mimet.2007.08.012 (2007).

47 Castro-Mejía, J. L. *et al.* Optimizing protocols for extraction of bacteriophages prior to metagenomic analyses of phage communities in the human gut. *Microbiome* **3**, 64, doi:10.1186/s40168-015-0131-4 (2015).

48 Thurber, R. V., Haynes, M., Breitbart, M., Wegley, L. & Rohwer, F. Laboratory procedures to generate viral metagenomes. *Nature protocols* **4**, 470-483, doi:10.1038/nprot.2009.10 (2009).

49 Brussaard, C. P., Marie, D. & Bratbak, G. Flow cytometric detection of viruses. *J Virol Methods* **85**, 175-182, doi:10.1016/s0166-0934(99)00167-6 (2000).

50 Andrews-Pfannkoch, C., Fadrosh, D. W., Thorpe, J. & Williamson, S. J. Hydroxyapatite-mediated separation of double-stranded DNA, single-stranded DNA, and RNA genomes from natural viral assemblages. *Appl Environ Microbiol* **76**, 5039-5045, doi:10.1128/aem.00204-10 (2010).

51 Džunková, M., D'Auria, G. & Moya, A. Direct sequencing of human gut virome fractions obtained by flow cytometry. *Frontiers in Microbiology* **6**, doi:10.3389/fmicb.2015.00955 (2015).

52 Dean, F. B. *et al.* Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci U S A* **99**, 5261-5266, doi:10.1073/pnas.082089499 (2002).

53 Kim, K.-H. & Bae, J.-W. Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Applied and environmental microbiology* **77**, 7663-7668, doi:10.1128/AEM.00289-11 (2011).

54 Lasken, R. S. & Stockwell, T. B. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol* **7**, 19, doi:10.1186/1472-6750-7-19 (2007).

55 Gregory, A. C. *et al.* The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut. *Cell Host & Microbe*, doi:10.1016/j.chom.2020.08.003 (2020).

56 Marine, R. *et al.* Evaluation of a transposase protocol for rapid generation of shotgun high-throughput sequencing libraries from nanogram quantities of DNA. *Appl Environ Microbiol* **77**, 8071-8079, doi:10.1128/aem.05610-11 (2011).

57 Hannigan, G. D. *et al.* The Human Skin Double-Stranded DNA Virome: Topographical and Temporal Diversity, Genetic Enrichment, and Dynamic Associations with the Host Microbiome. *mBio* **6**, e01578-01515, doi:doi:10.1128/mBio.01578-15 (2015).

58 Regnault, B. *et al.* Deep Impact of Random Amplification and Library Construction Methods on Viral Metagenomics Results. *Viruses* **13**, doi:10.3390/v13020253 (2021).

59 Shiwa, Y. *et al.* Evaluation of rRNA depletion methods for capturing the RNA virome from environmental surfaces. *BMC Research Notes* **16**, 142, doi:10.1186/s13104-023-06417-9 (2023).

60 Wetterstrand, K. A. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*, <https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data> (2023).

61 Zablocki, O. *et al.* VirION2: a short- and long-read sequencing and informatics workflow to study the genomic diversity of viruses in nature. *PeerJ* **9**, e11088, doi:10.7717/peerj.11088 (2021).

62 Cook, R. *et al.* Nanopore and Illumina sequencing reveal different viral populations from human gut samples. *Microb Genom* **10**, doi:https://doi.org/10.1099/mgen.0.001236 (2024).

63 Chen, J. *et al.* Efficient Recovery of Complete Gut Viral Genomes by Combined Short- and Long-Read Sequencing. *Advanced Science* **11**, 2305818, doi:https://doi.org/10.1002/advs.202305818 (2024).

64 Zhang, F. *et al.* Critical Assessment of Whole Genome and Viral Enrichment Shotgun Metagenome on the Characterization of Stool Total Virome in Hepatocellular Carcinoma Patients. *Viruses* **15**, doi:10.3390/v15010053 (2022).

65 Wu, L.-Y. *et al.* Benchmarking bioinformatic virus identification tools using real-world metagenomic data across biomes. *Genome Biology* **25**, 97, doi:10.1186/s13059-024-03236-4 (2024).

66 Roux, S., Hallam, S. J., Woyke, T. & Sullivan, M. B. Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *eLife* **4**, e08490, doi:10.7554/eLife.08490 (2015).

67 Bikel, S. *et al.* Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: towards a systems-level understanding of human microbiome. *Comput Struct Biotechnol J* **13**, 390-401, doi:10.1016/j.csbj.2015.06.001 (2015).

68 Poyet, M. *et al.* A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nature Medicine* **25**, doi:10.1038/s41591-019-0559-3 (2019).

69 Santiago-Rodriguez, T. M. *et al.* Transcriptome analysis of bacteriophage communities in periodontal health and disease. *BMC Genomics* **16**, 549, doi:10.1186/s12864-015-1781-0 (2015).

70 Santiago-Rodriguez, T. M. & Hollister, E. B. Multi 'omic data integration: A review of concepts, considerations, and approaches. *Seminars in Perinatology* **45**, 151456, doi:https://doi.org/10.1016/j.semperi.2021.151456 (2021).

71 Integrative, T. & Microbiome, H. The Integrative Human Microbiome Project. 1-8, doi:10.1038/s41586-019-1238-8.

72 Caporaso, J. G., Knight, R. & Kelley, S. T. Host-associated and free-living phage communities differ profoundly in phylogenetic composition. *PLoS One* **6**, e16900, doi:10.1371/journal.pone.0016900 (2011).

73 Aggarwala, V., Liang, G. & Bushman, F. D. Viral communities of the human gut: Metagenomic analysis of composition and dynamics. *Mobile DNA* **8**, 1-10, doi:10.1186/s13100-017-0095-y (2017).

74 Dutilh, B. E. Metagenomic ventures into outer sequence space. *Bacteriophage* **4**, e979664, doi:10.4161/21597081.2014.979664 (2014).

75 McCann, A. *et al.* Viromes of one year old infants reveal the impact of birth mode on microbiome diversity. *PeerJ* **6**, e4694-e4694, doi:10.7717/peerj.4694 (2018).

76 NCBI. *RefSeq Growth Statistics*, <https://www.ncbi.nlm.nih.gov/refseq/statistics/> (2020).

77 Brister, J. R., Ako-Adjei, D., Bao, Y. & Blinkova, O. NCBI Viral Genomes Resource. *Nucleic Acids Research* **43**, D571-D577, doi:10.1093/nar/gku1207 (2015).

78 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.e1099, doi:https://doi.org/10.1016/j.cell.2021.01.029 (2021).

79 Li, J., Yang, F., Xiao, M. & Li, A. Advances and challenges in cataloging the human gut virome. *Cell Host & Microbe* **30**, 908-916, doi:https://doi.org/10.1016/j.chom.2022.06.003 (2022).

80 Paez-Espino, D. *et al.* IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses. *Nucleic Acids Research* **45**, gkw1030, doi:10.1093/nar/gkw1030 (2017).

81 Camargo, A. P. *et al.* IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Research* **51**, D733-D743, doi:10.1093/nar/gkac1037 (2023).

82 Walters, W. A. *et al.* Longitudinal comparison of the developing gut virome in infants and their mothers. *Cell Host & Microbe* **31**, 187-198.e183, doi:https://doi.org/10.1016/j.chom.2023.01.003 (2023).

83 Tisza, M. J. & Buck, C. B. A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases. *Proceedings of the National Academy of Sciences* **118**, e2023202118, doi:10.1073/pnas.2023202118 (2021).

84 Nayfach, S. *et al.* Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nature Microbiology*, doi:10.1038/s41564-021-00928-6 (2021).

85 Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713-714, doi:10.1093/bioinformatics/btn025 (2008).

86 Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421, doi:10.1186/1471-2105-10-421 (2009).

87 Ziemann, M. Accuracy, speed and error tolerance of short DNA sequence aligners. *bioRxiv*, 053686-053686, doi:10.1101/053686 (2016).

88 Bush, S. J., Connor, T. R., Peto, T. E. A., Crook, D. W. & Walker, A. S. Evaluation of methods for detecting human reads in microbial sequencing datasets. *Microb Genom* **6**, doi:10.1099/mgen.0.000393 (2020).

89 Vancuren, S. J. & Hill, J. E. Update on cpnDB: a reference database of chaperonin sequences. *Database* **2019**, doi:10.1093/database/baz033 (2019).

