CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

| | |
|---|---|
| Title | Structural and biological insights into the most abundant phages of the human gut |
| Authors | Tobin, Ciara A. |
| Publication date | 2025 |
| Original Citation | Tobin, C. A. 2025. Structural and biological insights into the most abundant phages of the human gut. PhD Thesis, University College Cork. |
| Type of publication | Doctoral thesis |
| Rights | © 2025, Ciara Tobin. - https://creativecommons.org/licenses/by-nc-nd/4.0/ |
| Download date | 2026-11-03 19:38:15 |
| Item downloaded from | https://hdl.handle.net/10468/17931 |



UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh




# Structural and biological insights into the most abundant phages of the human gut

Thesis presented by

**Ciara Tobin, B.Sc.**

for the degree of

**Doctor of Philosophy**

School of Microbiology, University College Cork, Cork, Ireland

APC Microbiome Ireland, University College Cork, Cork, Ireland

**Supervisors:** Prof. Colin Hill, Prof. Andrey Shkoporov

**Head of School:** Prof. David Clarke

2025

# Table of Contents


Declaration ..... 5
Abbreviations ..... 6
**Thesis Abstract ..... 8**

**Chapter 1: Factors affecting variation in the human gut phageome ..... 11**

Abstract ..... 12
INTRODUCTION ..... 13
PHAGE-BACTERIAL HOST DYNAMICS ..... 14
BIOLOGICAL FACTORS CONTRIBUTING TO PHAGEOME VARIATION ..... 17
*Age ..... 17*
*Diet ..... 19*
*Geography ..... 20*
PHAGEOME VARIABILITY IN DISEASE ..... 22
*Inflammatory Bowel Disease ..... 22*
*Irritable Bowel Syndrome ..... 24*
*Colorectal Cancer ..... 26*
METHODOLOGICAL VARIATION ..... 28
CONCLUSIONS ..... 30
References ..... 32

**Chapter 2: The gut bacteriome and virome of patients with acute COVID is similar to those post recovery ..... 39**

Abstract ..... 40
Introduction ..... 41
Results ..... 42
*Gut virome composition in acute and convalescent COVID-19 patients ..... 42*
*No alterations in the gut bacteriome and virome between acute and convalescent patients ..... 45*
*Minimal shifts in bacterial species abundance between acute and convalescent patients ..... 48*
*Oral virome and bacteriome composition of COVID-19 patients ..... 51*
*Minimal alterations to the oral virome and bacteriome between COVID-19 patients ..... 53*
Discussion ..... 55
Methods ..... 58
*Study cohort ..... 58*
*Faecal VLP nucleic acid extraction ..... 59*
*Library preparation and sequencing ..... 59*
*Processing of faecal VLP shotgun sequencing data ..... 59*
*16S rRNA amplicon sequencing and bioinformatic analysis ..... 61*
*Saliva whole metagenomics ..... 62*
*Processing of saliva sequencing data ..... 62*
*Statistical analysis ..... 62*
*Data availability ..... 63*
Author Contributions ..... 63
Supplementary ..... 64
References ..... 67

**Chapter 3: Turning in silico phage into in vivo phage – the crAssphage story ..... 73**

Abstract ..... 74
Introduction ..... 74
In silico discovery ..... 75
Current classifications and general genomic features ..... 76
In vitro isolation and biological properties ..... 77
Structural model for the *Crassvirales* order ..... 79
Ecological models ..... 80

Conclusions ..... 81
References ..... 82
**Chapter 4: CRISPRi targeting of genes in *Crassvirales* phage DAC15 predicts multiple anti-CRISPR proteins ..... 85**
Abstract ..... 86
Introduction ..... 87
Results ..... 89
*CRISPRi targeting of phage genes ..... 91*
*CRISPRi targeting of TerL enriches for the production of empty phage capsids ..... 95*
*Prediction of five Acrs in the genome of DAC15 ..... 97*
*Prediction of Acrs in the genomes of CrAss001, CrAss002 and p-crAssphage ..... 98*
Discussion ..... 100
Materials and Methods ..... 103
*Bacterial strains, phage and culture conditions ..... 103*
*Plasmid design and transformation ..... 103*
*Conjugation ..... 104*
*CRISPRi experiments ..... 104*
*Transmission Electron Microscopy ..... 105*
*Acr prediction ..... 106*
Author contributions ..... 106
Supplementary ..... 107
References ..... 109
**Chapter 5: *Crassvirales* phage CrAss001 has three types of tailspikes with capsule depolymerase and esterase activity ..... 114**
Abstract ..... 115
Introduction ..... 116
Results ..... 117
*Three TF/TSP complexes identified by AlphaFold predictions and pull-down assays ..... 117*
*Structure and depolymerase activity of gp22_23 TSP ..... 119*
*Structure and esterase activity of gp25 TSP ..... 122*
*Predictions of outer TSP assembly ..... 125*
Discussion ..... 129
Methods ..... 133
*Cloning, protein expression and purification ..... 133*
*Electron Microscopy ..... 134*
*AlphaFold predictions ..... 134*
*Bacterial strains and culture conditions ..... 135*
*Enzymatic assays ..... 135*
Author Contributions ..... 135
Supplementary ..... 137
References ..... 140
**Chapter 6: Capsid structure of *Crassvirales* phage DAC15 ..... 145**
Abstract ..... 146
Introduction ..... 147
Results ..... 148
*Structure and interactions of MCP ..... 149*
*Structure and interactions of auxiliary protein ..... 152*
*Structural and sequence similarity to CrAss001 MCP and auxiliary protein ..... 153*
*Structure of HFD and comparison to CrAss001 HFD ..... 155*
*Structure of HFT and comparison to CrAss001 HFT ..... 156*
Discussion ..... 158
Methods ..... 160

*Virus propagation and purification* ..... *160*
*Electron microscopy* ..... *160*
*Atomic modelling and AlphaFold predictions* ..... *161*
Author contributions ..... 161
Supplementary ..... 162
References ..... 163
**Thesis summary and future work ..... 166**
References ..... 170
**Acknowledgements ..... 171**

# Declaration

This is to certify that the work I am submitting is my own and has not been submitted for another degree, either at University College Cork or elsewhere. All external references and sources are clearly acknowledged and identified within the contents. I have read and understood the regulations of University College Cork concerning plagiarism.

Student number: 116439372

Date: 28th March 2025

Signed: _____ Ciara Tobin _____

This research was conducted with the financial support of the Science Foundation Ireland (SFI) under Grant Number 12/RC/2273_P2 and Wellcome Trust Research Career Development Fellowship [220646/Z/20/Z].

## Abbreviations

| | |
|---|---|
| Aca | Acr associated protein |
| ACE2 | Angiotensin converting enzyme 2 |
| Acr | Anti-CRISPR protein |
| CD | Crohn's disease |
| Cas | CRISPR associated proteins |
| CBM | Carbohydrate binding module |
| crAss | cross assembly software |
| COVID-19 | Coronavirus disease 2019 |
| CPS | Capsular polysaccharide |
| CRC | Colorectal cancer |
| CRISPR | Clustered Regularly Interspaced Short Palindromic Repeats |
| CRISPRi | CRISPR interference |
| Cryo-EM | Cryogenic electron microscopy |
| FAB | Fastidious Anaerobe Broth |
| GI | Gastrointestinal |
| HFD | Head fibre dimer |
| HFT | Head fibre trimer |
| IBD | Inflammatory bowel disease |
| IBS | Irritable Bowel Syndrome |
| IQR | Interquartile range |
| LPS | Lipopolysaccharide |
| MCP | Major capsid protein |
| MDA | Multiple Displacement Amplification |
| MOI | Multiplicity of infection |
| p-crAssphage | Prototypical crAssphage |
| PERMANOVA | Permutational multivariate analysis of variance |
| PES | Polyethersulfone |
| PCoA | Principle coordinate analysis |

| | |
|---|---|
| Racrs | RNA based Acrs |
| RBP | Receptor binding protein |
| RNAP | RNA polymerase |
| SARS-CoV-2 | Severe acute respiratory syndrome coronavirus 2 |
| SCFA | Short chain fatty acid |
| SEC | Size exclusion chromatography |
| SpyCas9 | *Streptococcus pyogenes* Cas9 |
| TEM | Transmission electron microscopy |
| TerL | Terminase large subunit |
| TF | Tail fibre |
| TH | Tail hub |
| TSP | Tail spike protein |
| UC | Ulcerative colitis |
| VC | Viral cluster |
| VAL | Virion-associated lysin |
| VLP | Virus-like particle |
| vOTU | Viral operational taxonomic unit |

# Thesis Abstract

The importance of the gut microbiome in both health and disease states has emerged over recent years. However, the focus has mostly been on the bacterial portion of the human gut. As obligate predators of bacteria, bacteriophages (phages) have the potential to shape their microbial environment, through interactions with their bacterial hosts. The relationship between phages and their bacterial hosts is complex, each deriving various defence and counter-defence strategies in a bid to survive and persist.

In order to gain further insights into phages of the human gut (the gut phageome), this thesis explores two key areas of research. The first examines current literature on biological and methodological factors that can result in gut phageome variation in individuals. Biological factors include age, diet and geography, as well as disease states such as irritable bowel syndrome, irritable bowel disease, colorectal cancer, and infectious diseases such as severe acute respiratory syndrome coronavirus 2 (SARS-COV-2), the causative agent of coronavirus disease 2019 (COVID-19). This led to the compositional analysis of both the virome and bacteriome from the gut and oral cavity of 43 COVID-19 patients, during the acute and convalescence stages of infection. Overall, the gut and oral virome and bacteriome of acute patients was found to be similar to those in recovery.

The second aim of this thesis was to explore structural and biological aspects of the most abundant phages of the human gut, *Crassvirales* phages. These phages are remarkable for their ability to establish a long-term equilibrium with their bacterial hosts, without noticeable detrimental effects to the latter. CRISPR interference was used to target a number of genes predicted to encode for structural proteins and enzymes involved in various stages of the phage lifecycle in one *Crassvirales* phage, DAC15. Targeting of the terminase large subunit (TerL) resulted in an enrichment of the number of empty phage capsids, visualised by transmission electron microscopy (TEM), while still maintaining a large proportion of viable phages.

Minimal reductions in efficiency of plaquing (EOP) when targeting additional essential and non-essential genes led to the analysis of the DAC15 genome for potential defence strategies. A combined bioinformatic and structural homology approach identified five putative anti-CRISPR proteins (Acrs). This approach was extended to *Crassvirales* phages CrAss001, CrAss002 and p-crAssphage, identifying 16 more potential Acrs across the three genomes. This is the first instance of putative Acr genes being identified in *Crassvirales* phages.

A combination of structural, biochemical and bioinformatic approaches were used to elucidate the structure and function of the receptor binding proteins (RBP) of *Crassvirales* phage CrAss001. CrAss001 was determined to have three tail spike proteins (TSP), that make up inner and outer TSP assemblies. The structure of the inner TSP, gp22_23, was resolved to 3.4 Å and shown to act as a capsule depolymerase, producing zones of clearing on overlays of the bacterial host, *Bacteroides intestinalis*, that predominantly expresses capsule polysaccharide (CPS) PVR9. A second TSP that makes up part of the outer assembly, gp25, was resolved to 2.6 Å and was shown to exhibit esterase activity. The structures of the remaining two outer assembly proteins, gp24 and gp26, were predicted by AlphaFold. Gp24 is thought to be a linker protein for TSP's gp25 and gp26. This work provides the first mechanistic insights into the RBPs of a *Crassvirales* phage.

The final chapter of this thesis looks at the capsid structure of *Crassvirales* phage DAC15. This represents the second *Crassvirales* capsid to be resolved in full. The capsid was resolved to 2.95 Å and has a triangulation number of nine. Each asymmetric subunit consists of nine copies of the major capsid protein (MCP), nine copies of the auxiliary protein, two copies of the head fibre dimer (HFD) and one copy of the head fibre trimer (HFT). The MCP has the canonical HK97 fold and both the MCP and auxiliary protein are structural matches to the MCP and auxiliary protein of CrAss001. Extensive non-covalent interactions were identified between MCP subunits and between MCP and auxiliary proteins that are thought to lend stability to the capsid. While the HFD and HFT are structurally similar to their counterparts in CrAss001, there are some differences. The C-terminal domains of the HFD protrude upwards from the capsid, while the domains in the HFD of

CrAss001 point away from each other. The HFT of DAC15 has a 26-residue insert that allows it to extend several nanometres further from the capsid than the HFT of CrAss001.

Overall, the work in this thesis advances our understanding of several aspects of the gut phageome, primarily through structural and biological work focused on *Crassvirales* phages. This builds on over a decade of work elucidating the various aspects of *Crassvirales* phage biology in the human gut.

# Chapter 1: Factors affecting variation in the human gut phageome

This chapter was published as a review in Annual Review of Microbiology

Tobin CA, Hill C, Shkoporov AN. 2023. Factors affecting variation of the human gut phageome. Annu Rev Microbiol 77:363–379.

## Abstract

The gut microbiome is a dense and metabolically active consortium of microorganisms and viruses located in the lower gastrointestinal tract of the human body. Bacteria and their viruses (phages) are the most abundant members of the gut microbiome. Investigating their biology and the interplay between the two is important if we are to understand their roles in human health and disease. In this review, we summarize recent advances in resolving the taxonomic structure and ecological functions of the complex community of phages in the human gut— the gut phageome. We discuss how age, diet, and geography can all have a significant impact on phageome composition. We note that alterations to the gut phageome have been observed in several diseases such as inflammatory bowel disease, irritable bowel syndrome, and colorectal cancer, and we evaluate whether these phageome changes can directly or indirectly contribute to disease etiology and pathogenesis. We also highlight how lack of standardization in studying the gut phageome has contributed to variation in reported results.

## INTRODUCTION

The human gut microbiome is a complex and dynamic environment, estimated to be composed of ~$10^{14}$ microbial cells from all domains of cellular life (79). These inhabitants of the gastrointestinal tract play a role in various important metabolic processes and are vital for normal human gut physiology. Bacteria have received the greatest attention, likely because they have the largest biomass and account for ~93% of DNA in metagenomic studies (82). Numerous studies have highlighted their importance in both health and disease, while the viral, fungal, and protozoan components are often neglected. The viral component of the gut includes both eukaryotic viruses and bacteriophages (phages), the latter being the most abundant members of the gut virome. Phages, together with their bacterial hosts, are the most abundant components of the gut microbiome. In recent years, viral metagenomics has revolutionized research of complex viral communities and has enabled us to catalog hundreds of thousands of previously unknown phages of the gut (7, 12, 67). The expansion of catalogued phages has also resulted in numerous taxonomic changes ratified by the International Committee on Taxonomy of Viruses (ICTV) (47). One of the most recent changes has been the abolishment of the previous Siphoviridae, Myoviridae, and Podoviridae families as well as the Caudovirales order, with all members assigned to the class *Caudoviricetes* (96, 97). This is in recognition of the inaccuracy of using morphology-based taxonomy and moves instead toward genome-based classification.

Phages persist in the gut microbiome at ~$10^{9}$–$10^{10}$ viral-like particles (VLPs) per gram of feces. They are estimated to exist at an ~1:1 ratio with their bacterial hosts, although it has been suggested that viruses are outnumbered by bacteria in the gut, with ratios as low as 0.1:1 (82). As such, they can potentially exert considerable influence over bacteria through a variety of mechanisms such as lysis, lysogenic conversion, chronic infection, and phage transduction (horizontal gene transfer) (10). This can lead to both direct and indirect impacts on the human host in both health and disease. It has been noted that the complex community of gut phages, referred to as the gut phageome, is highly variable between individuals while being temporally stable within individuals (81). In this review, we analyze the state of the

art of phageome research, in particular focusing on phageome variation and its implications in both health and disease states (**Figure 1**).

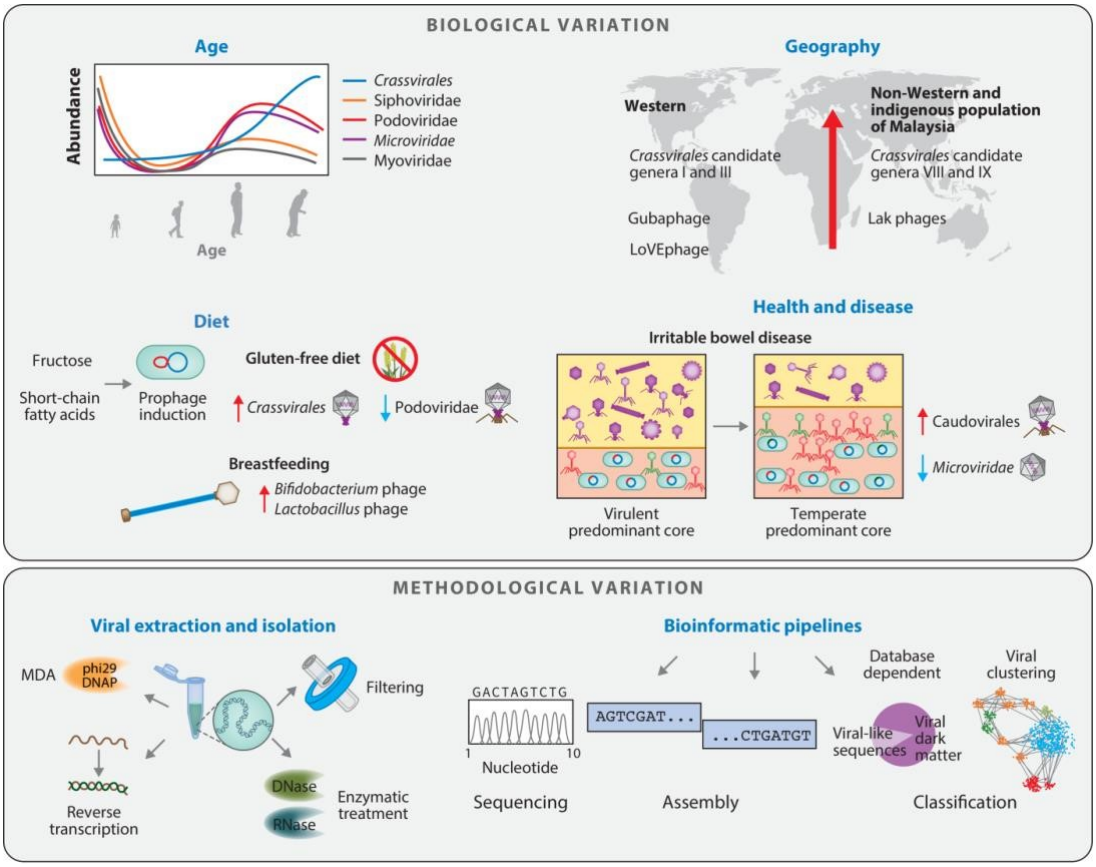



**Figure 1**. Biological and methodological factors contributing to phageome variation in the human gut. Biological factors include age, geography, diet, and various health and disease states. Methodological factors include viral extraction and isolation protocols and bioinformatic pipelines used. Abbreviations: DNAP, DNA polymerase; MDA, multiple displacement amplification. Figure adapted from images created with BioRender.com.

## PHAGE-BACTERIAL HOST DYNAMICS

Historically, both the lytic and lysogenic phage life cycles were expected to operate in the gut microbiome. Both involve a phage particle identifying and adsorbing to a receptor on the surface of the bacterial cell and injecting its genetic material into the susceptible host. In the lytic life cycle, carried out by virulent phages, the phage takes over the host cell machinery to replicate its own genetic material. Once mature virions are assembled, the bacterial cell is lysed to release new progeny. In lysogeny, carried out by temperate phages, the genetic material of the phage also has the possibility of integrating into the host chromosome as a prophage. This

prophage can be replicated for generations in tandem with bacterial reproduction until it is induced into the lytic life cycle, which is usually due to some environmental stressor but can also occur spontaneously (**Figure 2**). These are the main phage infection strategies observed to occur in the gut (21). Pseudolysogeny, carrier state, and chronic infection are the most common alternative phage infection strategies to be described. However, they are poorly characterized, with the terms pseudolysogeny and carrier state being used interchangeably in literature (58). As such, there have been efforts to define each more clearly (38, 58).

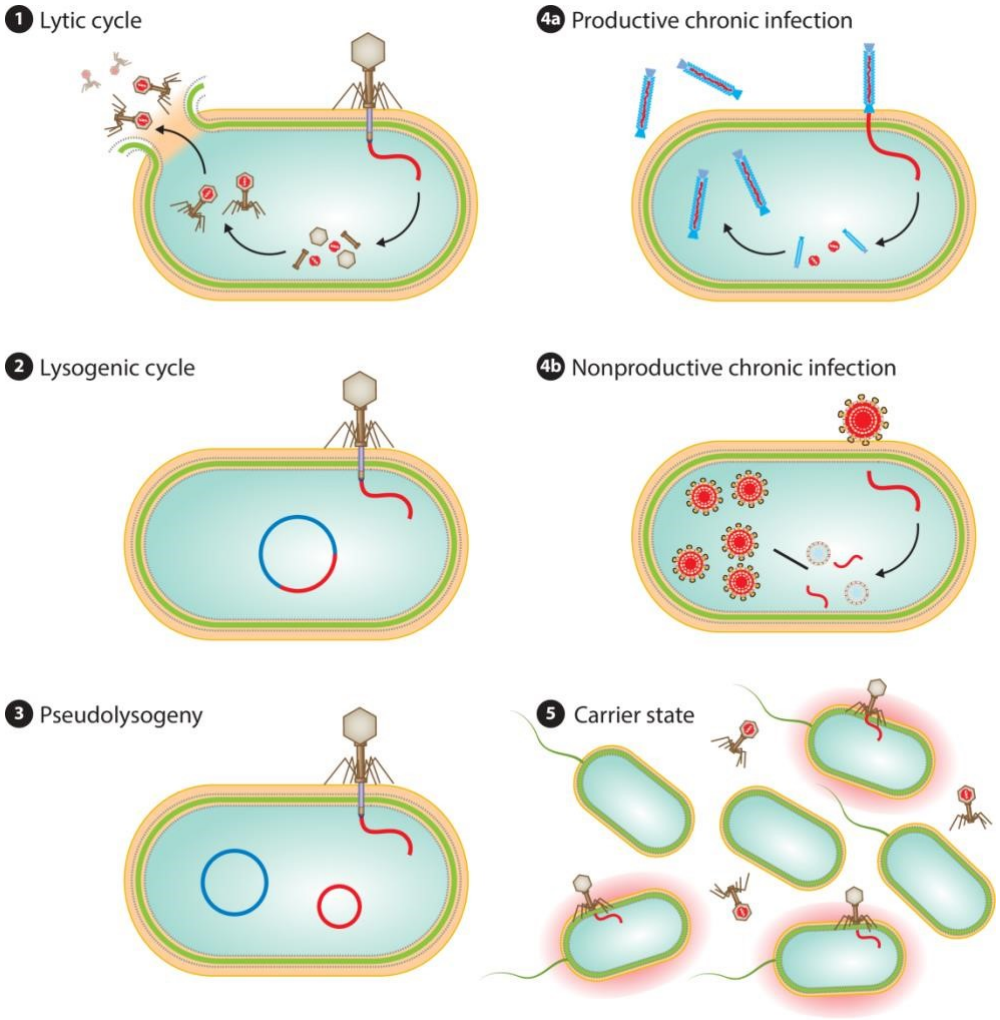



**Figure 2.** The principal life cycles that have been described for bacteriophages: (❶) lytic cycle, (❷) lysogenic cycle, (❸) pseudolysogeny, (❹) productive and nonproductive chronic infection, and (❺) carrier state. In the carrier state depicted here, a subpopulation of bacteria revert to resistance through loss of flagella, which can act as a phage receptor, and switching expression to a different capsular polysaccharide. Figure adapted from images created with BioRender.com.

Pseudolysogeny has been described as a form of delayed phage reproduction in which the phage genetic material remains dormant in the bacterial cell. It has been noted to occur in phages T4 (55) and P22 (14), which infect *Escherichia coli* and *Salmonella enterica* serovar Typhimurium. Chronic infection describes the ability of certain phage infections to continuously produce virions in the bacterial cell. In the productive form, these are excreted from the cell without the requirement to lyse it. In the nonproductive form, assembled phages accumulate in the cytoplasm of the bacterial cell but are not released. Productive chronic infection is most commonly described as occurring in filamentous phages, while nonproductive chronic infection has been reported in RNA phages (58).

While these four life cycles occur at the single-cell level, it has been suggested that the carrier state should be viewed at the population level, with both phages and their bacterial hosts persisting in equilibrium. A subpopulation of resistant bacteria serves to continuously maintain this balance. The carrier state has been described for *Campylobacter jejuni* and two of its phages (87), while the persistence of phages of the recently established order *Crassvirales* (88, 97) and their bacterial hosts has also been suggested to involve a population-level carrier state (83).

The temporal stability and persistence observed in the gut would suggest that temperate bacteria dominate the gut microbiome. Contrary to this, a large proportion of the gut phageome is composed of the virulent *Microviridae*, as well as viruses previously classified as Podoviridae, and *Crassvirales* (29, 31, 81). It is possible that alternative phage–bacterial host interactions play a larger role in the gut microbiome than currently anticipated, with benign interactions maintaining population stability. This shifts the view of phages as primarily predators of bacteria to participants in a more mutualistic relationship, in which both partners can benefit (85). In addition to this, spatial heterogeneity (56), host resistance, and phage counter-resistance mechanisms (32) and various abiotic and biotic factors along the gastrointestinal tract can all influence the maintenance of bacteria-phage populations and interactions.

## BIOLOGICAL FACTORS CONTRIBUTING TO PHAGEOME VARIATION

### Age

While the concept of a healthy core phageome has been examined in adults, it varies quite extensively over an entire life span. Phage populations in infants are quite unstable and dynamic, reflecting the early establishment of the gut bacteriome (the bacterial members of the gut microbiome). Epifluorescent microscopy analysis of meconium and stool samples from the first days of life reveal very few or no VLPs (**11**, **53**). In the first month of life, the infant virome undergoes a dramatic expansion, reaching up to $10^9$ VLPs per gram of stool, which is consistent with counts in healthy adults (**11**, **39**). Viral richness has in general been observed to increase over time from birth and strongly correlates with bacterial richness and diversity (**6**, **53**). However, Lim and colleagues (**54**) noted a steady decrease in both phage richness and diversity in fecal samples taken from infants and children over a period of two years. During the first year of life, the gut bacteriome has been described to go through three developmental phases, described as gut microbiota maturation stages. This has been characterized by shifts from *Escherichia* to *Bifidobacterium* and then to *Bacteroidetes* (**5**). It has been observed that over 60% of early infant phages associate with one of these gut microbiota maturation stages (**6**). Clostridia have also been noted to continuously increase in relative abundance over the first two years of life (**54**). Phages previously classified as Siphoviridae and Myoviridae, along with eukaryotic viruses, tend to dominate in infants but steadily decline in absolute abundance into childhood (**31**, **92**). Corresponding with the abundance of Siphoviridae phages, the vast majority of phages identified in infant stool samples have been predicted to have a temperate lifestyle (**6**). The relative abundance of these temperate phages also reduces over time, with mothers having a significantly lower level compared to their infants. This reduction, together with strong temporal correlations of phages and their bacterial hosts, has supported the hypothesis that the initial viral colonizers of the infant gut are prophages that excise from their bacterial hosts (**6**, **53**). Although lytic *Crassvirales* are the most abundant

phages in the adult gut, they are only present at a low abundance in the infant gut, often remaining absent in reads from fecal samples (**6**, **31**, **53**).

While several studies have focused on the dynamics of the infant gut virome, less attention has been given to the viromes of healthy children. In an analysis of gut viral richness across different age groups, using data sets from several studies, Gregory and colleagues (**31**) found that viral richness is lowest during childhood, before significantly increasing into adulthood. However, in a study on the gut virome of Danish children and adults, this significant increase in viral richness could not be recapitulated (**98**). Limitations in both papers, including uneven sample sizes (**31**) and a lack of data from young adults (**98**), may have contributed to the discrepancies in these results. During the transition from childhood to adulthood, the previously dominant phages classified as Siphoviridae and Myoviridae are outcompeted by Podoviridae, *Microviridae*, and *Crassvirales* (**31**). Bacterial composition remains relatively stable during this period, although slight shifts in the relative abundances of the phyla *Bacteroidetes* and *Actinobacteria*, with an increase in the former and a decrease in the latter, have been observed (**71**). A longitudinal study of ten healthy adults over a period of one year saw a similar persistence of phage families, although phages previously classified as Podoviridae were less abundant (**81**). In general, age appears to have a positive correlation with virome diversity, increasing from young adults to elderly adults in a Japanese cohort (**69**). However, in Western populations, this positive correlation is lost: Phage populations begin to decrease again as people reach old age. This is with the exception of *Crassvirales*, whose abundance continues to increase across the life span and appears to be at its highest in the elderly population (**31**). In an analysis of the gut viromes of over 4,000 Japanese individuals, age was shown to have the strongest association with the virome among 232 host and environmental factors analyzed. One hundred and seventy-six viral operational taxonomic units were significantly associated with age. Among these, phages of *Ruminococcus*, *Faecalibacterium*, and *Clostridium* were enriched in elderly people, while a viral cluster (VC) of *Bifidobacterium*-related phages were associated with young adults and steadily decreased with age (**69**).

Diet

It is well established that diet has a significant impact on the bacteriome. High-fat and Western diets, which have been associated with obesity, cardiovascular disease, and type 2 diabetes, have been shown to negatively impact bacterial gut diversity and function. Other diets associated with health, such as gluten-free, vegan, and vegetarian diets, have also resulted in differing bacterial compositions (**95**). Given the influence of diet on the abundance and diversity of phage hosts, it is a fair assumption that diet will exert similar influences on the phageome.

In infants, differences in phageome composition have been observed based on whether the infant was fed formula or breastfed. Specifically, temperate phages of *Bifidobacterium* and *Lactobacillus* are increased in breastfed infants, with a corresponding enrichment of their hosts (**53**). In an analysis of the viromes of 930 healthy Chinese adults, diet was found to have a significant effect on virome variation, with 22 dietary components contributing to significant variation. One hundred and thirty-seven correlations between dietary components and viral species could be made (**106**). Similarly, diet contributed to virome variation in Japanese individuals, with products such as milk, coffee, and fruit significantly associated (**69**). The interplay of ethnicity, diet, and virome variation was also highlighted in the former study (**106**). *Bacteroides* phages were noted to be enriched in the viromes of the Zang ethnic group as compared to the five other ethnic groups analyzed in this study. Butter milk tea and barley were predominant features in the diets of ethnic Zang, with barley significantly correlating with *Bacteroides* phages and butter milk tea being the primary contributor to bacterial variation (**106**). A limited number of studies involving deliberate dietary intervention have suggested that diet can have some effect on the composition of the virome. Implementation of a week-long gluten free diet in 11 healthy adults resulted in shifts in phage family abundance, decreasing phages previously classified as Podoviridae and increasing *Crassvirales*. At the viral genomic contig level, the number of shared contigs among individuals increased during a gluten free diet (**29**). A study of high- and low-fat dietary interventions also gave rise to

the idea that a shared diet can increase similarity in gut virome composition (**63**). A high-fat diet in mice significantly depletes phages previously classified as Siphoviridae (**77**). Concomitant with this is a significant reduction in integrase genes. Given that a high proportion of phages classified as Siphoviridae have a temperate lifestyle, these results suggest that a high-fat diet may shift the core virome toward expansion of lytic phages. However, this contrasts with findings by Kim & Bae (**44**), who observed an expansion of temperate phages previously classified as Caudovirales in mice on a high-fat, high-sucrose diet. Both fructose and short-chain fatty acids have been shown to activate prophage induction from *Lactobacillus reuteri* in the gut of mice (**72**). This induction appears to be partly due to the activation of stress responses in *L. reuteri*. In addition, many common dietary components were demonstrated to result in prophage induction from several gut bacteria (**9**), highlighting the potential influence that diet can exert on both the gut phageome and the microbiome as a whole.

Geography

Several recent studies have consistently noted differences in gut phageome composition based on geographical location (**12**, **65**, **98**, **106**). Two of the most prevalent phage genomes identified in a Danish cohort, a *Crassvirales* phage and a newly identified LoVEphage, were also recognized to be prolific worldwide. However, their abundance differed by location (**98**). The *Crassvirales* phage was found to be more abundant in healthy people in Asia and Europe than in healthy people in the United States and Africa. The LoVEphage was more prevalent in healthy Europeans and Americans than in healthy Asians and Africans. Thirty-seven other prevalent phage genomes were also significantly associated with geographical location, being found most commonly in Europeans (**98**). Similarly, another recently uncovered, highly abundant phage clade, Gubaphage, is most prolific in Europeans and absent in samples originating from South America (**12**). The relative abundance of ten candidate genera of *Crassvirales* varies significantly depending on geographical location. The most prevalent in Western populations is candidate genus I, which contains prototypical crAssphage (p-crAssphage) (named after the

cross-assembly software originally used to identify it), and candidate genus III (**12**, **33**). In Africa and South America, candidate genera VIII and IX are much more abundant. In the largest analysis of globally distributed p-crAssphage sequences, p-crAssphage was found to cluster based on geographical location (**26**). The gut viromes differed significantly between individuals living in two different regions of China, Hong Kong and Yunnan (**106**). All alpha diversity indices were significantly lower in the Hong Kong population: They were depleted for several phages previously classified as Caudovirales while being enriched for *Microviridae*. Urbanization also appears to play a role in variation, with significant differences in virome structure between individuals inhabiting urban and rural regions of Yunnan. The length of time spent in an urban region was also positively associated with several phages, including phages of lactococci and lactobacilli (**106**).

In a study examining diarrheal disease in an Indian cohort, individuals dwelling in the urban region of Nagpur had a less diverse virome than those living in the rural area of Melghat (**65**). It was also found that geographical location exerted more of an influence over virome variation than diarrheal disease did. It was also found to have an impact on gut virome alterations in obese patients from China. Obesity appeared to have more of an effect on the gut virome of individuals living in Hong Kong than it did on the Kunming cohort. There was a significant reduction in gut viral richness and diversity in obese Hong Kong individuals, but this significance was lost in the individuals from Kunming (**100**).

With geography demonstrated to have a considerable impact on the gut virome, it would be prudent to characterize the composition of healthy gut viromes across the world. Currently, gut virome studies are biased toward Europe and the United States, while coverage of Asia is dominated by studies focused on China and Japan. Studies focusing on South America and Africa are lacking. In recent efforts to characterize the viromes of indigenous populations in Malaysia, researchers found that their virome composition was markedly different from other viromes worldwide (**49**). A focus on analyzing the gut viromes of individuals from Bangladesh resulted in the discovery of Lak phages, termed megaphages due to their extremely large genomes (>540 kb) (**24**). *Prevotella* species dominated in

individuals in which these phages were identified, and these bacteria were identified as likely Lak phage hosts via CRISPR targeting. *Prevotella* is prevalent in indigenous and non-Western microbiomes, which is attributed to the consumption of a high-fiber, low-fat diet (**35**). This discovery highlights the importance of moving away from a Western-dominated view of the gut phageome.

## PHAGEOME VARIABILITY IN DISEASE

### Inflammatory Bowel Disease

Inflammatory bowel disease (IBD) is a term used to describe two chronic inflammatory conditions of the gut: ulcerative colitis (UC) and Crohn's disease. Although there is no known etiology, some causative factors have been linked, including genetics, diet, environmental exposures, and gut microbiome composition (**46**, **60**, **103**). A handful of studies have focused on changes to the gut virome in IBD patients, where alterations to the phageome have been observed, but the variations noted differ between studies. The most consistent observation is the increase in abundance of phages previously classified as Caudovirales in individuals with IBD, whether this be in relation to healthy controls (**70**, **99**, **105**) or in comparison across disease type (**27**), although no expansion of these phages has also been noted (**52**). An increase in the richness and diversity of phages previously classified as Caudovirales has also been reported (**17**, **70**), while Zuo and colleagues (**105**) observed a significant reduction in the richness and diversity of phages previously classified as Caudovirales. The viromes of mucosa samples taken from the rectum were examined here, rather than fecal samples, highlighting how sample sites can also contribute to variation. Multiple studies have also reported an altered ratio of the abundance of phages previously classified as Caudovirales to that of *Microviridae*: The relative abundance of phages previously classified as Caudovirales is significantly increased in IBD patients (**52**, **70**). A large quantity of phages previously classified as Caudovirales have been identified to be enriched

and associated with IBD. Most commonly these have been phages of *Lactococcus, Lactobacillus*, *Clostridium*, *Enterococcus*, and *Escherichia* (**40**, **52**, **70**, **105**). Imai and colleagues (**40**) observed a significant increase in abundance of *Crassvirales* in Crohn's disease patients. It should be noted that this pattern was observed in a cohort of only 19 patients. In contrast, an analysis of *Crassvirales* and their association with diseases found that 19 genus-level clusters of *Crassvirales* were significantly depleted in a cohort of >2,000 IBD patients (**34**). In addition, Clooney et al. (**17**) identified a depletion of a *Crassvirales* VC to be associated with a shift to IBD status.