90 Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res* **27**, 824-834, doi:10.1101/gr.213959.116 (2017).

91 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, doi:10.1093/bioinformatics/btv033 (2015).

92 Antipov, D., Raiko, M., Lapidus, A. & Pevzner, P. A. MetaviralSPAdes: assembly of viruses from metagenomic data. *Bioinformatics* **36**, 4126-4129, doi:10.1093/bioinformatics/btaa490 (2020).

93 Sutton, T. D. S., Clooney, A. G., Ryan, F. J., Ross, R. P. & Hill, C. Choice of assembly software has a critical impact on virome characterisation. *Microbiome* **7**, 12, doi:10.1186/s40168-019-0626-5 (2019).

94 Dutilh, B. E. *et al.* Reference-independent comparative metagenomics using cross-assembly: crAss. *Bioinformatics* **28**, 3225-3231, doi:10.1093/bioinformatics/bts613 (2012).

95 Nayfach, S. *et al.* CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nature Biotechnology* **39**, 578-585, doi:10.1038/s41587-020-00774-7 (2021).

96 Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter : mining viral signal from microbial genomic data. 1-20, doi:10.7717/peerj.985 (2015).

97 Li, Y. *et al.* VIP: an integrated pipeline for metagenomics of virus identification and discovery. *Sci Rep* **6**, 23774, doi:10.1038/srep23774 (2016).

98 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, doi:10.1186/s40168-020-00867-0 (2020).

99 Guo, J. *et al.* VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 37, doi:10.1186/s40168-020-00990-y (2021).

100 Roux, S., Tournayre, J., Mahul, A., Debroas, D. & Enault, F. Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinformatics* **15**, 76, doi:10.1186/1471-2105-15-76 (2014).

101 Wang, Q., Jia, P. & Zhao, Z. VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data. *PLoS One* **8**, e64465, doi:10.1371/journal.pone.0064465 (2013).

102 Wommack, K. E. *et al.* VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Standards in genomic sciences* **6**, 427-439, doi:10.4056/sigs.2945050 (2012).

103 Zhao, G. *et al.* VirusSeeker, a computational pipeline for virus discovery and virome composition analysis. *Virology* **503**, 21-30, doi:10.1016/j.virol.2017.01.005 (2017).

104 Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biology* **20**, 257, doi:10.1186/s13059-019-1891-0 (2019).

105 Antipov, D., Rayko, M., Kolmogorov, M. & Pevzner, P. A. viralFlye: assembling viruses and identifying their hosts from long-read metagenomics data. *Genome Biology* **23**, 57, doi:10.1186/s13059-021-02566-x (2022).

106 Li, W. & Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659, doi:10.1093/bioinformatics/btl158 (2006).

107 Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152, doi:10.1093/bioinformatics/bts565 (2012).

108 Roux, S. & Bolduc, B. Stampede-ClusterGenomes v. 2017-10-26 (MAVERICKLab, 2017).

109 Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* **35**, 1026-1028, doi:10.1038/nbt.3988 (2017).

110 Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473, doi:10.1186/s12859-019-3019-7 (2019).

111 Turner, D. *et al.* Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ICTV bacterial viruses subcommittee. *Archives of Virology* **168**, 74, doi:10.1007/s00705-022-05694-2 (2023).

112 Adriaenssens Evelien, M. Phage Diversity in the Human Gut Microbiome: a Taxonomist's Perspective. *mSystems* **6**, 10.1128/msystems.00799-00721, doi:10.1128/msystems.00799-21 (2021).

113 Kim, M.-S., Park, E.-J., Roh, S. W. & Bae, J.-W. Diversity and Abundance of Single-Stranded DNA Viruses in Human Feces. *Applied and Environmental Microbiology* **77**, 8062-8070, doi:doi:10.1128/AEM.06331-11 (2011).

114 Caetano-Anollés, G., Claverie, J.-M. & Nasir, A. A critical analysis of the current state of virus taxonomy. *Frontiers in Microbiology* **14**, doi:10.3389/fmicb.2023.1240993 (2023).

115 Sielemann, K., Hafner, A. & Pucker, B. The reuse of public datasets in the life sciences: potential risks and rewards. *PeerJ* **8**, e9954, doi:10.7717/peerj.9954 (2020).

116 Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**, 69-69, doi:10.1186/s40168-017-0283-5 (2017).

117 Zheng, T. *et al.* Mining , analyzing , and integrating viral signals from metagenomic data. 1-15 (2019).

118 Amgarten, D., Braga, L. P. P., da Silva, A. M. & Setubal, J. C. MARVEL, a Tool for Prediction of Bacteriophage Sequences in Metagenomic Bins. *Front Genet* **9**, 304, doi:10.3389/fgene.2018.00304 (2018).

119 Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119, doi:10.1186/1471-2105-11-119 (2010).

120 Zhu, W., Lomsadze, A. & Borodovsky, M. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Research* **38**, e132-e132, doi:10.1093/nar/gkq275 (2010).

121 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 Research* **45**, D491-D498, doi:10.1093/nar/gkw975 (2017).

122 CUBE, U. o. V. *VOGDB: Virus Orthologous Groups*, <https://vogdb.org/> (2021).

123 Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**, 27-30, doi:10.1093/nar/28.1.27 (2000).

124 Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Research* **49**, D412-D419, doi:10.1093/nar/gkaa913 (2020).

125 Versoza, C. J. & Pfeifer, S. P. Computational Prediction of Bacteriophage Host Ranges. *Microorganisms* **10**, doi:10.3390/microorganisms10010149 (2022).

126 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, doi:10.1093/bioinformatics/btx383 (2017).

127 Gao, N. L. *et al.* MVP: a microbe-phage interaction database. *Nucleic Acids Res* **46**, D700-d707, doi:10.1093/nar/gkx1124 (2018).

128 Rousseau, C., Gonnet, M., Le Romancer, M. & Nicolas, J. CRISPI: A CRISPR interactive database. *Bioinformatics* **25**, 3317-3318, doi:10.1093/bioinformatics/btp586 (2009).

129 Dion, M. B. *et al.* Streamlining CRISPR spacer-based bacterial host predictions to decipher the viral dark matter. *Nucleic Acids Research* **49**, 3127-3138, doi:10.1093/nar/gkab133 (2021).

130 Edgar, R. C. PILER-CR: Fast and accurate identification of CRISPR repeats. *BMC Bioinformatics* **8**, doi:10.1186/1471-2105-8-18 (2007).

131 Couvin, D. *et al.* CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Research* **46**, W246-W251, doi:10.1093/nar/gky425 (2018).

132 Hockenberry, A. J. & Wilke, C. O. BACPHLIP: predicting bacteriophage lifestyle from conserved protein domains. *PeerJ* **9**, e11396-e11396, doi:10.7717/peerj.11396 (2021).

133 Wishart, D. S. *et al.* PHASTEST: faster than PHASTER, better than PHAST. *Nucleic Acids Research* **51**, W443-W450, doi:10.1093/nar/gkad382 (2023).

134 Song, W. *et al.* Prophage Hunter: an integrative hunting tool for active prophages. *Nucleic Acids Research* **47**, W74-W80, doi:10.1093/nar/gkz380 (2019).

135 Khan Mirzaei, M. *et al.* Bacteriophages Isolated from Stunted Children Can Regulate Gut Bacterial Communities in an Age-Specific Manner. *Cell Host Microbe* **27**, 199-212.e195, doi:10.1016/j.chom.2020.01.004 (2020).

136 Hedžet, S., Rupnik, M. & Accetto, T. Broad host range may be a key to long-term persistence of bacteriophages infecting intestinal Bacteroidaceae species. *Scientific Reports* **12**, 21098, doi:10.1038/s41598-022-25636-x (2022).

137 Fernandes, M. A. *et al.* Enteric Virome and Bacterial Microbiota in Children With Ulcerative Colitis and Crohn Disease. *J Pediatr Gastroenterol Nutr* **68**, 30-36, doi:10.1097/mpg.0000000000002140 (2019).

138 Foulongne, V. *et al.* Human skin microbiota: high diversity of DNA viruses identified on the human skin by high throughput sequencing. *PLoS One* **7**, e38499, doi:10.1371/journal.pone.0038499 (2012).

139 Willner, D. *et al.* Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS ONE* **4**, doi:10.1371/journal.pone.0007370 (2009).

140 Megremis, S. *et al.* Respiratory eukaryotic virome expansion and bacteriophage deficiency characterize childhood asthma. *Scientific Reports* **13**, 8319, doi:10.1038/s41598-023-34730-7 (2023).

141 Salabura, A. *et al.* Urinary Tract Virome as an Urgent Target for Metagenomics. *Life (Basel)* **11**, doi:10.3390/life11111264 (2021).

142 Broecker, F., Russo, G., Klumpp, J. & Moelling, K. Stable core virome despite variable microbiome after fecal transfer. *Gut Microbes* **8**, 214-220, doi:10.1080/19490976.2016.1265196 (2017).

143 Garmaeva, S. *et al.* Studying the gut virome in the metagenomic era: challenges and perspectives. *BMC Biology* **17**, doi:10.1186/s12915-019-0704-y (2019).

144 Manrique, P. *et al.* Healthy human gut phageome. *Proceedings of the National Academy of Sciences* **113**, 10400, doi:10.1073/pnas.1601060113 (2016).

145 Stockdale, S. R. *et al.* Interpersonal variability of the human gut virome confounds disease signal detection in IBD. *Communications Biology* **6**, 221, doi:10.1038/s42003-023-04592-w (2023).

146 Zuo, T. *et al.* Human-Gut-DNA Virome Variations across Geography, Ethnicity, and Urbanization. *Cell Host & Microbe* **28**, 741-751.e744, doi:https://doi.org/10.1016/j.chom.2020.08.005 (2020).

147 Honap, T. P. *et al.* Biogeographic study of human gut-associated crAssphage suggests impacts from industrialization and recent expansion. *PLoS One* **15**, e0226930, doi:10.1371/journal.pone.0226930 (2020).

148 Bao, S. *et al.* Viral metagenomics of the gut virome of diarrheal children with Rotavirus A infection. *Gut Microbes* **15**, 2234653, doi:10.1080/19490976.2023.2234653 (2023).

149 Zhu, Y., Shang, J., Peng, C. & Sun, Y. Phage family classification under Caudoviricetes: A review of current tools using the latest ICTV classification framework. *Front Microbiol* **13**, 1032186, doi:10.3389/fmicb.2022.1032186 (2022).

150 Gulyaeva, A. *et al.* Diversity and Ecology of Caudoviricetes Phages with Genome Terminal Repeats in Fecal Metagenomes from Four Dutch Cohorts. *Viruses* **14**, 2305 (2022).

151 Mäntynen, S., Laanto, E., Oksanen, H. M., Poranen, M. M. & Díaz-Muñoz, S. L. Black box of phage-bacterium interactions: exploring alternative phage infection strategies. *Open Biol* **11**, 210188-210188, doi:10.1098/rsob.210188 (2021).

152 Sanchez-Torres, V., Kirigo, J. & Wood, T. K. Implications of lytic phage infections inducing persistence. *Curr Opin Microbiol* **79**, 102482, doi:10.1016/j.mib.2024.102482 (2024).

153 Sutcliffe, S. G., Reyes, A. & Maurice, C. F. Bacteriophages playing nice: Lysogenic bacteriophage replication stable in the human gut microbiota. *iScience* **26**, 106007, doi:10.1016/j.isci.2023.106007 (2023).

154 Dutilh, B. E. *et al.* A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nature Communications* **5**, doi:10.1038/ncomms5498 (2014).

155 Edwards, R. A. *et al.* Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. *Nature Microbiology* **4**, 1727-1736, doi:10.1038/s41564-019-0494-6 (2019).

156 Stachler, E. *et al.* Quantitative CrAssphage PCR Assays for Human Fecal Pollution Measurement. *Environmental Science & Technology* **51**, 9146-9154, doi:10.1021/acs.est.7b02703 (2017).

157 Park, G. W. *et al.* CrAssphage as a Novel Tool to Detect Human Fecal Contamination on Environmental Surfaces and Hands. *Emerg Infect Dis* **26**, 1731-1739, doi:10.3201/eid2608.200346 (2020).

158 Langeveld, J. *et al.* Normalisation of SARS-CoV-2 concentrations in wastewater: The use of flow, electrical conductivity and crAssphage. *Science of The Total Environment* **865**, 161196, doi:https://doi.org/10.1016/j.scitotenv.2022.161196 (2023).

159 Wilder, M. L. *et al.* Co-quantification of crAssphage increases confidence in wastewater-based epidemiology for SARS-CoV-2 in low prevalence areas. *Water Res X* **11**, 100100, doi:10.1016/j.wroa.2021.100100 (2021).

160 Greenwald, H. D. *et al.* Tools for interpretation of wastewater SARS-CoV-2 temporal and spatial trends demonstrated with data collected in the San Francisco Bay Area. *Water Res X* **12**, 100111, doi:10.1016/j.wroa.2021.100111 (2021).

161 Ramos-Barbero, M. D. *et al.* Characterization of crAss-like phage isolates highlights Crassvirales genetic heterogeneity and worldwide distribution. *Nature Communications* **14**, 4295, doi:10.1038/s41467-023-40098-z (2023).

162 Shkoporov, A. N. *et al.* Long-term persistence of crAss-like phage crAss001 is associated with phase variation in Bacteroides intestinalis. *BMC Biology* **19**, 163, doi:10.1186/s12915-021-01084-3 (2021).

163 Koonin, E. V. & Yutin, N. The crAss-like Phage Group: How Metagenomics Reshaped the Human Virome. *Trends in Microbiology* **28**, 349-359, doi:https://doi.org/10.1016/j.tim.2020.01.010 (2020).

164 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.e656, doi:10.1016/j.chom.2018.10.002 (2018).

165 Yutin, N. *et al.* Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features. *Nature Communications* **12**, 1044, doi:10.1038/s41467-021-21350-w (2021).

166 Bayfield, O. W. *et al.* Structural atlas of a human gut crassvirus. *Nature* **617**, 409-416, doi:10.1038/s41586-023-06019-2 (2023).

167 Guerin, E. *et al.* Isolation and characterisation of ΦcrAss002, a crAss-like phage from the human gut that infects Bacteroides xylanisolvens. *Microbiome* **9**, 89, doi:10.1186/s40168-021-01036-7 (2021).

168 Spencer, L., Olawuni, B. & Singh, P. Gut Virome: Role and Distribution in Health and Gastrointestinal Diseases. *Front Cell Infect Microbiol* **12**, 836706, doi:10.3389/fcimb.2022.836706 (2022).

169 Lecuit, M. & Eloit, M. The human virome: new tools and concepts. *Trends Microbiol* **21**, 510-515, doi:10.1016/j.tim.2013.07.001 (2013).

170 Lim, E. S. *et al.* Early life dynamics of the human gut virome and bacterial microbiome in infants. *Nature medicine* **21**, 1228-1234, doi:10.1038/nm.3950 (2015).

171 Liang, G. *et al.* Dynamics of the Stool Virome in Very Early-Onset Inflammatory Bowel Disease. *Journal of Crohn's and Colitis* **14**, 1600-1610, doi:10.1093/ecco-jcc/jjaa094 (2020).

172 Rivera-Gutiérrez, X. *et al.* The fecal and oropharyngeal eukaryotic viromes of healthy infants during the first year of life are personal. *Scientific Reports* **13**, 938, doi:10.1038/s41598-022-26707-9 (2023).

173 Liang, G. *et al.* The stepwise assembly of the neonatal virome is modulated by breastfeeding. *Nature* **581**, 470-474, doi:10.1038/s41586-020-2192-1 (2020).

174 Shah, S. A. *et al.* Expanding known viral diversity in the healthy infant gut. *Nature Microbiology* **8**, 986-998, doi:10.1038/s41564-023-01345-7 (2023).

175 Pannaraj, P. S. *et al.* Shared and Distinct Features of Human Milk and Infant Stool Viromes. *Frontiers in Microbiology* **9**, doi:10.3389/fmicb.2018.01162 (2018).

176 Wang, J. *et al.* Maternal and neonatal viromes indicate the risk of offspring's gastrointestinal tract exposure to pathogenic viruses of vaginal origin during delivery. *mLife* **1**, 303-310, doi:10.1002/mlf2.12034 (2022).

177 Maqsood, R. *et al.* Discordant transmission of bacteria and viruses from mothers to babies at birth. *Microbiome* **7**, 156, doi:10.1186/s40168-019-0766-7 (2019).

178 Zhang, Y. & Wang, R. The human gut phageome: composition, development, and alterations in disease. *Front Microbiol* **14**, 1213625, doi:10.3389/fmicb.2023.1213625 (2023).