Many studies use the viral-like sequences identified at the species level to determine virome-specific changes in disease. However, interpersonal gut virome variation can result in changes to the virome being masked at this level of analysis. In addition, often only those viral sequences that can be taxonomically assigned are used for analysis, despite this being as little as just 5% of the total in some studies (**76**). To address this, Clooney et al. (**17**) reanalyzed the data from Norman et al.'s study (**70**) and reduced taxonomic resolution by clustering viral sequences to family level (**8**). The majority of VCs found to be significantly increased in IBD patients were predicted to contain temperate phages, while the highly abundant VCs in controls were largely lytic in nature (**17**). Similarly, Sinha et al. (**86**) observed that an inoculum of pooled VLPs from UC patients contained a significantly higher proportion of temperate phages than that of healthy controls. These data suggest that in IBD patients, the core virome of virulent phages associated with healthy individuals (**81**) shifts to being predominantly temperate. It is likely that stressors associated with the inflamed environment of an IBD gut result in prophage induction. Importantly, this shift in composition was not detectable when analysis was performed at species level.

In addition to phageome variation, Zuo et al. (**105**) reported an increase in *Chrysochromulina ericina virus* and *Mimivirus*, two giant eukaryotic viruses of the family *Mimiviridae*, in their UC cohort. However, the use of 0.22-µm filters, followed by lysozyme and chloroform treatment during VLP extraction, means that the likelihood of *Mimivirus*, with a capsid size of 0.5 µm, ending up in the extracted

VLPs is extremely low (**89**). However, other studies that have used similar filter size and treatment steps have also reported the presence of *Mimiviridae* in the gut virome (**77**, **100**). The presence of these viruses likely reflects incorrect taxonomic assignment (**89**) and highlights one of the challenges of characterizing the human gut virome. It is also possible that virophages, which are known to infect *Mimiviridae* and carry *Mimiviridae*-like elements, are present and contribute to incorrect taxonomic assignment (**73**). Any correlation made between *Mimiviridae* and IBD or other diseases should be interpreted with caution.

Recent work in mice has hinted at a more direct role of altered phages in IBD. The addition of *E. coli* phages to the drinking water of specific pathogen-free mice was shown to exacerbate the severity of induced colitis. Phage interaction with the immune system, mediated by IFN-γ induction, resulted in this increase in inflammation. These in vivo findings were further validated when VLPs isolated from patients with active UC disease were shown to induce IFN-γ production from $CD4^+$ T cells (**30**). Similarly, mice colonized by bacterial communities from UC patients had increased proinflammatory cytokines and exacerbated colitis, when they underwent gavage with pooled VLPs from UC patients (**86**). In addition, UC VLPs were found to have an effect on the UC bacterial community, including reducing the relative abundance of *Eubacterium limosum* and increasing the relative abundance of *Escherichia* and *Shigella* species. *E. limosum* has previously been shown to improve experimental colitis symptoms (**42**), while pathobiont strains of *E. coli* are frequently implemented in IBD development (**64**). Taken together, these findings suggest a tripartite role of bacteria, phages, and the immune system in IBD development.

## Irritable Bowel Syndrome

Irritable bowel syndrome (IBS) is a common gastrointestinal disorder, estimated to affect up to 21% of the general population worldwide (**16**). Despite links to various host and environmental factors, there is no identified causative factor. Diagnosis is

based predominantly on symptoms and by exclusion of other inflammatory diseases. Rome IV criteria divide it into four clinical subtypes based on the primary symptoms: diarrhea (IBS-D), constipation (IBS-C), a mix of diarrhea and constipation (IBS-M), or unclassified (IBS-U) (**48**).

Four recent studies examining the gut virome in IBS patients have consistently reported variations to the gut phageome (**3**, **19**, **50**, **62**). In general, IBS patients have a less rich and diverse virome compared to controls (**3**, **19**), although Mihindukulasuriya et al. (**62**) did not find any difference in alpha diversity metrics. Coughlan and colleagues (**19**) reported that gut viromes of IBS-D and IBS-M but not IBS-C patients remained significantly less diverse than those of controls. Ansari and colleagues' (**3**) data set also contained these subtypes, while Mihindukulasuriya et al. had a data set of IBS-C and IBS-D patients only. Differences in subtypes may account for the differing diversity metrics reported and highlight how variation, even within the subtypes of disease, can influence reported results.

Coughlan et al. (**19**) identified ten VCs that were differentially abundant between IBS patients and controls. Seven were of reduced abundance in IBS patients, including two VCs previously classified as Siphoviridae and one previously classified as Podoviridae, while one VC previously classified as Podoviridae and one previously classified as Siphoviridae were increased in the IBS cohort (**19**). At the viral species level, Mihindukulasuriya et al. (**62**) also identified variation in abundance of ten phages classified as Siphoviridae, Myoviridae, Podoviridae, and *Microviridae*. They noted that an increase in these phages could also be stratified by IBS subtype. Li et al. (**50**) reported that the enrichment of viruses in IBS patients correlated with an enrichment of bacteria and that the same pattern occurred for the depletion of viruses and bacteria, signaling a possible interlinked role for the two in IBS. Unlike the latter three studies, Ansari et al. (**3**) did not find any significant alterations to the phageome. Instead, they noted an increase in abundance of several large eukaryotic viruses such as *Pandoravirus salinus* and *Herpesviridae* and a decreased abundance of several eukaryotic viruses that typically infect hosts other than humans, such as plants and amoebae.

The shift from a dominant virulent core to one that is predominantly temperate that is noted to occur in the gut phageome of patients with IBD (**17**) was not observed for patients with IBS (**19**). This emphasizes that despite IBD and IBS being somewhat similar symptomatically, there are likely different alterations and etiological factors at play. A couple of these studies were limited by a relatively small cohort of IBS patients, so that splitting by IBS subtype was not feasible. Given that there have been several indications that the IBS subtype contributes to variation, future studies should consider inclusion and analysis of all subtypes. Investigation into the link between the gut virome and IBS is still in its infancy, and future studies will hopefully elucidate whether changes are a causative factor of IBS or occur simply as a consequence of the disease.

Colorectal Cancer

Colorectal cancer (CRC) is the third-most common cancer diagnosis. Development of this cancer is complex and believed to be multifactorial, with risks including genetic and environmental factors and lifestyle choices (**23**, **59**). Alterations to the gut microbiome are increasingly recognized as an important environmental risk factor. Significant deviation to overall virome structure has been observed in CRC patients (**28**, **80**). In general, viral diversity and richness appear to increase in CRC patients (**66**, **80**, **107**), although in some studies alpha diversity metrics were not able to detect any differences (**28**, **36**). Despite this, every gut virome study of CRC patients has identified virome signatures associated with CRC, the majority of which are a result of changes in phage composition. In general, enrichment of phages classified as Siphoviridae, Myoviridae, Podoviridae, *Drexlerviridae*, and *Inoviridae* was identified to be driving this signature (**36**, **66**, **107**). Nakatsu and colleagues (**66**) reported that phages from the genera *Inovirus* and *Tunalikevirus* were the greatest discriminators between people with CRC and healthy controls. Gao et al. (**28**) identified an increase in abundance in phages infecting *Escherichia* and *Salmonella* in CRC patients, while p-crAssphage and *Enterobacteria* phages were depleted in these patients when compared to those in healthy controls. Enrichment of a *Parabacteroides* phage was also associated with a mutated KRAS

gene, a frequently mutated gene in CRC that is linked with poorer prognosis and recovery (**28**, **61**). *Herelleviridae*, a relatively newly identified family of phages that infect *Bacillota* (formerly Firmicutes), are also noted to be diminished in CRC patients (**4**, **107**).

Whether these shifts in phage composition are directly involved in CRC etiology is unknown. Enrichment of *Inoviridae* in CRC is intriguing, as they have been linked with biofilm development (**78**), a factor that has been associated with progression of CRC (**22**). Changes in a number of key bacterial species have been linked to CRC, some of which have proposed mechanisms for CRC progression. Enrichment of *Fusobacterium nucleatum* has been most commonly identified in several studies (**37**). This bacterium has the ability to promote tumorigenesis through the invasion of damaged gut epithelial cells and interacting with Toll-like receptors in the tumor cell (**45**). Expansion of enterotoxigenic *Bacteroides fragilis*, *E. coli* carrying the polyketide synthase gene, *Parvimonas micra*, and bacteria of the genera *Porphyromonas* and *Peptostreptococcus* has also been commonly reported (**15**). These bacteria are potential hosts to several of the phages previously noted to be enriched in CRC (**28**, **66**). In addition, Shen et al. (**80**) identified phages of *F. nucleatum* and *P. micra* as potential biomarkers of CRC. This suggests a possible interlink of bacteria and phages in the development of CRC. On the other hand, a low correlation in the abundance of the identified influential bacteria and phages has been noted (**36**). This has led to one hypothesis that alterations to the phageome indirectly impact CRC development. This would be through altering the previously stable bacterial community, providing the opportunity for the expansion of bacterial drivers of CRC (**36**). In a study analyzing gut virome changes during induced tumor formation in mice, phages of the genera *Brunovirus* and *Hpunavirus* were positively associated with tumor growth, hinting at a possible more direct involvement in CRC development (**51**). While these alterations in the phageome highlight the possible use of these phages as CRC biomarkers, further studies are required to elucidate the extent of the role of the phageome in CRC development and progression.

## METHODOLOGICAL VARIATION

It is evident that there is enormous variation in the human gut phageome of individuals. However, many metagenomic studies focus their attention on changes that occur in the identifiable fraction of the gut virome. The remaining unidentified component, termed “viral dark matter” (**2**), is likely to have a large impact on many facets of the gut microbiome and as demonstrated by Clooney et al. (**17**) can change the outcome of studies when it is included in analysis. Unlike 16S rRNA or the *chaperonin-60* (*cpn60*) gene in bacteria, viruses have no universal phylogenetic marker, leaving researchers to rely on sequence homology searches of reference databases. The primary reference database used, NCBI RefSeq, has only 11,598 viral reference genomes (as of September 2022). Rapid mutation rates and mosaicism also lead to many highly divergent sequences unlikely to be present in databases, further hindering identification. To overcome this, many studies incorporate de novo assembly of reads, which investigators can then attempt to annotate and classify through a combination of de novo and reference-based steps. The choice of assembler has also been shown to impact virome identification and classification (**90**).

Perhaps the best example of the requirement to move away from using only reference-based identification was the discovery of *Crassvirales*. They remained a component of the large unidentified mass of viral reads until 2014, when the use of cross-assembly software (crAss) led to the identification of p-crAssphage (**25**). Since then, a plethora of *Crassvirales* have been identified from diverse environments (**26**, **101**, **102**). Highly prolific and widespread, they can account for over 86% of viral reads from some individuals and are a core component of the virome in healthy adults (**101**). Recent efforts based on de novo assembly from globally distributed metagenomes have led to the establishment of several human gut virome databases, expanding our knowledge on both the abundance and diversity of gut virome genomes (**12**, **31**, **67**, **93**). The establishment of both these and future databases will be beneficial for alignment-based analysis. The assembly of a catalog of gut virome databases has also led to the identification of several new phages.

The highly prevalent LoVEphage, named as such because it has lots of viral elements, was discovered through the assembly of a Danish gut viral catalog (**98**). The assembly of the Gut Phage Database, the largest amalgamation of gut phage genomes so far, led to identification of a clade of phage named Gubaphage. With many similar features to *Crassvirales*, they were also found to be highly ubiquitous, appearing in samples from all continents except for South America (**12**).

Although suggestions of standard protocols have been put forward (**18**, **84**), there is currently no standardized approach to studying the gut phageome. This leads to inevitable variation in results, as researchers adopt different protocols. It also hinders accurate comparisons across studies. There are two main methods through which viral sequences are isolated: (*a*) viral extraction from stool and enrichment and (*b*) a bulk metagenomics approach. While bulk metagenomics has the advantages of a simpler isolation protocol and capturing temperate phages, rare phages are less likely to be recovered (**75**). Recent gut virome studies have seen an increase in the use of bulk metagenomics (**3**, **20**, **65**, **94**), and comparison between the two methods has demonstrated that a greater number of viral sequences are detected from bulk metagenomics (**31**). Continuous improvements to identification of viral sequences with bulk metagenomics (**41**) further increase the appeal of this method. However, comparison of the two methods also shows only ~10% of identified viral sequences shared between enriched and bulk viromes (**31**), highlighting how the two methods select for very different portions of the virome. This also suggests that an integrated approach of both VLP enrichment and bulk metagenomics should be used to capture greater viral diversity and representation. In the enrichment and isolation of VLPs, there is often a trade-off between maximizing the number of VLPs obtained and minimizing the amount of bacterial contamination. Filters with a pore size of 0.22–0.45 μm are commonly used to reduce bacterial contamination but at the loss of large and enveloped viruses. Chloroform can provide further purification from bacterial cells but also results in the loss of some chloroform-sensitive viruses. The use of RNase to remove free nucleic acids could potentially have a negative impact on RNA viruses, leading to further bias toward enrichment of DNA viruses (**1**). In addition to this, many studies

do not include a reverse transcription step to include RNA viruses and focus on the DNA portion alone. The recent expansion of metagenomic sequences predicted to be RNA phages further suggests that the number of gut RNA phages is being underestimated (**68**). Extraction and enrichment of nucleic acids often result in a low yield and require an amplification step. Multiple displacement amplification is most often used in studies (**31**), despite being well documented to introduce bias via amplification of ssDNA viruses. This is due to the preference of the enzyme used, polymerase φ29, to propagate ssDNA (**43**). It is likely to be a contributing factor as to why some studies report a high relative abundance of *Microviridae.*

The choice of sequencing platform is also important, as sequencing depth, chemistry, and read length can all further complicate the analysis of virome variation. This can be seen particularly in some of the earlier IBD studies that were restricted by the sequencing depth of 454 pyrosequencing and therefore found limited changes to the gut virome (**74**, **99**). Due to much lower DNA concentrations being required, most gut virome studies use short-read sequencing platforms. Short-read sequencing also has the advantage of allowing deep sequencing but can hinder downstream analysis, as short reads can often lead to fragmented assemblies (**91**). The use of long-read sequencing could overcome these assembly inaccuracies, but the requirement of a high starting concentration of DNA limits its application in gut virome analysis. Recently, Lee et al. (**49**) used an optimized PEG precipitation protocol to achieve on average 20 ng/μL of DNA from fecal samples of indigenous Malaysians. This high starting concentration could be used for long-read sequencing, which despite a lower sequencing depth was shown to improve the breadth of coverage compared to short-read sequencing.

## CONCLUSIONS

Many other biological and health and disease states have also been associated with alterations to the gut phageome. Sex significantly correlates with the gut phageome, with males enriched in phages that infect *Megamonas* and *Prevotella*

while females have increased abundances of *Faecalibacterium* and *Ruminococcus* phages (**69**). *Crassvirales* have been found to be significantly depleted in patients with two autoimmune diseases, rheumatoid arthritis and systemic lupus erythematosus (**94**). Along with other phageome alterations, *Crassvirales* were significantly reduced in patients infected with SARS-CoV-2 (**13**, **57**, **104**). Phage VCs were significantly altered in individuals with metabolic syndrome (**20**). At the same time, an analysis of 232 host and environmental factors contributed only 0.6% of variance to the gut virome in a cohort of healthy Japanese individuals. This was much lower than the 5% of variance of the bacteriome that could be attributed to these factors (**69**), suggesting that there are still many unknown components contributing to virome variation between individuals.

To understand alterations to the gut virome in disease states, it is vital that we have a clear understanding of virome composition in healthy individuals. Currently our view of a healthy phageome is predominantly based on Western adults. Data from children and elderly people are lacking, as well as data from people in large portions of Africa, Asia, and South America. It is important that future work focus on incorporating data from these under-sampled groups in order to form as complete an understanding as possible of the gut phageome. This understanding is also necessary because of the increasing use of phage therapy and other medical applications such as fecal microbiota/viral transplantation.

Standardized protocols are required to enable cross study comparisons. At the same time, this should not dissuade researchers from pursuing novel techniques, as these are likely to further advance the field of gut virome research. Despite great progress in the last couple of years, gut phageome research is still in its infancy, particularly compared to the efforts made in deciphering the gut bacteriome. This is a dynamic field of research, and as methods improve and studies adapt to improve gaps in gut phageome knowledge, it is likely that current views will change.

## References


1. Adriaenssens EM, Farkas K, Harrison C, Jones DL, Allison HE, McCarthy AJ. **2018**. Viromic analysis of wastewater input to a river catchment reveals a diverse assemblage of RNA viruses. *mSystems* 3:e00025–18

2. Aggarwala V, Liang G, Bushman FD. **2017**. Viral communities of the human gut: metagenomic analysis of composition and dynamics. *Mob. DNA* 8:12

3. Ansari MH, Ebrahimi M, Fattahi MR, Gardner MG, Safarpour AR et al. **2020**. Viral metagenomic analysis of fecal samples reveals an enteric virome signature in irritable bowel syndrome. *BMC Microbiol.* 20:123

4. Barylski J, Kropinski AM, Alikhan NF, Adriaenssens EM, ICTV Rep. Consort **2020**. ICTV virus taxonomy profile: *Herelleviridae*. *J. Gen. Virol.* 101:362–63

5. Beller L, Deboutte W, Falony G, Vieira-Silva S, Tito RY et al. **2021**. Successional stages in infant gut microbiota maturation. *mBio*12:e0185721

6. Beller L, Deboutte W, Vieira-Silva S, Falony G, Tito RY et al. **2022**. The virota and its transkingdom interactions in the healthy infant gut. *PNAS* 119:e2114619119

7. Benler S, Yutin N, Antipov D, Rayko M, Shmakov S et al. **2021**. Thousands of previously unknown phages discovered in whole-community human gut metagenomes. *Microbiome* 9:78

8. Bin Jang H, Bolduc B, Zablocki O, Kuhn JH, Roux S et al. **2019**. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* 37:632–39

9. Boling L, Cuevas DA, Grasis JA, Kang HS, Knowles B et al. **2020**. Dietary prophage inducers and antimicrobials: toward landscaping the human gut microbiome. *Gut Microbes* 11:721–34

10. Borodovich T, Shkoporov AN, Ross RP, Hill C. **2022**. Phage-mediated horizontal gene transfer and its implications for the human gut microbiome. *Gastroenterol. Rep.* 10:goac012

11. Breitbart M, Haynes M, Kelley S, Angly F, Edwards RA et al. **2008**. Viral diversity and dynamics in an infant gut. *Res. Microbiol.* 159:367–73

12. Camarillo-Guerrero LF, Almeida A, Rangel-Pineros G, Finn RD, Lawley TD. **2021**. Massive expansion of human gut bacteriophage diversity. *Cell* 184:1098–109.e9

13. Cao J, Wang C, Zhang Y, Lei G, Xu K et al. **2021**. Integrated gut virome and bacteriome dynamics in COVID-19 patients. *Gut Microbes*13:1887722

14. Cenens W, Makumi A, Mebrhatu MT, Lavigne R, Aertsen A. **2013**. Phage-host interactions during pseudolysogeny: lessons from the Pid/dgo interaction. *Bacteriophage* 3:e25029

15. Cheng Y, Ling Z, Li L. **2020**. The intestinal microbiota and colorectal cancer. *Front. Immunol.* 11:615056

16. Chey WD, Kurlander J, Eswaran S. **2015**. Irritable bowel syndrome: a clinical review. *JAMA* 313:949–58

17. Clooney AG, Sutton TDS, Shkoporov AN, Holohan RK, Daly KM et al. **2019**. Whole-virome analysis sheds light on viral dark matter in inflammatory bowel disease. *Cell Host Microbe* 26:764–78.e5

18. Conceição-Neto N, Zeller M, Lefrère H, De Bruyn P, Beller L et al. **2015**. Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. *Sci. Rep.* 5:16532

19. Coughlan S, Das A, O'Herlihy E, Shanahan F, O'Toole PW, Jeffery IB. **2021**. The gut virome in Irritable Bowel Syndrome differs from that of controls. *Gut Microbes* 13:1

20. de Jonge PA, Wortelboer K, Scheithauer TPM, van den Born BH, Zwinderman AH et al. **2022**. Gut virome profiling identifies a widespread bacteriophage family associated with metabolic syndrome. *Nat. Commun.* 13:3594

21. De Sordi L, Lourenço M, Debarbieux L. **2019**. The battle within: interactions of bacteriophages and bacteria in the gastrointestinal tract. *Cell Host Microbe* 25:210–18

22. Dejea CM, Wick EC, Hechenbleikner EM, White JR, Mark Welch JL et al. **2014**. Microbiota organization is a distinct feature of proximal colorectal cancers. *PNAS* 111:18321–26

23. Dekker E, Tanis PJ, Vleugels JLA, Kasi PM, Wallace MB. **2019**. Colorectal cancer. *Lancet* 394:1467–80

24. Devoto AE, Santini JM, Olm MR, Anantharaman K, Munk P et al. **2019**. Megaphages infect *Prevotella* and variants are widespread in gut microbiomes. *Nat. Microbiol.* 4:693– 700

25. Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GG et al. **2014**. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* 5:4498

26. Edwards RA, Vega AA, Norman HM, Ohaeri M, Levi K et al. **2019**. Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. *Nat. Microbiol.* 4:1727–36

27. Fernandes MA, Verstraete SG, Phan TG, Deng X, Stekol E et al. **2019**. Enteric virome and bacterial microbiota in children with ulcerative colitis and Crohn disease. *J. Pediatr. Gastroenterol. Nutr.* 68:30–36

28. Gao R, Zhu Y, Kong C, Xia K, Li H et al. **2021**. Alterations, interactions, and diagnostic potential of gut bacteria and viruses in colorectal cancer. *Front. Cell Infect. Microbiol.* 11:657867

29. Garmaeva S, Gulyaeva A, Sinha T, Shkoporov AN, Clooney AG et al. **2021**. Stability of the human gut virome and effect of gluten-free diet. *Cell Rep.* 35:109132

30. Gogokhia L, Buhrke K, Bell R, Hoffman B, Brown DG et al. **2019**. Expansion of bacteriophages is linked to aggravated intestinal inflammation and colitis. *Cell Host Microbe* 25:285–99.e8

31. Gregory AC, Zablocki O, Zayed AA, Howell A, Bolduc B, Sullivan MB. **2020**. The Gut Virome

Database reveals age-dependent patterns of virome diversity in the human gut. *Cell Host Microbe* 28:724–40.e8

32. Guerin E, Hill C. **2020**. Shining light on human gut bacteriophages. *Front. Cell Infect. Microbiol.* 10:481

33. Guerin E, Shkoporov A, Stockdale SR, Clooney AG, Ryan FJ et al. **2018**. Biology and taxonomy of crAss-like bacteriophages, the most abundant virus in the human gut. *Cell Host Microbe* 24:653–64.e6

34. Gulyaeva A, Garmaeva S, Ruigrok R, Wang D, Riksen NP et al. **2022**. Discovery, diversity, and functional associations of crAss-like phages in human gut metagenomes from four Dutch cohorts. *Cell Rep.* 38:110204

35. Gupta VK, Paul S, Dutta C. **2017**. Geography, ethnicity or subsistence-specific variations in human microbiome composition and diversity. *Front. Microbiol.* 8:1162

36. Hannigan GD, Duhaime MB, Ruffin MT 4th, Koumpouras CC, Schloss PD. **2018**. Diagnostic potential and interactive dynamics of the colorectal cancer virome. *mBio* 9:6e02248–18

37. Hashemi Goradel N, Heidarzadeh S, Jahangiri S, Farhood B, Mortezaee K et al. **2019**. *Fusobacterium nucleatum* and colorectal cancer: a mechanistic overview. *J. Cell Physiol.* 234:2337–44

38. Hobbs Z, Abedon ST. **2016**. Diversity of phage infection types and associated terminology: the problem with 'Lytic or lysogenic. '. *FEMS Microbiol. Lett.* 363:7fnw047

39. Hoyles L, McCartney AL, Neve H, Gibson GR, Sanderson JD et al. **2014**. Characterization of virus-like particles associated with the human faecal and caecal microbiota. *Res. Microbiol.* 165:10803–12

40. Imai T, Inoue R, Nishida A, Yokota Y, Morishima S et al. **2022**. Features of the gut prokaryotic virome of Japanese patients with Crohn's disease. *J. Gastroenterol.* 57:559–70

41. Johansen J, Plichta DR, Nissen JN, Jespersen ML, Shah SA et al. **2022**. Genome binning of viral entities from bulk metagenomics data. *Nat. Commun.* 13:965

42. Kanauchi O, Fukuda M, Matsumoto Y, Ishii S, Ozawa T et al. **2006**. *Eubacterium limosum* ameliorates experimental colitis and metabolite of microbe attenuates colonic inflammatory action with increase of mucosal integrity. *World J. Gastroenterol.* 12:71071–77

43. Kim KH, Bae JW. **2011**. Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Appl. Environ. Microbiol.* 77:7663–68

44. Kim MS, Bae JW. **2016**. Spatial disturbances in altered mucosal and luminal gut viromes of diet-induced obese mice. *Environ. Microbiol.* 18:1498–510

45. Kostic AD, Chun E, Robertson L, Glickman JN, Gallini CA et al. **2013**. *Fusobacterium nucleatum* potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host Microbe* 14:207–15

46. Krela-Kaźmierczak I, Zakerska-Banaszak O, Skrzypczak-Zielińska M, Łykowska-Szuber L, Szymczak-Tomczak A et al. **2022**. Where do we stand in the behavioral pathogenesis of

inflammatory bowel disease? The Western dietary pattern and microbiota—a narrative review. *Nutrients* 14:122520

47. Krupovic M, Turner D, Morozova V, Dyall-Smith M, Oksanen HM et al. **2021**. Bacterial Viruses Subcommittee and Archaeal Viruses Subcommittee of the ICTV: update of taxonomy changes in 2021. *Arch. Virol.* 166:3239–44

48. Lacy BE, Patel NK. **2017**. Rome criteria and a diagnostic approach to irritable bowel syndrome. *J. Clin. Med.* 6:1199

49. Lee CZ, Zoqratt M, Phipps ME, Barr JJ, Lal SK et al. **2022**. The gut virome in two indigenous populations from Malaysia. *Sci. Rep.*12:1824

50. Li M, Wang C, Guo Q, Xu C, Xie Z et al. **2022**. More positive or more negative? Metagenomic analysis reveals roles of virome in human disease-related gut microbiome. *Front. Cell Infect. Microbiol.* 12:846063

51. Li Y, Zhang F, Zheng H, Kalasabail S, Hicks C et al. **2022**. Fecal DNA virome is associated with the development of colorectal neoplasia in a murine model of colorectal cancer. *Pathogens* 11:4457

52. Liang G, Conrad MA, Kelsen JR, Kessler LR, Breton J et al. **2020**. Dynamics of the stool virome in very early-onset inflammatory bowel disease. *J. Crohn's Colitis* 14:1600–10

53. Liang G, Zhao C, Zhang H, Mattei L, Sherrill-Mix S et al. **2020**. The stepwise assembly of the neonatal virome is modulated by breastfeeding. *Nature* 581:470–74

54. Lim ES, Zhou Y, Zhao G, Bauer IK, Droit L et al. **2015**. Early life dynamics of the human gut virome and bacterial microbiome in infants. *Nat. Med.* 21:1228–34

55. Los M, Wegrzyn G, Neubauer P. **2003**. A role for bacteriophage T4 *rI* gene function in the control of phage development during pseudolysogeny and in slowly growing host cells. *Res. Microbiol.* 154:547–52

56. Lourenço M, Chaffringeon L, Lamy-Besnier Q, Pédron T, Campagne P et al. **2020**. The spatial heterogeneity of the gut limits predation and fosters coexistence of bacteria and bacteriophages. *Cell Host Microbe* 28:390–401.e5

57. Lu ZH, Zhou HW, Wu WK, Fu T, Yan M et al. **2021**. Alterations in the composition of intestinal DNA virome in patients with COVID-19. *Front. Cell Infect. Microbiol.* 11:790422

58. Mäntynen S, Laanto E, Oksanen HM, Poranen MM, Díaz-Muñoz SL. **2021**. Black box of phage-bacterium interactions: exploring alternative phage infection strategies. *Open. Biol.* 11:210188

59. Mármol I, Sánchez-de-Diego C, Pradilla Dieste A, Cerrada E, Rodriguez Yoldi MJ. **2017**. Colorectal carcinoma: a general overview and future perspectives in colorectal cancer. *Int. J. Mol. Sci.* 18:1197

60. McGovern DP, Kugathasan S, Cho JH. **2015**. Genetics of inflammatory bowel diseases. *Gastroenterology* 149:1163–76.e2

61. Meng M, Zhong K, Jiang T, Liu Z, Kwan HY, Su T. **2021**. The current understanding on the impact of KRAS on colorectal cancer. *Biomed. Pharmacother.* 140:111717

62. Mihindukulasuriya KA, Mars RAT, Johnson AJ, Ward T, Priya S et al. **2021**. Multi-omics analyses show disease, diet, and transcriptome interactions with the virome. *Gastroenterology* 161:1194–207.e8

63. Minot S, Sinha R, Chen J, Li H, Keilbaugh SA et al. **2011**. The human gut virome: interindividual variation and dynamic response to diet. *Genome Res.* 21:1616–25

64. Mirsepasi-Lauridsen HC, Vallance BA, Krogfelt KA, Petersen AM. **2019**. *Escherichia coli* pathobionts associated with inflammatory bowel disease. 322e00060–18

65. Monaghan TM, Sloan TJ, Stockdale SR, Blanchard AM, Emes RD et al. **2020**. Metagenomics reveals impact of geography and acute diarrheal disease on the Central Indian human gut microbiome. *Gut Microbes* 12:1752605

66. Nakatsu G, Zhou H, Wu WKK, Wong SH, Coker OO et al. **2018**. Alterations in enteric virome are associated with colorectal cancer and survival outcomes. *Gastroenterology* 155:529–41.e5

67. Nayfach S, Páez-Espino D, Call L, Low SJ, Sberro H et al. **2021**. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat. Microbiol.* 6:960–70

68. Neri U, Wolf YI, Roux S, Camargo AP, Lee B et al. **2022**. Expansion of the global RNA virome reveals diverse clades of bacteriophages. *Cell* 185:214023–37.e18

69. Nishijima S, Nagata N, Kiguchi Y, Kojima Y, Miyoshi-Akiyama T et al. **2022**. Extensive gut virome variation and its associations with host and environmental factors in a population level cohort. *Nat. Commun.* 13:5252

70. Norman JM, Handley SA, Baldridge MT, Droit L, Liu CY et al. **2015**. Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* 160:447–60

71. Odamaki T, Kato K, Sugahara H, Hashikura N, Takahashi S et al. **2016**. Age-related changes in gut microbiota composition from newborn to centenarian: a cross-sectional study. *BMC Microbiol.* 16:90

72. Oh JH, Alexander LM, Pan M, Schueler KL, Keller MP et al. **2019**. Dietary fructose and microbiota-derived short-chain fatty acids promote bacteriophage production in the gut symbiont *Lactobacillus reuteri*. *Cell Host Microbe* 25:273–84.e6

73. Paez-Espino D, Zhou J, Roux S, Nayfach S, Pavlopoulos GA et al. **2019**. Diversity, evolution, and classification of virophages uncovered through global metagenomics. *Microbiome* 7:157

74. Pérez-Brocal V, García-López R, Vázquez-Castellanos JF, Nos P, Beltrán B et al. **2013**. Study of the viral and microbial communities associated with Crohn's disease: a metagenomic approach. *Clin. Transl. Gastroenterol.* 4:e36

75. Roux S, Adriaenssens EM, Dutilh BE, Koonin EV, Kropinski AM et al. **2019**. Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nat. Biotechnol.* 37:29–37

76. Roux S, Hallam SJ, Woyke T, Sullivan MB. **2015**. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *eLife* 4:e08490

77. Schulfer A, Santiago-Rodriguez TM, Ly M, Borin JM, Chopyk J et al. **2020**. Fecal viral community responses to high-fat diet in mice. *mSphere* 5:1e00833–19

78. Secor PR, Sweere JM, Michaels LA, Malkovskiy AV, Lazzareschi D et al. **2015**. Filamentous bacteriophage promote biofilm assembly and function. *Cell Host Microbe* 18:549–59

79. Sender R, Fuchs S, Milo R. **2016**. Revised estimates for the number of human and bacteria cells in the body. *PLOS Biol.* 14:e1002533

80. Shen S, Huo D, Ma C, Jiang S, Zhang J. **2021**. Expanding the colorectal cancer biomarkers based on the human gut phageome. *Microbiol. Spectr.* 9:e0009021

81. Shkoporov AN, Clooney AG, Sutton TDS, Ryan FJ, Daly KM et al. **2019**. The human gut virome is highly diverse, stable, and individual specific. *Cell Host Microbe* 26:527–41.e5

82. Shkoporov AN, Hill C. **2019**. Bacteriophages of the human gut: the "known unknown" of the microbiome. *Cell Host Microbe* 25:195–209

83. Shkoporov AN, Khokhlova EV, Stephens N, Hueston C, Seymour S et al. **2021**. Long-term persistence of crAss-like phage crAss001 is associated with phase variation in *Bacteroides intestinalis*. *BMC Biol.* 19:163

84. Shkoporov AN, Ryan FJ, Draper LA, Forde A, Stockdale SR et al. **2018**. Reproducible protocols for metagenomic analysis of human faecal phageomes. *Microbiome* 6:68

85. Shkoporov AN, Turkington CJ, Hill C. **2022**. Mutualistic interplay between bacteriophages and bacteria in the human gut. *Nat. Rev. Microbiol.* 20:12737–49

86. Sinha A, Li Y, Mirzaei MK, Shamash M, Samadfam R et al. **2022**. Transplantation of bacteriophages from ulcerative colitis patients shifts the gut bacteriome and exacerbates the severity of DSS colitis. *Microbiome* 10:105

87. Siringan P, Connerton PL, Cummings NJ, Connerton IF. **2014**. Alternative bacteriophage life cycles: the carrier state of *Campylobacter jejuni*. *Open. Biol.* 4:130200

88. Smith L, Goldobina E, Govi B, Shkoporov AN. **2023**. Bacteriophages of the order *Crassvirales*: What do we currently know about this keystone component of the human gut virome?. *Biomolecules* 13:4584

89. Sutton TDS, Clooney AG, Hill C. **2020**. Giant oversights in the human gut virome. *Gut* 69:71357–58

90. Sutton TDS, Clooney AG, Ryan FJ, Ross RP, Hill C. **2019**. Choice of assembly software has a critical impact on virome characterisation. *Microbiome* 7:12

91. Sutton TDS, Hill C. **2019**. Gut bacteriophage: current understanding and challenges. *Front. Endocrinol.* 10:784

92. Taboada B, Morán P, Serrano-Vázquez A, Iša P, Rojas-Velázquez L et al. **2021**. The gut virome of healthy children during the first year of life is diverse and dynamic. *PLOS ONE* 16:e0240958

93. Tisza MJ, Buck CB. **2021**. A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases. *PNAS* 118:23e2023202118

94. Tomofuji Y, Kishikawa T, Maeda Y, Ogawa K, Nii T et al. **2022**. Whole gut virome analysis of 476 Japanese revealed a link between phage and autoimmune disease. *Ann. Rheum. Dis.* 81:278–88

95. Tomova A, Bukovsky I, Rembert E, Yonas W, Alwarith J et al. **2019**. The effects of vegetarian and vegan diets on gut microbiota. *Front. Nutr.* 6:47

96. Turner D, Kropinski AM, Adriaenssens EM. **2021**. A roadmap for genome-based phage taxonomy. *Viruses* 13:3506

97. Turner D, Shkoporov AN, Lood C, Millard AD, Dutilh BE et al. **2023**. Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ICTV bacterial viruses subcommittee. *Arch. Virol.* 168:74

98. Van Espen L, Bak EG, Beller L, Close L, Deboutte W et al. **2021**. A previously undescribed highly prevalent phage identified in a Danish enteric virome catalog. *mSystems* 6:e0038221

99. Wagner J, Maksimovic J, Farries G, Sim WH, Bishop RF et al. **2013**. Bacteriophages in gut samples from pediatric Crohn's disease patients: metagenomic analysis using 454 pyrosequencing. *Inflamm. Bowel Dis.* 19:1598–608

100. Yang K, Niu J, Zuo T, Sun Y, Xu Z et al. **2021**. Alterations in the gut virome in obesity and type 2 diabetes mellitus. *Gastroenterology* 161:1257–69.e13

101. Yutin N, Benler S, Shmakov SA, Wolf YI, Tolstoy I et al. **2021**. Analysis of metagenome eassembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features. *Nat. Commun.* 12:1044

102. Yutin N, Makarova KS, Gussow AB, Krupovic M, Segall A et al. **2018**. Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nat. Microbiol.* 3:38–46

103. Zhao M, Feng R, Ben-Horin S, Zhuang X, Tian Z et al. **2022**. Systematic review with metaanalysis: environmental and dietary differences of inflammatory bowel disease in Eastern and Western populations. *Aliment. Pharmacol. Ther.* 55:266–76

104. Zuo T, Liu Q, Zhang F, Yeoh YK, Wan Y et al. **2021**. Temporal landscape of human gut RNA and DNA virome in SARS-CoV-2 infection and severity. *Microbiome* 9:91

105. Zuo T, Lu XJ, Zhang Y, Cheung CP, Lam S et al. **2019**. Gut mucosal virome alterations in ulcerative colitis. *Gut*68:1169–79

106. Zuo T, Sun Y, Wan Y, Yeoh YK, Zhang F et al. **2020**. Human-gut-DNA virome variations across geography, ethnicity, and urbanization. *Cell Host Microbe* 28:741–51.e4

107. Zuo W, Michail S, Sun F. **2022**. Metagenomic analyses of multiple gut datasets revealed the association of phage signatures in colorectal cancer. *Front. Cell Infect. Microbiol.* 12:918010

# Chapter 2: The gut bacteriome and virome of patients with acute COVID is similar to those post recovery

Ciara A Tobin, Junhui Li, Manasi Nadkarni, Lorraine A Draper, Andrey N Shkoporov and Colin Hill

## Abstract

The COVID-19 pandemic caused by SARS-COV-2 infection has impacted the lives of millions of people worldwide. Here, we characterised both the virome and bacteriome of faecal and saliva samples obtained from 43 COVID-19 patients in acute and recovery stages of the disease. We identified no significant alterations to the gut virome and bacteriome between acute and convalescent patients, characterised by no significant changes to alpha and beta diversity metrics. Eight bacterial species were discovered to be differentially abundant between acute and convalescent patients. This included a loss of *Coprococcus eutactus* and *Roseburia faecis* and an enrichment of *Enterococcus* and *Lactobacillus* in acute patients. There was an overall trend of reducing bacterial α-diversity with increasing disease severity. Additionally, a viral cluster (VC) with a predicted *Eubacterium* host was found to be differentially abundant in patients who had critical COVID infection. While there was a significant separation of the oral bacteriome between acute and convalescent patients, no other differences could be identified between the oral virome and bacteriome of COVID patients. Overall, the gut and oral viromes and bacteriomes of acute and convalescent patients were quite similar. Recent studies have identified consistent perturbations to the gut bacteriome many months post COVID infection that have been linked to the development of Long COVID. Follow up studies with matched healthy controls would be required to determine if that is also the case in the gut virome.