179 Garmaeva, S. *et al.* Stability of the human gut virome and effect of gluten-free diet. *Cell Reports* **35**, 109132, doi:https://doi.org/10.1016/j.celrep.2021.109132 (2021).

180 Johansen, J. *et al.* Centenarians have a diverse gut virome with the potential to modulate metabolism and promote healthy lifespan. *Nature Microbiology* **8**, 1064-1078, doi:10.1038/s41564-023-01370-6 (2023).

181 Sabbaghian, M., Gheitasi, H., Shekarchi, A. A., Tavakoli, A. & Poortahmasebi, V. The mysterious anelloviruses: investigating its role in human diseases. *BMC Microbiol* **24**, 40, doi:10.1186/s12866-024-03187-7 (2024).

182 Cao, Z. *et al.* The gut virome: A new microbiome component in health and disease. *eBioMedicine* **81**, doi:10.1016/j.ebiom.2022.104113 (2022).

183 Pavia, G., Marascio, N., Matera, G. & Quirino, A. Does the Human Gut Virome Contribute to Host Health or Disease? *Viruses* **15**, 2271 (2023).

184 Dery, K. J., Górski, A., Międzybrodzki, R., Farmer, D. G. & Kupiec-Weglinski, J. W. Therapeutic Perspectives and Mechanistic Insights of Phage Therapy in Allotransplantation. *Transplantation* **105** (2021).

185 Virgin, H. W. The virome in mammalian physiology and disease. *Cell* **157**, 142-150, doi:10.1016/j.cell.2014.02.032 (2014).

186 Gogokhia, L. *et al.* Expansion of Bacteriophages Is Linked to Aggravated Intestinal Inflammation and Colitis. *Cell Host Microbe* **25**, 285-299.e288, doi:10.1016/j.chom.2019.01.008 (2019).

187 Tetz, G. V. *et al.* Bacteriophages as potential new mammalian pathogens. *Sci Rep* **7**, 7043, doi:10.1038/s41598-017-07278-6 (2017).

188 Popescu, M., Van Belleghem, J. D., Khosravi, A. & Bollyky, P. L. Bacteriophages and the Immune System. *Annual Review of Virology* **8**, 415-435, doi:https://doi.org/10.1146/annurev-virology-091919-074551 (2021).

189 Pargin, E. *et al.* The human gut virome: composition, colonization, interactions, and impacts on human health. *Frontiers in Microbiology* **14**, doi:10.3389/fmicb.2023.963173 (2023).

190 Shkoporov, A. N. & Hill, C. Review Bacteriophages of the Human Gut : The ‘‘ Known Unknown ’’ of the Microbiome. *Cell Host and Microbe* **25**, 195-209, doi:10.1016/j.chom.2019.01.017 (2019).

191 Silveira, C. B. & Rohwer, F. L. Piggyback-the-Winner in host-associated microbial communities. *npj Biofilms and Microbiomes* **2**, 16010, doi:10.1038/npjbiofilms.2016.10 (2016).

192 Kirsch, J. M. *et al.* Bacteriophage-Bacteria Interactions in the Gut: From Invertebrates to Mammals. *Annu Rev Virol* **8**, 95-113, doi:10.1146/annurev-virology-091919-101238 (2021).

193 Lourenço, M. *et al.* The Spatial Heterogeneity of the Gut Limits Predation and Fosters Coexistence of Bacteria and Bacteriophages. *Cell Host & Microbe* **28**, 390-401.e395, doi:10.1016/j.chom.2020.06.002 (2020).

194 Sutcliffe, S. G., Shamash, M., Hynes, A. P. & Maurice, C. F. Common Oral Medications Lead to Prophage Induction in Bacterial Isolates from the Human Gut. *Viruses* **13**, 455 (2021).

195 Górska, A. *et al.* Dynamics of the human gut phageome during antibiotics treatment. *Computational Biology and Chemistry* **74**, 420-427, doi:10.1016/j.compbiolchem.2018.03.011 (2018).

196 Łusiak-Szelachowska, M., Weber-Dąbrowska, B., Żaczek, M., Borysowski, J. & Górski, A. The Presence of Bacteriophages in the Human Body: Good, Bad or Neutral? *Microorganisms* **8**, doi:10.3390/microorganisms8122012 (2020).

197 Zhao, G. *et al.* Intestinal virome changes precede autoimmunity in type I diabetes-susceptible children. *Proceedings of the National Academy of Sciences* **114**, E6166-E6175, doi:10.1073/pnas.1706359114 (2017).

198 Nakatsu, G. *et al.* Alterations in Enteric Virome Are Associated With Colorectal Cancer and Survival Outcomes. *Gastroenterology* **155**, 529-541.e525, doi:10.1053/j.gastro.2018.04.018 (2018).

199 Imai, T. *et al.* Features of the gut prokaryotic virome of Japanese patients with Crohn's disease. *Journal of Gastroenterology* **57**, 559-570, doi:10.1007/s00535-022-01882-8 (2022).

200 Chen, C. *et al.* Alterations of the gut virome in patients with systemic lupus erythematosus. *Frontiers in Immunology* **13**, doi:10.3389/fimmu.2022.1050895 (2023).

201 Massimino, L., Lovisa, S., Antonio Lamparelli, L., Danese, S. & Ungaro, F. Gut eukaryotic virome in colorectal carcinogenesis: Is that a trigger? *Comput Struct Biotechnol J* **19**, 16-28, doi:10.1016/j.csbj.2020.11.055 (2021).

202 Rasmussen, T. S. *et al.* Faecal virome transplantation decreases symptoms of type 2 diabetes and obesity in a murine model. *Gut*, gutjnl-2019-2320, doi:10.1136/gutjnl-2019-320005 (2020).

203 Coward, S. *et al.* The 2023 Impact of Inflammatory Bowel Disease in Canada: Epidemiology of IBD. *J Can Assoc Gastroenterol* **6**, S9-s15, doi:10.1093/jcag/gwad004 (2023).

204 Kaplan, G. G. The global burden of IBD: from 2015 to 2025. *Nature Reviews Gastroenterology & Hepatology* **12**, 720-727, doi:10.1038/nrgastro.2015.150 (2015).

205 McDowell, C., Farooq, U. & Haseeb, M. *Inflammatory Bowel Disease*, <https://www.ncbi.nlm.nih.gov/books/NBK470312/> (2023).

206 Xavier, R. J. & Podolsky, D. K. Unravelling the pathogenesis of inflammatory bowel disease. *Nature* **448**, 427-434, doi:10.1038/nature06005 (2007).

207 Shouval, D. S. & Rufo, P. A. The Role of Environmental Factors in the Pathogenesis of Inflammatory Bowel Diseases. *JAMA Pediatrics* **171**, doi:10.1001/jamapediatrics.2017.2571 (2017).

208 Sartor, R. B. & Wu, G. D. Roles for Intestinal Bacteria, Viruses, and Fungi in Pathogenesis of Inflammatory Bowel Diseases and Therapeutic Approaches. *Gastroenterology* **152**, 327-339.e324, doi:10.1053/j.gastro.2016.10.012 (2017).

209 Li, J., Butcher, J., Mack, D. & Stintzi, A. Functional impacts of the intestinal microbiome in the pathogenesis of inflammatory bowel disease. *Inflammatory Bowel Diseases* **21**, 139-153, doi:10.1097/MIB.0000000000000215 (2015).

210 Sheehan, D., Moran, C. & Shanahan, F. The microbiota in inflammatory bowel disease. *Journal of Gastroenterology* **50**, 495-507, doi:10.1007/s00535-015-1064-1 (2015).

211 Lepage, P. *et al.* Dysbiosis in inflammatory bowel disease: a role for bacteriophages? *Gut* **57**, 424-425, doi:10.1136/gut.2007.134668 (2008).

212 Wagner, J. *et al.* Bacteriophages in gut samples from pediatric Crohn's disease patients: Metagenomic analysis using 454 pyrosequencing. *Inflammatory Bowel Diseases* **19**, 1598-1608, doi:10.1097/MIB.0b013e318292477c (2013).

213 Zuo, T. *et al.* Gut mucosal virome alterations in ulcerative colitis. 1-11, doi:10.1136/gutjnl-2018-318131 (2019).

214 Jansen, D. *et al.* Community Types Of The Human Gut Virome Are Associated With Endoscopic Outcome In Ulcerative Colitis. *Journal of Crohn's and Colitis*, jjad061, doi:10.1093/ecco-jcc/jjad061 (2023).