## Introduction

Severe acute respiratory syndrome-coronavirus-2 (SARS-COV-2), the causative agent of coronavirus disease 2019 (COVID-19), has been reported to have been involved in more than seven million deaths globally to date[1]. Although primarily a respiratory pathogen, gastrointestinal (GI) symptoms such as nausea, vomiting and diarrhea are commonly reported by patients[2] and are associated with increased disease severity[3]. One of the key receptors for SARS-COV-2 is Angiotensin converting enzyme 2 (ACE2)[4,5] that is highly expressed in the GI tract[6]. SARS-COV-2 viral RNA can be detected in faecal samples from patients and can persist even after respiratory swabs become negative[7]. This suggests that alterations to the consortium of microorganisms that inhabit the GI tract, collectively referred to as the gut microbiome, may also result from SARS-COV-2 infection. Coinciding with this are reports of reduced bacterial diversity, depletion of beneficial bacteria such as short chain fatty acid (SCFA) producers and increased abundance of opportunistic bacterial pathogens in patients in the acute stage of disease, compared to healthy individuals[8,9,10,11,12,13]. In addition to bacterial changes, alterations to the viral[14,15,16] and fungal[17,18,19] components of the gut microbiome have also been reported.

The oral microbiome is the second most diverse microbiome (after that of the gut) in the human body[20]. The oral cavity acts as one of the main routes of entry for SARS-COV-2 infection, with SARS-COV-2 commonly present in saliva during infection[21]. Similar to the gut microbiome, significant alterations to the oral microbial community have been observed following SARS-COV-2 infection[22,23,24]. This includes the enrichment of pro-inflammatory genera such as *Prevotella* and *Veillonella* and bacterial genera linked with the development of periodontitis[25,26].

Post COVID-19 condition, commonly referred to as “Long COVID”, is defined by the World Health Organisation as the continuation or development of new symptoms three months post-infection and these symptoms continuing for at least another two months, with no other explanation[27]. It is a multisystemic condition, comprising

over 200 symptoms and can be quite debilitating[28]. Estimates of the incidence rate vary, but it is thought to occur in up to 30% of non-hospitalised patients[28]. Multiple mechanisms have been proposed as potential causative factors for Long COVID, including immune system dysregulation[29,30], SARS-COV-2 viral RNA persistence[31,32] and perturbations to the gut microbiome[32,33,34]. In Long COVID patients, alterations to the gut bacterial community have been reported up to one year post-infection[35]. However, both the gut and oral virome of convalescent patients have yet to be examined.

In this study, we obtained a total of 84 faecal samples from 43 individuals diagnosed with COVID-19. Sixteen of these samples were from patients during the acute stage of infection and 27 samples from patients post recovery (convalescent). The remaining 40 faecal samples were provided as follow up samples from three of the acute patients and 21 convalescent. These were used to characterise both the bacteriome and virome and to identify any changes due to the stage of disease or severity of COVID infection. Additionally, two acute and 13 convalescent patients provided saliva samples, which allowed the compositional analysis of the oral bacteriome and virome.

## Results

### Gut virome composition in acute and convalescent COVID-19 patients

To study changes to the gut microbiome both during and post-recovery, faecal samples were obtained from 43 individuals hospitalised with SARS-COV-2 infection (median age 57, 47% male, **Fig S1a**). Faecal samples were obtained from 16 individuals ≤ 19 days post infection, referred to as acute samples. The remaining 27 samples were provided 2-5 months post infection (convalescent samples). In addition, 24 patients provided a second faecal sample six months post infection and a further 16 patients provided a third faecal sample seven months post infection, resulting in a total of 84 faecal samples (**Fig S1b**).

The final catalogue of viral sequences included 59,240 nonredundant partial and complete genomes, as well as genome fragments. The majority of these (46,752) were identified by displaying a degree of sequence similarity (≥ 50% identity over ≥90% of contig length in BLASTn) with previously reported genomes. The remaining genomes were identified using a multi-classifier approach employed by VirSorter2[36]. Only short (200-350 bp) contigs aligning to the SARS-COV-2 genome could be identified in 3 acute patients, who were 9-15 days post diagnosis.

To compensate for inter-individual variation, contigs > 3Kb were clustered to approximately genus level with vContact2[37]. This resulted in 2,446 viral clusters (VC's), comprised of 11,244 contigs. Contigs classified as either singletons or outliers by vContact2 were also added as individual VC's, bringing the total number of VC's to 8,318. Taxonomic classification of these VC's identified 17 different viral families. The nine most abundant families are shown in **Figure 1a**, with the remaining less abundant families grouped as "Other". Bacteriophages (phages) that were previously classified in the now abolished "Podoviridae", "Siphoviridae" and "Myoviridae" families[38,39] were manually assigned as "Unclassified *Caudoviricetes*." Approximately 53% of VC's in acute patients and 57% in convalescent patients could not be classified. The next largest grouping was of unclassified *Caudoviricetes*, with 25% of VC's in acute patients and 24% of VC's in convalescent patients classified here. Additionally, acute patients had increased relative abundance of *Suoliviridae* and *Microviridae* compared to convalescent patients (**Fig 1b**).

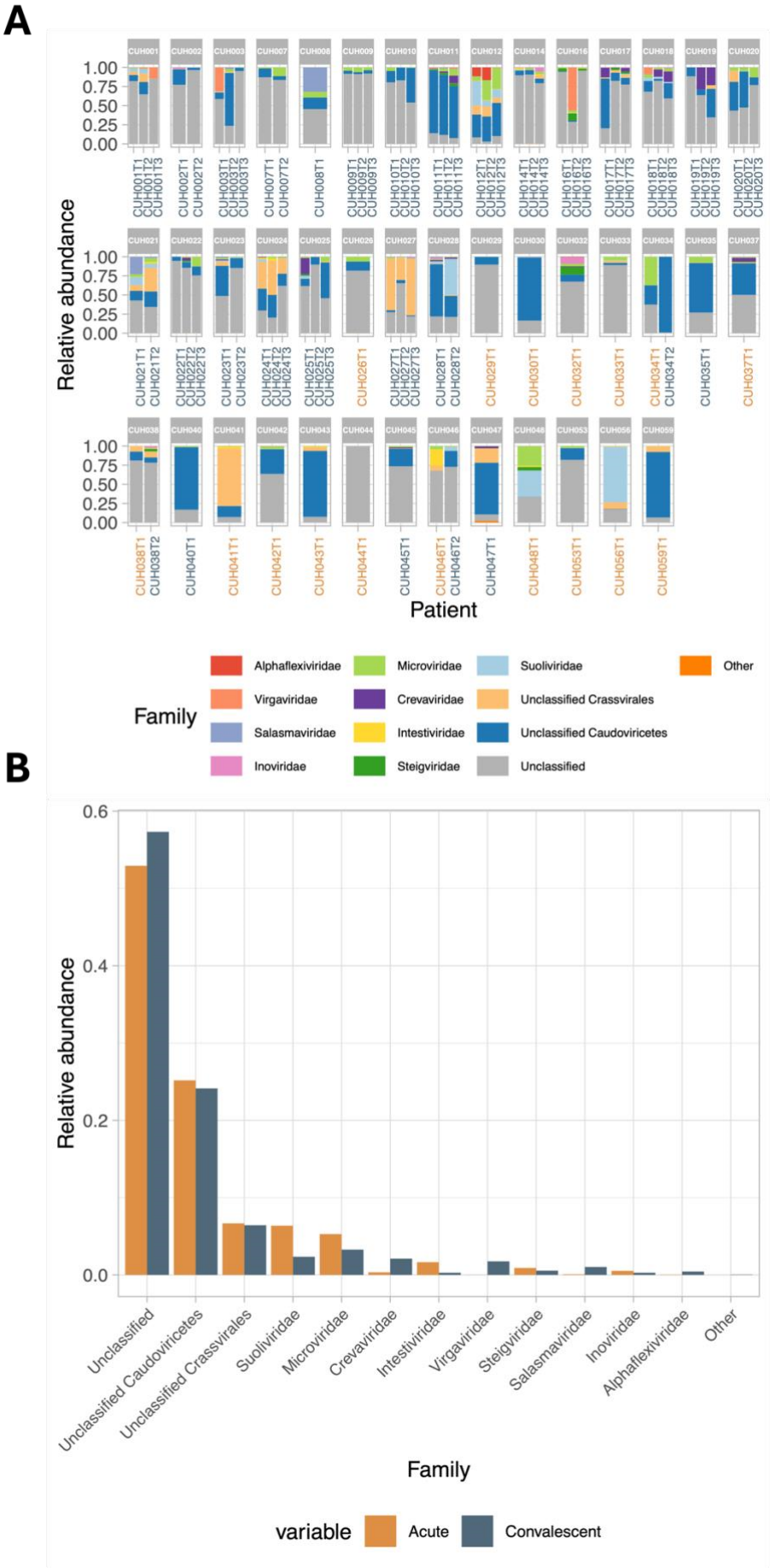

A
Relative abundance
1.00
0.75
0.50
0.25
0.00
Patient
Family
Alphaflexiviridae
Virgaviridae
Salasmaviridae
Inoviridae
Microviridae
Crevaviridae
Intestiviridae
Steigviridae
Suoliviridae
Unclassified Crassvirales
Unclassified Caudoviricetes
Unclassified
Other
B
Relative abundance
0.6
0.4
0.2
0.0
Unclassified
Unclassified Caudoviricetes
Unclassified Crassvirales
Suoliviridae
Microviridae
Crevaviridae
Intestiviridae
Virgaviridae
Steigviridae
Salasmaviridae
Inoviridae
Alphaflexiviridae
Other
Family
variable
Acute
Convalescent

**Figure 1. Family level comparison of the gut virome between COVID-19 patients. (a)** Barplots showing the relative viral composition, at the family level, of each patient and timepoint. Samples taken during acute stage of infection are labelled orange and convalescent samples labelled navy. **(b)** Barplots showing the difference in relative abundance of each viral family between acute and convalescent patients.

## No alterations in the gut bacteriome and virome between acute and convalescent patients

Comparison at the VC level showed no significant differences in diversity (Wilcoxon test, $p = 0.56$, Shannon diversity), evenness (Wilcoxon test, $p = 0.42$, Pielou's J for evenness) or richness (Wilcoxon test, $p = 0.12$, rarefied richness) between acute and convalescent patients (**Fig 2a**). For convalescent patients that provided samples at multiple timepoints, no significant changes to VC α-diversity metrics were observed over time (**Fig S2a**). Overall severity of SARS-COV-2 infection was also found to have no effect on α-diversity metrics (**Fig 2b**). Beta diversity analysis, measured by Bray-Curtis dissimilarity of VC abundance, showed no significant separation of acute and convalescent patients (Permutational multivariate analysis of variance (PERMANOVA) R^2 = 0.0234, $p = 0.525$, **Fig 2c**). A number of other factors were included in the adonis2 model (severity of disease, co-morbidities and medications), none of which were found to significantly contribute to variation between samples (**Fig S2b**).

No significant differences in bacterial diversity (Wilcoxon test, $p = 0.62$, Shannon diversity), evenness (Wilcoxon test, $p = 0.42$, Pielou's J for evenness) or richness (Wilcoxon test, $p = 0.17$, rarefied richness) were identified between acute and convalescent patients (**Fig 2d**). In convalescent patients, there was no significant changes to bacterial α-diversity over time (**Fig S2c**). Furthermore, there was a general trend of decreasing bacterial α-diversity with increasing severity of disease, with the difference between mild and severe cases approaching significance

(Wilcoxon test, *p* = 0.054, Shannon diversity, **Fig 2e**). Principle coordinate analysis (PCoA) based on Bray-Curtis distance showed no significant difference in the bacteriome of acute and convalescent patients (PERMANOVA, R^2 = 0.0284, *p* = 0.297) (**Fig 2f**). No other metadata factor was found to significantly contribute to variation (**Fig S2d**).

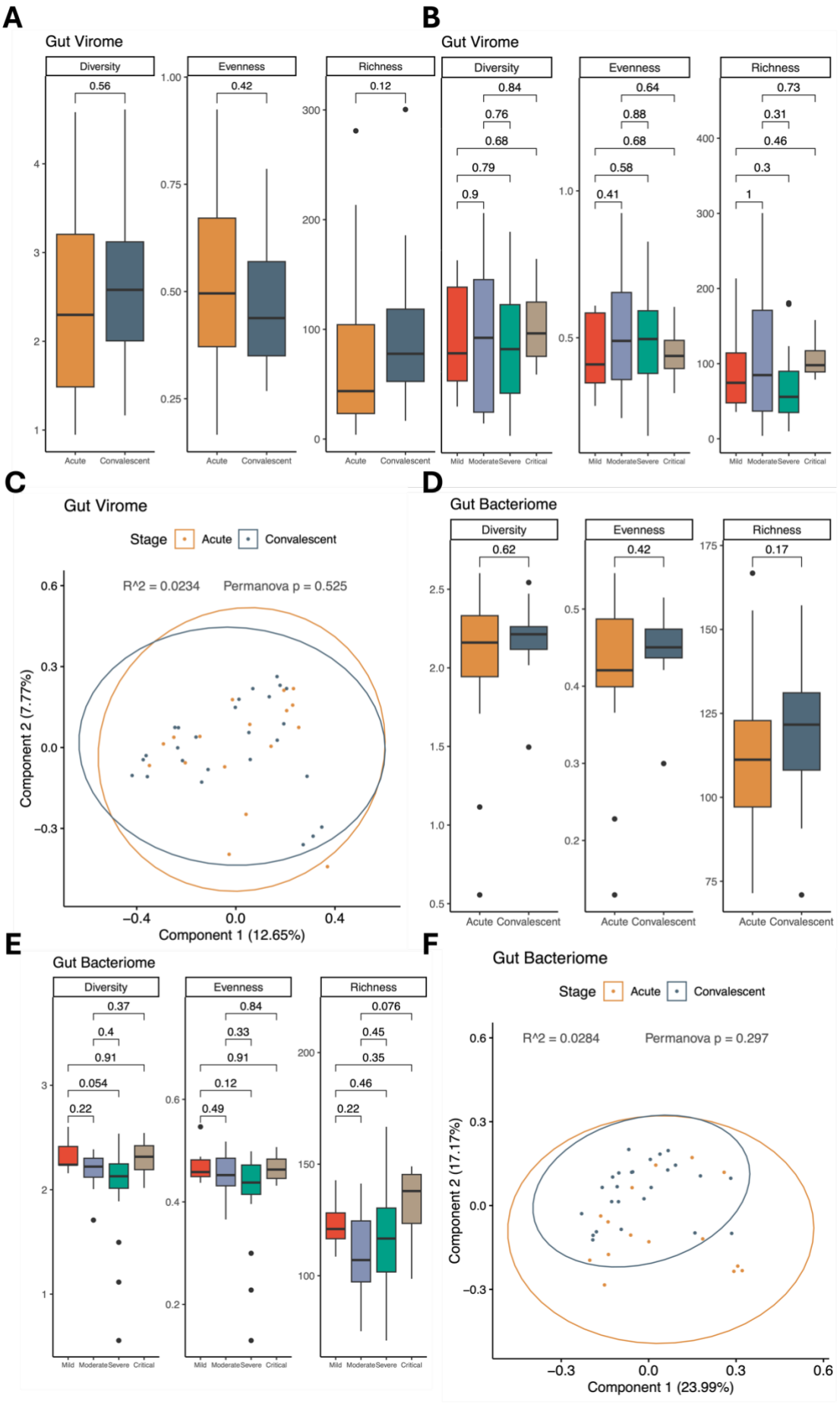

A
Gut Virome
Diversity
Evenness
Richness
0.56
0.42
0.12
Acute
Convalescent
B
Gut Virome
0.84
0.76
0.68
0.79
0.9
0.64
0.88
0.68
0.58
0.41
0.73
0.31
0.46
0.3
1
Mild
Moderate
Severe
Critical
C
Gut Virome
Stage
Acute
Convalescent
R^2 = 0.0234
Permanova p = 0.525
Component 2 (7.77%)
Component 1 (12.65%)
D
Gut Bacteriome
0.62
0.42
0.17
E
Gut Bacteriome
0.37
0.4
0.91
0.054
0.22
0.84
0.33
0.91
0.12
0.49
0.076
0.45
0.35
0.46
0.22
F
Gut Bacteriome
R^2 = 0.0284
Permanova p = 0.297
Component 2 (17.17%)
Component 1 (23.99%)

**Figure 2. Gut bacteriome and virome α and β-diversity analysis between acute and convalescent patients. A)** Shannon α-diversity barplots showing no significant reduction in diversity, evenness and richness of acute (n = 16) patient virome samples compared to convalescent (n = 29). Statistical significance was determined by Wilcoxon's rank sum test. The boxplots represent the 25$^{th}$, 50$^{th}$ (median) and 75$^{th}$ interquartile range (IQR) percentiles, the whiskers encompass values within 1.5 times the IQR. **B)** Shannon α-diversity barplots showing no significant changes in diversity, evenness and richness between mild (n = 8), moderate (n = 10), severe (n = 21) and critical (n = 4) patients. Statistical parameters were the same as in **A). C)** Principle coordinates analysis (PCoA) of Bray-Curtis dissimilarity, depicting no significant separation between the gut virome of acute and convalescent patients. PERMANOVA *p* = 0.793, R^2 = 0.0196. **D)** Shannon α-diversity barplots showing no significant differences in diversity, evenness and richness of acute (n = 15) bacteriome samples compared to convalescent (n = 25). Statistical parameters were the same as in **A). E)** Shannon α-diversity barplots showing an overall trend of reducing bacterial α-diversity between mild (n = 6), moderate (n = 10) and severe (n = 20). **F)** Principle coordinates analysis (PCoA) of Bray-Curtis dissimilarity, depicting no separation between the gut bacteriome of acute and convalescent patients. PERMANOVA *p* = 0.297, R^2 = 0.0284.

## Minimal shifts in bacterial species abundance between acute and convalescent patients

MaAsLin2 was utilised to identify any gut viral or bacterial signatures of acute or convalescent stages of disease. No differentially abundant VC's between acute and convalescent patients could be identified, while eight bacterial species were differentially abundant between the two groups (**Fig 3a**). Acute patients were enriched in four bacterial species compared to convalescent patients, including the opportunistic pathogens *Enterococcus faecium* and *Enterococcus faecalis*, as well as *Lactobacillus casei* and *Lactobacillus rhamnosus*. Convalescent patients saw a significant increase in the abundance of four bacterial species compared to acute patients, including *Clostridium ruminantium* and the nosocomial pathogen *Clostridioides difficile*, as well as butyrate producing *Roseburia faecis* and *Coprococcus eutactus*. Additionally, one VC with a predicted *Eubacterium* host was found to significantly associate with the critical stage of disease (**Fig S3**).

The relative abundance of these eight bacterial species were traced in 18 convalescent patients who provided follow up samples, as well as three acute patients who provided a follow up (convalescent) sample. *Enterococcus faecium*, *Enterococcus faecalis*, *Lactobacillus rhamnosus* and *Lactobacillus casei* remained at a low abundance across convalescent timepoints. In one acute patient, there was a noticeable decrease in the relative abundance of *Enterococcus faecalis* between timepoint one (12 days post infection) and timepoint two (34 days post infection, **Fig 3b**). For the four bacterial species enriched in convalescent patients, different patterns were observed. The relative abundance of *Coprococcus eutactus* and *Clostridium ruminantium* were most variable between timepoints, while *Clostridioides difficile* and *Roseburia faecis* tended to persist with the same relative abundance across timepoints (**Fig 3c**).

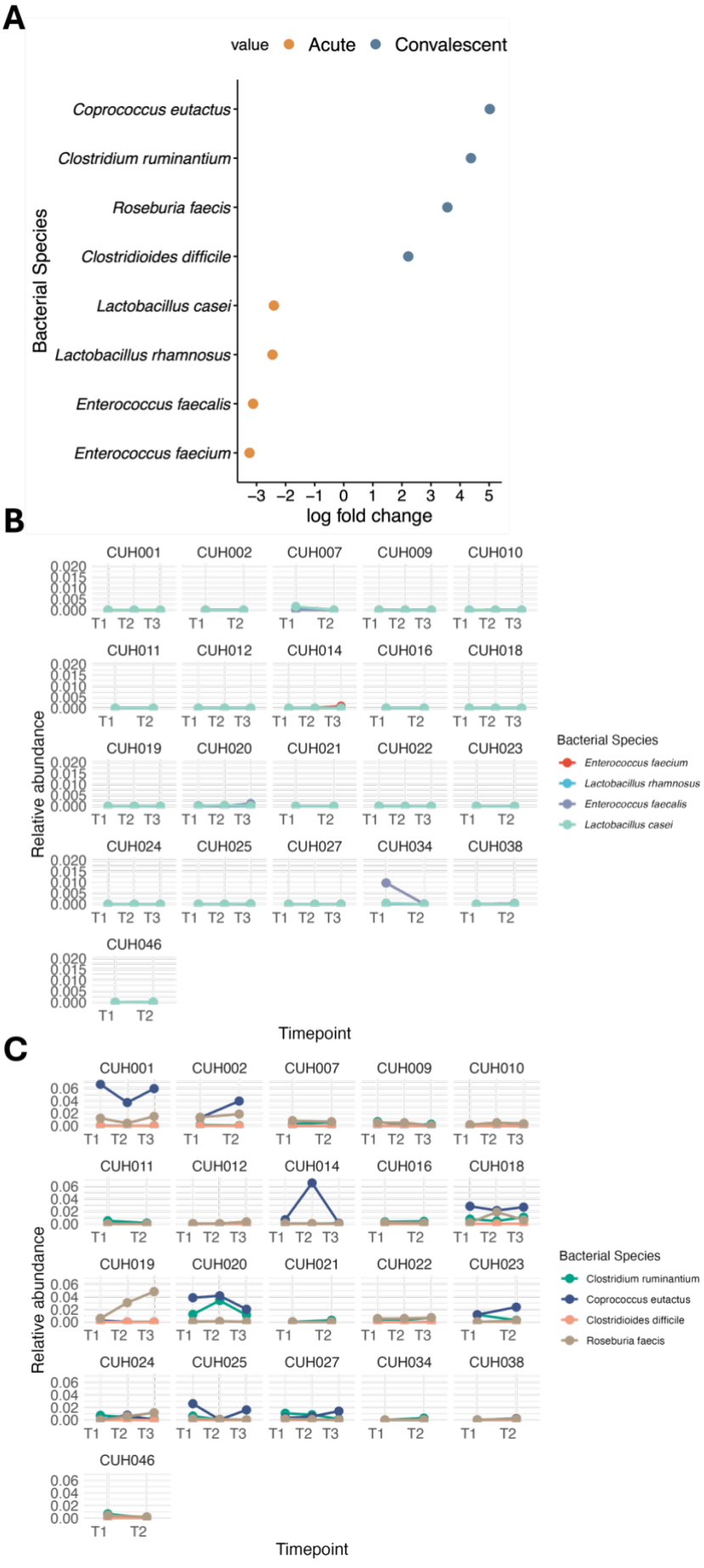

A
value
Acute
Convalescent
Coprococcus eutactus
Clostridium ruminantium
Roseburia faecis
Clostridioides difficile
Lactobacillus casei
Lactobacillus rhamnosus
Enterococcus faecalis
Enterococcus faecium
Bacterial Species
−3 −2 −1 0 1 2 3 4 5
log fold change
B
CUH001 CUH002 CUH007 CUH009 CUH010
CUH011 CUH012 CUH014 CUH016 CUH018
CUH019 CUH020 CUH021 CUH022 CUH023
CUH024 CUH025 CUH027 CUH034 CUH038
CUH046
0.020
0.015
0.010
0.005
0.000
T1 T2 T3
Relative abundance
Timepoint
Bacterial Species
Enterococcus faecium
Lactobacillus rhamnosus
Enterococcus faecalis
Lactobacillus casei
C
CUH001 CUH002 CUH007 CUH009 CUH010
CUH011 CUH012 CUH014 CUH016 CUH018
CUH019 CUH020 CUH021 CUH022 CUH023
CUH024 CUH025 CUH027 CUH034 CUH038
CUH046
0.06
0.04
0.02
0.00
T1 T2 T3
Relative abundance
Timepoint
Bacterial Species
Clostridium ruminantium
Coprococcus eutactus
Clostridioides difficile
Roseburia faecis

**Figure 3. Significant changes to the relative abundance of eight bacterial species between acute and convalescent patients. A)** Convalescent patients are enriched in *Coprococcus eutactus*, *Clostridium ruminantium*, *Roseburia faecis* and *Clostridioides difficile* compared to acute patients, while acute patients are enriched in *Lactobacillus casei*, *Lactobacillus rhamnosus*, *Enterococcus faecalis* and *Enterococcus faecium*. Differential bacterial species were identified using MaAslin2, $q < 0.05$ and log fold change > 2. **B)** Depleted bacterial species in convalescent patients remain at a low relative abundance across timepoints. **C)** Enriched bacterial species in convalescent patients show different patterns of relative abundance across timepoints.

## Oral virome and bacteriome composition of COVID-19 patients

A subset of patients (2 acute and 13 convalescent) enrolled in this study also provided saliva samples for oral bacterial and viral analysis. Whole metagenomic sequencing resulted in a final catalogue of 18,668 viral sequences, representing the DNA virome. These were clustered by vContact2 into 3,022 VC's, comprised of 3,837 contigs. Taxonomic classification of VC's identified 13 different families and the six most abundant families are shown in **Figure 4a**. The majority of viruses across both acute and convalescent patients were unclassified, with acute having 83% and convalescent patients having 80% unclassified (**Fig 4b**). This was followed by unclassified *Caudoviricetes*, with 16% in acute patients and 19% in convalescent. The remaining 1% was composed of viral families appearing in certain individuals, such as *Orlajensenviridae* in CUH041S and *Orthoherpesviridae* and *Suoliviridae* in CUH001S, highlighting the inter-individual variation of the human virome.

Compositional profiling of the bacterial community by metaphlan identified 48 different bacterial families. The 23 most abundant families are represented in **Figure 4c**, with the remaining families grouped as other. In general, the most abundant bacterial families identified were *Streptococcaceae, Prevotellaceae, Veillonellaceae* and *Actinomycetaceae*. There was a general trend of increasing relative abundance of *Streptococcaceae, Neisseriaceae* and *Lachnospiraceae,* moving from acute to convalescent patients. There was an increase in relative

abundance in *Carnobacteriaceae* and *Selenomonadaceae* in acute patients due to blooms of *Carnobacteriaceae* in CUH041S and *Selenomonadaceae* in CUH053S (**Fig 4d**).

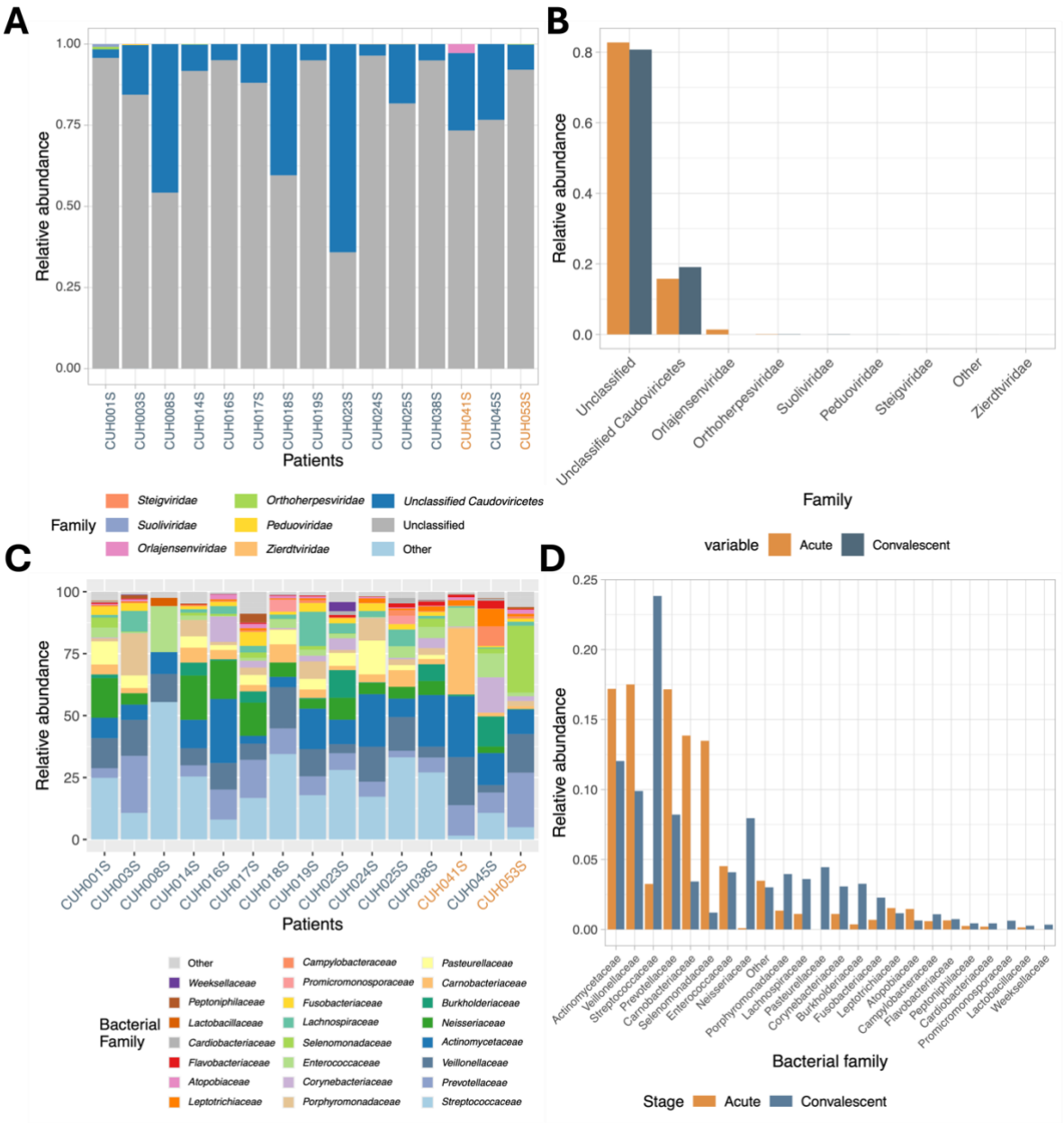



**Figure 4. Family level comparison of the oral virome and bacteriome between acute and convalescent patients. (A)** Barplots showing the viral composition, at the family level, of each acute and convalescent patient. Samples taken during acute stage of infection are labelled orange and convalescent samples labelled navy. **(B)** Barplots showing the difference in relative abundance of

each viral family, between acute (n = 2) and convalescent (n = 13) patients. **(C)** Barplots displaying the 23 most abundant bacterial families in each acute and convalescent patient. Samples taken during acute stage of infection are labelled orange and convalescent samples labelled navy. **(D)** Barplots showing the relative abundance of each bacterial family in acute (n = 2) and convalescent (n = 13) patients.

## Minimal alterations to the oral virome and bacteriome between COVID-19 patients

Comparisons at the VC level displayed no significant changes to α-diversity (Wilcoxon test, $p = 0.38$, Shannon diversity), evenness (Wilcoxon test, $p = 0.8$, Pielou's J for evenness), richness (Wilcoxon test $p = 0.38$, rarefied richness) between acute and convalescent patients (**Fig 5a**). Additionally, α-diversity was not significantly impacted by disease severity between patients (**Fig 5b)**. Disease severity and stage of disease, as well as co-morbidity factors were not found to significantly contribute to β-diversity variation between patients (**Fig 5c**). There were also no significant changes in bacterial α-diversity (Wilcoxon test, $p = 0.23$, Shannon diversity) and evenness (Wilcoxon test, $p = 0.11$, Pielou's J for evenness) between acute and convalescent patients (**Fig 5d**). Similar to the oral virome, disease severity did not have a significant impact on α-diversity metrics (**Fig 5e**). Stage of disease was found to have a significant effect on oral bacteriome variation, contributing over 10% of variation between acute and convalescent patients (PERMANOVA, $p = 0.022$, R^2 = 0.1066). Additionally, three comorbidities, asthma (PERMANOVA, $p = 0.03$, R^2 = 0.1117), hypertension (PERMANOVA, $p = 0.019$, R^2 = 0.1113) and obesity (PERMANOVA, $p = 0.018$, R^2 = 0.1109) were all found to have a significant effect on oral bacteriome variation (**Fig 5f**). No differentially abundant VC's or bacterial species could be identified between acute and convalescent stages of disease, or between different levels of disease severity.

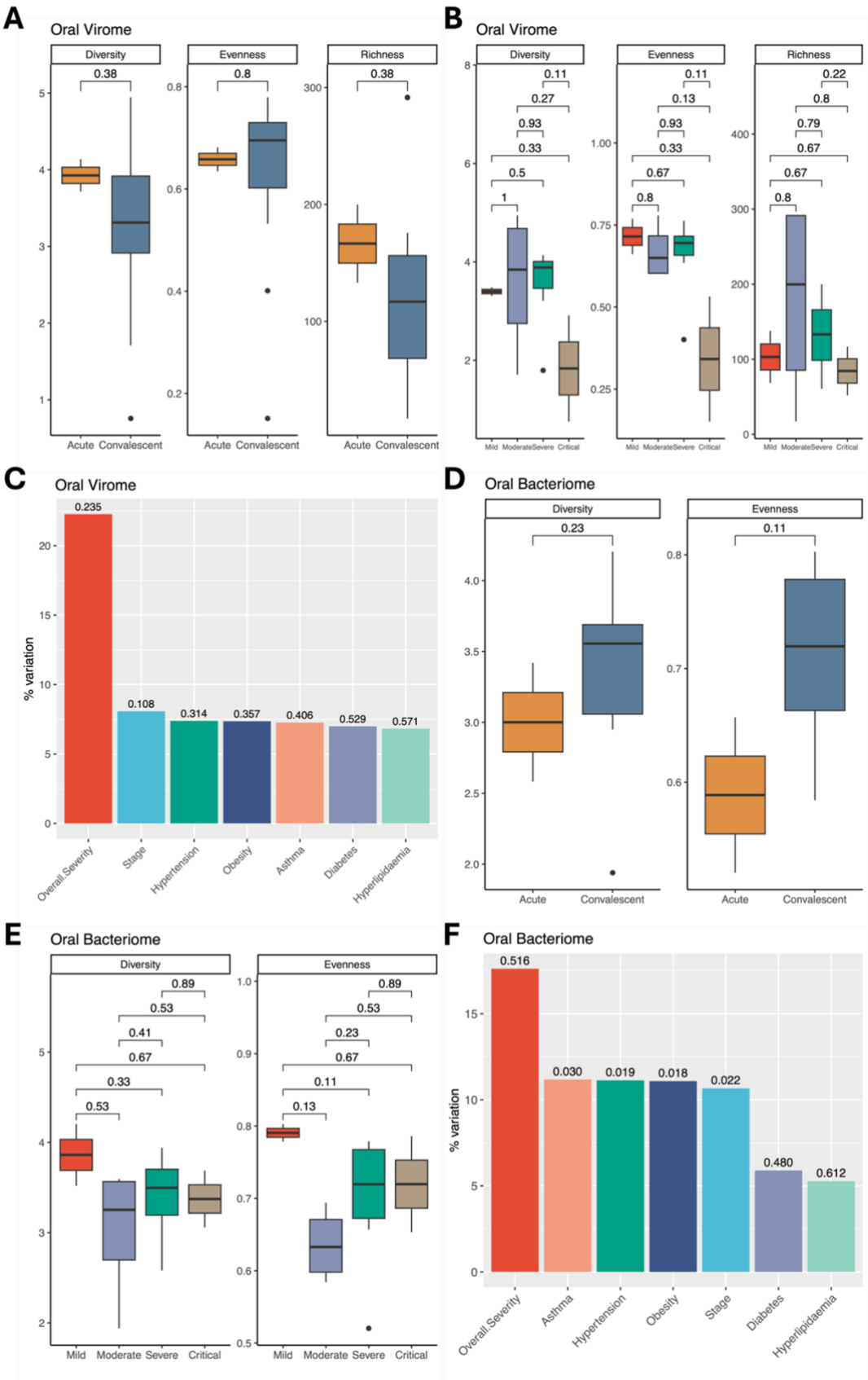



**Figure 5. Minimal changes to the viral and bacterial components of the oral microbiome between SARS-COV-2 patients. A)** Shannon α-diversity barplots showing no significant changes in diversity,

evenness and richness of the oral virome, between acute and convalescent patients. Statistical parameters were the same as in Fig 2a. **B)** Shannon α-diversity barplots showing no significant changes in diversity, evenness and richness in patients who had mild (n = 2), moderate (n = 4), severe (n = 7) and critical (n = 2) COVID infection. Statistical parameters were the same as in Fig 2a. **C)** Stage and severity of disease, as well as comorbidities, did not significantly contribute to variation in patients, as determined by PERMANOVA. **D)** Shannon α-diversity barplots showing no significant changes in diversity and evenness of the oral bacteriome, between acute and convalescent patients. Statistical parameters were the same as in Fig 2a**. E)** Shannon α-diversity barplots showing no significant changes in diversity, evenness and richness of the oral bacteriome in patients who had mild (n = 2), moderate (n = 4), severe (n = 7) and critical (n = 2) COVID infection. Statistical parameters were the same as in Fig 2a. **E)** Oral bacteriome variation explained by metadata factors, as determined by PERMANOVA. Hypertension, obesity and asthma all significantly contributed to variation between SARS-COV-2 patients.

## Discussion

In this study, both the bacteriome and virome from the oral cavity and the gut were examined, during acute and recovery stages of SARS-COV-2 infection. The virome and bacteriome from the gut of acute patients were similar to those in recovery from COVID infection, noted by a lack of change in alpha and beta diversity metrics. Significant shifts in the relative abundance of eight bacteria were the only observed differences between the two groups. An altered gut bacteriome in acute patients has been consistently reported across multiple studies, often characterised by a loss of abundance of genera such as *Roseburia*[9,40], *Coprococcus*[25,41], *Eubacterium*[42,43] and *Faecalibacterium*[44,45] and an enrichment of *Lactobacillus*[10,25,41,46], *Rothia* and *Enterococcus*[25,47]. Consistent with this was the observation of an enrichment of *Enterococcus faecalis, Enterococcus faecium*, *Lactobacillus casei* and *Lactobacillus rhamnosus*, as well as a depletion of *Coprococcus eutactus* and *Roseburia faecis* in acute compared to convalescent patients. Additionally, a number of studies have reported persistent perturbations to the gut bacteriome many months post SARS-COV-2 infection[32,50,51] which has been linked to the development of Long COVID[33,34]. It is possible that the lack of changes observed here between acute and

convalescent patients are a reflection of the continued impact of SARS-COV-2 infection. However, comparisons with matched healthy controls would be required to verify this and represents a limitation to this study.

The impact of disease severity on the gut bacteriome and virome was also assessed amongst patients. There was a trend of decreasing bacterial α-diversity across mild, moderate and severe patients, with the difference between mild and severe patients nearing statistical significance. While there were no observable trends in viral α-diversity, one phage with a predicted *Eubacterium* host did significantly correlate with the critical stage of disease.

Cao *et al* previously reported an increase in a cucumber RNA virus of the *Virgaviridae* family in the gut virome of COVID-19 patients[13], while Zuo *et al* reported a significant reduction in a pepper RNA virus, also belonging to *Virgaviridae*[14]. We did not identify any significant changes in *Virgaviridae* VC's in this study. Two convalescent patients (CUH003 and CUH016) had a spike in *Virgaviridae* abundance at timepoint 1 and timepoint 2 respectively, which subsequently disappeared at the following timepoint. This transient detection may reflect the involvement of the diet in gut virome variation[48,49].

Similar to the gut microbiome, the oral virome and bacteriome of acute patients was similar to those post-recovery. Beta diversity analysis did identify a significant separation in the oral bacteriome between acute and convalescent patients. However, no differentially abundant bacteria could be identified between the two groups. It should be noted that the low number of saliva samples and uneven distribution between acute (n = 2) and convalescent (n = 13) impacts the ability to identify significant changes between the two groups.

Approximately 82% of the VC's identified in the oral virome could not be classified, compared to 53% of the gut virome. Uncharted viral composition and diversity is a well-known issue in gut virome studies[56]. This problem appears to be further amplified in the oral virome, where viruses are likely further divergent from the majority of gut virome databases used for identification. This was highlighted by the recent establishment of the Oral Virus Database[57] and the Cenote Human Virus

Database[58], where the majority of oral viruses had not been identified in previous databases. Concerted efforts to create both further oral virome specific databases and taxonomic families for the current large number of unclassified *Caudoviricetes* are required in order to identify further meaningful changes between various oral health and disease states.

The mechanisms by which SARS-COV-2 infection culminates in shifts in the gut microbiome, and whether this plays a role in COVID progression, is not fully understood. The invasion of SARS-COV-2 results in the activation of various proinflammatory cytokines[59]. It has been hypothesized that this immune response can subsequently disrupt the gut microbiome by decreasing the abundance of commensals, allowing for the enrichment of opportunistic pathogens. This altered gut microbiota can in turn disrupt the natural gut barrier, leading to enhanced gut permeability[11,60] and allowing opportunistic pathogens and toxins to translocate into the bloodstream. In addition, the presence of opportunistic pathogens may be further recognized by the innate immune system, further exacerbating gut inflammation[8].

Over five years on from the start of the COVID pandemic, it appears SARS-COV-2 may be here for the long term. With that has come the increasing recognition of the detrimental impacts Long COVID can have on a significant portion of the population and the requirement to unravel potential causes in the search for solutions. Continued loss of microbial diversity, including depletion of anti-inflammatory bacteria such as *Faecalibacterium prausnitzii*[8,45] *Bifidobacterium adolescentis*[12,61] and *Eubacterium rectale*[8,62] may sustain gut inflammation. Further studies are required to identify molecular mechanisms that may underpin this.

## Methods

### Study cohort

The study cohort consisted of 43 individuals hospitalized with SARS-COV-2 infection at Cork University Hospital, Ireland, between March 2020 and April 2021. COVID-19 severity was classified as (i) mild, if there were no radiographic indications of pneumonia; (ii) moderate, if pneumonia with fever and respiratory tract symptoms were detected; (iii) severe, if respiratory rate was ≥ 30 breaths per minute, oxygen saturation ≤ 93% when breathing ambient air or PaO2/FiO2 ≤ 300mm Hg; (iv) critical, if there was respiratory failure requiring mechanical ventilation or organ failure requiring intensive care. Informed consent was obtained from all patients. This research was approved by the Cork Research Ethics Committee and all research was performed in accordance with relevant guidelines/regulations. Faecal samples were taken from 16 patients ≤ 19 days post infection (acute stage of disease). The remaining patients had samples taken at 2-5 months post infection (convalescent stage). In addition, 24 patients provided a second faecal sample six months post infection and a further 16 patients provided a third faecal sample seven months post infection, resulting in a total of 84 faecal samples. Clinical characteristics at the time of sampling, including comorbidities and treatments used, are summarized in **Fig S1a**.

Saliva was collected from patients using Salivette tubes (Sarstedt), following manufacturer's instructions and frozen at –80°C. Two patients provided saliva samples during the acute stage of disease and thirteen provided saliva samples between 2-7 months post infection.

Faecal VLP nucleic acid extraction

Faecal VLP fractions were prepared from 0.25g of faeces in 650 µL of SM buffer and nucleic acids were extracted as described previously[63]. The concentration of DNA purified was measured using the Qubit dsDNA HS kit and the Qubit 3 fluorometer (Invitrogen/Thermo Fisher Scientific). Reverse transcription was carried out using the SuperScript IV First-Strand Synthesis System (Invitrogen/Thermo Fisher Scientific). Starting template DNA was 22 µL and the protocol was carried out as per manufacturer's manual, scaling up accordingly.

Library preparation and sequencing

Library preparation was achieved using Accel-NGS 1S Plus kit (IDT). For sonication, nuclease free water (QIAGEN) was used to adjust 40 µL reverse transcription product to 55 µL. Shearing was performed on the M220 focused ultrasonicator (Covaris) with the following settings; peak power of 50W, duty factor of 20%, 200 cycles per burst, total duration of 35 seconds. The resulting DNA/cDNA was purified with the Genomic DNA Clean and Concentrator kit (Zymo Research), according to the manufacturer's protocol and eluted in a 15 µL volume. The remaining steps were carried out according to the manufacturer's manual for an input of 350 bp. An additional SPRI clean up step with a ratio of 1:1 was added post PCR. The quality of the library was checked using the Agilent 2100 Bioanalyzer, following the Agilent High Sensitivity DNA protocol. Libraries were pooled and sequenced using Illumina Novaseq (2 X 150 bp) at Genewiz (Leipzig, Germany).

Processing of faecal VLP shotgun sequencing data

The quality of raw reads was accessed using FastQC v0.11.9. Illumina adapter sequences were removed using cutadapt v2.8. To remove low quality reads, trimmomatic v0.39 was used with the following parameters; SLIDINGWINDOW:

4:20, MINLEN:60, HEADCROP:15. Contaminating human reads were removed using classification with Kraken v 1.1.1. Levels of bacterial contamination in the resulting reads were assessed by alignment to a database of bacterial *cpn*60 genes. Reads were assembled into contigs using SPAdes v3.13.1 in metagenomic mode. All contigs > 1kb were pooled and redundancy was removed by identifying contig pairs with 90% identity over 90% of the length of the shorter contig in each pair, discarding the shorter of the two in each instance. This resulted in a total of 262,774 nonredundant contigs.

To extract viral sequences, several criteria were used. Contigs were aligned using BLASTn to viral sequences in the NCBI refseq database, Gut Virome Database[64], MGV database[65], Gut Phage Database[66], IMG/VR database v4, as well as in-house *Crassvirales* and SARS-COV-2 databases. Contigs that aligned with an e-value ≤ 1x $10^{10}$ and at least 50% identity over 90% of contig length were deemed viral. Contigs identified by Virsorter2[36] and with a score of ≥ 0.9 or ≥ 0.7 with at least one viral hallmark protein coding gene present were also added. The level of completeness of contigs was assessed using CheckV[67] and any Virsorter2 identified contigs marked as prophages by CheckV were subsequently removed. This resulted in a final catalogue of 59,240 viral contigs.

Circular contigs were identified using VRCA. G+C content was calculated using EMBOSS geecee. Viral protein coding genes were predicted using prodigal[68]. Translated protein sequences were searched against viral refseq database (e-value cut-off of ≤ $1 \times 10^{-10}$), PHROGS database[69] using hmmsearch (HMMER v3.3, e-value cut-off of ≤ $1 \times 10^{-5}$) and a *Crassvirales* in-house protein database using BlastP v2.10.0 with an e-value cut-off of ≤ $1 \times 10^{-10}$.

Viral contigs > 3kb in length were clustered to approximately genus level viral clusters (VC's) using vContact2[37] with the following optional parameters; --relmode Diamond --pcs-mode MCL --vcs-mode ClusterONE. Taxonomic classification of VC's was performed using *graphanalyzer*[70] and the inphared database[71]. Human and fungal viruses assigned to a VC were classified based on RefSeq identification.

Phage lifestyles, either temperate or virulent, were predicted using BACPHLIP[72], phages with a score ≥ 0.8 were deemed temperate.

In order to predict the bacterial hosts of phages, host predictions for high confidence genomes of the IMG/VR v4 database were used, where the contig aligned with an e-value ≤ 1x $10^{10}$ and at least 50% identity over 90% of contig length. If no IMG/VR prediction was available, blastn was used to search an inhouse CRISPR dataset (compiled of bacterial RefSeq and HMP reference genomes) as described previously[35]. Finally, blastn similarity of viral contigs to prophages in bacterial genomes (≥90% identity over ≥85% of viral contig length to Refseq database of bacterial genomes and HMP reference genomes) were used when the first two assignment methods failed. In the case of VC's which had multiple hosts assigned, the host that was assigned to the majority was kept.

Quality filtered reads were aligned to the final catalogue of viral contigs on a per sample basis, using bowtie2 v2.3.5.1[73] in *–end-to-end* mode. A counts table was generated using SAMTools v1.1[74]. SAMTools 'mpileup' command was used to calculate sequence coverage, per nucleotide position per contig per sample. Read counts which featured a breadth of coverage of less than 1 X 75% of a contig length were set to zero.

16S rRNA amplicon sequencing and bioinformatic analysis

Total community DNA was extracted from stool samples by a combined repeat beat beading – QIAGEN DNA extraction method and the V3-V4 region of the 16S rRNA gene was amplified and sequenced. The uniquely barcoded amplicons were sequenced on an Illumina MiSeq platform (Illumina, California, USA) utilising 2 x 300 bp chemistry.

The raw Illumina reads obtained were quality-filtered using trimmomatic[75]. The quality filtered reads were then taxonomically classified to the species level using SPINGO[76], under default settings.

## Saliva whole metagenomics

Due to low saliva sample volumes acquired from patients, a nucleic acid extraction protocol was developed in order to capture both the bacteriome and virome. Frozen saliva samples were thawed on ice and 1mL of sample was transferred into 1mL of SM buffer. Twenty microlitres of ZymoBIOMICS spike-in (Zymo Research) was added (2 X $10^7$ cells of *Imtechella halotolerans* and 2 X $10^7$ cells of *Allobacillus halotolerans).* Cells were harvested by centrifuging for 10 minutes at 5000 x g. The resulting pellet was resuspended in 180 µL of enzymatic lysis buffer (20 mM Tris·Cl, pH 8.0; 2 mM sodium EDTA; 1.2% Triton® X-100. Lysozyme added to 20 mg/mL immediately before use).

Nucleic acids were precipitated using QIAGEN Blood and Tissue Kit as per manufacturer's instructions, with the following modifications; after addition of proteinase K and Buffer AL, samples were incubated at 56°C for 30 minutes. Samples were eluted in a final volume of 30 µL. Library preparation and sequencing was carried out as previously described.

## Processing of saliva sequencing data

Quality check, trimming and filtering of reads was carried out as described for faecal data. Metaphlan[77] v4.0.6 was used for taxonomic compositional analysis of the saliva bacteriome. For virome analysis, assembly of contigs was carried out as described previously. Contigs were aligned to the same databases, with the addition of the Oral Virus Database[57] (OVD) and the Cenote Human Virus Database[58] (CHVD), to improve identification of saliva specific viruses. The remaining processing was carried out as described for the faecal viral data.

## Statistical analysis

All statistical analysis was carried out in R environment v4.2.2. The α-diversity metrics presented are the Shannon index for diversity, Pielou's J for evenness and rarified richness for species richness, using package vegan v2.6.4. Analysis of β-diversity through PERMANOVA was performed with Bray-Curtis dissimilarity, using phyloseq v1.42.0 and through vegan adonis2 function. In addition to stage of disease, all other available metadata factors were included in the adonis2 model and the marginal sum of squares was applied. Differentially abundant bacterial species were identified using Maaslin2 v1.12, with Benjamini-Hochberg corrections. All visualisations were created using ggplot2 v3.4.4 and ggscatter from ggpubr v0.6.0.

Data availability

The gut virome and oral microbiome sequencing data is available through NCBI BioProject code PRJNA1155319, 16S and gut bacteriome data at PRJEB55909.

## Author Contributions

C.A.T, L.A.D, A.N.S and C.H contributed to the conceptual development of this study. C.A.T and M.N. performed extractions and library preparation for sequencing. C.A.T, A.N.S and J.L performed bioinformatic analysis. C.A.T performed data analysis. C.A.T, A.N.S and C.H interpreted the results. L.A.D supervised the project. C.H conceived the study and secured funding. C.A.T wrote the manuscript with editorial contributions from J.L, A.N.S and C.H.

Supplementary

**A**

| **Clinical characteristics of the 43 COVID-19 patients** | | |
|---|---|---|
| Female | 23 | 53.5% |
| Male | 20 | 46.5% |
| Median age | 57 | 49-64 (IQR) |
| Presence of any comorbidities | 28 | 65% |
| **Symptoms at time of admission** | | |
| Dyspnea | 34 | 79% |
| Cough | 31 | 72% |
| Fever | 27 | 63% |
| Diarrhea | 12 | 28% |
| Anosmia | 9 | 21% |
| Sore throat | 7 | 16% |
| Rhinorrhea | 1 | 2% |
| **Types of comorbidities** | | |
| Hypertension | 16 | 37% |
| Obesity | 12 | 28% |
| Hyperlipidemia | 9 | 21% |
| Asthma | 7 | 16% |
| Diabetes | 6 | 14% |
| Cardiac Disease | 6 | 14% |
| **Severity of COVID-19** | | |
| Mild | 8 | 19% |
| Moderate | 10 | 23% |
| Severe | 20 | 49% |
| Critical | 4 | 9% |
| **Treatments for COVID-19** | | |
| Antibiotics | 37 | 86% |
| Anticoagulants | 36 | 84% |
| Immunomodulators | 26 | 60% |
| Antivirals | 2 | 5% |
| **Symptoms at 4-5 months post discharge (n = 21)** | | |
| Fatigue | 14 | 66% |
| Dyspnea | 8 | 38% |
| Night sweats | 3 | 14% |
| Myalgia | 3 | 14% |
| Chest pains | 2 | 10% |
| Nausea | 2 | 10% |
| Brain fog | 2 | 10% |
| Anosmia | 1 | 5% |
| Dizziness | 1 | 5% |
| Hypoxia | 1 | 5% |
| Heart palpitations | 1 | 5% |

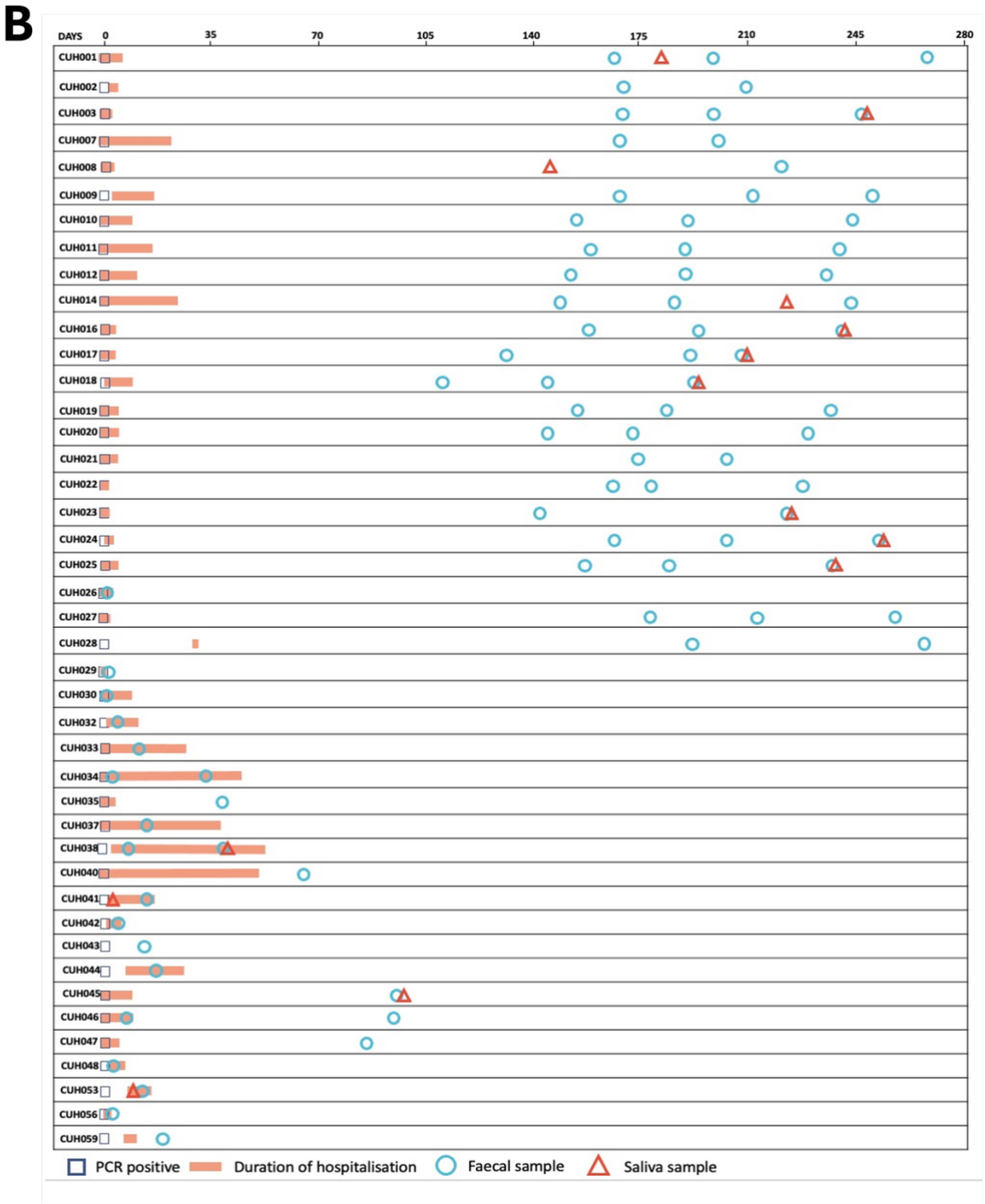



**Figure S1. Clinical characteristics and timeline of SARS-COV-2 samples. A)** Clinical characteristics of the 43 COVID-19 patients, including symptoms at the time of admission, types of comorbidities, severity of SARS-COV-2 infection, treatments used and symptoms at 4-5 months post-discharge. **B)** Timeline for each of the 43 COVID patients, showing when a positive PCR test was received, duration of hospitalisation, and timepoints when faecal and saliva samples were taken.

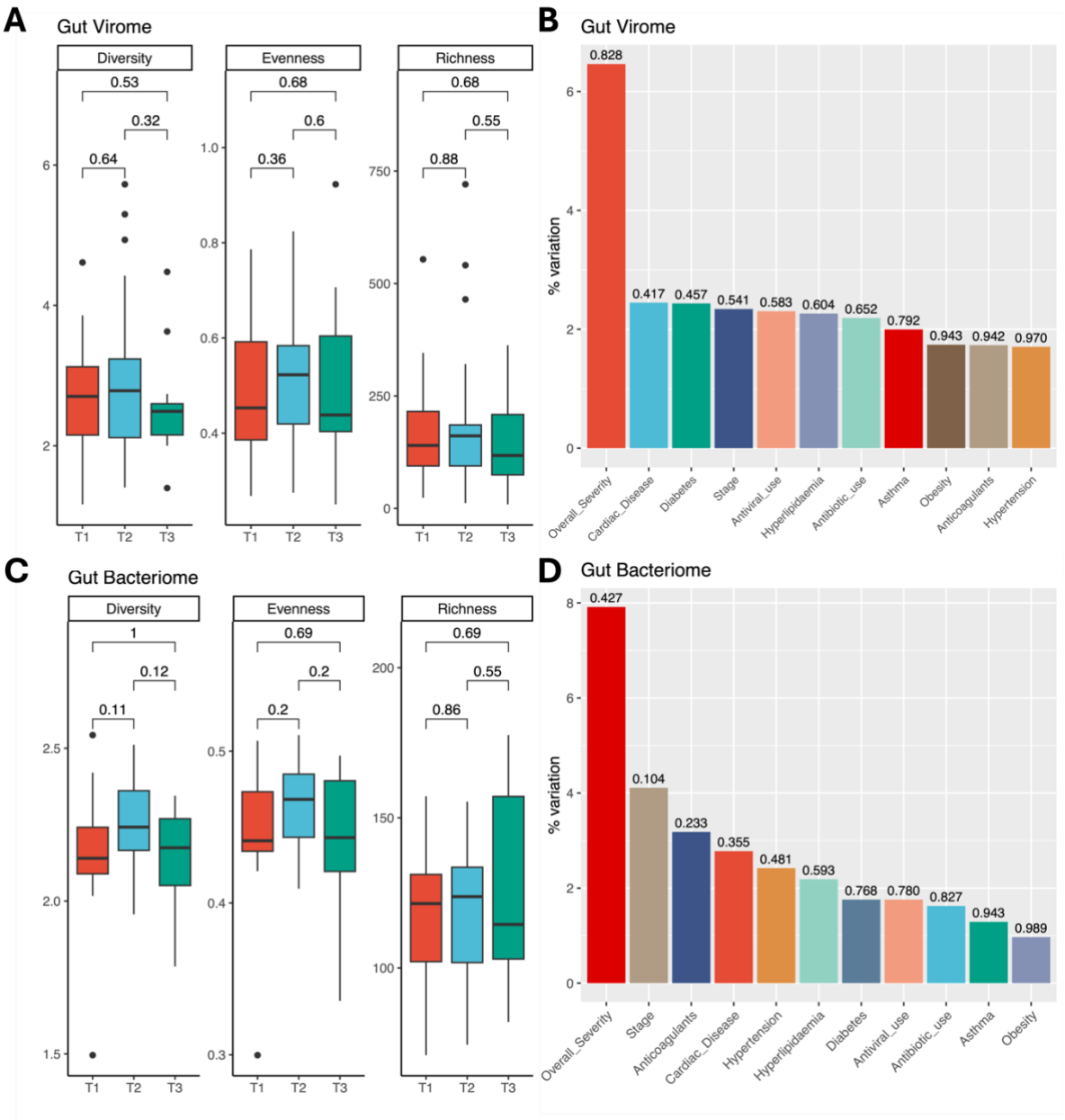



**Figure S2. No significant changes to gut virome and bacteriome of convalescent patients over time. A)** Shannon alpha diversity barplots showing no significant changes in diversity, evenness and richness of the gut virome in convalescent patients timepoint 1 (n = 21), timepoint 2 (n = 21) and timepoint 3 (n = 16). The mean number of days post infection was 159, 198 and 239 respectively. Statistical significance was determined by Wilcoxon's rank sum test. The boxplots represent the 25$^{th}$, 50$^{th}$ (median) and 75$^{th}$ interquartile range (IQR) percentiles, the whiskers encompass values within 1.5 times the IQR **B)** Metadata factors do not significantly contribute to gut virome variation, as determined by PERMANOVA. **C)** Shannon alpha diversity barplots showing no significant changes in diversity, evenness and richness of the gut bacteriome in convalescent patients timepoint 1 (n =

18), timepoint 2 (n = 18) and timepoint 3 (n = 12). The mean number of days post infection was 99, 181 and 239 respectively. Statistical parameters were the same as in **A**. **D)** Metadata factors do not significantly contribute to gut bacteriome variation, as determined by PERMANOVA.

| **VC** | NODE_196_length_35346 |
|---|---|
| **Predicted bacterial host** | *Eubacterium* |
| **Metadata association** | Overall Severity |
| **value** | Critical |
| **coef** | 4.244530491 |
| **stderr** | 0.582421429 |
| **N** | 43 |
| **N.not.0** | 9 |
| **pval** | 8.66E-09 |
| **qval** | 1.50E-05 |

**Fig S3. One bacterial VC significantly associates with the critical stage of infection**. Differential bacterial species were identified using MaAslin2, q < 0.05 and log fold change > 2.

## References


1. World Health Organization. *COVID-19 Weekly Epidemiological Update, 16 February 2024.* (2024).
2. Villapol, S. Gastrointestinal symptoms associated with COVID-19: impact on the gut microbiome. *Translational Research* vol. 226 57–69 Preprint at https://doi.org/10.1016/j.trsl.2020.08.004 (2020).
3. Zeng, W. *et al.* Gastrointestinal symptoms are associated with severity of coronavirus disease 2019: A systematic review and meta-analysis. *Eur J Gastroenterol Hepatol* **34**, (2022).
4. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, (2020).
5. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, (2020).
6. Xu, J. *et al.* Digestive symptoms of COVID-19 and expression of ACE2 in digestive tract organs. *Cell Death Discov* **6**, (2020).

7. Wu, Y. *et al.* Prolonged presence of SARS-CoV-2 viral RNA in faecal samples. *The Lancet Gastroenterology and Hepatology* vol. 5 Preprint at https://doi.org/10.1016/S2468-1253(20)30083-2 (2020).

8. Zhang, F. *et al.* Gut microbiota in COVID-19: key microbial changes, potential mechanisms and clinical applications. *Nat Rev Gastroenterol Hepatol* **20**, (2023).

9. Zuo, T. *et al.* Alterations in Gut Microbiota of Patients With COVID-19 During Time of Hospitalization. *Gastroenterology* **159**, (2020).

10. Zhou, T. *et al.* SARS-CoV-2 triggered oxidative stress and abnormal energy metabolism in gut microbiota. *MedComm (Beijing)* **3**, (2022).

11. Sun, Z. *et al.* Gut microbiome alterations and gut barrier dysfunction are associated with host immune homeostasis in COVID-19 patients. *BMC Med* **20**, (2022).

12. Yeoh, Y. K. *et al.* Gut microbiota composition reflects disease severity and dysfunctional immune responses in patients with COVID-19. *Gut* **70**, (2021).

13. Cao, J. *et al.* Integrated gut virome and bacteriome dynamics in COVID-19 patients. *Gut Microbes* **13**, (2021).

14. Zuo, T. *et al.* Temporal landscape of human gut RNA and DNA virome in SARS-CoV-2 infection and severity. *Microbiome* **9**, (2021).

15. Lu, Z. H. *et al.* Alterations in the Composition of Intestinal DNA Virome in Patients With COVID-19. *Front Cell Infect Microbiol* **11**, (2021).

16. Zuo, T. *et al.* Alterations in Fecal Fungal Microbiome of Patients With COVID-19 During Time of Hospitalization until Discharge. *Gastroenterology* **159**, (2020).

17. Lv, L. *et al.* Gut mycobiota alterations in patients with COVID-19 and H1N1 infections and their associations with clinical features. *Commun Biol* **4**, (2021).

18. Kusakabe, T. *et al.* Fungal microbiota sustains lasting immune activation of neutrophils and their progenitors in severe COVID-19. *Nat Immunol* **24**, (2023).

19. Wade, W. G. The oral microbiome in health and disease. *Pharmacological Research* vol. 69 Preprint at https://doi.org/10.1016/j.phrs.2012.11.006 (2013).

20. Huang, N. *et al.* SARS-CoV-2 infection of the oral cavity and saliva. *Nat Med* **27**, (2021).

21. Ren, Z. *et al.* Alterations in the human oral and gut microbiomes and lipidomics in COVID-19. *Gut* **70**, (2021).

22. Hu, X. *et al.* Oral Fungal Alterations in Patients with COVID-19 and Recovered Patients. *Advanced Science* **10**, (2023).

23. Soffritti, I. *et al.* Oral Microbiome Dysbiosis Is Associated With Symptoms Severity and Local Immune/Inflammatory Response in COVID-19 Patients: A Cross-Sectional Study. *Front Microbiol* **12**, (2021).

24. Iebba, V. *et al.* Profiling of Oral Microbiota and Cytokines in COVID-19 Patients. *Front Microbiol* **12**, (2021).

25. Wu, Y. *et al.* Altered oral and gut microbiota and its association with SARS-CoV-2 viral load in COVID-19 patients during hospitalization. *NPJ Biofilms Microbiomes* **7**, (2021).

26. Soriano, J. B., Murthy, S., Marshall, J. C., Relan, P. & Diaz, J. V. A clinical case definition of post-COVID-19 condition by a Delphi consensus. *The Lancet Infectious Diseases* vol. 22 Preprint at https://doi.org/10.1016/S1473-3099(21)00703-9 (2022).

27. Davis, H. E., McCorkell, L., Vogel, J. M. & Topol, E. J. Long COVID: major findings, mechanisms and recommendations. *Nature Reviews Microbiology* vol. 21 Preprint at https://doi.org/10.1038/s41579-022-00846-2 (2023).

28. Liew, F. *et al.* Large-scale phenotyping of patients with long COVID posthospitalization reveals mechanistic subtypes of disease. *Nat Immunol* **25**, 607–621 (2024).

29. Phetsouphanh, C. *et al.* Immunological dysfunction persists for 8 months following initial mild-to-moderate SARS-CoV-2 infection. *Nat Immunol* **23**, (2022).

30. Zuo, W. *et al.* The persistence of SARS-CoV-2 in tissues and its association with long COVID symptoms: a cross-sectional cohort study in China. *Lancet Infect Dis* (2024) doi:10.1016/S1473-3099(24)00171-3.

31. Tejerina, F. *et al.* Post-COVID-19 syndrome. SARS-CoV-2 RNA detection in plasma, stool, and urine in patients with persistent symptoms after COVID-19. *BMC Infect Dis* **22**, (2022).

32. Chen, Y. *et al.* Six-month follow-up of gut microbiota richness in patients with COVID-19. *Gut* vol. 71 Preprint at https://doi.org/10.1136/gutjnl-2021-324090 (2022).

33. Liu, Q. *et al.* Gut microbiota dynamics in a prospective cohort of patients with post Acute COVID-19 syndrome. *Gut* **71**, (2022).

34. Zhan, D. *et al.* Gut Microbiota Dysbiosis Correlates With Long COVID-19 at One-Year After Discharge. *J Korean Med Sci* **38**, (2023).

35. Shkoporov, A. N. *et al.* The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. *Cell Host Microbe* **26**, (2019).

36. Guo, J. *et al.* VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, (2021).

37. Bin Jang, H. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat Biotechnol* **37**, (2019).

38. Turner, D., Kropinski, A. M. & Adriaenssens, E. M. A roadmap for genome-based phage taxonomy. *Viruses* **13**, (2021).

39. Turner, D. *et al.* Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ICTV bacterial viruses subcommittee. *Arch Virol* **168**, (2023).

40. Khan, M. *et al.* Gut dysbiosis and il-21 response in patients with severe covid-19. *Microorganisms* **9**, (2021).

41. Gaibani, P. *et al.* The Gut Microbiota of Critically Ill Patients With COVID-19. *Front Cell Infect Microbiol* **11**, (2021).

42. Liu, Y. *et al.* Lower gut abundance of Eubacterium rectale is linked to COVID-19 mortality. *Front Cell Infect Microbiol* **13**, (2023).

43. Li, S. *et al.* Microbiome Profiling Using Shotgun Metagenomic Sequencing Identified Unique Microorganisms in COVID-19 Patients With Altered Gut Microbiota. *Front Microbiol* **12**, (2021).

44. Hazan, S. *et al.* Lost microbes of COVID-19: *Bifidobacterium*, *Faecalibacterium* depletion and decreased microbiome diversity associated with SARS-CoV-2 infection severity. *BMJ Open Gastroenterol* **9**, e000871 (2022).

45. Miquel, S. *et al.* Identification of metabolic signatures linked to anti-inflammatory effects of Faecalibacterium prausnitzii. *mBio* **6**, (2015).

46. Tao, W. *et al.* Analysis of the intestinal microbiota in COVID-19 patients and its correlation with the inflammatory factor IL-18. *Medicine in Microecology* **5**, (2020).

47. Tang, L. *et al.* Clinical Significance of the Correlation between Changes in the Major Intestinal Bacteria Species and COVID-19 Severity. *Engineering* **6**, (2020).

48. Tobin, C. A., Hill, C. & Shkoporov, A. N. Factors Affecting Variation of the Human Gut Phageome. *Annual Review of Microbiology* vol. 77 Preprint at https://doi.org/10.1146/annurev-micro-032421-105754 (2023).

49. Nishijima, S. *et al.* Extensive gut virome variation and its associations with host and environmental factors in a population-level cohort. *Nat Commun* **13**, (2022).

50. Tian, Y. *et al.* Gut Microbiota May Not Be Fully Restored in Recovered COVID-19 Patients After 3-Month Recovery. *Front Nutr* **8**, (2021).

51. Upadhyay, V. *et al.* Mild SARS-CoV-2 infection results in long-lasting microbiota instability. *mBio* **14**, (2023).

52. Nobre, J. G. *et al.* Gut microbiota profile of COVID-19 patients: Prognosis and risk stratification (MicroCOVID-19 study). *Front Microbiol* **13**, (2022).

53. Gogokhia, L. *et al.* Expansion of Bacteriophages Is Linked to Aggravated Intestinal Inflammation and Colitis. *Cell Host Microbe* **25**, (2019).

54. Zuo, T. *et al.* Gut mucosal virome alterations in ulcerative colitis. *Gut* **68**, (2019).

55. Wei, N. *et al.* Characterization of oral bacterial and fungal microbiome in recovered COVID-19 patients. *BMC Microbiol* **23**, (2023).

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

57. Li, S. *et al.* A catalog of 48,425 nonredundant viruses from oral metagenomes expands the horizon of the human oral virome. *iScience* **25**, (2022).

58. Tisza, M. J. & Buck, C. B. A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases. *Proc Natl Acad Sci U S A* **118**, (2021).

59. Merad, M. & Martin, J. C. Pathological inflammation in patients with COVID-19: a key role for monocytes and macrophages. *Nature Reviews Immunology* vol. 20 Preprint at https://doi.org/10.1038/s41577-020-0331-4 (2020).