215 Clooney, A. G. *et al.* Comparing apples and oranges?: Next generation sequencing and its impact on microbiome analysis. *PLoS ONE* **11**, 1-16, doi:10.1371/journal.pone.0148028 (2016).

216 The promise of phages. *Nature Biotechnology* **41**, 583-583, doi:10.1038/s41587-023-01807-7 (2023).

217 Pirnay, J.-P. *et al.* Personalized bacteriophage therapy outcomes for 100 consecutive cases: a multicentre, multinational, retrospective observational study. *Nature Microbiology* **9**, 1434-1453, doi:10.1038/s41564-024-01705-x (2024).

218 Al-Anany, A. M. *et al.* Phage Therapy in the Management of Urinary Tract Infections: A Comprehensive Systematic Review. *Phage (New Rochelle)* **4**, 112-127, doi:10.1089/phage.2023.0024 (2023).

219 Hyman, P. Phages for Phage Therapy: Isolation, Characterization, and Host Range Breadth. *Pharmaceuticals (Basel)* **12**, doi:10.3390/ph12010035 (2019).

220 Fujimoto, K. *et al.* Metagenome Data on Intestinal Phage-Bacteria Associations Aids the Development of Phage Therapy against Pathobionts. *Cell Host Microbe* **28**, 380-389 e389, doi:10.1016/j.chom.2020.06.005 (2020).

221 Febvre, H. P. *et al.* PHAGE Study: Effects of Supplemental Bacteriophage Intake on Inflammation and Gut Microbiota in Healthy Adults. *Nutrients* **11**, 666 (2019).

222 Nood, E. v. *et al.* Duodenal Infusion of Donor Feces for Recurrent Clostridium difficile. *New England Journal of Medicine* **368**, 407-415, doi:doi:10.1056/NEJMoa1205037 (2013).

223 Colman, R. J. & Rubin, D. T. Fecal microbiota transplantation as therapy for inflammatory bowel disease: A systematic review and meta-analysis. *Journal of Crohn's and Colitis* **8**, 1569-1581, doi:10.1016/j.crohns.2014.08.006 (2014).

224 Draper, L. A. *et al.* Long-term colonisation with donor bacteriophages following successful faecal microbial transplantation. 1-9 (2018).

225 Ott, S. J. *et al.* Efficacy of Sterile Fecal Filtrate Transfer for Treating Patients With Clostridium difficile Infection. *Gastroenterology* **152**, 799-811, doi:10.1053/j.gastro.2016.11.010 (2017).

226 Rasmussen, T. S. *et al.* Fecal virome transfer improves proliferation of commensal gut Akkermansia muciniphila and unexpectedly enhances the fertility rate in laboratory mice. *Gut Microbes* **15**, 2208504, doi:10.1080/19490976.2023.2208504 (2023).

227 Manrique, P. *et al.* Gut bacteriophage dynamics during fecal microbial transplantation in subjects with metabolic syndrome. *Gut Microbes* **13**, 1897217, doi:10.1080/19490976.2021.1897217 (2021).

228 Duan, Y., Young, R. & Schnabl, B. Bacteriophages and their potential for treatment of gastrointestinal diseases. *Nature Reviews Gastroenterology & Hepatology* **19**, 135-144, doi:10.1038/s41575-021-00536-z (2022).

229 Moye, Z. D., Woolston, J. & Sulakvelidze, A. Bacteriophage Applications for Food Production and Processing. *Viruses* **10**, doi:10.3390/v10040205 (2018).

230 Raeisi, H. *et al.* Emerging applications of phage therapy and fecal virome transplantation for treatment of Clostridioides difficile infection: challenges and perspectives. *Gut Pathogens* **15**, 21, doi:10.1186/s13099-023-00550-3 (2023).

231 Shkoporov, A. N., Turkington, C. J. & Hill, C. Mutualistic interplay between bacteriophages and bacteria in the human gut. *Nature Reviews Microbiology* **20**, 737-749, doi:10.1038/s41579-022-00755-4 (2022).

232 Barr, J. J. *et al.* Bacteriophage adhering to mucus provide a non–host-derived immunity. *Proceedings of the National Academy of Sciences* **110**, 10771, doi:10.1073/pnas.1305923110 (2013).

233 Barr, J. J. *et al.* Subdiffusive motion of bacteriophage in mucosal surfaces increases the frequency of bacterial encounters. *Proc Natl Acad Sci U S A* **112**, 13675-13680, doi:10.1073/pnas.1508355112 (2015).

234 Almeida, G. M. F., Laanto, E., Ashrafi, R. & Sundberg, L.-R. Bacteriophage Adherence to Mucus Mediates Preventive Protection against Pathogenic Bacteria. *mBio* **10**, e01984-01919, doi:10.1128/mBio.01984-19 (2019).

235 Townsend, E. M. *et al.* The Human Gut Phageome: Origins and Roles in the Human Gut Microbiome. *Frontiers in Cellular and Infection Microbiology* **11**, doi:10.3389/fcimb.2021.643214 (2021).

236 Eckburg, P. B. *et al.* Diversity of the human intestinal microbial flora. *Science* **308**, 1635-1638, doi:1110591 [pii] 10.1126/science.1110591 (2005).

237 Lepage, P. *et al.* Biodiversity of the mucosa-associated microbiota is stable along the distal digestive tract in healthy individuals and patients with IBD. *Inflamm Bowel Dis* **11**, 473-480, doi:10.1097/01.mib.0000159662.62651.06 (2005).

238 Zoetendal, E. G., Rajilić-Stojanović, M. & De Vos, W. M. High-throughput diversity and functionality analysis of the gastrointestinal tract microbiota. *Gut* **57**, 1605-1615, doi:10.1136/gut.2007.133603 (2008).

239 Durbán, A. *et al.* Assessing Gut Microbial Diversity from Feces and Rectal Mucosa. *Microbial Ecology* **61**, 123-133, doi:10.1007/s00248-010-9738-y (2011).

240 Zoetendal Erwin, G. *et al.* Mucosa-Associated Bacteria in the Human Gastrointestinal Tract Are Uniformly Distributed along the Colon and Differ from the Community Recovered from Feces. *Applied and Environmental Microbiology* **68**, 3401-3407, doi:10.1128/AEM.68.7.3401-3407.2002 (2002).

241 Lavelle, A. *et al.* Spatial variation of the colonic microbiota in patients with ulcerative colitis and control volunteers. *Gut* **64**, 1553, doi:10.1136/gutjnl-2014-307873 (2015).

242 Li, X. *et al.* A metaproteomic approach to study human-microbial ecosystems at the mucosal luminal interface. *PLoS One* **6**, e26542, doi:10.1371/journal.pone.0026542 (2011).

243 Starr, A. E. *et al.* Proteomic analysis of ascending colon biopsies from a paediatric inflammatory bowel disease inception cohort identifies protein biomarkers that differentiate Crohn's disease from UC. *Gut* **66**, 1573-1583, doi:10.1136/gutjnl-2015-310705 (2017).

244 Alipour, M. *et al.* Mucosal Barrier Depletion and Loss of Bacterial Diversity are Primary Abnormalities in Paediatric Ulcerative Colitis. *Journal of Crohn's and Colitis* **10**, 462-471, doi:10.1093/ecco-jcc/jjv223 (2016).

245 Jervis-Bardy, J. *et al.* Deriving accurate microbiota profiles from human samples with low bacterial content through post-sequencing processing of Illumina MiSeq data. *Microbiome* **3**, 19, doi:10.1186/s40168-015-0083-8 (2015).

246 Adiliaghdam, F. *et al.* Human enteric viruses autonomously shape inflammatory bowel disease phenotype through divergent innate immunomodulation. *Science immunology* **7**, eabn6660-eabn6660, doi:10.1126/sciimmunol.abn6660 (2022).

247 Mottawea, W. *et al.* The mucosal–luminal interface: an ideal sample to study the mucosa-associated microbiota and the intestinal microbial biogeography. *Pediatric Research*, doi:10.1038/s41390-019-0326-7 (2019).

248 Presley, L. L. *et al.* Host–Microbe Relationships in Inflammatory Bowel Disease Detected by Bacterial and Metaproteomic Analysis of the Mucosal–Luminal Interface. *Inflammatory Bowel Diseases* **18**, 409-417, doi:10.1002/ibd.21793 (2011).