60. Tsounis, E. P., Triantos, C., Konstantakis, C., Marangos, M. & Assimakopoulos, S. F. Intestinal barrier dysfunction as a key driver of severe COVID-19. *World J Virol* **12**, (2023).

61. Fan, L. *et al.* B. adolescentis ameliorates chronic colitis by regulating Treg/Th2 response and gut microbiota remodeling. *Gut Microbes* **13**, (2021).

62. Lu, H. *et al.* Butyrate-producing Eubacterium rectale suppresses lymphomagenesis by alleviating the TNF-induced TLR4/MyD88/NF-κB axis. *Cell Host Microbe* **30**, (2022).

63. Shkoporov, A. N. *et al.* Viral biogeography of the mammalian gut and parenchymal organs. *Nat Microbiol* **7**, (2022).

64. Gregory, A. C. *et al.* The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut. *Cell Host Microbe* **28**, (2020).

65. Nayfach, S. *et al.* Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat Microbiol* **6**, (2021).

66. Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D. & Lawley, T. D. Massive expansion of human gut bacteriophage diversity. *Cell* **184**, (2021).

67. Nayfach, S. *et al.* CheckV assesses the quality and completeness of metagenome assembled viral genomes. *Nat Biotechnol* **39**, (2021).

68. Hyatt, D. *et al.* Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, (2010).

69. Terzian, P. *et al.* PHROG: Families of prokaryotic virus proteins clustered using remote homology. *NAR Genom Bioinform* **3**, (2021).

70. Pandolfo, M., Telatin, A., Lazzari, G., Adriaenssens, E. M. & Vitulo, N. MetaPhage: an Automated Pipeline for Analyzing, Annotating, and Classifying Bacteriophages in Metagenomics Sequencing Data. *mSystems* **7**, (2022).

71. Cook, R. *et al.* INfrastructure for a PHAge REference Database: Identification of Large-Scale Biases in the Current Collection of Cultured Phage Genomes. *PHAGE: Therapy, Applications, and Research* **2**, (2021).

72. Hockenberry, A. J. & Wilke, C. O. BACPHLIP: Predicting bacteriophage lifestyle from conserved protein domains. *PeerJ* **9**, (2021).

73. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, (2012).

74. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, (2009).

75. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

76. Allard, G., Ryan, F. J., Jeffery, I. B. & Claesson, M. J. SPINGO: A rapid species classifier for microbial amplicon sequences. *BMC Bioinformatics* **16**, (2015).

77. Blanco-Míguez, A. *et al.* Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat Biotechnol* **41**, (2023).

# Chapter 3: Turning in silico phage into in vivo phage – the crAssphage story

This chapter was originally printed in the 34th edition of the Old Herborn University Seminar Monograph (2022). Edits have been made to include current relevant literature since the time of printing.

## Abstract

Phages of the order *Crassvirales* were originally identified as highly abundant and ubiquitous members of the gut microbiome from studies involving metagenomic sequencing and cross-assembly of reads from multiple individuals (crAss - Cross Assembly). More recently, it has been established that *Crassvirales* phages infect bacteria of the phylum *Bacteroidetes*. One of the most interesting facets to *Crassvirales* phages is their ability to persist at high numbers over extended periods in the laboratory and in the gut without significantly impacting the abundance of their bacterial hosts. Here we recount what is known about *Crassvirales* phages so far, from their in-silico discovery in 2014 and the subsequent identification of unique genomic features, to the first isolation of CrAss001 and elucidation of various biological characteristics resulting from the study of phage-host pairs in vivo. A large amount of information has been gleaned in a relatively short period of time but it is clear that *Crassvirales* phage research is still in its infancy. Future research lies in further in vivo work with phage-host pairs, coupled with the isolation of further *Crassvirales* phages from more divergent groups.

## Introduction

*Crassvirales* phages are an intriguing member of the human gut microbiome. They are both extremely prolific and widespread, and account for over 86% of gut viral genomes[1]. They have been identified in faecal samples from all over the globe and are found in all age groups from infants to the elderly[2]. While *Crassvirales* phages are rarely a component of the newborn microbiome, they become increasingly prevalent within the first year of life. It has been suggested that vertical transmission leads to this initial colonisation[3,4]. They have also been shown to be transferred and to stably engraft in recipients of faecal microbiota transplants[5]. *Crassvirales* phages have been identified in both remote and traditional hunter-gatherer populations, albeit at a reduced prevalence to that of industrialised populations[2,6]. *Crassvirales* phages have also been identified in sequences

recovered from old and new world primates, suggesting an ancient association of *Crassvirales* phages with these primates[2]. *Crassvirales* phages are generally viewed as benign viruses of the gut. Correlations between *Crassvirales* phages and a number of different health and disease factors have been examined, but no significant associations have been identified[6]. At the same time, one *Crassvirales* phage, termed IAS virus, was identified in stool samples from HIV-1 patients with unexplained diarrhoea. While no correlations between IAS virus and diarrhoea were established, it was significantly associated with advanced stages of HIV-1 disease[7].

## In silico discovery

The founding member of the *Crassvirales* phages was originally identified in silico in 2014[8]. Despite being both highly abundant and ubiquitous, its divergence from other bacteriophage sequences in viral databases allowed it to remain a component of the 'viral dark matter' for decades. Identification of the ~ 97 Kb dsDNA genome of this prototypical crAssphage (p-crAssphage) came from the reanalysis of publicly available viral metagenomes using cross-assembly software (crAss), from which crAssphage derives its name. A reference independent method, crAss cross-assembles all reads from input samples and counts the number of shared contigs[9]. At the time of assembly, fewer than 20 of the 80 predicted ORF's could be assigned a function. While the phage had never been grown in the laboratory, CRISPR spacers and co-occurrence analysis predicted that *Bacteroides* was the likely host of this phage.

Subsequent studies identified variations to the p-crAssphage genome[10,11]. This was expanded on by Yutin and colleagues[12], in which the use of sensitive protein analysis methods revealed several hundreds of putative members of a wider crAss-like family, found not only in the human gut but also in diverse environments such as groundwater, marine sediments and hypersaline brine. Annotation of a number of putative tail proteins with homologs to phage P22 led to the prediction that

crAss-like phages encode for short tails, typical of that of podophages. This was later supported by transmission electron microscopy (TEM) images of a faecal filtrate rich in crAss-like phages that was dominated by podophages[13].

## Current classifications and general genomic features

Hundreds of additional crAss-like phage genomes (each assembled from sequences derived from a single individual) were further classified into four subfamilies (Alphacrassvirinae, Betacrassvirinae, Gammacrassvirinae and Deltacrassvirinae) and ten genera (I - X), based on the percentage of protein coding genes shared between phages. CrAss-like phages sharing > 40% of protein coding genes were classified at the genus level, while crAss-like phages sharing 20 - 40% were classified at the subfamily level[13]. Further analysis by Yutin and colleagues[1] identified almost 600 crAss-like genomes and expanded classification to include two more groups, epsilon and zeta. This resulted in the establishment of the current *Crassvirales* order, which is comprised of four families (*Intestiviridae*, *Crevaviridae*, *Suoliviridae* and *Steigviridae*), eleven subfamilies, 42 genera and 73 species[14]. This was based on phylogenetic analysis of highly conserved proteins, including large terminase subunit, major capsid protein and DNA primase[15].

In general, *Crassvirales* phages have a genome size of ~ 100 Kb, although they can range from as large as ~150 - 200 Kb in the case of groups epsilon and zeta. Their genomes are divided roughly in two modules, with opposite gene orientation. One side harbours genes responsible for replication and nucleotide metabolism while the other encodes for structural genes[12,16,17]. At the junction of the two opposite modules 2-3 genes encode a very large RNA polymerase (RNAP) of up to 6000 amino acids in total. The genes encoding these subunits are sometimes on opposite coding strands. These subunits are highly divergent both from each other and other known RNAP proteins. They have been confirmed to be structurally most similar to eukaryotic RNAP that are involved in RNA interference[18]. Analysis of the first *Crassvirales* phage to have been grown in culture confirms that this large RNAP is a

component of the virion and are almost certainly involved in transcription of early phage genes[16,18]. Why *Crassvirales* phages encode such abnormally large genes for these subunits is currently unclear.

In addition to these large RNAP genes, *Crassvirales* phages possess a number of other intriguing genomic features. Most encode a DNA polymerase from either the A or B family and switching can occur between the two, even within the same subfamily[12,1]. A large number of *Crassvirales* phages, mainly from the beta, zeta and epsilon groups, appear to use alternative genetic codes[1,13]. Typically, this is by the reassignment of stop codon TAG to glutamine (genetic code 15) or TGA to tryptophan (genetic code 4) and seems to mainly occur in the late phage genes. The use of genetic code 15 was experimentally confirmed through liquid chromatography with tandem mass spectrometry (LC-MS/MS) in two *Crassvirales* phages[19]. The use of stop-codon recoding is widespread in phages[20]. In the human gut, this is predominantly in *Crassvirales* phages and Lak phages[1,20,21]. It has been proposed that phages may use stop-codon recoding as a regulatory mechanism to prevent premature expression of structural and lytic genes[20]. Furthermore, some *Crassvirales* phages, particularly of the zeta group, are characterised by an abnormally high density of introns and inteins[1]. Whether this spread of mobile genetic elements is uncontrollable due to some aspect of the phage lifestyle or is an adaptive approach used by the phage is so far unknown. Of those *Crassvirales* phages isolated to date, p-crAssphage and others used in comparative studies have lacked these features, leading to relatively simple annotation.

## In vitro isolation and biological properties

To date, 42 *Crassvirales* phages have been isolated with their hosts from two families, *Steigviridae* and *Intestiviridae*. *Crassvirales* phage 14:2, which infects *Cellulophaga baltica*, was first isolated from seawater in 2013, a year before the in-silico discovery of p-crAssphage[22]. It was originally classified in the now abolished

Podoviridae family, before later being reclassified as *Crassvirales* phage *Akihdevirus balticus* in the family *Steigviridae*[12]. *Kehishuvirus primarius* (CrAss001) was the first *Crassvirales* phage to be isolated from the human gut[16]. It also belongs to the family *Steigviridae* and infects *Bacteroides intestinalis*. *Jahgtovirus secundus* (CrAss002) was the first *Intestiviridae* phage isolated and infects *Bacteroides xylanisolvens*[17]. DAC15 and DAC17 (*Steigviridae*) both infect *Bacteroides thetaiotaomicron*[23]. Fourteen *Crassvirales* phages were isolated on *Bacteroides cellulosilyticus,* twelve belonging to *Steigviridae* and two belonging to *Intestiviridae*[24]. Finally, a further 25 *Steigviridae* phages were isolated on *Bacteroides intestinalis* (termed crAssBcn phages)[25]. Our current knowledge of the biological properties of *Crassvirales* phages comes from the studies of these phage-host pairs.

*Crassvirales* phages appear to establish an equilibrium with their hosts in liquid culture, in which they can propagate stably over many subcultures, without causing lysis of their hosts. This has also been demonstrated in a mouse model, where CrAss001 can maintain total counts between approximately $10^6$ -$10^8$ pfu/mL over a period of 136 days, without causing a reduction in *B. intestinalis* counts[26]. This corresponds with a longitudinal study of ten human volunteers, where *Crassvirales* phages were able to stably persist within individuals over a period of one year[27]. *Bacteroides* typically encode a number of alternative capsular polysaccharides (CPS) within their genomes, some which seem to be permissive to phage infection, while others are likely to be neutral or protective[28]. Phase variation of these CPS's appear to help to maintain the bacterial population in the presence of phages by creating a resistant subpopulation. The switch to a resistant subtype can occur even in the absence of phages[26]. DAC15 and DAC17 can infect the wild type version of *B. thetaiotaomicron*, which produces eight different CPS's. On the wild type strain they produce hazy plaques and can be maintained in liquid culture at $10^7$-$10^8$ pfu/mL. When infecting an engineered strain that produces only CPS3, clear plaques are formed and phage counts are maintained at a level one log higher in liquid culture[26]. Interestingly, lysis of the cell culture still does not occur, suggesting that other mechanisms, such as phase variation of outer membrane lipoproteins, may also contribute to the dynamic equilibrium between both partners[28].

A second mechanism, likely to be working in parallel to phase variation, is an apparent delayed release of phage progeny from infected cells. In classic one step growth curves at a multiplicity of infection (MOI) of one, CrAss001, DAC15 and DAC17 all generate very small burst sizes of between 2.5-8 pfu per infected cell[16,26]. In the case of CrAss001, it has been observed that a second small burst occurs around 90 minutes after the first. Despite this, more than 50 phage particles per cell can be observed in TEM's of CrAss001 spotted onto lawns of *B. intestinalis*, and at least 20 copies of CrAss001 genome per copy of the *B. intestinalis* genome are produced within 90 minutes of infection. Intriguingly, this mechanism of delayed release can be overcome by increasing the MOI in *B. theta* but the same cannot be observed for *B. intestinalis*[26].

While CrAss001, DAC15 and DAC17 have all been demonstrated to behave quite similarly, CrAss002 has a number of additional interesting properties. It is unable to form plaques on its host and can only form opaque spots that are barely visible when a highly concentrated phage stock is spotted onto a lawn of its sensitive bacterial host[17]. Furthermore, propagation in liquid culture only occurs after a minimum of three days subculturing and it appears that it is *B. xylanisolvens* that adapts to allow phage infection. Why adaptation by the host is delayed is unknown. Phase variation is a possible contributor here too, with *B. xylanisolvens* possessing many recombination hotspots[17].

## Structural model for the *Crassvirales* order

A major development in our structural and functional understanding of *Crassvirales* phages came from the cryogenic electron microscopy (cryo-EM) reconstruction of CrAss001, which allowed for atomic modelling of the majority of the proteins that make up the virion[29]. Additionally, comparative sequence analysis across the *Crassvirales* order revealed extensive conservation of the majority of the structural proteins. The major capsid protein (MCP) was found to have a HK97-like fold[30], which is the classical fold of phages in the class *Caudoviricetes*[31]. Additionally, the capsid is composed of an auxiliary protein, whose structure appears to be unique to *Crassvirales* phages, as well as dimeric and trimeric head fibre proteins[29]. The

current function of these head fibre proteins is unknown, but they may be involved in binding to surfaces of the bacterial host, or assist in persistence, through binding to surfaces or molecules in the gut, such as mucus[32].

The tail barrel is composed of five stacked dodecameric rings[29]. The presence of multiple rings appears to be unique to *Crassvirales* phages, with other podophage possessing one such ring. Tail hub (TH) proteins A and B form a continuous collar around the tail barrel. THB was shown to dock the N-terminal domain of one tail fibre, gp22, to the tail barrel, while the remaining tail fibre proteins could not be resolved due to flexibility[29]. At the end of the tail is a muzzle protein which contains a domain with a unique fold, referred to as the crass fold. Areas within both the capsid and tail were identified to carry proteins, referred to as cargo proteins, which are thought to be ejected from the tail before phage DNA and assist in the formation of a transmembrane channel through to the bacterial host's cytoplasm[29].

## Ecological models

A number of different models have been described to explain bacteriophage lifestyles and their interactions with their hosts in their environment. While often used to describe phage-host dynamics in the ocean, they can be extrapolated to illustrate interactions within the human gut. The "piggyback the winner" model has traditionally been viewed as a switch from a lytic to temperate lifestyle, in order to keep host levels at a high abundance[33]. However, recent work has suggested that a healthy human virome is dominated by a lytic core of *Crassvirales* phages and *Microviridae*[27]. It is possible that the use of mechanisms such as phase variation and delayed release of phage progeny by *Crassvirales* phages correspond with alternative mechanisms of benign infection that align with the piggyback the winner model. These less aggressive means of replication may allow the bacterial population to persist to a high abundance, therefore enabling *Crassvirales* phages persistence.

It is also possible that other models are at play, such as "Royal Family" dynamics. This is where the "kill the winner" model operates at the strain level, such that any

fluctuations in phages or host populations would then go undetected[34]. Recently, enrichment of nonsynonymous variants has been described in tail protein encoding genes of *Crassvirales* phages[35,4]. As tail proteins are often important mediators in bacterial host recognition, it is possible that *Crassvirales* phages use these tail protein variants to switch hosts at the strain level, in line with Royal family dynamics. Possibly, *Crassvirales* phages use a number of different mechanisms from different models in combination, to enable them to persist stably, without affecting host abundance.

## Conclusions

In just over a decade, *Crassvirales* phages have gone from a viral unknown, to being recognised as the most abundant phages of the human gut, with sequences isolated all over the world. In silico analysis have revealed an array of interesting genomic features but many of their biological purposes remain unclear. Isolation of the first few phage-host pairs has allowed the elucidation of some of the biological characteristics of these phages (Figure 1). However, 39 of the 42 isolated *Crassvirales* phages are from the family *Steigviridae* (previously Betacrassvirinae group), which have been shown to make up just < 1.5% of the *Crassvirales* phages dominating the human gut[1]. Families *Intestiviridae* and *Crevaviridae* (previously Alphacrassvirinae and Gammacrassvirinae groups) dominate the gut virome, combined they make up just under half of all *Crassvirales* phages[1]. Of the three isolated *Intestiviridae*, one (CrAss002), exhibits a number of biological mechanisms that are different from those of the family *Steigviridae*. It is likely that *Crassvirales* phages employ a number of different mechanisms that allow them to persist so stably within the human gut. Isolation of members from other families will be required in order to garner further insights into these mechanisms. This will not be an easy task given that the hosts of these phages are *Bacteroidetes*, strict anaerobes that are difficult to isolate in their own right. The use of alternative genetic codes and the presence of introns and inteins in some of the *Crassvirales*

phages also add to the difficulty of correct annotation. It is vital that we continue to culture new members of this extensive order. Only by growing these phages in the laboratory will we begin to truly understand their role in the gut microbiota. As the most abundant phages of the human gut, it is befitting that they should also become the most studied.

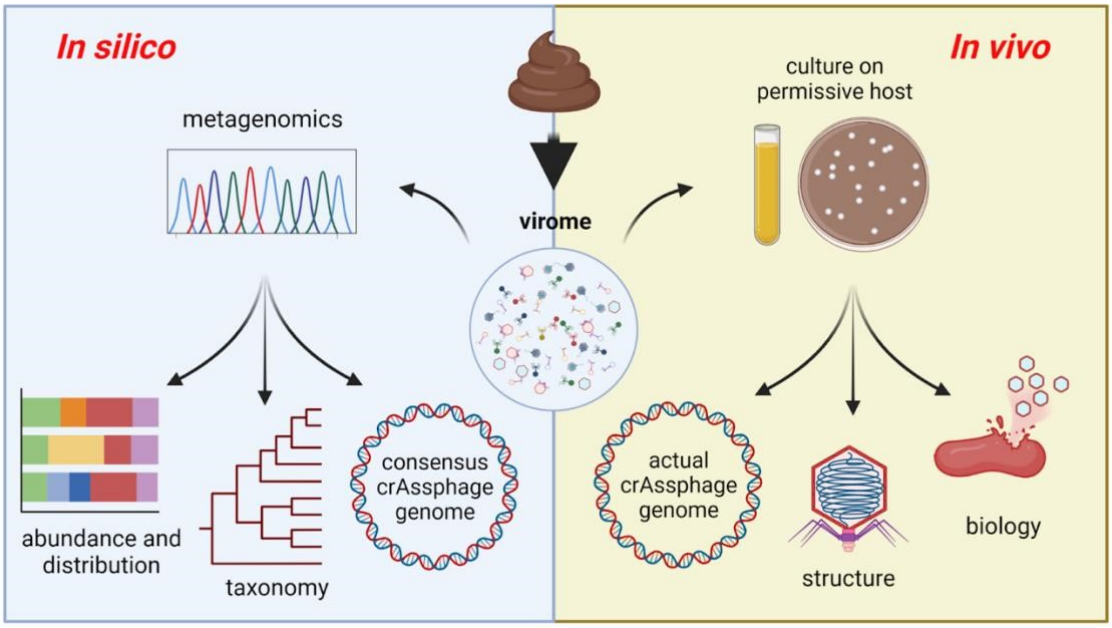



**Figure 1.** Both in silico and in vivo approaches can provide important information about *Crassvirales* phages, but a true understanding of these highly abundant viruses (especially in terms of their biology, structure, and genomic features) will depend on our ability to grow them in the laboratory.

## References


1. Yutin, N. *et al.* Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features. *Nat Commun* **12**, (2021).

2. Edwards, R. A. *et al.* Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. *Nat Microbiol* **4**, (2019).

3. McCann, A. *et al.* Viromes of one year old infants reveal the impact of birth mode on microbiome diversity. *PeerJ* **6**, e4694 (2018).

4. Siranosian, B. A., Tamburini, F. B., Sherlock, G. & Bhatt, A. S. Acquisition, transmission and strain diversity of human gut-colonizing crAss-like phages. *Nat Commun* **11**, 280 (2020).

5. Draper, L. A. *et al.* Long-term colonisation with donor bacteriophages following successful faecal microbial transplantation. *Microbiome* **6**, 220 (2018).

6. Honap, T. P. *et al.* Biogeographic study of human gut-associated crAssphage suggests impacts from industrialization and recent expansion. *PLoS One* **15**, e0226930 (2020).

7. Oude Munnink, B. B. *et al.* Unexplained diarrhoea in HIV-1 infected individuals. *BMC Infect Dis* **14**, 22 (2014).

8. Dutilh, B. E. *et al.* A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat Commun* **5**, (2014).

9. Dutilh, B. E. *et al.* Reference-independent comparative metagenomics using cross-assembly: crAss. *Bioinformatics* **28**, 3225–3231 (2012).

10. Manrique, P. *et al.* Healthy human gut phageome. *Proceedings of the National Academy of Sciences* **113**, 10400–10405 (2016).

11. LIANG, Y. Y., ZHANG, W., TONG, Y. G. & CHEN, S. P. crAssphage is not associated with diarrhoea and has high genetic diversity. *Epidemiol Infect* **144**, 3549–3553 (2016).

12. Yutin, N. *et al.* Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nat Microbiol* **3**, 38–46 (2018).

13. Guerin, E. *et al.* Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. *Cell Host Microbe* **24**, 653-664.e6 (2018).

14. Turner, D. *et al.* Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ICTV bacterial viruses subcommittee. *Arch Virol* **168**, (2023).

15. Shkoporov AN *et al.* Create one new order (Crassvirales) including four new families, ten new subfamilies, 42 new genera and 73 new species (Caudoviricetes). *https://ictv.global/ictv/proposals/2021.022B.R.Crassvirales.zip* (2021).

16. Shkoporov, A. N. *et al.* ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. *Nat Commun* **9**, (2018).

17. Guerin, E. *et al.* Isolation and characterisation of ΦcrAss002, a crAss-like phage from the human gut that infects Bacteroides xylanisolvens. *Microbiome* **9**, (2021).

18. Drobysheva, A. V. *et al.* Structure and function of virion RNA polymerase of a crAss-like phage. *Nature* **589**, 306–309 (2021).

19. Peters, S. L. *et al.* Experimental validation that human microbiome phages use alternative genetic coding. *Nat Commun* **13**, (2022).

20. Borges, A. L. *et al.* Widespread stop-codon recoding in bacteriophages may regulate translation of lytic genes. *Nat Microbiol* **7**, 918–927 (2022).

21. Devoto, A. E. *et al.* Megaphages infect Prevotella and variants are widespread in gut microbiomes. *Nat Microbiol* **4**, 693–700 (2019).

22. Holmfeldt, K. *et al.* Twelve previously unknown phage genera are ubiquitous in global oceans. *Proc Natl Acad Sci U S A* **110**, 12798–12803 (2013).

23. Hryckowian, A. J. *et al.* Bacteroides thetaiotaomicron-Infecting Bacteriophage Isolates Inform Sequence-Based Host Range Predictions. *Cell Host Microbe* **28**, (2020).

24. Papudeshi, B. *et al.* Host interactions of novel Crassvirales species belonging to multiple families infecting bacterial host, Bacteroides cellulosilyticus WH2. *Microb Genom* **9**, (2023).

25. Ramos-Barbero, M. D. *et al.* Characterization of crAss-like phage isolates highlights Crassvirales genetic heterogeneity and worldwide distribution. *Nat Commun* **14**, (2023).

26. Shkoporov, A. N. *et al.* Long-term persistence of crAss-like phage crAss001 is associated with phase variation in Bacteroides intestinalis. *BMC Biol* **19**, (2021).

27. Shkoporov, A. N. *et al.* The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. *Cell Host Microbe* **26**, (2019).

28. Porter, N. T. *et al.* Phase-variable capsular polysaccharides and lipoproteins modify bacteriophage susceptibility in Bacteroides thetaiotaomicron. *Nat Microbiol* **5**, 1170–1181 (2020).

29. Bayfield, O. W. *et al.* Structural atlas of a human gut crassvirus. *Nature* **617**, (2023).

30. Helgstrand, C. *et al.* The Refined Structure of a Protein Catenane: The HK97 Bacteriophage Capsid at 3.44 Å Resolution. *J Mol Biol* **334**, 885–899 (2003).

31. Duda, R. L. & Teschke, C. M. The amazing HK97 fold: versatile results of modest differences. *Curr Opin Virol* **36**, 9–16 (2019).

32. Barr, J. J. *et al.* Bacteriophage adhering to mucus provide a non–host-derived immunity. *Proceedings of the National Academy of Sciences* **110**, 10771–10776 (2013).

33. Knowles, B. *et al.* Lytic to temperate switching of viral communities. *Nature* **531**, 466–470 (2016).

34. Breitbart, M., Bonnain, C., Malki, K. & Sawaya, N. A. Phage puppet masters of the marine microbial realm. *Nat Microbiol* **3**, 754–766 (2018).

35. Brown, B. P. *et al.* crAssphage genomes identified in fecal samples of an adult and infants with evidence of positive genomic selective pressure within tail protein genes. *Virus Res* **292**, 198219 (2021).

# Chapter 4: CRISPRi targeting of genes in *Crassvirales* phage DAC15 predicts multiple anti-CRISPR proteins

Ciara A Tobin, Oliver W Bayfield, Andrey N Shkoporov and Colin Hill

## Abstract

CRISPR interference (CRISPRi) is a powerful tool for suppressing gene expression that has successfully been applied to prokaryotes and eukaryotes. We used CRISPRi to target nine genes predicted to code for structural proteins and enzymes involved in different stages of the life cycle in a *Crassvirales* phage DAC15 infecting *Bacteroides thetaiotaomicron*. The functionality of the CRISPRi system was confirmed using a host chromosomal gene as a target. For phage genes encoding essential functions such as DNA polymerase and terminase large subunit (TerL), we observed reductions in efficiency of plaquing (< 50% reduction). Suppression of TerL resulted in an increase in the number of empty phage capsids, while still maintaining a large proportion of viable phages. The incomplete suppression of targeted phage genes, in contrast to high levels of suppression of a chromosomal target, suggested that DAC15 may deploy a defensive strategy to overcome CRISPR. Further bioinformatic analysis led to the predictions of five putative anti-CRISPR proteins (Acrs). Similar Acr-like genes could be identified in the genomes of other *Crassvirales* phages, including CrAss001, CrAss002 and p-crAssphage. Our results show that CRISPRi may be compromised when used to target phage genes, particularly in phages that may have as yet unidentified anti-CRISPR defensive strategies.

## Introduction

Viral metagenomic studies have highlighted the abundance and diversity of bacteriophages (phages) across various environments including the gut[1], soil[2] and ocean[3]. Estimated at a global abundance of $10^{31}$ particles[4], it is postulated that phages can significantly shape their microbial environments through infection of their bacterial hosts. Bacteria have evolved a number of defence systems to overcome phage predation[5], including adaptive immune systems such as Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)- CRISPR associated proteins (Cas)[6], toxin-antitoxins[7] and abortive infection[8], in addition to innate systems such as restriction modification[9]. CRISPR-Cas systems protect prokaryotic hosts through storage of foreign DNA fragments in the form of CRISPR spacers, as a record of past infection events. When that foreign DNA is encountered again, the spacer is transcribed, processed and can then direct Cas effectors to the foreign DNA and inactivate it[10]. CRISPR systems are diverse, spanning two classes, six types and 33 sub-types[11]. Unsurprisingly, phages have responded to this evolutionary arms race with several anti-defence systems[12], including anti-CRISPR proteins (Acrs). First discovered in prophage of *Pseudomonas aeruginosa* in 2013[13], over 100 Acrs have since been identified in mobile genetic elements[14]. These proteins can have various modes of action but they typically involve either direct interaction with Cas proteins to prevent target binding or cleavage, or degradation of cyclic oligonucleotide signalling molecules[15].

Phage genome diversity results in a large number of genes with unknown function, sometimes up to 70%[16], as they have no sequence homology to genes with known functions. This impedes our understanding of the roles that phages play in microbial communities. While classical genetic methods have provided a detailed insight into the structure and function of a number of canonical phages[17,18,19], these methods are labour and resource intensive and it is not feasible to extend this approach to

the rapidly increasing number of phage genomes generated in silico as a result of high throughput sequencing.

CRISPR interference (CRISPRi) has emerged as a useful tool in functional genomics. Adapted from CRISPR gene editing, CRISPRi employs the use of a catalytically inactive version of Cas9 (dCas9[20]) or Cas12a (dCas12a[21]), coupled with a single guide RNA (sgRNA). The inactive or "dead" Cas protein is no longer able to introduce a double-stranded DNA break and instead sterically hinders RNA polymerase, leading to a reduction in transcription. Importantly, this is a reversible process and doesn't involve any modifications to the genome, making it an attractive tool for probing gene functions. CRISPRi has been used to confirm the essentiality of selected genes in bacteria[22,23,24] and eukaryotes[25,26], as well as confirming roles for noncoding RNA's in bacteria[27]. Recently, CRISPRi has also been extended to phages, with dCas9 used to induce prophages λ and P2[28], while essential and non-essential genes were examined in λ and P1 using dCas12a[29].

The in-silico discovery of the first member of the order *Crassvirales*, *Carjivirus communis* (prototypical or p-crAssphage), occurred just over a decade ago[30]. Since then, a plethora of *Crassvirales* phages have been identified, predominantly from the human gut, but they have also been identified from aquatic and terrestrial sources[31]. Despite being highly abundant and prevalent in human gut microbiomes across the globe[32,33], only a handful of *Crassvirales* phages have been successfully isolated in pure culture against members of the *Bacteroidetes*[34,35,36,37,38,39]. Studies with these cultured phages have revealed bacterial defence mechanisms such as phase variation[37] and delayed lysis of infected bacterial cells[40], which provides some insight into how these phages persist at such high levels in tandem with their hosts. Metagenomic analysis has also identified a number of intriguing genomic features in these phages, including the use of alternative genetic codes in late genes in several genera[41,42], a high density of self-splicing introns and inteins[41], and gene(s) encoding large RNA-polymerase subunits that are packaged within the virion[43,44]. The recent cryogenic electron microscopy reconstruction of *Kehishuvirus primarius* (CrAss001) resulted in the functional assignment of the majority of the structural proteins in this phage, which could be extrapolated by sequence homology to other

members of the *Crassvirales*[44]. However, there are still a large number of genes in *Crassvirales* phages whose functions are unknown.

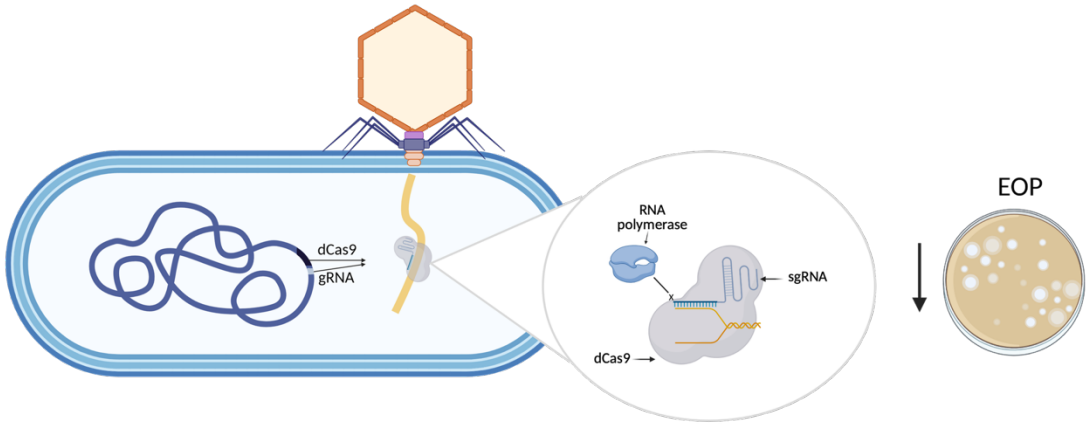
In this study, we sought to use CRISPRi to target presumed essential and nonessential genes in *Crassvirales*. Mimee *et al* designed a dCas9 plasmid vector with the necessary genetic elements to stably integrate into the genome of *Bacteroides thetaiotaomicron* (*B. theta*)[45] and successfully used this system to suppress gene expression. As *B. theta* has two cognate *Crassvirales* phages, DAC15 (*Wulfhauvirus bangladeshii*) and DAC17 (an unclassified *Wulfhauvirus*), we decided to use this system to target genes in DAC15. We designed targets to nine predicted structural and functional genes. After observing only a minimal impact on gene expression, we analysed the DAC15 genome for potential CRISPR defensive strategies and identified five putative Acr candidates. This analysis was further expanded to include CrAss001, CrAss002 (*Jahgtovirus secundus*) and p-crAssphage, highlighting how CRISPRi technology may not always be effective for phage genomes, particularly where defensive strategies may be encoded in the genome.

## Results

The integrative plasmid designed by Mimee *et al* contains both dCas9 and the gRNA sequence, under the control of an IPTG inducible promoter (**Fig1a**). To test whether this system could be programmed to target DAC15 genes during phage infection, we designed gRNA sequences to target four predicted tail fibre genes (*gp099*, *gp100*, *gp101* and *gp102*), a trimeric head fibre gene *(gp103*), two subunits of RNA polymerase (*gp076* and *gp078*), DNA polymerase (*gp063*) and terminase large subunit (*gp106*, TerL, **Fig1b**). In addition, a gRNA nonsense sequence (NS) in pMM710 that has no targets in either *B. theta* or DAC15 was used as a negative control. A gRNA targeting *BT1854*, a host chromosomal gene required for resistance to cationic microbial peptides in *B. theta*, was used as a positive control. The gRNA sequences were designed to target the non-template strand and PAM sequences located at the start of the gene were chosen. A gRNA to target the third RNA polymerase subunit, *gp077*, was originally designed but was discovered to have an

off-target effect on *B. theta*, halting growth, and so was not included in any further analysis (**FigS1**).

**a**




**b**

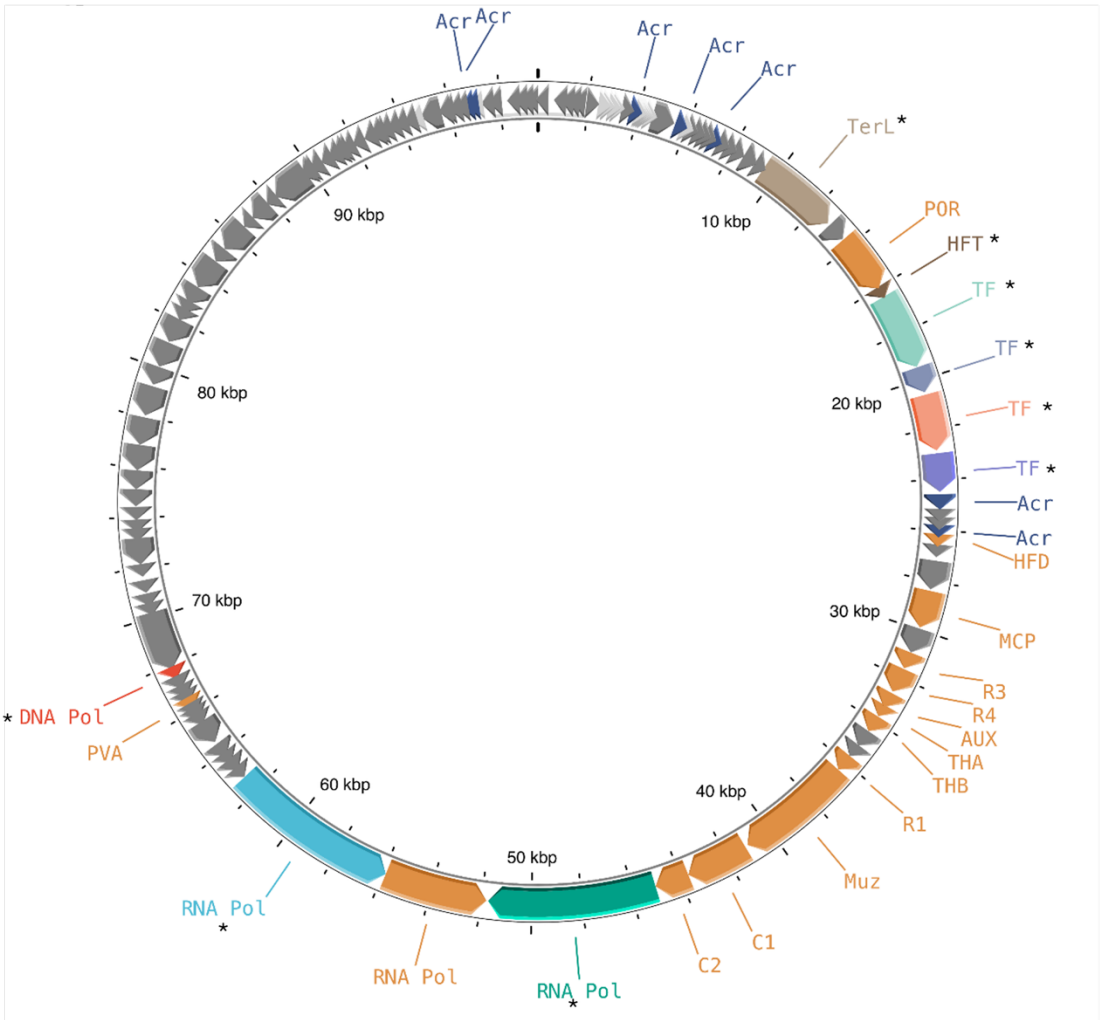


**Figure 1. Schematic overview of CRISPRi system and genes targeted in DAC15 genome. a)** The dCas9 and gRNA genes stably persist in the *B. theta* genome, due to the integration of the pNBU2 based plasmid. Under the control of a lac operon, expression of dCas9 and gRNA are induced in the presence of IPTG. When DAC15 infects *B. theta*, dCas9 is guided to the gene of interest by the gRNA and blocks the ability of RNA polymerase to proceed with transcription. If an essential gene is targeted, this would be expected to translate to a reduction in EOP. **b)** Gene organisation of DAC15. Genes targeted during this study are colour coded and marked with an asterisk, while genes with predicted structural function are coloured in orange and predicted Acrs during this study are coloured purple. Genes with unknown function are coloured in grey. POR, portal protein; HFT, head fibre trimer; TF, tail fibre; HFD, head fibre dimer; MCP, major capsid protein; R1-4, rings 1-4; THA/B, tail hub proteins A/B; Muz, muzzle protein, C1/C2, cargo proteins; PVA, portal vertex auxiliary capsid protein.

## CRISPRi targeting of phage genes

The impact of CRISPRi on each of the gene targets was first assessed by plate assays, where serial dilutions of DAC15 were spotted on induced overlay cultures. Non-significant reductions in efficiency of plaquing (EOP) was observed for one RNA polymerase subunit *gp076* (54%, paired *t*-test, $p = 0.0686$) and tail fibres *gp099* (46%, paired *t*-test, $p = 0.0748$), *gp101* (29%, paired *t*-test, $p = 0.1843$) and *gp100* (20%, paired *t*-test, $p = 0.4573$) (**Fig2a**). Targeting TerL only resulted in a 9% reduction in EOP (paired *t*-test, 0.0798). Knockdown of targeted genes was also tested in liquid assays, where DAC15 was propagated on the induced strains containing the integrated plasmids. Samples were taken after three hours and 24 hours and serial dilutions of the resulting lysate were spotted on a derivative of *B. theta* VPI-5482, expressing a single capsular polysaccharide CPS3 ($CPS3^+$). Overall, liquid propagation resulted in a greater decrease in phage titres as compared to plate assays. After three hours, a significant reduction was observed when RNA polymerase subunit *gp076* (*71%*, paired *t*-test, $p = 0.006$) was targeted (**Fig 2b**). A large but non-significant reduction was also observed when tail fibres *gp099* (60%, paired *t*-test, $p = 0.0819$) and *gp100* (59%, paired *t*-test, $p = 0.1032$), were targeted (**Fig2b**). There was also an overall 40% reduction for tail fibre *gp102* (paired *t*-test, $p = 0.1673$), 32% reduction for both head fibre *gp103* (paired *t*-test, $p = 0.1686$) and

TerL *gp106* (paired *t*-test, *p* = 0.1531) and a 30% reduction for tail fibre *gp101* (paired *t*-test, *p* = 0.2359), but there was a large degree of variation between replicates for these four genes. Continuing phage propagation over 24 hours resulted in a greater reduction in phage titre for the majority of the knockdowns (**Fig2c**). Two of the tail fibres had a significant reduction in titre (*gp100*, 69%, paired *t*-test, *p* = 0.0130 and *gp099*, 50%, paired *t*-test, *p* = 0.0105). Tail fibre *gp101* also had a 50% reduction in phage numbers (paired *t*-test, *p* = 0.0961) but there was greater variation between replicates. Similarly, tail fibre *gp102*, which had a 21% reduction in phage titre, was also highly variable between replicates (paired *t*-test, *p* = 0.4167). Targeting DNA polymerase *gp062* and RNA polymerase subunit *gp076* both led to a significant reduction of 51% (paired *t*-test, *p* = 0.0481 and *p* = 0.0294), while RNA polymerase subunit *gp078* and TerL *gp106* were reduced by 39% and 40% respectively (paired *t*-test, p = 0.1513 and *p* = 0.0977). Overall, CRISPRi resulted in significant reductions in liquid assays when targeting DNA polymerase, one RNA polymerase subunit and two tail fibres. However, this gene repression was limited to 50-70% reduction in phage titre.

When designing gRNA targets for CRISPRi with dCas9 in prokaryotes, it has generally been found that targeting closer to the 5' end of the gene leads to improved repression[46,47]. With this in mind, we identified PAM sequences located at the start of the gene of interest and designed the gRNA to target the subsequent 20 base pairs. To examine the effect that the gRNA position has on phage gene repression, we designed three gRNA targets to repress DNA polymerase *gp062*, located approximately 3% (gp062-1), 6% (gp062-2) and 15% (gp062-3) of the total gene length starting from the 5' end of the gene (**Fig2d**). When assessed by plate assays, only gp062-3 showed a reduction in EOP (14%, paired *t*-test, *p* = 0.0668). There were no reductions in titre after three hours of liquid assay propagation while after 24 hours, gp062-2 was reduced by 51% (paired *t*-test, *p* = 0.0481) and gp062-1 by 35% (paired *t*-test, 0.1182). Overall, reduction of EOP and phage titre was variable when targeting within the first 15% of gp062, suggesting that targets should be designed for further along the gene to assess the importance of gRNA positioning for phage gene suppression.

The limited reduction in EOP when targeting TerL and DNA polymerase expression was unexpected, given their essentiality in DNA packaging and genome replication, respectively. A greater than 99.99% reduction would be expected based on similar essential genes in λ and P1[29]. As a positive control, we also included pMM704, which has the same design as pMM710, but contains a gRNA that targets *BT1854* in *B. theta*[45]. *BT1854* encodes LpxF, which is required for resistance to cationic antimicrobial peptides, such as polymyxin B[48]. Targeted repression of *BT1854* should therefore result in a reduction of the minimum inhibitory concentration (MIC) of polymyxin B required to prevent growth of *B. theta*[45]. In *B. theta* containing the NS sequence target, bacterial growth, measured by changes in $OD_{595nm}$, was unaffected when polymyxin B was added at concentrations between 32-1024 µg/mL (**Fig 2e**). Targeting of *BT1854* resulted in inhibition of growth between 128-1024 µg/mL of polymyxin B. Intermediate growth was observed with the addition of 64 µg/mL and growth was minimally affected with 32 µg/mL (**Fig 2e**). This demonstrates that the CRISPRi study design is effective when targeting bacterial genes.

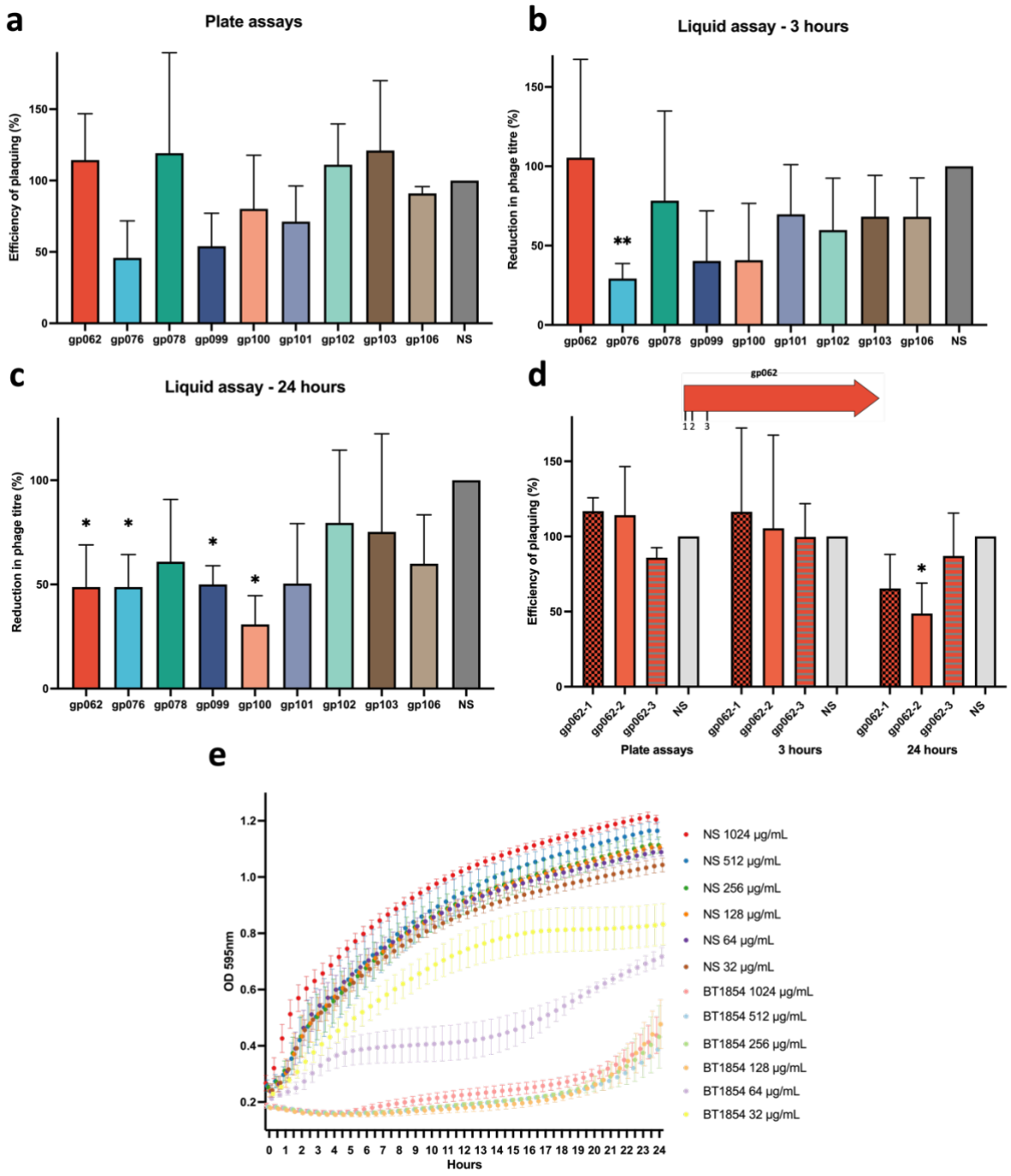



**Figure 2. Reduction in EOP is most consist in tail fibres and one RNA polymerase subunit. A)** EOP reduction in nine targeted genes, when assessed by plate assays. **B)** Reduction in phage titres in nine targeted genes, when assessed three hours after liquid assay propagation. **C)** Reduction in phage titres in nine targeted genes, when assessed 24 hours after liquid assay propagation. **D)** Position of gRNA on the target gp062 shows no preference within the first 15% of the gene. When assessed by plate assays, targeting 15% from the 5' end (gp062-3) shows a 14% reduction in EOP, while targeting 6% from the 5' end (gp062-2) and 3% from the 5' end (gp062-1) shows 51% and 35% reduction after 24 hours of liquid propagation. All EOP and phage titre results were calculated by dividing the pfu/mL of DAC15 on the engineered strain by the pfu/mL on the strain containing the NS target and multiplying by 100 for percentage reduction. The *P*-value of each target was determined by the two-

tailed paired *t*-test, compared to the NS target. **E)** CRISPRi is effective when targeting *B. theta* genome. Targeting of *BT1854*, which is required for resistance to polymyxin B, inhibits growth between 128-1024 μg/mL of polymyxin B. All error bars represent the standard deviation of three biological replicates, prepared on separate days.

## CRISPRi targeting of TerL enriches for the production of empty phage capsids

Effective gene repression when DAC15 is not present suggests that something is interfering with the system in the presence of phage resulting in suboptimal knockdown of DAC15 genes. However, despite this, we hypothesized that CRISPRi may have resulted in certain structural defects in progeny phages, such as missing tail fibres or the formation of empty phage capsids when TerL was targeted. To examine the potential formation of empty capsids, DAC15 samples were taken after 24 hours of propagation on *B. theta* CPS3$^{+}$ containing either the TerL or the NS target and were purified by caesium chloride step gradients. Three bands, one lower phage band and two upper bands, appeared for each sample **(Fig 3)**. Upper band 1 was mainly composed of small heterogeneous particles, possibly cell debris and lipids, in both the TerL knockdown and the NS control **(Fig 3a-b)**. Upper band 2 in the TerL knockdown contained empty capsids, while the control mainly contained more small heterogeneous particles **(Fig 3c-d)**. There were a small number of empty capsids present in upper band 2 of the NS control but empty capsid formation was enriched in the TerL knockdown. Viable phages appeared in the lower band of both the TerL knockdown and NS control **(Fig 3e-f)**. Taken together, these results suggest that while there was only a small reduction in EOP and phage titre, targeting of TerL resulted in an accumulation of empty phage capsids in at least some of the cells, or a small portion of newly formed phage heads could not be filled with DNA in every cell, due to a lower TerL availability, which could be the rate-limiting factor in the capsid filling process.

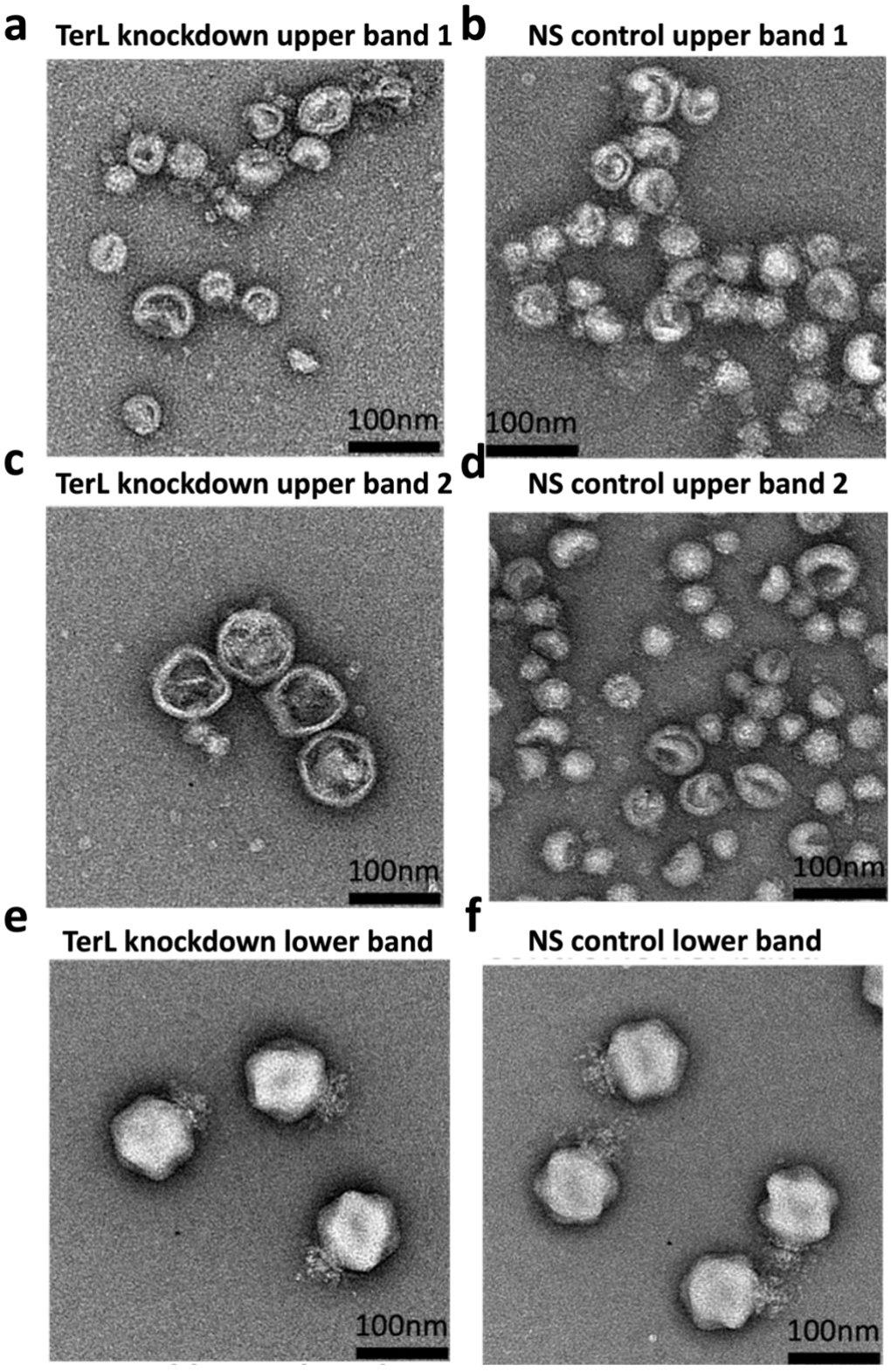



**Figure 3. Transmission Electron Microscopy (TEM) of uranyl acetate negatively contrasted DAC15 post propagation on TerL knockdown *B. theta* CPS3$^+$ strain and NS control. A-B)** TEM's of first upper bands small heterogeneous particles for both TerL knockdown and control. **C-D)** Upper band 2 in TerL knockdown is enriched in empty phage capsids while upper band 2 of the control contains mainly small heterogeneous particles and a small number of empty capsids. **E-F)** Lower band in both TerL knockdown and control contains viable phages. Scale bar is 100 nm.

Prediction of five Acrs in the genome of DAC15

One possibility that might explain the incomplete suppression of targeted genes could be the presence of a defensive mechanism against CRISPR located on the DAC15 genome, such as genes encoding Acrs. In order to identify potential Acrs in DAC15, all proteins of between 30-300 amino acids in length were passed through a machine learning based predictor, PaCRISPR[49]. Thirty proteins in total were initially predicted to be Acrs, scoring greater than the threshold of 0.5 (**Table S1**). Pairwise structural alignment was then performed between these putative DAC15 Acrs and all experimentally verified Acrs. Five of the predicted Acrs shared structural similarities with published Acrs, of which four had high PaCRISPR scores (**Fig 4**). Gp013 and gp117 had structural similarities to Acrs that target class I CRISPR-Cas systems (**Fig 4a-b)**. Gp013 is composed of three inner anti-parallel β-strands, surrounded by three α-helices, which is the hallmark structure of AcrIC9[50]. Gp117 shared structural features with AcrIF15, an Acr that has broad inhibitory breadth, as it can inhibit type I-F CRISPR-Cas systems across distinct species[51]. The remaining three predicted Acrs, gp014, gp112, and gp119 all shared similar structural topology to AcrIIC4 (**Fig 4c-e**). Initially identified in *Haemophilus influenzae*[52], AcrIIC4 inhibits Cas9 of the type II-C CRISPR-Cas system through binding and preventing Cas9 activation[53]. Each of the three DAC15 proteins possess the long flexible helix that is characteristic of AcrIIC4[53]. A further two DAC15 proteins (gp095 and gp098) had structural similarity to AcrIIC4 **(Table S1)**. However, as Acrs are required to inhibit CRISPR-Cas systems before the phage is destroyed, they are usually one of the earliest genes to be expressed upon infection[54]. Gp095 and gp098 are located in between tail fibre, head fibre and major capsid genes **(Fig 1b)** and are located in the gene expression/capsid modules that are transcribed during the middle or late stages of infection[43]. They are therefore less likely to be true Acrs than the other five proteins, which are located in clusters of short genes of unknown function, likely part of the replicative module and therefore transcribed during early stages of infection[43]. Identification of putative Acrs on the DAC15

genome is one possible explanation for the minimal repression observed when using CRISPRi to repress DAC15 genes.

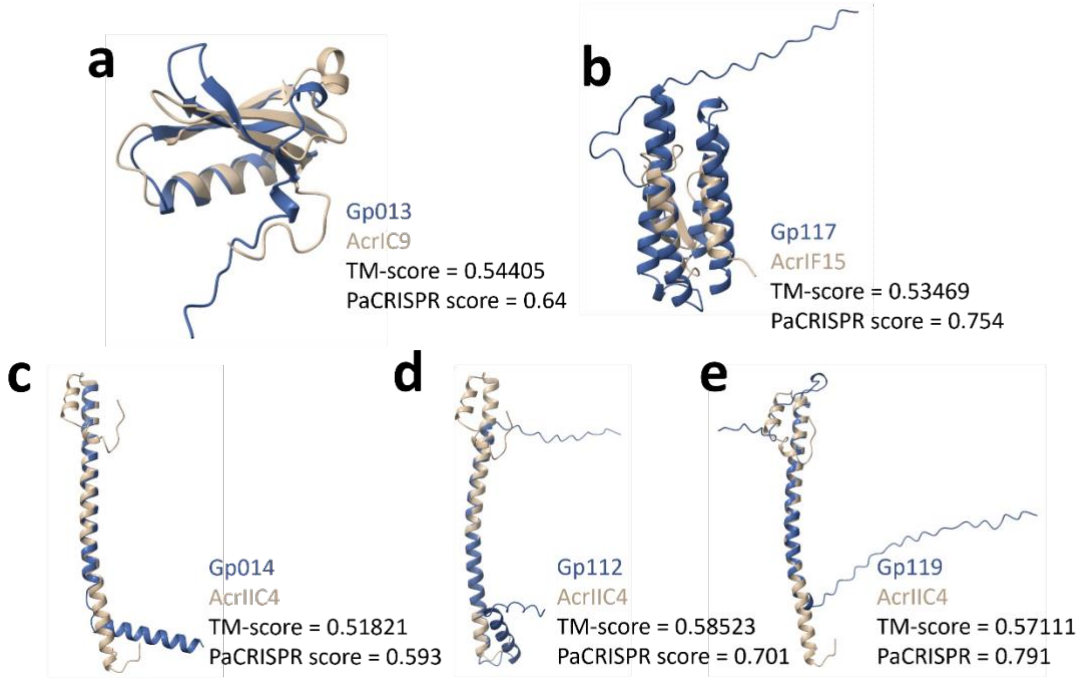



**Figure 4. Predicted Acrs in DAC15 with structural similarity to experimentally determined Acrs.** Five DAC15 proteins share the same global topology with previously published Acrs. **A)** gp013 shares structural similarities to AcrIC9. **B)** gp117 is structurally similar to AcrIF15. **C-E)** Three of these proteins (gp014, gp112, and gp119) are all structurally similar to AcrIIC4. TM-scores range from 0-1, with 1 indicating an identical structural match and ≥ 0.5 suggests shared structural features.

## Prediction of Acrs in the genomes of CrAss001, CrAss002 and p-crAssphage

We further extended our Acr analysis to the culturable *Crassvirales* CrAss001 and CrAss002, as well as the uncultured p-crAssphage. Four putative Acrs were identified in CrAss002, six in CrAss001, as well as six in p-crAssphage **(Fig 5)**. Five of these had structural similarity to AcrIE6, which targets the type I-E CRISPR Cas system[48], while a further seven had structural homology to AcrIIC4. The remaining structural similarities were to Acrs targeting type II-A systems **(Fig 5h and 5p)** and type I-C systems **(Fig 5k)**. Predictions of Acrs across multiple *Crassvirales* genomes suggests that this may be a common anti-defence system employed by these phages.

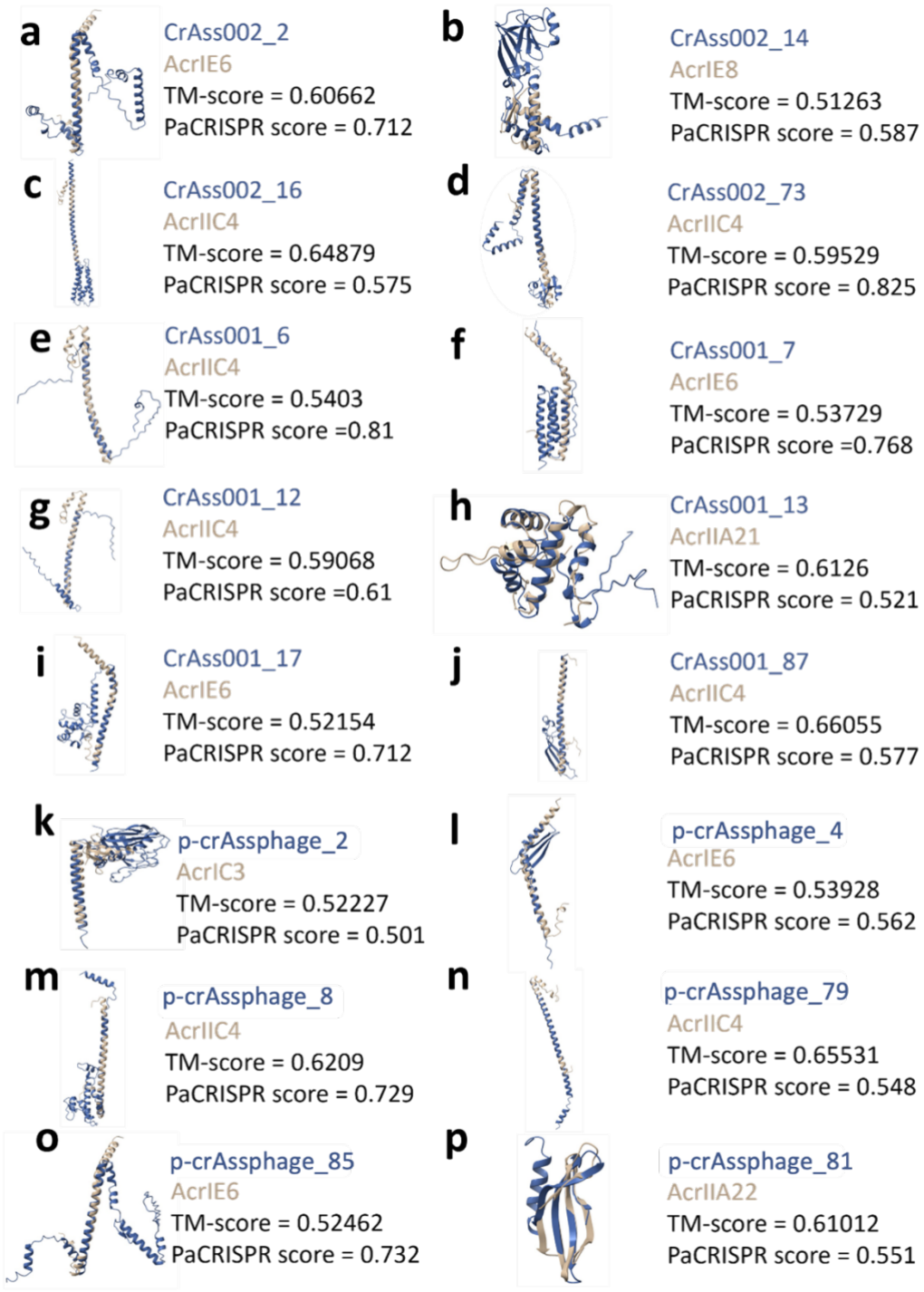



**Figure 5. Predicted Acrs in CrAss002, CrAss001 and p-crAssphage. A-J)** Four Acrs were predicted in in the CrAss002 genome while six were predicted in the CrAss001 genomes. **K-P)** Six Acrs were predicted in p-crAssphage genome. Most structural hits were to AcrIE6 and AcrIIC4, with a minority of hits to Acrs targeting type II-A and type I-C systems. TM-scores range from 0-1, with 1 indicating an identical structural match and ≥ 0.5 suggests shared structural features.

## Discussion

Here, we used CRISPRi technology to suppress the expression of a number of putative essential and non-essential genes in DAC15. Reductions in EOP and phage titre varied from 20-69% in the four tail fibre genes and across plate and liquid assay propagation. This is consistent with reductions in EOP noted for non-essential tail fibre genes in λ and P1[29]. A reduction in EOP and phage titre of 9-40% when targeting TerL gene coincided with an enrichment of empty phage capsids together with viable phage particles, visible under TEM. This partial reduction of infective particles counts and increased yields of empty phage capsids can be explained in two different ways: a) CRISPRi targeting TerL resulted in full inhibition only in some cells, leading to stalled phage assembly and accumulation of empty phage capsids; b) a limited reduction of TerL expression in every cell by CRISPRi led to the lower TerL availability, which could be the rate-limiting factor in the capsid filling process, and accumulation of empty capsids. This limited efficiency in targeting TerL and DNA polymerase was unexpected, given the > 99.99% reduction typically observed for essential genes in λ and P1[29]. There could be a number of explanations for this. The first is that positioning of the gRNA target may not have been fully optimised. Targets closest to the 5' end of the gene were chosen for this study as this has generally been observed to be successful in prokaryotes when carrying out interference with dCas9. We did extend targeting to 15% from the 5' end for one essential gene, DNA polymerase, but saw no improvement in repression. However, it is possible that when targeting phage genomes, optimal repression may occur even further from the 5' end of the gene and additional targets need to be tested to confirm this. A second possible explanation is the presence of a defence strategy encoded on the DAC15 genome. Phages can modify their genomes to escape CRISPR-Cas systems by a number of strategies, such as DNA glycosylation by T4 phages to avoid interference by type I-E and II-A CRISPR-Cas systems[56]. Certain phages can also compartmentalise their DNA to avoid CRISPR-Cas systems, such as Jumbo phage ΦKZ that infects *Pseudomonas aeruginosa,* which creates a

proteinaceous nucleus compartment around its DNA that protects it from DNA-cleaving enzymes[57]. However, the most commonly identified defence strategy to CRISPR appears to be Acrs, with over 100 Acrs experimentally verified, in particular those with activity against type I and type II-A and II-C systems[14].

It has previously been hypothesized that *Crassvirales* phages encode Acrs, due to the blocks of short genes on their genomes that have no homology outside of the *Crassvirales* order[41]. Identification of new Acrs can be challenging, due to their low sequence similarity. Bioinformatic approaches have typically relied on methods such as 'guilt by association' or self-targeting. Guilt by association involves searching for more highly conserved helix-turn-helix Acr associated proteins (Aca), that are typically encoded downstream from Acrs[58]. Self-targeting looks for prokaryotic genomes that contain CRISPR spacers that target its own genome, because in that scenario Acrs would also be required to prevent cell death[59]. A number of machine learning models have been created that have greatly increased the number of predicted Acrs[49,60,61], but the use of one such machine learning tool failed to identify any Acrs in *Crassvirales* phages[41].

Recently, Duan *et al* used structural similarities to identify the first Acr that is produced by a *Bacteroidota* phage and inhibits Cas12a[14]. We employed a similar approach and identified five potential Acrs in the genome of DAC15. Three of these had structural similarity to AcrIIC4, targeting the type II-C CRISPR system. While *B. theta* VPI-5482 does not appear to encode any CRISPR systems, closely related strains such as *B. theta* CL11T00C24 do possess a type II-C CRISPR system[62]. Phages often encode more than one Acr, which typically target the same CRISPR subtype, suggesting they have evolved to infect a narrow range of hosts[54]. Where phages carry genes for multiple CRISPR subtypes, it is thought that they may have a broader host range[54]. DAC15 also had two proteins with structural similarity to Acrs targeting type I-C and type I-F systems. While it has previously been postulated that *Crassvirales* phages may be narrow specialists[35], a recent preprint reports that p-crAssphage can replicate in three different bacterial species, across two different genera, suggesting that at least some *Crassvirales* may have a broad host range[63].

The Cas9 used in these CRISPRi experiments was the *Streptococcus pyogenes* Cas9 (SpyCas9) of the type II-A CRISPR system. While some type II Acrs are specific to only one Cas9, others can be more broad, having varying activity against Cas9 homologs spanning across both the type II-A and type II-C systems[64]. AcrIIC4 has varying activity against a number of type II-C Cas9 proteins but is not active against SpyCas9[64]. Nevertheless, we cannot rule out that an Acr with structural similarity to AcrIIC4 could also interact with SpyCas9.

We further extended our Acr analysis to the genomes of three other *Crassvirales* and identified several putative Acrs in each. Across the four phages, there were a number of structural homologs to AcrIE6 and AcrIIC4. Both these Acrs are characterised by the presence of a long alpha helix. For some of these putative Acrs, the primary structural match was to this alpha helix (for example gp17 from CrAss001 and AcrIE6) (**Fig 5i**), leading to a possibly artificially high TM-score. Acrs targeting type I-C, I-F and type II-A systems have more distinctive folds and so gp013 and gp117 from DAC15, gp013 from CrAss001 and gp02 and gp81 from p-crAssphage would be of interest for future experimental verification.

While there has generally been good success using CRISPRi to target chromosomal and plasmid genes in bacteria and eukaryotes, phage genomes appear to add another level of complexity that can sometimes lead to minimal repression. This highlights the importance of characterising phage genes of unknown function. It has also recently been reported that phages can possess small non-coding RNA based Acrs (Racrs) that are also capable of inhibiting CRISPR systems, adding another weapon to the arsenal protecting phages from inactivation[65]. Recently, Adler *et al* sought to circumvent phage DNA protection strategies by targeting translation, using RNA-guided RNA-targeting dRfxCas13d (CRISPRi-ART)[66]. CRISPRi-ART was able to successfully probe gene essentiality across four phage genomes, including T4, which has the aforementioned DNA defence strategy of glycosylation. This suggests that CRISPRi-ART may also be a promising tool to explore *Crassvirales* and other phages, where DNA targeting approaches were less effective.

## Materials and Methods

### Bacterial strains, phage and culture conditions

All CRISPRi experiments were carried out on a mutant of *Bacteroides thetaiotaomicron* VPI-5482 Δtdk (*B. theta*), designed to express one capsular polysaccharide only ($CPS3^+$)[67]. Cultures were grown in Fastidious Anaerobe Broth (FAB, Oxoid) anaerobically at 37°C. For conjugation of plasmids, *Escherichia coli* S17-1 λ*pir (E. coli*) was used, which was maintained in Luria Bertoni broth (LB, Difco) aerobically at 37°C. High titre stocks of DAC15 were prepared on *B. theta* $CPS3^+$, using the plate lysis method[68], on FAB plates and using 0.45% FAB top agar.

### Plasmid design and transformation

All plasmids used for targeting DAC15 genes were derivative of pMM710 (Addgene; plasmid #68893), which was designed with the genetic parts necessary to integrate into the chromosome of *B. theta*[45]. All gRNA sequences were designed to target the non-template strand and were synthesized and cloned into plasmids at Genewiz. The gRNA sequences used are listed in supplementary table 2. The nonsense sequence (NS) target, whose sequence did not target anything in either the *B. theta* $CPS3^+$ or DAC15 genomes, was used as a negative control. As a positive control, pMM704 (Addgene; plasmid #68891), targeting *B. theta* gene *BT1854* was used. All plasmids were introduced into *E. coli* cells via electroporation. One µL of plasmid DNA was added to a 40 µL aliquot of electrocompetent *E. coli* cells on ice and electroporated at 2.5 kV using an Eppendorf Eporator®. The cells were immediately resuspended in 1 mL of LB media and incubated aerobically for one hour, at 200 rpm and 37°C. The cells were plated in 200 µL and 20 µL aliquots on LB agar, supplemented with 100 µg/mL of ampicillin and incubated aerobically at 37°C overnight.

Conjugation

Plasmids were introduced into *B. theta* CPS3$^+$ cells via conjugation, using an adapted protocol from Goodman *et al*[69]. Briefly, *E. coli* donor cells and *B. theta* CPS3$^+$ recipient cells were both pelleted at 4,700 g for 15 minutes and resuspended in phosphate buffer saline solution (PBS). Donor and recipient cells were mixed in a ratio of 10:1 and centrifuged at 4,700 rpm for 15 minutes. The resulting supernatant was removed and the pellet was resuspended in the remaining medium. This suspension was plated in 100 µL aliquots on BHI agar, supplemented with 10% sheep blood. The plates were incubated overnight at 37°C aerobically. The following day, the 100 µL aliquots were scrapped off the plates and resuspended in PBS. Ten fold serial dilutions were prepared and plated on FAB agar, supplemented with 50 µg/mL kanamycin and 25 µg/mL erythromycin. The plates were incubated for 48 hours at 37°C anaerobically. Resulting colonies were confirmed to be *B. theta* cells and to contain the integrated plasmid by colony PCR.

CRISPRi experiments

For plate assays, 200 µL of the bacterial strain containing the integrated plasmid was added to 3.5 mL of 0.45% FAB top agar and IPTG to a final concentration of 500 µM. Ten fold dilutions of DAC15 stock were spotted in 10 µL volumes. After drying, the plates were inverted and incubated at 37°C anaerobically for 24 hours. Efficiency of plaquing (EOP) was calculated by dividing the plaque forming units per ml (pfu/mL) of DAC15 on the engineered strain by the pfu/mL on the strain containing the NS target and multiplying by 100 for percentage reduction.