249 Watt, E. *et al.* Extending colonic mucosal microbiome analysis—assessment of colonic lavage as a proxy for endoscopic colonic biopsies. *Microbiome* **4**, 61, doi:10.1186/s40168-016-0207-9 (2016).

250 Schultsz, C., Van Den Berg, F. M., Ten Kate, F. W., Tytgat, G. N. & Dankert, J. The intestinal mucus layer from patients with inflammatory bowel disease harbors high numbers of bacteria compared with controls. *Gastroenterology* **117**, 1089-1097, doi:10.1016/s0016-5085(99)70393-8 (1999).

251 Mottawea, W. *et al.* Altered intestinal microbiota–host mitochondria crosstalk in new onset Crohn's disease. *Nature Communications* **7**, 13419-13419, doi:10.1038/ncomms13419 (2016).

252 McHardy, I. H. *et al.* Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships. *Microbiome* **1**, 17, doi:10.1186/2049-2618-1-17 (2013).

253 Zhang, X. *et al.* MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. *Microbiome* **4**, 31, doi:10.1186/s40168-016-0176-z (2016).

254 Deeke, S. A. *et al.* Mucosal–luminal interface proteomics reveals biomarkers of pediatric inflammatory bowel disease-associated colitis. *The American Journal of Gastroenterology*, 1-12, doi:10.1038/s41395-018-0024-9 (2018).

255 Zhang, X. *et al.* Metaproteomics reveals associations between microbiome and intestinal extracellular vesicle proteins in pediatric inflammatory bowel disease. *Nature Communications* **9**, 2873, doi:10.1038/s41467-018-05357-4 (2018).

256 Dethlefsen, L., Huse, S. M., Sogin, M. L. & Relman, D. A. The Pervasive Effects of an Antibiotic on the Human Gut Microbiota, as Revealed by Deep 16S rRNA Sequencing. *PLoS Biology* **6**, 2383-2400, doi:10.1371/journal.pbio.0060280 (2008).

257 Yan, A., Butcher, J., Mack, D. & Stintzi, A. Virome Sequencing of the Human Intestinal Mucosal–Luminal Interface. *Frontiers in Cellular and Infection Microbiology* **10**, doi:10.3389/fcimb.2020.582187 (2020).

258 Yan, A., Butcher, J., Schramm, L., Mack, D. R. & Stintzi, A. Multiomic spatial analysis reveals a distinct mucosa-associated virome. *Gut Microbes* **15**, 2177488, doi:10.1080/19490976.2023.2177488 (2023).

259 Belkaid, Y. & Timothy. Role of the Microbiota in Immunity and Inflammation. *Cell* **157**, 121-141, doi:10.1016/j.cell.2014.03.011 (2014).

260 Carmody, R. N. & Turnbaugh, P. J. Host-microbial interactions in the metabolism of therapeutic and diet-derived xenobiotics. *Journal of Clinical Investigation* **124**, 4173-4181, doi:10.1172/JCI72335 (2014).

261 Boulangé, C. L., Neves, A. L., Chilloux, J., Nicholson, J. K. & Dumas, M.-E. Impact of the gut microbiota on inflammation, obesity, and metabolic disease. *Genome Medicine* **8**, 42-42, doi:10.1186/s13073-016-0303-2 (2016).

262 NCBI. *Genome - NCBI*, <https://www.ncbi.nlm.nih.gov/genome/> (2020).

263 Lopes, S. *et al.* Looking into Enteric Virome in Patients with IBD: Defining Guilty or Innocence? *Inflammatory bowel diseases* **23**, 1-1, doi:10.1097/MIB.0000000000001167 (2017).

264 Hannigan, G. D., Duhaime, M. B., Koutra, D. & Schloss, P. D. Biogeography and environmental conditions shape bacteriophage-bacteria networks across the human microbiome. *PLOS Computational Biology* **14**, e1006099-e1006099, doi:10.1371/journal.pcbi.1006099 (2018).

265 Lin, D. M. *et al.* Transplanting Fecal Virus-Like Particles Reduces High-Fat Diet-Induced Small Intestinal Bacterial Overgrowth in Mice. *Frontiers in Cellular and Infection Microbiology* **9**, doi:10.3389/fcimb.2019.00348 (2019).

266 Hayes, S., Mahony, J., Nauta, A. & Van Sinderen, D. Metagenomic Approaches to Assess Bacteriophages in Various Environmental Niches. *Viruses* **9**, 127, doi:10.3390/v9060127 (2017).

267 Norman, J. M., Handley, S. A. & Virgin, H. W. Kingdom-agnostic metagenomics and the importance of complete characterization of enteric microbial communities. *Gastroenterology* **146**, 1459-1469, doi:10.1053/j.gastro.2014.02.001 (2014).

268 Martinez-Guryn, K., Leone, V. & Chang, E. B. Regional Diversity of the Gastrointestinal Microbiome. *Cell Host & Microbe* **26**, 314-324, doi:10.1016/j.chom.2019.08.011 (2019).

269 Galley, J. D. *et al.* The structures of the colonic mucosa-associated and luminal microbial communities are distinct and differentially affected by a prolonged murine stressor. **5**, 748-760, doi:10.4161/19490976.2014.972241 (2014).

270 Kim, M. S. & Bae, J. W. Spatial disturbances in altered mucosal and luminal gut viromes of diet-induced obese mice. *Environmental Microbiology* **18**, 1498-1510, doi:10.1111/1462-2920.13182 (2016).

271 Zhao, G. *et al.* Virome biogeography in the lower gastrointestinal tract of rhesus macaques with chronic diarrhea. *Virology* **527**, 77-88, doi:10.1016/j.virol.2018.10.001 (2019).

272 Hoyles, L. *et al.* Characterization of virus-like particles associated with the human faecal and caecal microbiota. *Research in Microbiology* **165**, 803-812, doi:10.1016/j.resmic.2014.10.006 (2014).

273 Ungaro, F. *et al.* Metagenomic analysis of intestinal mucosa revealed a specific eukaryotic gut virome signature in early-diagnosed inflammatory bowel disease. *Gut Microbes*, 1-10, doi:10.1080/19490976.2018.1511664 (2018).

274 Kropinski, A. M. *et al.* Genome and Proteome of Campylobacter jejuni Bacteriophage NCTC 12673. *Applied and Environmental Microbiology* **77**, 8265-8271, doi:10.1128/aem.05562-11 (2011).

275 Li, H. seqtk Toolkit for processing sequences in FASTA/Q formats (2012).

276 Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research* **27**, 849-864, doi:10.1101/gr.213611.116 (2017).

277 Li, H. *et al.* The Sequence Alignment / Map format and SAMtools. **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

278 Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**, 59-60, doi:10.1038/nmeth.3176 (2014).

279 Huson, D. H. *et al.* MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Computational Biology* **12**, 1-12, doi:10.1371/journal.pcbi.1004957 (2016).

280 Eddy, S. R. Accelerated Profile HMM Searches. *PLOS Computational Biology* **7**, e1002195, doi:10.1371/journal.pcbi.1002195 (2011).

281 Galperin, M. Y., Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Research* **43**, D261-D269, doi:10.1093/nar/gku1223 (2015).

282 Bin Jang, H. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nature Biotechnology* **37**, 632-639, doi:10.1038/s41587-019-0100-8 (2019).

283 feargalr. Demovir: Taxonomic classification of viruses at Order and Family level. (2019).

284 McMurdie, P. J. & Holmes, S. phyloseq : An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. **8**, doi:10.1371/journal.pone.0061217 (2013).

285 Wickham, H. Reshaping Data with the reshape Package (2007).

286 Wickham, H. ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag New York, 2016).

287 Arnold, J. B. *et al.* ggthemes: Extra Themes, Scales and Geoms for 'ggplot2' (2019).

288 Kassambara, A. ggpubr: 'ggplot2' Based Publication Ready Plots (2020).

289 Campitelli, E. ggnewscale: Multiple Fill and Colour Scales in 'ggplot2' (2019).

290 Harrell Jr, F. E. Hmisc: Harrell Miscellaneous (2020).

291 Wei, T. & Simko, V. R package "corrplot": Visualization of a Correlation Matrix (2017).

292 Marotz, C. A. *et al.* Improving saliva shotgun metagenomics by chemical host DNA depletion. *Microbiome* **6**, doi:10.1186/s40168-018-0426-3 (2018).