For liquid assays, IPTG to a final concentration of 500 µM was added to the engineered strain and allowed to grow to an $OD_{600}$ of 0.2. DAC15 was added at a multiplicity of infection (MOI) of one. After three hours (roughly one round of phage infection) and 24 hours, a 1 mL aliquot was taken and centrifuged at 4,500 g for 15 minutes. The supernatant was filtered through a 0.45 µM polyethersulfone (PES) membrane filter and 10-fold dilutions were prepared. Ten microlitre volumes were spotted on *B. theta* CPS3$^+$ overlays and incubate at 37°C for 24 hours

anaerobically. Reduction in phage titre was calculated in the same manner as EOP for plate assays.

For targeting of *B. theta* gene *BT1854*, overnight cultures of *B. theta* with integrated pMM704 were subcultured to a starting $OD_{600}$ of 0.15 in a 96-well plate, with the addition of IPTG, at a final concentration of 500 µM. Two-fold serial dilutions of polymyxin B, ranging from 1024 µg/mL down to 32 µg/mL were added and incubated in a microplate reader anaerobically at 37°C for 24 hours. Changes in $OD_{595}$ were recorded every 30 minutes. Results were visualised using Graphpad Prism software.

Transmission Electron Microscopy

DAC15 was propagated on *B. theta* $CPS3^{+}$ targeting TerL and the NS target, as previously described for the 24-hour liquid assays. One litre of lysate was produced for each of the targets and concentrated by polyethylene glycol precipitation. PEG-8000 at 10% w/v and 0.5M NaCl were added to the lysate and incubated overnight at 4°C. The solution was centrifuged at 5,000*g* at 4°C for 40 minutes and the pellet resuspended in a minimum volume of SM buffer (100 mM NaCl, 8.5 mM $MgSO_4$ and 50 mM Tris-HCl, pH 7.5). Empty capsids were obtained by using a caesium chloride step gradient, using solutions of densities 1.25 g/mL, 1.37 g/mL and 1.54 g/mL in 20 mM Tris-HCL pH 7.5, 10 mM $MgSO_4$. Gradients were centrifuged at 40,000 rpm for 3 hours at 4°C. Two upper bands, containing empty capsids, debris and lipids and one lower band, containing viable phages, were obtained for both samples. All six bands were extracted using a sterile syringe. The samples were then dialysed against SM buffer overnight at 4°C. For negative staining, carbon-formvar coated copper grids (Agar Scientific) were plasma cleaned in a PELCO easiGlow for 60 s at 0.38 mbar (air) and 20 mA. Sample solution (5 µl) was applied to a grid for 1 min then blotted with filter paper and stained with 5 µl of 2% w/v uranyl acetate. Micrographs were recorded using a JEOL 2100 + microscope at 25k x magnification.

Acr prediction

All DAC15, CrAss002, CrAss001 and p-crAssphage protein sequences between 30-300 amino acids in length were passed through PaCRISPR[49] with a threshold of 0.5. The structures of all computationally predicted Acrs were predicted with AlphaFold3[70]. US-align[71] was used for pairwise structural alignment between predicted and currently published Acrs, where a TM-score of ≥ 0.5 means the proteins share the same global topology. Alphafold3 was used for published Acrs whose structure had not been experimentally determined. Superposed structures were visualised with ChimeraX[72].

## Author contributions

CAT, ANS and CH contributed to the conceptual development of this study. CAT performed CRISPRi experiments, Acr predictions and purified phages for TEM. OWB obtained TEM images. CAT analysed the data and wrote the manuscript with editorial contributions from ANS and CH.

Supplementary

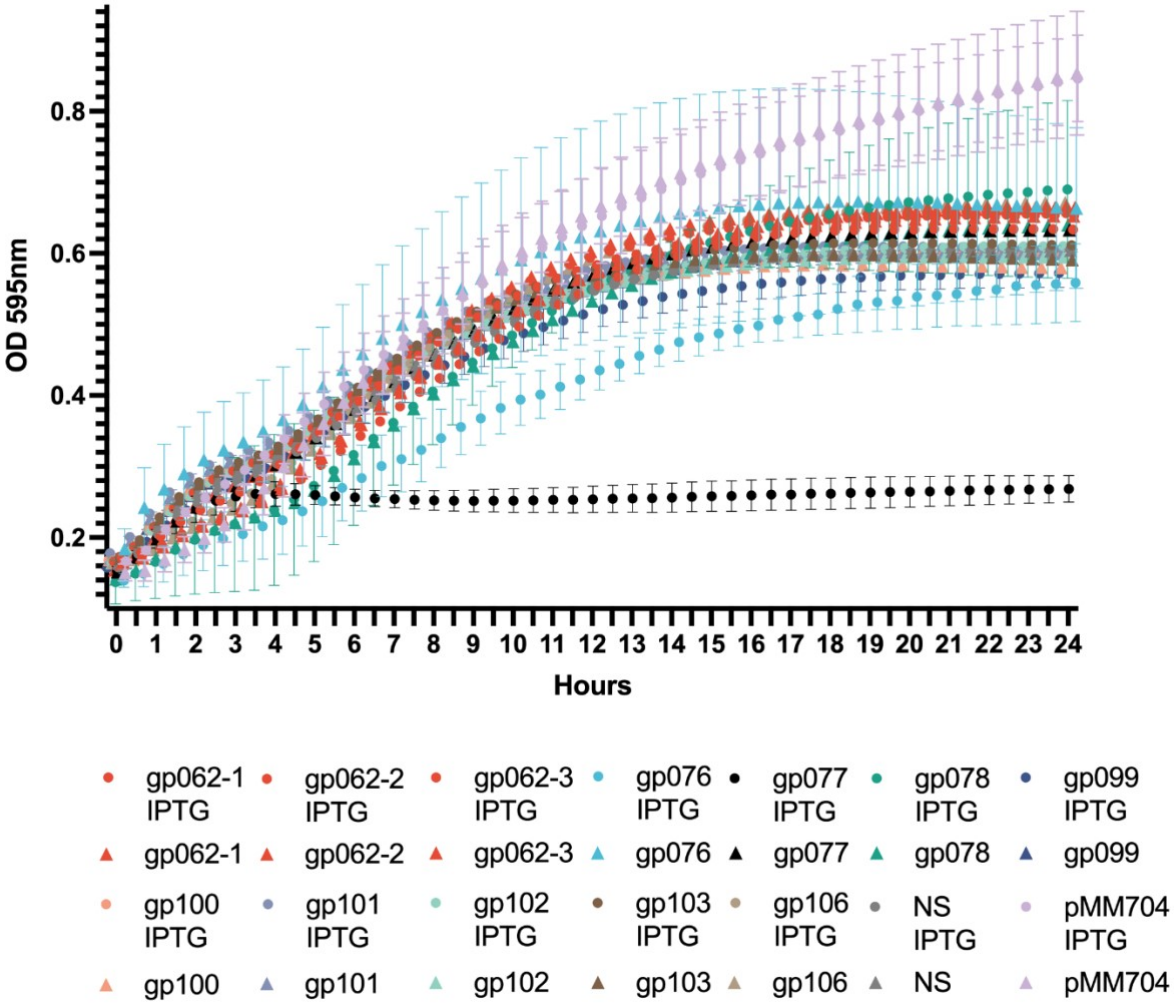



**Figure S1. Growth curves of *B. theta* strains with integrated pMM710 and pMM704 plasmids and gRNA target in the presence and absence of IPTG.** The growth of all engineered strains, except for *B. theta* containing gp077 target, are unaffected by the gRNA targeting DAC15. The gp077 target appears to have an off-target effect on the *B. theta* genome, halting growth. All error bars represent the standard deviation of three biological replicates, prepared on separate days.

| | | TM-score | | |
|---|---|---|---|---|
| **DAC15 gene product** | **PaCRISPR score** | **AcrIIC4** | **AcrIF15** | **AcrIC9** |
| gp054 | 0.811 | | | |
| gp056 | 0.792 | | | |
| gp119 | 0.791 | 0.57111 | | |
| gp117 | 0.754 | | 0.53469 | |
| gp086 | 0.744 | | | |
| gp065 | 0.704 | | | |
| gp112 | 0.701 | 0.58523 | | |
| gp098 | 0.69 | 0.53583 | | |
| gp107 | 0.69 | | | |
| gp068 | 0.668 | | | |
| gp013 | 0.649 | | | 0.54405 |
| gp075 | 0.64 | | | |
| gp026 | 0.629 | | | |
| gp094 | 0.618 | | | |
| gp027 | 0.612 | | | |
| gp072 | 0.599 | | | |
| gp052 | 0.597 | | | |
| gp014 | 0.593 | 0.51821 | | |
| gp110 | 0.584 | | | |
| gp083 | 0.56 | | | |
| gp005/gp007 | 0.557 | | | |
| gp115 | 0.536 | | | |
| gp011/gp001 | 0.533 | | | |
| gp063 | 0.527 | | | |
| gp096 | 0.522 | | | |
| gp095 | 0.52 | 0.58811 | | |
| gp019 | 0.515 | | | |
| gp020 | 0.508 | | | |

**Table S1. List of 30 DAC15 genes that scored positively as an Acr, ranked by PaCRISPR.** A protein was considered a putative Acr when a PaCRISPR score was ≥ 0.5. Seven putative Acrs had similar

structural topology to experimentally confirmed Acrs, AcrIIC4, AcrIF15 and AcrIC9. Similar structural topology was considered when a TM-score ≥ 0.5 was achieved.

| Name | Target gene | Predicted function | gRNA sequence |
|---|---|---|---|
| gp062-1 | gp062 | DNA polymerase | TGTTCCCCATAATATACTAT |
| gp062-2 | gp062 | DNA polymerase | TCAACTGAGATACATTTGTA |
| gp062-3 | gp062 | DNA polymerase | TAAAGAGCCTGTCACTTTCA |
| gp076 | gp076 | RNA polymerase | ACATCATCCAGACCTTTCTC |
| gp078 | gp078 | RNA polymerase | TCCCAATTAGCTCTATCTTC |
| gp099 | gp099 | Tail fibre | TCCTGTACTGTGCTTCTACT |
| gp100 | gp100 | Tail fibre | AGCTCCTAAGTCTATATCCT |
| gp101 | gp101 | Tail fibre | ATTGTCTCATCACCATTGAG |
| gp102 | gp102 | Tail fibre | ATCCATAACTTGTCAGTGCT |
| gp103 | gp103 | Head fibre | CACAAGTCCATTGTGGATGC |
| gp106 | gp106 | Terminase large subunit | TCTAATAGTTCATCTGTAAT |

**Table S2.** List of gRNA sequences used in this study.

## References


1. Nayfach, S. et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. Nat Microbiol 6, 960–970 (2021).
2. Graham, E. B. *et al.* A global atlas of soil viruses reveals unexplored biodiversity and potential biogeochemical impacts. *Nat Microbiol* **9**, 1873–1883 (2024).
3. Warwick-Dugdale, J. *et al.* Long-read powered viral metagenomics in the oligotrophic Sargasso Sea. *Nat Commun* **15**, (2024).
4. Mushegian, A. R. Are there 1031virus particles on earth, or more, or fewer? *J Bacteriol* **202**, (2020).
5. Georjon, H. & Bernheim, A. The highly diverse antiphage defence systems of bacteria. *Nat Rev Microbiol* **21**, 686–700 (2023).
6. Rath, D., Amlinger, L., Rath, A. & Lundgren, M. The CRISPR-Cas immune system: Biology, mechanisms and applications. *Biochimie* **117**, 119–128 (2015).

7. Kelly, A., Arrowsmith, T. J., Went, S. C. & Blower, T. R. Toxin–antitoxin systems as mediators of phage defence and the implications for abortive infection. *Curr Opin Microbiol* **73**, 102293 (2023).

8. Lopatina, A., Tal, N. & Sorek, R. Annual Review of Virology Abortive Infection: Bacterial Suicide as an Antiviral Immune Strategy. (2020) doi:10.1146/annurev-virology-011620.

9. Dupuis, M. È., Villion, M., Magadán, A. H. & Moineau, S. CRISPR-Cas and restriction-modification systems are compatible and increase phage resistance. *Nat Commun* **4**, (2013).

10. Garneau, J. E. *et al.* The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67–71 (2010).

11. Makarova, K. S. *et al.* Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat Rev Microbiol* **18**, 67–83 (2020).

12. Murtazalieva, K., Mu, A., Petrovskaya, A. & Finn, R. D. The growing repertoire of phage anti-defence systems. *Trends Microbiol* (2024) doi:10.1016/j.tim.2024.05.005.

13. Bondy-Denomy, J., Pawluk, A., Maxwell, K. L. & Davidson, A. R. Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. *Nature* **493**, (2013).

14. Duan, N., Hand, E., Pheko, M., Sharma, S. & Emiola, A. Structure-guided discovery of anti-CRISPR and anti-phage defense proteins. *Nat Commun* **15**, (2024).

15. Jia, N. & Patel, D. J. Structure-based functional mechanisms and biotechnology applications of anti-CRISPR proteins. *Nat Rev Mol Cell Biol* **22**, 563–579 (2021).

16. Hatfull, G. F. Actinobacteriophages: Genomics, Dynamics, and Applications. *Annu Rev Virol* **7**, 37–61 (2020).

17. Miller, E. S. *et al.* Bacteriophage T4 Genome. *Microbiology and Molecular Biology Reviews* **67**, 86–156 (2003).

18. Casjens, S. R. & Hendrix, R. W. Bacteriophage lambda: Early pioneer and still relevant. *Virology* **479–480**, 310–330 (2015).

19. Łobocka, M. B. *et al.* Genome of bacteriophage P1. *J Bacteriol* **186**, 7032–7068 (2004).

20. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-γuided platform for sequence-specific control of gene expression. *Cell* **152**, 1173–1183 (2013).

21. Zhang, X. *et al.* Multiplex gene regulation by CRISPR-ddCpf1. *Cell Discov* **3**, (2017).

22. Peters, J. M. *et al.* A comprehensive, CRISPR-based functional analysis of essential genes in bacteria. *Cell* **165**, 1493–1506 (2016).

23. Lee, H. H. *et al.* Functional genomics of the rapidly replicating bacterium Vibrio natriegens by CRISPRi. *Nat Microbiol* **4**, 1105–1113 (2019).

24. Liu, X. *et al.* High-throughput CRISPRi phenotyping identifies new essential genes in Streptococcus pneumoniae . *Mol Syst Biol* **13**, (2017).

25. Gilbert, L. A. *et al.* XCRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442 (2013).

26. Lawhorn, I. E. B., Ferreira, J. P. & Wang, C. L. Evaluation of sgRNA target sites for CRISPR-mediated repression of TP53. *PLoS One* **9**, (2014).

27. Prezza, G., Liao, C., Reichardt, S., Beisel, C. L. & Westermann, A. J. CRISPR-based screening of small RNA modulators of bile susceptibility in Bacteroides thetaiotaomicron. *Proc Natl Acad Sci U S A* **121**, (2024).

28. Cornuault, J. K. & Moineau, S. Induction and Elimination of Prophages Using CRISPR Interference. *CRISPR Journal* **4**, 549–557 (2021).

29. Piya, D. *et al.* Systematic and scalable genome-wide essentiality mapping to identify nonessential genes in phages. *PLoS Biol* **21**, (2023).

30. Dutilh, B. E. *et al.* A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat Commun* **5**, (2014).

31. Yutin, N. *et al.* Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nat Microbiol* **3**, 38–46 (2018).

32. Edwards, R. A. *et al.* Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. *Nat Microbiol* **4**, (2019).

33. Honap, T. P. *et al.* Biogeographic study of human gut-associated crAssphage suggests impacts from industrialization and recent expansion. *PLoS One* **15**, (2020).

34. Holmfeldt, K. *et al.* Twelve previously unknown phage genera are ubiquitous in global oceans. *Proc Natl Acad Sci U S A* **110**, 12798–12803 (2013).

35. Shkoporov, A. N. *et al.* ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. *Nat Commun* **9**, (2018).

36. Guerin, E. *et al.* Isolation and characterisation of ΦcrAss002, a crAss-like phage from the human gut that infects Bacteroides xylanisolvens. *Microbiome* **9**, (2021).

37. Porter, N. T. *et al.* Phase-variable capsular polysaccharides and lipoproteins modify bacteriophage susceptibility in Bacteroides thetaiotaomicron. *Nat Microbiol* **5**, 1170–1181 (2020).

38. Papudeshi, B. *et al.* Host interactions of novel Crassvirales species belonging to multiple families infecting bacterial host, Bacteroides cellulosilyticus WH2. *Microb Genom* **9**, (2023).

39. Ramos-Barbero, M. D. *et al.* Characterization of crAss-like phage isolates highlights Crassvirales genetic heterogeneity and worldwide distribution. *Nat Commun* **14**, (2023).

40. Shkoporov, A. N. *et al.* Long-term persistence of crAss-like phage crAss001 is associated with phase variation in Bacteroides intestinalis. *BMC Biol* **19**, (2021).

41. Yutin, N. *et al.* Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features. *Nat Commun* **12**, (2021).

42. Peters, S. L. *et al.* Experimental validation that human microbiome phages use alternative genetic coding. *Nat Commun* **13**, (2022).

43. Drobysheva, A. V. *et al.* Structure and function of virion RNA polymerase of a crAss-like phage. *Nature* **589**, 306–309 (2021).

44. Bayfield, O. W. *et al.* Structural atlas of a human gut crassvirus. *Nature* **617**, (2023).

45. Mimee, M., Tucker, A. C., Voigt, C. A. & Lu, T. K. Programming a Human Commensal Bacterium, Bacteroides thetaiotaomicron, to Sense and Respond to Stimuli in the Murine Gut Microbiota. *Cell Syst* **1**, (2015).

46. Hawkins, J. S., Wong, S., Peters, J. M., Almeida, R. & Qi, L. S. Targeted transcriptional repression in bacteria using CRISPR interference (CRISPRi). *Methods in Molecular Biology* **1311**, 349–362 (2015).

47. Wang, T. *et al.* Pooled CRISPR interference screening enables genome-scale functional genomics study in bacteria with superior performance-net. *Nat Commun* **9**, (2018).

48. Cullen, T. W. *et al.* Antimicrobial peptide resistance mediates resilience of prominent gut commensals during inflammation. *Science (1979)* **347**, 170–175 (2015).

49. Wang, J. *et al.* PaCRISPR: A server for predicting and visualizing anti-CRISPR proteins. *Nucleic Acids Res* **48**, W348–W357 (2020).

50. Kang, Y. J. *et al.* The structure of AcrIC9 revealing the putative inhibitory mechanism of AcrIC9 against the type IC CRISPR-Cas system. *IUCrJ* **10**, 624–634 (2023).

51. Pinilla-Redondo, R. *et al.* Discovery of multiple anti-CRISPRs highlights anti-defense gene clustering in mobile genetic elements. *Nat Commun* **11**, (2020).

52. Lee, J. *et al.* Potent Cas9 Inhibition in Bacterial and Human Cells by AcrIIC4 and AcrIIC5 Anti-CRISPR Proteins. *mBio* **9**, (2018).

53. Sun, W. *et al.* AcrIIC4 inhibits type II-C Cas9 by preventing R-loop formation. *Proc Natl Acad Sci U S A* **120**, (2023).

54. Pons, B. J., van Houte, S., Westra, E. R. & Chevallereau, A. Ecology and evolution of phages encoding anti-CRISPR proteins. *J Mol Biol* **435**, 167974 (2023).

55. Marino, N. D. *et al.* Discovery of widespread type I and type V CRISPR-Cas inhibitors. *Science (1979)* **362**, 240–242 (2018).

56. Vlot, M. *et al.* Bacteriophage DNA glucosylation impairs target DNA binding by type I and II but not by type V CRISPR–Cas effector complexes. *Nucleic Acids Res* **46**, 873–885 (2018).

57. Mendoza, S. D. *et al.* A bacteriophage nucleus-like compartment shields DNA from CRISPR nucleases. *Nature* **577**, 244–248 (2020).

58. Borges, A. L., Davidson, A. R. & Bondy-Denomy, J. The Discovery, Mechanisms, and Evolutionary Impact of Anti-CRISPRs. *Annu Rev Virol* **4**, 37–59 (2017).

59. Rauch, B. J. *et al.* Inhibition of CRISPR-Cas9 with Bacteriophage Proteins. *Cell* **168**, 150158.e10 (2017).

60. Gussow, A. B. *et al.* Machine-learning approach expands the repertoire of anti-CRISPR protein families. *Nat Commun* **11**, (2020).

61. Li, Y. *et al.* AcrNET: predicting anti-CRISPR with deep learning. *Bioinformatics* **39**, (2023).

62. Pourcel, C. *et al.* CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays and cas genes from complete genome sequences, and tools to download and query lists of repeats and spacers. *Nucleic Acids Res* **48**, D535–D544 (2020).

63. Schmidtke, D. T., Hickey, A. S., Liachko, I., Sherlock, G. & Bhatt, A. S. Analysis and culturing of the prototypic crAssphage reveals a phage-plasmid lifestyle. Preprint at https://doi.org/10.1101/2024.03.20.585998 (2024).

64. Garcia, B. *et al.* Anti-CRISPR AcrIIA5 Potently Inhibits All Cas9 Homologs Used for Genome Editing. *Cell Rep* **29**, 1739-1746.e5 (2019).

65. Camara-Wilpert, S. *et al.* Bacteriophages suppress CRISPR–Cas immunity using RNA-based anti-CRISPRs. *Nature* **623**, 601–607 (2023).

66. Adler, B. A. *et al.* CRISPRi-ART enables functional genomics of diverse bacteriophages using RNA-binding dCas13d. *Nat Microbiol* 10, 694–709 (2025).

67. Hickey, C. A. *et al.* Colitogenic Bacteroides thetaiotaomicron antigens access host immune cells in a sulfatase-dependent manner via outer membrane vesicles. *Cell Host Microbe* **17**, 672–680 (2015).

68. Bonilla, N. *et al.* Phage on tap-a quick and efficient protocol for the preparation of bacteriophage laboratory stocks. *PeerJ* **2016**, (2016).

69. Goodman, A. L. *et al.* Identifying Genetic Determinants Needed to Establish a Human Gut Symbiont in Its Habitat. *Cell Host Microbe* **6**, (2009).

70. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

71. Zhang, C., Shine, M., Pyle, A. M. & Zhang, Y. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nat Methods* **19**, 1109–1115 (2022).

72. Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Science* **30**, 70–82 (2021).

# Chapter 5: *Crassvirales* phage CrAss001 has three types of tailspikes with capsule depolymerase and esterase activity

Ciara A Tobin, Oliver W Bayfield, Christian Cambillau, Bianca Govi, Alfred A Antson, Colin Hill and Andrey N Shkoporov

## Abstract

*Crassvirales* phages are a highly abundant order that are capable of persisting with their *Bacteroidetes* hosts in the human gut. A recent cryogenic electron microscopy (cryo-EM) reconstruction of *Crassvirales* phage *Kehishuvirus primarius* (CrAss001) allowed atomic models to be built for the majority of proteins involved in virion structure, assembly and infection. While the tailspikes (TSP)/tail fibres (TF) were thought to be composed of five proteins, these structures remained unmodelled, due to their flexibility. Here, we present structural and biochemical data demonstrating interactions between these five proteins, with complex formation observed between gp22_23, gp24_25 and gp24_26. The structure of the gp22_23 complex was determined by cryo-EM to a resolution of 3.4 Å, which was shown to contain central β-helix domains typical of TSP's. Gp22_23 TSP was demonstrated to exhibit capsule depolymerase activity on sensitive cultures of *Bacteroides intestinalis* that predominately express capsular polysaccharide (CPS) PVR9. The structure of gp25 trimeric assembly was also determined, to a resolution of 2.6 Å, and exhibited esterase activity likely *via* its SGNH domain. AlphaFold predictions of gp24 and gp26 suggest that gp24 is a long, fibrous protein, while gp26 possesses a virion-associated lysin (VAL) domain that may locally degrade peptidoglycan during phage infection. Based on analysis of the reconstruction of the CrAss001 virion, gp22_23 TSP makes up the inner TSP assembly, while gp25 and gp26 TSP's are likely to make up an outer assembly, with gp24 acting as a linker to the tail barrel for both. This work provides biochemical, structural and mechanistic insights into receptor binding proteins (RBP's) of a *Crassvirales* phage and adds to our understanding of how these phages are successfully persisting alongside their bacterial hosts.

## Introduction

Bacteriophages (phages) are the most abundant entities on earth. Through interactions with their bacterial hosts, they play an important role in shaping various bacterial rich environments, such as soil[1], ocean[2] and the human gut[3]. This is mainly through their natural predation on bacteria, and they can also contribute to horizontal gene transfer via transduction[4]. Within the human gut, phages in the recently proposed families "Quimbyviridae"[5], "Flandersviridae"[5], the order *Crassvirales*[6,7] and the viral clade Gubaphage[8] are all highly abundant. *Crassvirales* are estimated to account for over 86% of gut virome sequences[9] and with no significant correlations to human health or disease factors[10,11], they are thought to be a benign component of the gut microbiome. *Carjivirus communis* (prototypical or p-crAssphage) was the first member of this order to be identified through cross-assembly from multiple metagenomes in 2014[12]. Since then, 42 *Crassvirales* have been grown in the laboratory, three belonging to the family *Intestiviridae*[13,14] and 39 from the family *Steigviridae*[15,16,17], while over 700 have been identified in metagenomic sequences.

Receptor binding proteins (RBP's) are critical for phage infection, as they mediate the initial interaction through specific binding to a receptor on the bacterial host. Phage tail fibres (TF), tail spikes (TSP) and tail tips are typically used as RBP's and they can bind to bacterial cell surface exposed molecules such as lipopolysaccharide[18] (LPS), capsular polysaccharide[19] (CPS), teichoic acids[20] and various proteins[21]. TF's are generally long, fibrous proteins that lack enzymatic activity, while TSP's often adopt more compact structures and exhibit enzymatic activity for degrading the bacterial cell wall and facilitate infection[22]. Bacteriophage particles (virions) commonly have six or twelve tail appendages[21], due to the 12-fold or six-fold symmetry of the attachment site on the tail. Bacteriophages may encode one type of TF or TSP[23,24], but can also encode two or more types of TF's or TSP's[25,26].

The recent cryo-EM reconstruction of *Kehishuvirus primarius* (CrAss001) has resolved both the structure and function of the majority of the viral proteins,

serving as a structural model for the *Crassvirales* order[27]. This includes the storage of cargo proteins within both the capsid and the tail, and the identification of a novel crass fold within the muzzle protein at the end of tail, which likely acts as a gatekeeper to the cargo proteins and phage DNA[27]. Five putative TF/TSP genes—*gp022* to *gp026*—were identified in the genome[17]. Inner and outer TF/TSP assemblies were observed in the virion reconstruction, at a low resolution, while the N-terminal residues (1-26) of TF/TSP gp22 was modelled interacting with tail hub protein B, docking it to the tail barrel. Due to their flexibility relative to the main capsid and tail, the remaining TF's/TSP's were unmodelled[27]. In this study, a combination of structural, biochemical and bioinformatic techniques were employed, in order to determine the structure and function of CrAss001 TF/TSP's. The inner TSP assembly was determined to be composed of two proteins, gp22 and gp23. A complex of gp22 and gp23 was prepared by co-expression, and the structure determined by cryo-EM. The gp22_23 TSP was also demonstrated to possess capsule depolymerase activity against a *B. intestinalis* strain expressing CPS PVR9. A cryo-EM reconstruction was also determined for gp25, and this protein was shown to exhibit esterase activity. AlphaFold predictions of gp24 indicate that it is a linker protein, aiding assembly of gp25 and gp26, while gp26 appears to be a third TSP with a virion-associated lysin (VAL) domain. The outer TSP is likely composed of gp25 and gp26, with gp24 docking them both to the tail barrel. The identification of three types of TSP's in a podovirus is unusual, which may be another mechanism that allows CrAss001 to persist with its bacterial host in the gut.

## Results

### Three TF/TSP complexes identified by AlphaFold predictions and pull-down assays

In order to investigate which TF's/TSP's interact to form both the inner and outer TF/TSP assemblies of CrAss001, we co-expressed five predicted TF/TSP genes (*gp022-gp026*) in consecutive and non-consecutive pairs. The leading TF/TSP gene was $His_6$-tagged, while the second TF/TSP gene was untagged. Purification by Ni-

affinity chromatography, followed by size-exclusion chromatography (SEC) resulted in the co-elution of gp22 and gp23 (**Fig 1a**), gp24 and gp25 (**Fig 1b**) and gp25 and gp26 (**Fig 1c**). Purification of the remaining combinations of TF/TSP proteins resulted in elution of the $His_6$-tagged protein only (**Fig 1d**). AlphaFold predictions are consistent with these interactions, and indicate that each protein complex is a hetero-hexamer composed of two homo-trimers (**Fig 1a-c**)

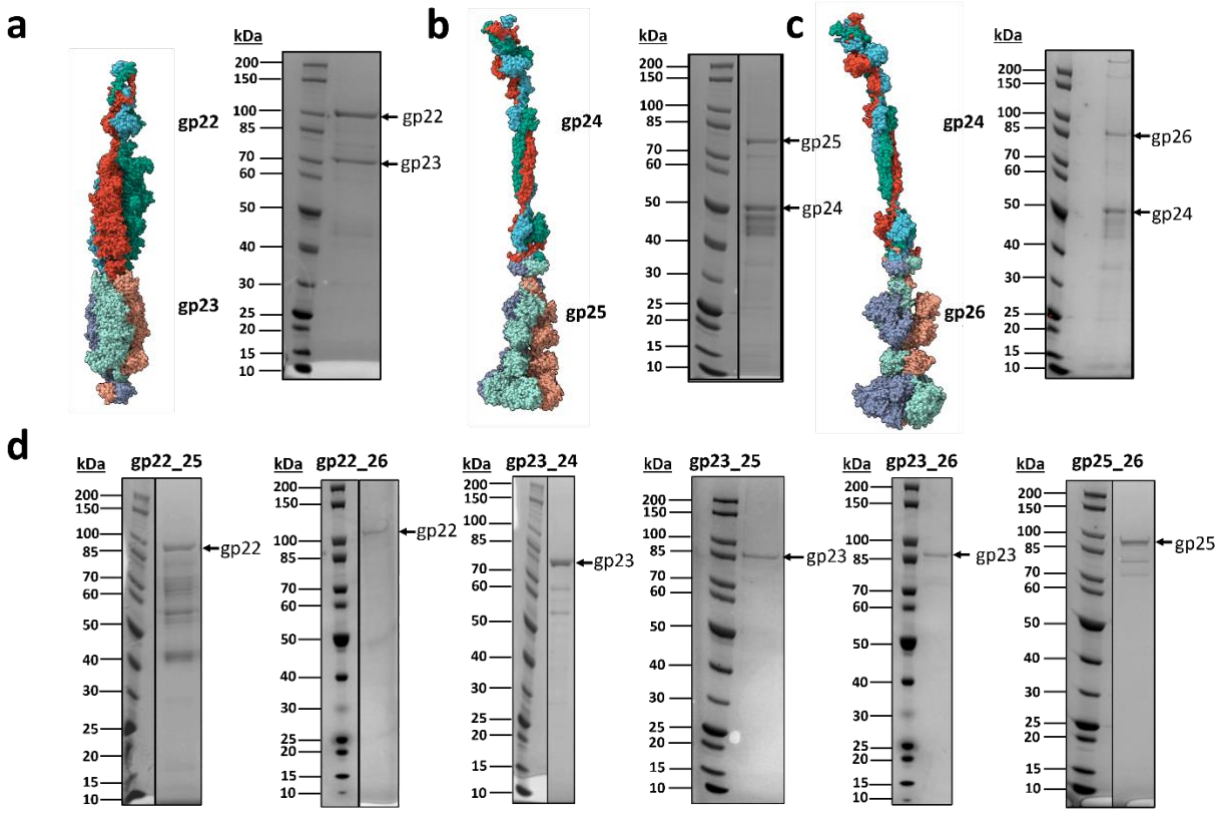



**Figure 1. AlphaFold predictions and pull-down assays identifies three TF/TSP complexes. A)** An AlphaFold model and an SDS-PAGE gel showing that a homo-trimer of 22 and a homo-trimer of 23 interact to form a 22-23 TF/TSP complex. **B)** An AlphaFold model and an SDS-PAGE gel showing that a homo-trimer of 24 and a homo-trimer of 25 interact to form a 24-25 TF/TSP complex. **C)** An AlphaFold model and an SDS-PAGE gel showing that a homo-trimer of 24 and a homo-trimer of 26 interact to form a 25-26 TF/TSP complex. **D)** SDS-PAGE gels showing that the remaining combinations of TF/TSP proteins results in the elution of the $His_6$-tagged protein only. Each purification was by Ni-affinity chromatography, followed by SEC for further purity. SDS-PAGE gels shown are from SEC and are representative of two independent experiments.

Structure and depolymerase activity of gp22_23 TSP

We attempted to obtain negatively stained micrographs of each of the three complexes but only the gp22_23 complex could be observed. A high level of background was observed for gp24_25 and gp24_26 complexes, likely due to issues with degradation when purifying gp24 and gp26. The gp22_23 complex was imaged by cryo-EM and a reconstruction determined by imposing $C_3$ symmetry to a resolution of 3.4 Å. The gp22_23 trimeric complex possesses an elongated structure approximately 31 nm in length and is charactered by extensive β-helix domains in both gp22 and gp23 (**Fig 2a-b**). The first 137 residues of gp22 were not resolved in the reconstruction, likely due to the flexibility of this region, and was not modelled. Each monomer of gp22 can be divided into three domains (**Fig 2c**): the N-terminal domain (residues 138-295), composed of two β-hairpins and three β-sheets; the β-helix domain (residues 296-769), consisting of a 17-rung, righthanded, parallel β-helix fold, interspersed with loops, with α-helices at both ends; and the C-terminal domain (residues 770-832), composed of a partial α-helix and an anti-parallel β-sheet. The C-terminal domain of gp22 is terminated by a single β-strand, which forms a β-sheet with the N-terminal domain of gp23. Each monomer of gp23 also consists of three domains. The N-terminal domain (residues 2-97) consists of two parallel β-sheets and a partial α-helix. The β-helix domain of gp23 (residues 98-446) is shorter than the β-helix of gp22, containing 11 rungs. It is also interspersed with longer loops than gp22 and has no terminal α-helix. The C-terminal domain (residues 447-688) is primarily composed of a β-sandwich, interspersed with loops and partial α-helices. The N-terminal and β-helix domains of gp22 and gp23 are structurally similar, with a DALI Z-score of 13.4 and r.m.s.d of 2.9 Å (N-terminal domain) and DALI Z-score of 22.5 and r.m.s.d of 2.3 (β-helix domain). The three domains observed in gp22 and gp23 are a conserved feature of TSP's[21,28] and DALI searches revealed structural similarities to TSP's from several different phages, including capsule depolymerase DpK2 from *Klebsiella* phage RAD2[29] (**Fig S1**). Additionally, structural similarities of the β-helical domains to glycoside hydrolases (5zru-A; *Z*-value 26.2 and r.m.s.d 2.5 Å (gp22), *Z*-value 30.1 and r.m.s.d 2.7 Å (gp23);

7jwf-A; *Z*-value 24 and r.m.s.d 2.9 Å (gp22), *Z*-value 30.1 and r.m.s.d 3 Å (gp23)), is indicative of enzymatic activity. Overall, the gp22_23 complex has structural similarity to a number of phage TSPs, with central β-helix domains that suggest enzymatic activity.

*Bacteroides intestinalis* (the host of CrAss001) has five phase variable CPS biosynthetic loci, termed PVR7, PVR8, PVR9, PVR11 and PVR12, with PVR9 potentially acting as a phage receptor[30,31]. To test for depolymerase activity, each of the five individual purified proteins, as well as the gp22_23 complex, were spotted on three clones of *B. intestinalis*: a *B. intestinalis* wild type (WT), that predominantly expresses PVR9; *B. intestinalis* 8T, a clone resistant to phage infection that primarily expresses PVR7; and *B. intestinalis* 8W, a sensitive revertant sub-clone of 8T, that again predominantly expresses PVR9[30]. Gp22_23 TSP generated hazy zones of clearing with a minimum concentration of 12.5 ng on lawns of WT and 8W but not 8T, confirming that gp22_23 TSP degrades PVR9 (**Fig 2d**). Gp22 alone could also produce these hazy zones on WT and 8W, but its activity was less potent, requiring a minimum concentration of 50 ng to produce hazy zones. Curiously, gp23 alone does not appear to be enzymatically active, as no zones of clearing were observed for this protein. No depolymerase activity was observed with gp24, gp25 or gp26 (**Fig S2**).