293 Pereira-Marques, J. *et al.* Impact of Host DNA and Sequencing Depth on the Taxonomic Resolution of Whole Metagenome Sequencing for Microbiome Analysis. *Frontiers in Microbiology* **10**, doi:10.3389/fmicb.2019.01277 (2019).

294 Hatfull, G. F. Bacteriophage genomics. *Current Opinion in Microbiology* **11**, 447-453, doi:10.1016/j.mib.2008.09.004 (2008).

295 Sender, R., Fuchs, S. & Milo, R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLOS Biology* **14**, e1002533, doi:10.1371/journal.pbio.1002533 (2016).

296 Kim, K. H. & Bae, J. W. Amplification Methods Bias Metagenomic Libraries of Uncultured Single-Stranded and Double-Stranded DNA Viruses. **77**, 7663-7668, doi:10.1128/aem.00289-11 (2011).

297 Zhang, C. *et al.* Identification of low abundance microbiome in clinical samples using whole genome sequencing. *Genome Biology* **16**, doi:10.1186/s13059-015-0821-z (2015).

298 Lewis, J. D. *et al.* Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease. *Cell Host and Microbe* **18**, 489-500, doi:10.1016/j.chom.2015.09.008 (2015).

299 Sutton, T. D. S., Clooney, A. G. & Hill, C. Giant oversights in the human gut virome. *Gut* **69**, 1357-1358, doi:10.1136/gutjnl-2019-319067 (2020).

300 Garretto, A., Hatzopoulos, T. & Putonti, C. virMine: automated detection of viral sequences from complex metagenomic samples. *PeerJ* **7**, e6695, doi:10.7717/peerj.6695 (2019).

301 Hatzopoulos, T., Watkins, S. C. & Putonti, C. PhagePhisher: a pipeline for the discovery of covert viral sequences in complex genomic datasets. doi:10.1099/mgen.0.000053 (2016).

302 Paez-Espino, D., Pavlopoulos, G. A., Ivanova, N. N. & Kyrpides, N. C. Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data. *Nature Protocols* **12**, 1673-1682, doi:10.1038/nprot.2017.063 (2017).

303 Roux, S. *et al.* Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. *PeerJ* **4**, e2777, doi:10.7717/peerj.2777 (2016).

304 Zhang, T. *et al.* RNA Viral Community in Human Feces: Prevalence of Plant Pathogenic Viruses. *PLoS Biology* **4**, e3, doi:10.1371/journal.pbio.0040003 (2005).

305 Tokarz, R. *et al.* Characterization of Stool Virome in Children Newly Diagnosed With Moderate to Severe Ulcerative Colitis. *Inflammatory Bowel Diseases*, doi:10.1093/ibd/izz099 (2019).

306 Johansen, J. *et al.* Genome binning of viral entities from bulk metagenomics data. *Nature Communications* **13**, 965, doi:10.1038/s41467-022-28581-5 (2022).

307 Callanan, J. *et al.* Expansion of known ssRNA phage genomes: From tens to over a thousand. *Science Advances* **6**, eaay5981, doi:10.1126/sciadv.aay5981 (2020).

308 Owen, S. V. *et al.* A window into lysogeny: revealing temperate phage biology with transcriptomics. *Microb Genom* **6**, e000330, doi:10.1099/mgen.0.000330 (2020).

309 Leskinen, K., Blasdel, B. G., Lavigne, R. & Skurnik, M. RNA-Sequencing Reveals the Progression of Phage-Host Interactions between φR1-37 and Yersinia enterocolitica. *Viruses* **8**, 111-111, doi:10.3390/v8040111 (2016).

310 Shkoporov, A. N. *et al.* Viral biogeography of the mammalian gut and parenchymal organs. *Nature Microbiology* **7**, 1301-1311, doi:10.1038/s41564-022-01178-w (2022).

311 Jimenez-Rivera, C., Haas, D., Boland, M., Barkey, J. L. & Mack, D. R. Comparison of Two Common Outpatient Preparations for Colonoscopy in Children and Youth. *Gastroenterology Research and Practice* **2009**, 518932, doi:10.1155/2009/518932 (2009).

312 Stahl, M. *et al.* L-Fucose utilization provides *Campylobacter jejuni* with a competitive advantage. *Proceedings of the National Academy of Sciences* **108**, 7194-7199, doi:doi:10.1073/pnas.1014125108 (2011).

313 Clarke, E. L. *et al.* Sunbeam : an extensible pipeline for analyzing metagenomic sequencing experiments. 1-13 (2019).

314 Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* **14**, R36, doi:10.1186/gb-2013-14-4-r36 (2013).

315 Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211-3217, doi:10.1093/bioinformatics/bts611 (2012).

316 von Meijenfeldt, F. A. B., Arkhipova, K., Cambuy, D. D., Coutinho, F. H. & Dutilh, B. E. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biology* **20**, 217, doi:10.1186/s13059-019-1817-x (2019).

317 Seemann, T. Prokka : rapid prokaryotic genome annotation. **30**, 2068-2069, doi:10.1093/bioinformatics/btu153 (2014).

318 Haft, D. H., Selengut, J. D. & White, O. The TIGRFAMs database of protein families. *Nucleic Acids Research* **31**, 371-373, doi:10.1093/nar/gkg128 (2003).

319 Seemann, T. Snippy: Rapid haploid variant calling and core genome alignment (version 4.6.0). (2020).

320 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

321 Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930, doi:10.1093/bioinformatics/btt656 (2013).

322 Jiang, P., Lai, S., Wu, S., Zhao, X.-M. & Chen, W.-H. Host DNA contents in fecal metagenomics as a biomarker for intestinal diseases and effective treatment. *BMC Genomics* **21**, 348, doi:10.1186/s12864-020-6749-z (2020).

323 Schroeder, K. W., Tremaine, W. J. & Ilstrup, D. M. Coated Oral 5-Aminosalicylic Acid Therapy for Mildly to Moderately Active Ulcerative Colitis. *New England Journal of Medicine* **317**, 1625-1629, doi:10.1056/nejm198712243172603 (1987).

324 Lau, M. C. Y. *et al.* Taxonomic and Functional Compositions Impacted by the Quality of Metatranscriptomic Assemblies. *Frontiers in Microbiology* **9**, doi:10.3389/fmicb.2018.01235 (2018).

325 Parras-Moltó, M., Rodríguez-Galet, A., Suárez-Rodríguez, P. & López-Bueno, A. Evaluation of bias induced by viral enrichment and random amplification protocols in metagenomic surveys of saliva DNA viruses. *Microbiome* **6**, 119, doi:10.1186/s40168-018-0507-3 (2018).

326 Vaga, S. *et al.* Compositional and functional differences of the mucosal microbiota along the intestine of healthy individuals. *Scientific Reports* **10**, 14977, doi:10.1038/s41598-020-71939-2 (2020).

327 Van den Abbeele, P., Van de Wiele, T., Verstraete, W. & Possemiers, S. The host selects mucosal and luminal associations of coevolved gut microorganisms: a novel concept. *FEMS Microbiology Reviews* **35**, 681-704, doi:10.1111/j.1574-6976.2011.00270.x (2011).

328 Moraru, C., Varsani, A. & Kropinski, A. M. VIRIDIC—A Novel Tool to Calculate the Intergenomic Similarities of Prokaryote-Infecting Viruses. *Viruses* **12**, 1268 (2020).

329 Gulyaeva, A. *et al.* Discovery, diversity, and functional associations of crAss-like phages in human gut metagenomes from four Dutch cohorts. *Cell Reports* **38**, 110204, doi:https://doi.org/10.1016/j.celrep.2021.110204 (2022).

330 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**, doi:10.1038/s41467-019-14103-3 (2020).

331 Yutin, N. *et al.* Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nature Microbiology* **3**, 38-46, doi:10.1038/s41564-017-0053-y (2018).

332 Fogg, P. C. M., Rigden, D. J., Saunders, J. R., McCarthy, A. J. & Allison, H. E. Characterization of the relationship between integrase, excisionase and antirepressor activities associated with a superinfecting Shiga toxin encoding bacteriophage. *Nucleic acids research* **39**, 2116-2129, doi:10.1093/nar/gkq923 (2011).

333 Balding, C., Bromley, S. A., Pickup, R. W. & Saunders, J. R. Diversity of phage integrases in Enterobacteriaceae: development of markers for environmental analysis of temperate phages. *Environmental Microbiology* **7**, 1558-1567, doi:https://doi.org/10.1111/j.1462-2920.2005.00845.x (2005).

334 Nami, Y., Imeni, N. & Panahi, B. Application of machine learning in bacteriophage research. *BMC Microbiology* **21**, 193, doi:10.1186/s12866-021-02256-5 (2021).

335 Song, K. Classifying the Lifestyle of Metagenomically-Derived Phages Sequences Using Alignment-Free Methods. *Frontiers in Microbiology* **11**, doi:10.3389/fmicb.2020.567769 (2020).

336 Collier, A. M. *et al.* Initiation of RNA Polymerization and Polymerase Encapsidation by a Small dsRNA Virus. *PLOS Pathogens* **12**, e1005523, doi:10.1371/journal.ppat.1005523 (2016).

337 Manichanh, C., Borruel, N., Casellas, F. & Guarner, F. The gut microbiota in IBD. *Nature Reviews Gastroenterology & Hepatology* **9**, 599-608, doi:10.1038/nrgastro.2012.152 (2012).

338 Halfvarson, J. *et al.* Dynamics of the human gut microbiome in inflammatory bowel disease. *Nature Microbiology* **2**, 17004-17004, doi:10.1038/nmicrobiol.2017.4 (2017).

339 Pérez-Brocal, V. *et al.* Metagenomic Analysis of Crohn's Disease Patients Identifies Changes in the Virome and Microbiome Related to Disease Status and Therapy, and Detects Potential Interactions and Biomarkers. *Inflammatory Bowel Diseases* **0**, 1-1, doi:10.1097/MIB.0000000000000549 (2015).

340 Jansen, D. & Matthijnssens, J. The Emerging Role of the Gut Virome in Health and Inflammatory Bowel Disease: Challenges, Covariates and a Viral Imbalance. *Viruses* **15**, 173 (2023).

341 Kong, C., Liu, G., Kalady, M. F., Jin, T. & Ma, Y. Dysbiosis of the stool DNA and RNA virome in Crohn's disease. *Journal of Medical Virology* **95**, e28573, doi:https://doi.org/10.1002/jmv.28573 (2023).

342 Smith, L., Goldobina, E., Govi, B. & Shkoporov, A. N. Bacteriophages of the Order Crassvirales: What Do We Currently Know about This Keystone Component of the Human Gut Virome? *Biomolecules* **13** (2023).

343 Sinha, A. *et al.* Transplantation of bacteriophages from ulcerative colitis patients shifts the gut bacteriome and exacerbates the severity of DSS colitis. *Microbiome* **10**, 105, doi:10.1186/s40168-022-01275-2 (2022).

344 Tian, X. *et al.* Gut virome-wide association analysis identifies cross-population viral signatures for inflammatory bowel disease. *Microbiome* **12**, 130, doi:10.1186/s40168-024-01832-x (2024).

345 Levine, A. *et al.* ESPGHAN revised porto criteria for the diagnosis of inflammatory bowel disease in children and adolescents. *J Pediatr Gastroenterol Nutr* **58**, 795-806, doi:10.1097/mpg.0000000000000239 (2014).

346 Levine, A. *et al.* Pediatric modification of the Montreal classification for inflammatory bowel disease: The Paris classification. *Inflammatory Bowel Diseases* **17**, 1314-1321, doi:10.1002/ibd.21493 (2011).

347 Hyams, J. *et al.* Evaluation of the pediatric crohn disease activity index: a prospective multicenter experience. *Journal of pediatric gastroenterology and nutrition* **41**, 416-421, doi:10.1097/01.mpg.0000183350.46795.42 (2005).

348 Turner, D. *et al.* Development, validation, and evaluation of a pediatric ulcerative colitis activity index: a prospective multicenter study. *Gastroenterology* **133**, 423-432, doi:10.1053/j.gastro.2007.05.029 (2007).

349 Daperno, M. *et al.* Development and validation of a new, simplified endoscopic activity score for Crohn's disease: the SES-CD. *Gastrointest Endosc* **60**, 505-512, doi:10.1016/s0016-5107(04)01878-4 (2004).

350 Lobatón, T. *et al.* The Modified Mayo Endoscopic Score (MMES): A New Index for the Assessment of Extension and Severity of Endoscopic Activity in Ulcerative Colitis Patients. *Journal of Crohn's and Colitis* **9**, 846-852, doi:10.1093/ecco-jcc/jjv111 (2015).

351 Turner, D. *et al.* Mathematical weighting of the pediatric Crohn's disease activity index (PCDAI) and comparison with its other short versions. *Inflamm Bowel Dis* **18**, 55-62, doi:10.1002/ibd.21649 (2012).

352 Harris, P. A. *et al.* Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* **42**, 377-381, doi:10.1016/j.jbi.2008.08.010 (2009).

353 Mack, D. R. *et al.* Canadian Association of Gastroenterology Clinical Practice Guideline for the Medical Management of Pediatric Luminal Crohn's Disease. *Gastroenterology* **157**, 320-348, doi:10.1053/j.gastro.2019.03.022 (2019).

354 Turner, D. *et al.* Management of Paediatric Ulcerative Colitis, Part 1: Ambulatory Care-An Evidence-based Guideline From European Crohn's and Colitis Organization and European Society of Paediatric Gastroenterology, Hepatology and Nutrition. *J Pediatr Gastroenterol Nutr* **67**, 257-291, doi:10.1097/mpg.0000000000002035 (2018).

355 Camargo, A. P. *et al.* You can move, but you can't hide: identification of mobile genetic elements with geNomad. *bioRxiv*, 2023.2003.2005.531206, doi:10.1101/2023.03.05.531206 (2023).

356 Wickham, H. in *Journal of Statistical Software; Vol 1, Issue 12 (2007)* (2007).

357 Wickham, H. (Springer-Verlag New York, 2016).

358 Arnold, J. B. *et al.* (2019).

359 Kassambara, A. (2020).

360 Moreno-Gallego, J. L. *et al.* Virome Diversity Correlates with Intestinal Microbiome Diversity in Adult Monozygotic Twins. *Cell Host & Microbe* **25**, 261-272.e265, doi:https://doi.org/10.1016/j.chom.2019.01.019 (2019).

361 Papudeshi, B. *et al.* Host interactions of novel Crassvirales species belonging to multiple families infecting bacterial host, Bacteroides cellulosilyticus WH2. *Microb Genom* **9**, doi:10.1099/mgen.0.001100 (2023).

362 Bloom, S. M. *et al.* Commensal Bacteroides species induce colitis in host-genotype-specific fashion in a mouse model of inflammatory bowel disease. *Cell Host Microbe* **9**, 390-403, doi:10.1016/j.chom.2011.04.009 (2011).

363 Mills, R. H. *et al.* Multi-omics analyses of the ulcerative colitis gut microbiome link Bacteroides vulgatus proteases with disease severity. *Nat Microbiol* **7**, 262-276, doi:10.1038/s41564-021-01050-3 (2022).

364 Da Silva Morais, E., Grimaud, G. M., Warda, A., Stanton, C. & Ross, P. Genome plasticity shapes the ecology and evolution of Phocaeicola dorei and Phocaeicola vulgatus. *Sci Rep* **14**, 10109, doi:10.1038/s41598-024-59148-7 (2024).

365 Nishijima, S. *et al.* Extensive gut virome variation and its associations with host and environmental factors in a population-level cohort. *Nat Commun* **13**, 5252, doi:10.1038/s41467-022-32832-w (2022).

366 Zuo, T. *et al.* Temporal landscape of human gut RNA and DNA virome in SARS-CoV-2 infection and severity. *Microbiome* **9**, 91, doi:10.1186/s40168-021-01008-x (2021).

367 Massimino, L. *et al.* Gut virome-colonising Orthohepadnavirus genus is associated with ulcerative colitis pathogenesis and induces intestinal inflammation in vivo. *Gut*, gutjnl-2022-328375, doi:10.1136/gutjnl-2022-328375 (2023).

368 Rothschild-Rodriguez, D., Hedges, M., Kaplan, M., Karav, S. & Nobrega, F. L. Phage-encoded carbohydrate-interacting proteins in the human gut. *Front Microbiol* **13**, 1083208, doi:10.3389/fmicb.2022.1083208 (2022).

369 Nagata, N. *et al.* Effects of bowel preparation on the human gut microbiome and metabolome. *Scientific Reports* **9**, 4042, doi:10.1038/s41598-019-40182-9 (2019).