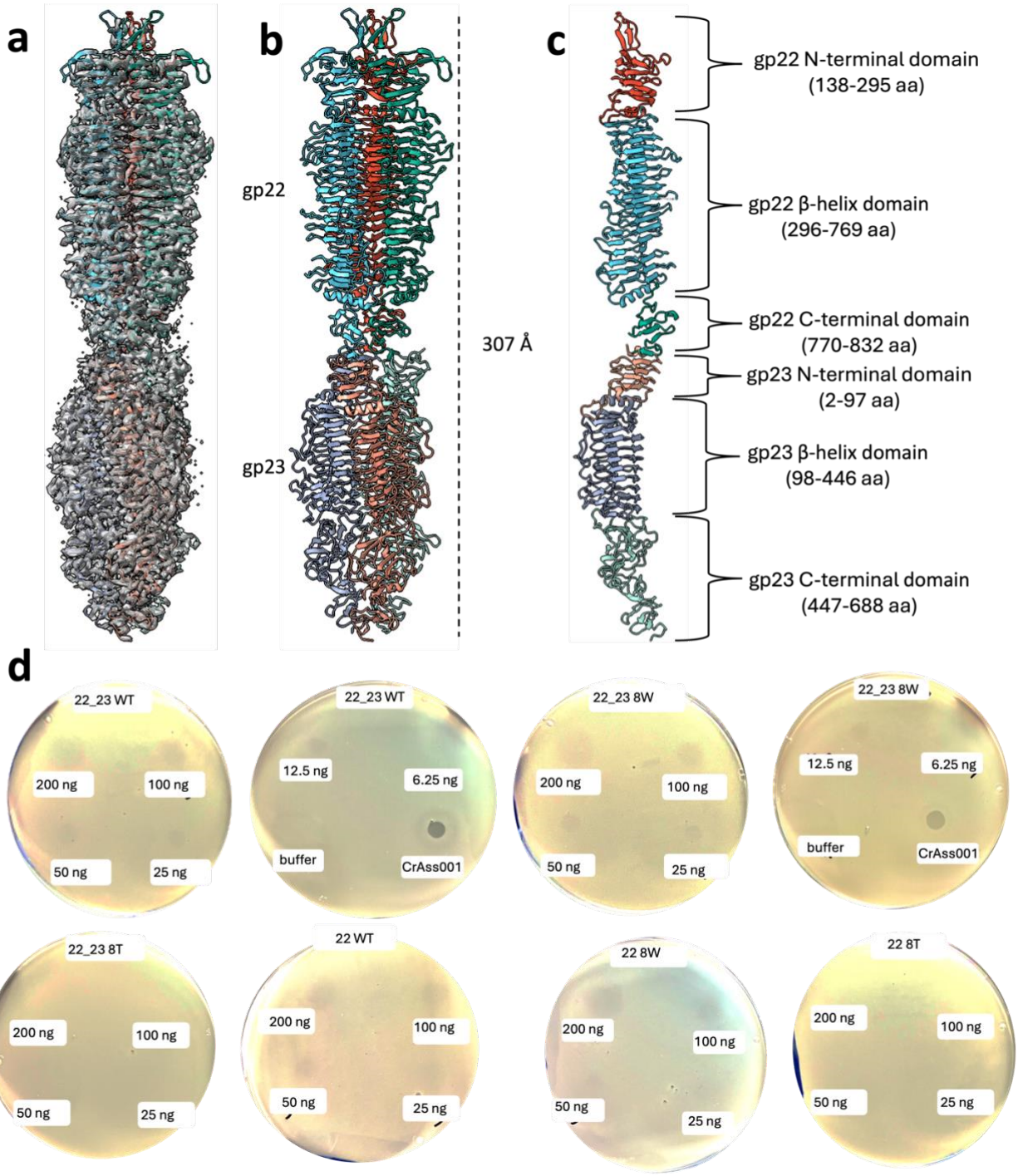



**Figure 2. Structure and depolymerase activity of gp22_23. A)** Atomic model of gp22_23 complex fit to the cryo-EM electron density map, resolved to 3.4 Å. **B)** High resolution structure of gp22_23 hetero-hexamer. Each individual monomer of gp22 is coloured blue, red and green and each monomer of gp23 is coloured purple, orange and light green. **C)** Structural domains of gp22 and gp23 monomers: gp22 N-terminal domain (residues 138-295, red), gp22 β-helix domain (residues 296-769, blue), gp22 C-terminal domain (residues 770-832, green), gp23 N-terminal domain

(residues 2-97, orange), gp23 β-helix domain (residues 98-446, purple) and gp23 C-terminal domain (residues 447-688, light green). **D)** Spot tests of gp22_23 TSP and gp22 alone on *B. intestinalis* WT, 8W and 8T. Hazy zones of clearing are observed down to 12.5 ng of gp22_23 TSP on WT and 8W, while no depolymerase activity is observed when spot on 8T. Gp22 alone is less active, producing hazy zones at dilutions down to 50 ng on WT and 8W. Each plate image has been edited to help with visibility of hazy zones. Each plate image is a representative of three biological replicates.

## Structure and esterase activity of gp25 TSP

Gp25 was reconstructed imposing $C_3$ symmetry, followed by symmetry expansion and masked classification ($C_1$ symmetry) to align the N-terminal domain which does not align with the $C_3$ axis. The resulting resolution was 2.6 Å (**Fig 3a**). Gp25 is approximately 16 nm in length, with a narrow N-terminal domain that gradually broadens towards the C-terminus (**Fig 3b**). Each monomer of gp25 can be divided into five globular domains plus an N-terminal domain (**Fig 3c**). The N-terminal domain (residues 2-78) is composed of a partial α-helix and an anti-parallel β-sheet, which connects to the following domain through an extended loop and α-helix. It is structurally similar to the N-terminal connector domain of HylP1, a tail fibre of a *Streptococcus pyogenes* prophage[32] (DALI *Z*-score 7.4, r.m.s.d 4.2 Å, 2c3f-A). The β-hairpin domain (residues 79-167) is primarily composed of two β-hairpins, interspersed with loops. A DALI search of this domain did not identify any significantly similar structures. The β-sandwich domain 1 (residues 168-286) has structural similarity to the C-terminal domain of TSP K2-2, from *Klebsiella* phage VLC6[19] (DALI *Z*-score 6.3, r.m.s.d 2.7 Å, 8iqe-C). The top DALI hit for the β-sandwich domain 2 (287-409 residues) was to the N-terminal baseplate binding domain of DpK2 from *Klebsiella* phage RAD2[29] (*Z*-score 6.6, r.m.s.d 3.2 Å, 7lzj-A). The CBM66-like domain (residues 410-597) also has β-sandwich topology and contains six strands in one β-sheet and seven strands on the other. It is structurally similar to a noncatalytic carbohydrate binding module (CBM), *Bs*CBM66, from *Bacillus subtilis*[33] (DALI *Z*-score 17.9, r.m.s.d 2.5 Å, 4b1m-C). *Bs*CBM66 is linked to a catalytic domain, SacC, which hydrolyses levan and the presence of *Bs*CBM66 enhances the activity of SacC[33]. The SGNH-like domain has an α-β-α fold, consisting of five α-helices

surrounding a central β-sheet. This structure is characteristic of SGNH hydrolases, which are a diverse family of carbohydrate modifying enzymes with a highly conserved catalytic core, composed of Asp, His and Ser residues[34]. Based on structural comparisons to other SGNH domains, this catalytic triad is formed by S612-H759-D756 in the SGNH-like domain (**Fig 3c**). The SGNH-like domain is structurally similar to an SGNH hydrolase from *Prolixibacter bellariivorans*, which is an acetyl xylan esterase[35] (DALI *Z*-score 21.6, r.m.s.d 2.1 Å, 7toi-A). It is also structurally similar to SGNH domains from two TSP's with esterase activity: gp82 of Moo19, a *Shigella flexneri* phage[36] (DALI *Z*-score 14.3, r.m.s.d 3 Å, 9d81-A) and gp63.1 of G7C, an *E. coli* phage[37] (DALI *Z*-score 13.4, r.m.s.d 3.2 Å, 4qn1-A). To explore whether gp25 may also exhibit esterase activity, we incubated gp25 with 4-nitrophenyl acetate at varying concentrations (6.25–100 ng protein to 100 µM of subtrate) under aerobic conditions and at room temperature, and observed absorbance at 405 nm, indicative of esterase activity (**Fig 3d**). No change in absorbance was observed for gp22_23 TSP, as well as for gp24 and gp26 (**Fig S3**). This demonstrates that gp25 possesses esterase activity, likely through its SGNH-like domain.

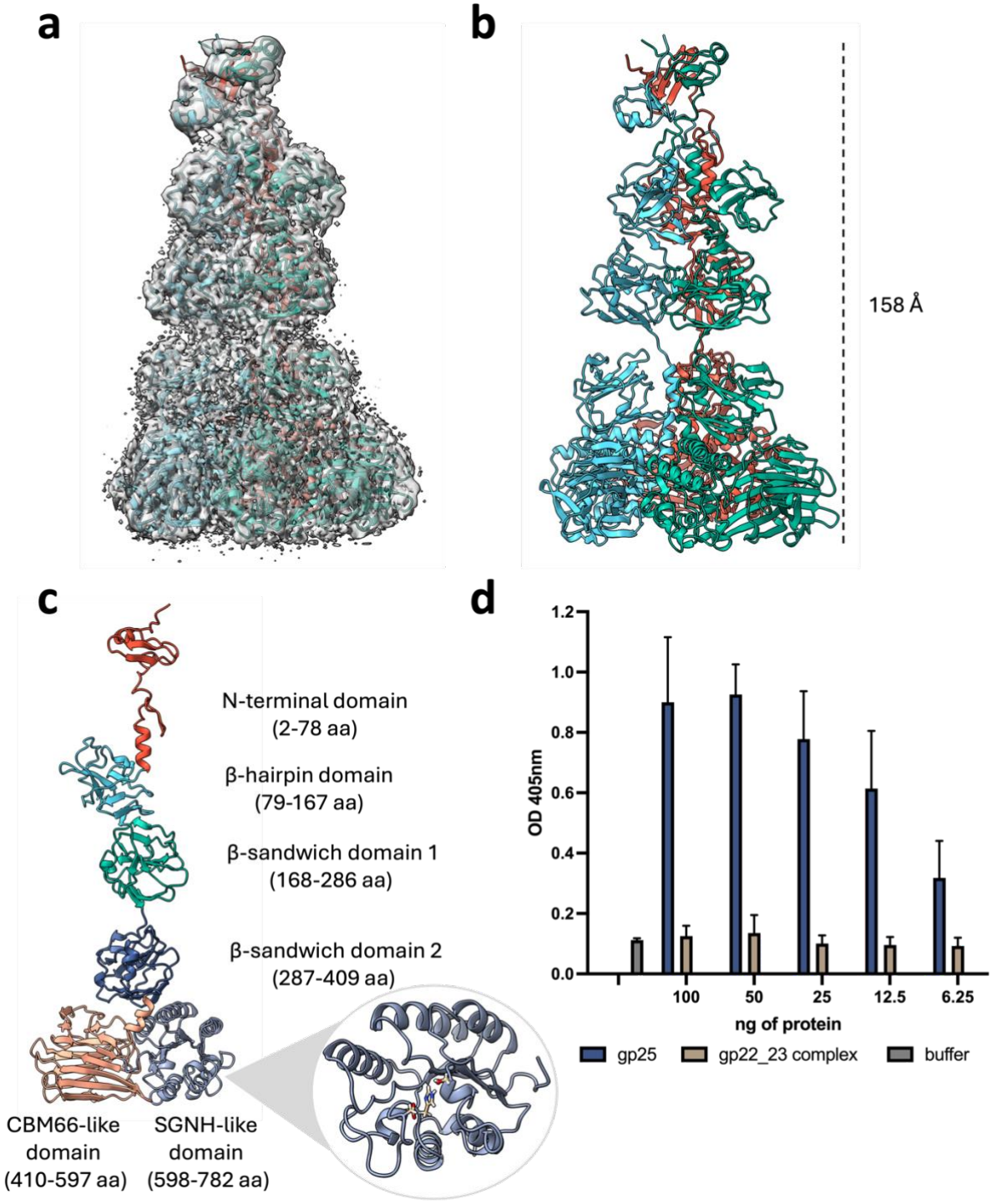



**Figure 3. Structure and activity of gp25. A)** Atomic model of gp25 fit to the cryo-EM electron density map, resolved to 2.6 Å. **B)** High resolution structure of gp25 trimer. Each individual monomer of gp25 is coloured blue, red and green. **C)** Structural domains of gp25 monomers: N-terminal domain (residues 2-78, red), β-hairpin domain (residues 79-167, blue), β-sandwich domain 1 (residues 168-286, green), β-sandwich domain 2 (residues 287-409, dark blue), CBM66-like domain (residues 410-597, orange) and SGNH-like domain (residues 598-782, purple). The catalytic triad formed by S612-H759-D756 is shown on the right. **D)** Esterase activity of gp25 on 4-nitrophenyl acetate at

concentrations ranging from 6.25-100 ng of protein. Absorbance was measured at $OD_{405nm}$, two minutes post 4-nitrophenyl acetate exposure. Error bars represent the standard deviation of three independent replicates.

## Predictions of outer TSP assembly

Gp26 is predicted by AlphaFold to be an elongated structure approximately 22 nm long and with eight domains, four of which are clustered at the C-terminal end **(Fig 4a-b).** This overall morphology is similar to tail fibre III of phage GP4 and Mu phage tail fibre[25,38]. The N-terminal domain (residues 1-94) consists of a β-sheet that connects to an α-helix through an extended loop. It has some structural similarity to the N-terminal β-prism domain of the major tropism determinant, RBP of *Bordetella* phage (DALI *Z*-score 4.1, r.m.s.d 8.8, 1yu0-A). The β-sandwich domain 1 (residues 95-213) consists of two antiparallel four-stranded β-sheets and is structurally similar to the β-sandwich domain 1 of gp25 (DALI *Z*-score 8.7, r.m.s.d, 2.8 Å). The β-sandwich domain 1 is connected to two RBP-like domains (residues 214-322 and residues 323-434) which are structurally similar to each other (DALI *Z*-score 10.5, r.m.s.d 3.9 Å). Both have structural similarity to the head domain of the RBP of *Lactococcus lactis* P2, which harbours the receptor binding site[39]. The gp26 domain consists of two α-helices and an anti-parallel β-sheet, interspersed with long loops. It has no similarity to any current structurally characterised proteins. The VAL like domain (residues 539-647) consists of three α-helices with a central anti-parallel β-sheet containing five strands. It is structurally similar to the catalytic CHAP domain of endolysin Ly7917, from *Streptococcus suis* temperate phage phi7917[40] (DALI *Z*-score 7.1, r.m.s.d 3.1, 5d76-A). The Fn3-like domain is primarily composed of a β-sandwich fold, consisting of five strands on one sheet and three strands on the other. It is structurally similar to the Fn3 domain of *Vibrio cholerae* enzyme VC0615, whose catalytic domain has exo-glucosidase activity[41] (DALI *Z*-score 7.4, r.m.s.d 2.4, 6gdt-E). The second β-sandwich domain (residues 757-850) contains two α-helices with a central anti-parallel β-sheet with three strands. It is structurally similar to the assembly domain II of the serine peptidase Vpr from *Bacillus amyloiquefaciens,* which is mainly responsible for inter-subunit

interactions[42] (DALI *Z*-score 8.4, r.m.s.d 2.8, 8jmw-D). Altogether, predictions of multiple RBP domains and a VAL-like domain suggests that gp26 may be able to locally degrade peptidoglycan.

Gp24 is predicted by AlphaFold to be an elongated, rod-like structure, approximately 29 nm long and 4 nm wide **(Fig 4c)**. Each monomer is composed of three domains. The N-terminal domain (residues 1-142) is composed of a β-hairpin and an α-helix, leading to a β-sheet surrounded by partial α-helices. Residues 44-106 of the N-terminal domain have structural similarity to the particle binding domain of K2-2 TSP[19] (DALI *Z*-score 3.2, r.m.s.d 2.6, 8iq5-D). A long linker leads to the rod-like domain (residues 143-244), which consists of three repeats of triple stranded anti-parallel β-sheets. This domain is structurally similar to the B region of accumulation associated protein (Aap, DALI *Z*-score 3.4, r.m.s.d 5.5, 8deo-A) and Staphylococcal surface protein G (SasG, DALI *Z*-score 3, r.m.s.d 4.5, 4wve-A). These long rod-like structures allow Aap and SasG to extend from the bacterial cell surface to bind to host cells and to be involved in biofilm formation[43]. The core of the C-terminal domain (residues 249-312) is composed of a β-sheet with four strands and a partial α-helix. It is structurally similar to the head domain of the TSP from the *Geobacillus* phage GVE2[44] (DALI *Z*-score 5.3, r.m.s.d 2.8, 7chu-B). Overall, gp24 is predicted to be a long, fibrous protein, whose three domains do not appear to have enzymatic activity.

In the absence of experimental models for gp24_25 and gp24_26 complexes, the AlphaFold models of gp24 (residues 20-312) and gp26, as well as the atomic model of gp25 were fitted into a locally reconstructed map of the outer TSP assembly **(Fig 4d)**. This shows the three proteins in a V-shaped formation. Gp24 and gp25 are in close proximity with possible contact at the following points; the N-terminal domain of gp25 and the rod-like domain of gp24, β-sandwich domain 1 of gp25 and the rod-like domain of gp24 and CBM66-like domain of gp25 and the C-terminal domain of gp24. Gp26 is at an angle from gp24 and gp25 and has just one point of contact between its N-terminal domain and the C-terminal domain of gp24 **(Fig 4e)**. CrAss001 has two tail hub (TH) proteins, THA and THB, which exist as $A_2B$ trimers encircling the tail barrel at 12 points[27]. THB has previously been shown to interact

with the N-terminal domain of gp22, docking the gp22_23 TSP to the tail barrel[27]. As gp24 appears to aid in assembly of a gp25 and gp26 complex, we hypothesised that the outer TSP assembly is docked to the tail barrel through interactions between the N-terminal domain of gp24 and the THA dimer. AlphaFold indeed predicts an interaction between gp24 and the THA dimer **(Fig 4f)**. The predicted outer assembly was combined with the atomic models for the inner TSP, capsid and tail barrel assembly to give a proposed model of the TSP's (**Fig 4g**).

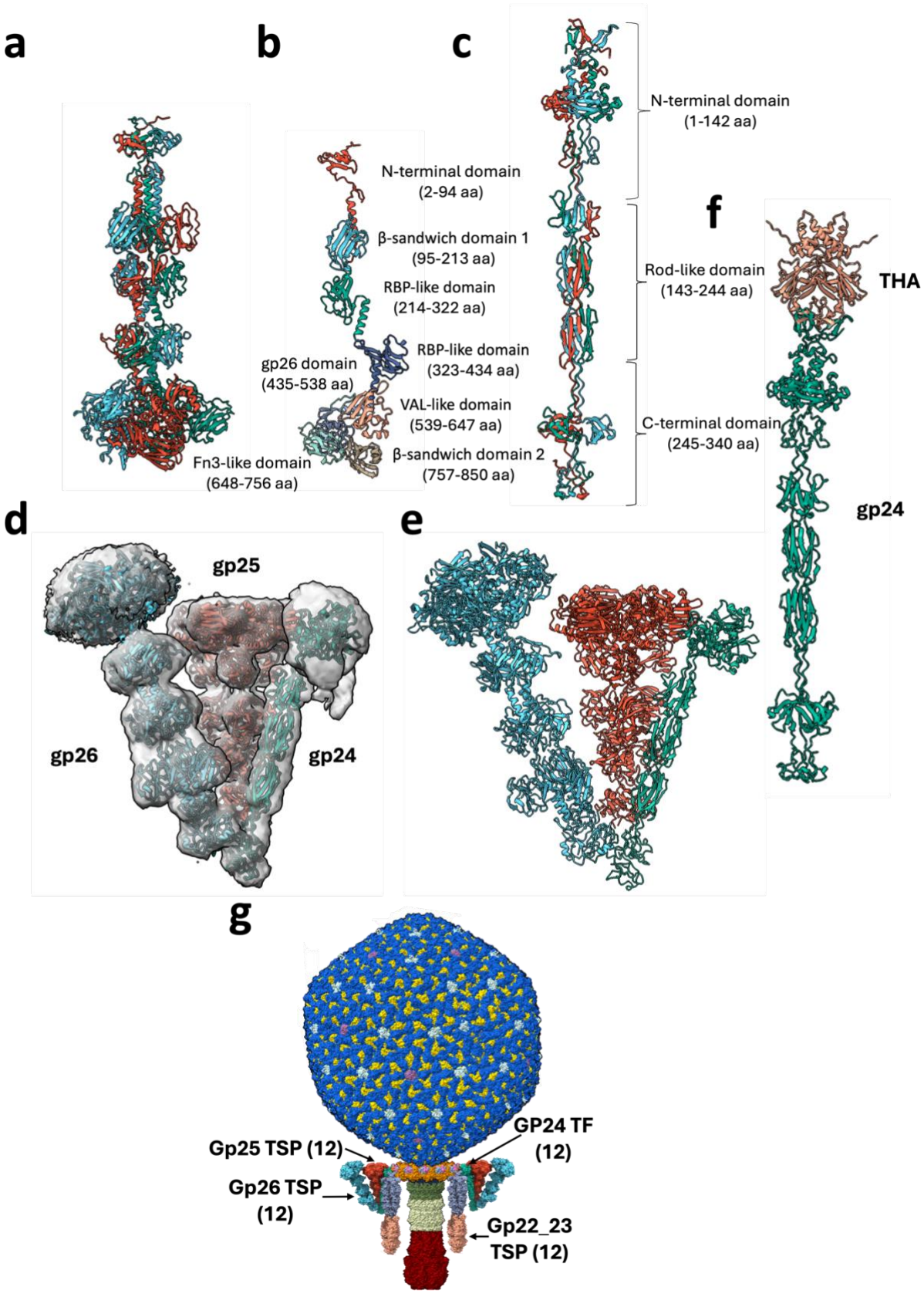



**Figure 4. AlphaFold predictions of outer TSP assembly. A)** AlphaFold prediction of gp26 trimer. Each individual monomer of gp26 is coloured blue, red and green. **B)** Structural domains of gp26

monomers: N-terminal domain (residues 2-94, red), β-sandwich domain (residues 95-213, blue), RBP-like domain (residues 214-322, green), RBP-like domain (residues 323-434, dark blue), gp26 domain (residues 435-538, purple), VAL-like domain (residues 539-647, orange), Fn3-like domain (residues 648-756, light green), β-sandwich domain (residues 757-850, beige). **C)** Alphafold prediction of gp24 trimer. Each individual monomer of gp24 is coloured blue, red and green. Gp24 consists of three domains: an N-terminal domain (residues 1-142), a rod-like domain (residues 143-244) and a C-terminal domain (residues 245-340). **D)** Alphafold models gp24 (residues 20-312) and gp26 and atomic model gp25 fit to a low density cryo-EM map of CrAss001 outer TSP assembly. **E)** Predicted outer TSP assembly. The C-terminal domain of gp24 makes contact with the N-terminal domains of gp25 and gp26. **F)** AlphaFold prediction of gp24 trimer interacting with THA dimer. **G)** Proposed assembly of inner and outer TSP's. Atomic models of capsid (PDB-7QOF) and tail barrel assembly (PDB-7QOJ) combined with atomic models of gp22_23 TSP and gp25 and AlphaFold models gp24 and gp26 to give a proposed structure of the phage. For simplicity, two copies out of 12 of each TSP are shown.

## Discussion

The long, thin and highly flexible nature of tail fibres and spikes poses a significant challenge in Cryo-EM studies of phages, where they are often averaged out during image processing[45]. To address this issue in CrAss001, the predicted TF's/TSP's were co-expressed with one protein tagged, and interactions confirmed by pulldown assays. This indicated interaction between gp22 and gp23, as well as between gp24 and gp25, and between gp24 and gp26. The gp22_23 complex was structurally characterized by cryo-EM and determined to 3.4 Å resolution. Both gp22 and gp23 proteins contain a modular, triple domain structure with a central β-helix domain that is typical of TSP's. Gp22_23 TSP was further demonstrated to be a capsule depolymerase, forming zones of clearing on overlays of a *B. intestinalis* strain that predominantly expresses PVR9. Gp22_23 was, however, unable to form zones of clearing on a *B. intestinalis* strain that primarily expresses PVR7, suggesting that PVR9 is a receptor for gp22_23 TSP and a target for its depolymerase activity. Curiously, this depolymerase activity appears to originate from gp22 only, as shown by zones of clearance, with gp23 alone unable to form such zones. This is despite gp23 also possessing a β-helix domain that would ordinarily be indicative of a

catalytic TSP. In the absence of gp23, the depolymerase activity of gp22 appeared reduced. The C-terminal domains of TSP's are generally thought to play a role in trimer assembly and stabilisation of the TSP[46,47]. It is likely that when gp22 is not in complex with gp23, the absence of the C-terminal domain of gp23 can result in a destabilised gp22, reducing its efficiency at capsule depolymerisation. Accordingly, gp22 alone is prone to degradation during purification, an issue that is resolved when purified in complex with gp23. The incorporation of gp23 may also be to assist with length: at approximately 31 nm, the gp22_23 complex is significantly longer than other characterised TSP's. The tail barrel of CrAss001 is likewise much longer than those observed in other podoviruses, due to the presence of five stacked ring proteins[27]. Therefore, the length of the hetero-hexamer assembly of gp22_23 likely assists in accessing the CPS receptor.

Gp25, which makes up part of the outer TSP assembly, was characterized by cryo-EM and a structure determined to 2.6 Å resolution. The SGNH-like domain and CBM66-like domain of gp25 are located together at the C-terminal end and are likely to work in tandem, with the CBM66-like domain determining substrate binding and possibly enhancing enzymatic activity of the SGNH-like domain. Moreover, gp25 was demonstrated to exhibited esterase activity by a colorimetric assay, suggesting at least one of these domains is indeed catalytic. Gp25's overall domain architecture and morphology is similar to that of the gp63.1 TSP of podovirus G7C[37]. Gp63.1 acts as an esterase by deacetylation of the O-antigen of *E. coli* 4's LPS, and was shown to be essential for G7C infection. Additionally, gp25's SGNH-like domain had structural similarity to gp82, a TSP of Moo19[36]. Gp82 also acts as an esterase and is dependent on specific O-antigen modifications of LPS. It is conceivable that gp25 also acts on LPS O-antigen. However, SGNH-like hydrolases are a diverse group of enzymes, with a number of known substrates including sialic acid and peptidoglycan[34].

Issues with protein degradation during purification of both gp24 and gp26 meant that the structural analysis of these proteins was based on AlphaFold predictions only. Gp24 is predicted to be a long, fibrous protein that lacks enzymatic activity.

Gp26 is predicted to have eight domains, one of which has structural similarity to a CHAP domain of endolysin Ly7917[40]. Fitting of the experimental structure of gp25 and AlphaFold predictions of gp24 and gp26 into the map of the outer TSP assembly in the CrAss001 virion showed that both gp25 and gp26 appeared to make contact with gp24, with few if any interactions between gp25 and gp26. While this agrees with pull-down assay results, it makes for an unusual arrangement where one gp24 seemingly docks two TSP's to the phage. Additionally, a conformational change must occur during phage infection to allow the C-terminal, enzymatically active domains of gp25 and gp26 access to their substrates. AlphaFold predictions further suggest that gp24 trimer interacts with THA dimer. However, it should be noted that AlphaFold is also able to predict two gp24 trimers interacting with each monomer of THA (**Fig S4**). This suggests a possible alternative conformation, where there are a total of 24 copies of gp24 trimers, each interacting with a THA monomer and docking either a gp25 trimer or gp26 trimer to the tail barrel. Further structural analysis is required to decipher the mode of interaction between gp24 and the tail hub proteins.

Structural analyses of several podoviruses have revealed the presence of either six[23,24,48,49] or twelve[36] TF's or TSP's and usually of only one type, matching the symmetry of 6-fold symmetric tail or 12-fold symmetric portal, respectively. The recent cryo-EM reconstruction of *Ralstonia solanacearum* phage GP4 was the first report of a podovirus possessing three types of TF's[25]. CrAss001 is unusual in having three TSP's that all appear to play active but separate roles during infection, and that there appears to be 12 copies of each. Such complexity may be necessary in a *Bacteroides* phage, where various mechanisms such as phase variation of CPS in *Bacteroides*[30,31,50] and delayed lysis by the phage[30], have been identified that allow both phage and bacterial host to co-exist in the gut. In the absence of CPS, *Bacteroides thetaiotaomicron* can persist in the presence of phages through phase variation of various surface proteins and lipoproteins[50]. This suggests that the use of multiple RBP's may be another mechanism employed by *Bacteroides* phages to assist in phage persistence, through targeting of alternative receptors.

The combined experimental analysis and structural predictions here have provided insight into how these TSP's interact with *B. intestinalis* to initiate phage infection (**Fig 5**). PVR9 CPS is likely to be a primary receptor, which gp22_23 TSP binds to and degrades (**Fig 5a**), allowing gp25 access to hydrolyse its receptor (possibly LPS, **Fig 5b**). These combined actions should provide access to the outer membrane for docking of the tail (**Fig 5c**) Finally, ejection of cargo proteins from the tail occurs, allowing the formation of a transmembrane channel into the cytoplasm of *B. intestinalis*[27]**(Fig 5d)**. The role gp26 plays in this process is unclear. The presence of a predicted VAL domain would suggest that it locally degrades peptidoglycan, but this would require gp26 to transverse the outer membrane. Additionally, while spotting of gp26 cultures of *B. intestinalis* did not produce zones of clearing, incubation of this protein with CPS extracted from WT *B. intestinalis* did show the production of some breakdown products. This suggests gp26 may possess activity against alternative CPS (possibly PVR8, PVR11 or PVR12) as a mechanism of phage persistence. While similar modes of action may be expected of other *Crassvirales* phages, it should be noted that there is a high level of diversity in both the number and type of TF/TSP genes, even within the same family[51]. Additionally, TF's/TSP's of *Crassvirales* are subjected to rapid evolution[52], likely as a selective pressure to co-evolve with their bacterial hosts.

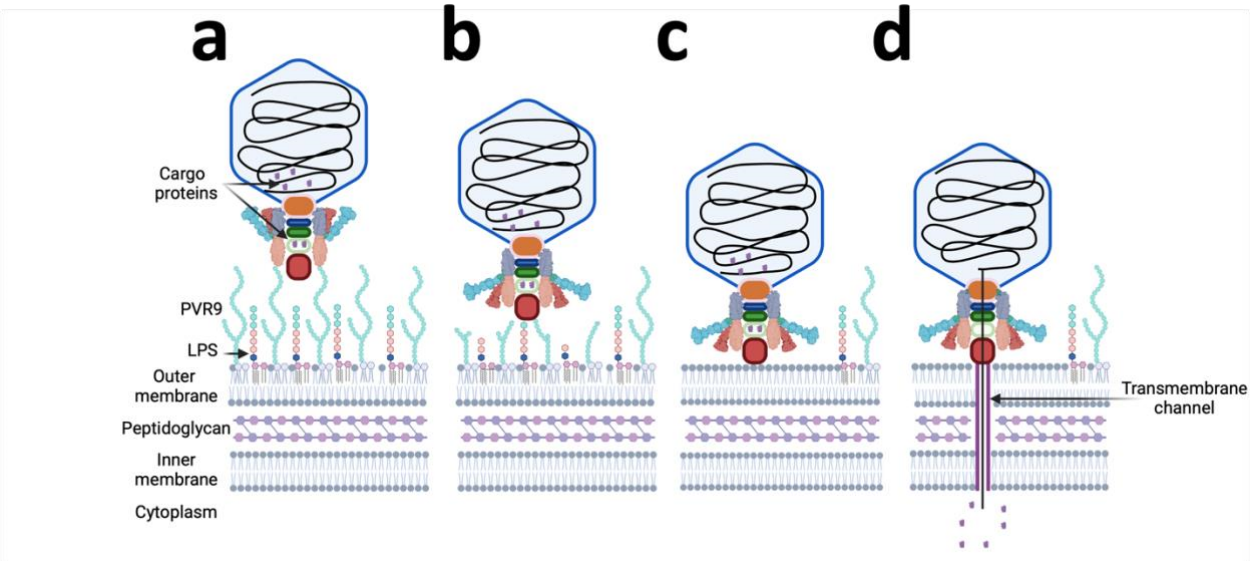


**Figure 5. Schematic of initial stages of CrAss001 infection, facilitated by TSP's. A)** PVR9 is the first phage receptor, which is recognized and degraded by gp22_23 inner TSP. **B)** A conformational change occurs to the outer TSP assembly, allowing gp25 to extend towards its receptor. **C)** Hydrolysis of LPS by gp25 correctly positions the tail of CrAss001 to dock to the outer membrane of *B. intestinalis*. **D)** A transmembrane channel into the cytoplasm of *B. intestinalis* is formed. Cargo proteins are ejected first, followed by DNA.

## Methods

### Cloning, protein expression and purification

Consecutive gene pairs were PCR amplified from CrAss001 genomic DNA and cloned into pET SUMO vector (Thermo Fisher Scientific). Primers used for amplification can be found in supplementary table 2. Non-consecutive gene pairs were synthesized at Azenta Life Sciences. Plasmids were transformed into *E. coli* Bl21(DE3) competent cells (Thermo Fisher Scientific). For protein expression, cells were grown in Luria–Bertani (LB) medium and 1% w/v glucose at 37°C for 2 hours, shaking at 200rpm. The culture was induced with 0.5 mM isopropyl-β-D-thiogalactopyranoside (IPTG) and incubated at 19°C for 24 hours.

The cells were pelleted by centrifugation at 4,500g for 30 minutes and pellets were resuspended in binding buffer (500 mM NaCl, 20 mM HEPES pH 8.2, 40 mM Imidazole, 10% v/v glycerol, 1mM DTT, 4 µg/ml Dnase I, and 1 x Pierce protease inhibitor EDTA-free tablet). The cells were lysed through the addition of lysozyme (0.1 mg/mL) and sonication. Lysates were then centrifuged at 10,000 g for 40 minutes.

Protein purification was carried out by Ni-Affinity Chromatography, using ÄKTA start protein purification system (Cytiva Life Sciences) and HisTrap Fast Flow columns. Wash and elution buffers were the same as the binding buffer, minus Dnase I and protease inhibitor and the elution buffer contained 500 mM Imidazole. Fractions containing the target protein were buffer exchanged into 300 mM NaCl, 20 mM

HEPES pH 8.2, 10% v/v glycerol and 1mM DTT (SEC buffer) using Zeba Spin Desalting Columns (Thermo Fisher Scientific). Protein was then further purified by Size Exclusion Chromatography, using a Hiprep 16/60 Sephacryl S-300 HR column. Molecular weight calibration of the column was performed using high molecular weight standards (Cytiva). Eluted fractions were analysed by SDS-PAGE using Bolt Bis-Tris 4-12% gels (Thermo Fisher Scientific), staining with Imperial Protein stain and visualised with Invitrogen iBright CL1500 Imaging System.

Electron Microscopy

Gp22_23 complex was purified as detailed above and concentrated to approximately 1.5 mg/ mL. Data was collected on a Glacios TEM instrument operating at 200kV. Data processing was performed in RELION 5.0[53]. Motion correction of movies was performed using RELION and contrast transfer functions were estimated using CTFFIND-4.1[54], with dose-weighting. Initial 2D classifications resulted in 6 good classes, composed of 46,988 particles. These particles were used for initial 3D refinements, imposing $C_3$ symmetry. This was followed by CTF-refinement of per-particle defocus values and astigmatism, as well as magnification anisotropy estimation. Per-particle, beam induced motion correction was performed through Bayesian Polishing. An initial atomic model was built using ModelAngelo[55], which was further manually edited in Coot[56] and realspace refined using Phenix[57]. Gp25 was processed in a similar manner, with the addition of imposing C1 symmetry to resolve the N-terminal asymmetrical domain. Refinement statistics are available in **Table S1**. Images were generated using ChimeraX[58].

AlphaFold predictions

Individual tail fibres and some complex structures were predicted using AlphaFold2[59] on the Jean-Zay server at IDRIs, while both the gp22_23 complex and $THA_2B$_gp22_gp24 complex were predicted using AlphaFold3[60] on Google servers at https://golgi.sandbox.google.com. For the gp22_23 complex, three copies of the C-terminal domain of gp22 (residues 781-832) were predicted in complex with three copies of gp23. The full length gp22 was then superimposed on the partial complex

with Coot[56] and the overlapping segments were deleted. The final predicted domain structures were submitted to the Dali server[61] to identify structurally similar proteins in the PDB. Visual representations of the structures were prepared in ChimeraX[58].

Bacterial strains and culture conditions

The host strain of CrAss001, *B. intestinalis* APC 919/174, a clone of *B. intestinalis* resistant to phage infection, 8T, and a sensitive revertant subclone of 8T, 8W, were used for depolymerase spot assays. Cultures were grown in Fastidious Anaerobe Broth (FAB, Oxoid) anaerobically at 37°C.

Enzymatic assays

For capsule depolymerase spot tests, a bacterial overlay was created with 200 µL of WT, 8T and 8W, which were added to 3.5 mL of 0.45% FAB top agar. Two-fold dilutions of 100 ng of protein were spot in 2 µL volumes. After drying, the plates were inverted and incubated at 37°C for 24 hours. The appearance of hazy zones of clearing was indicative of capsule digestion. Spot tests of each protein were repeated in triplicate. Esterase assays were carried out in 96-well plates, with a final volume of 200 µL. Each well contained SEC buffer, 100 µM 4-nitrophenyl acetate and protein (ranging from 6.25-100 ng). SEC buffer and 100 µM 4-nitrophenyl acetate alone was used as a negative control. Absorbance was read at room temperature at 405 nm two minutes after the addition of 4-nitrophenyl acetate using a microplate reader, with vigorous shaking for five seconds before reading. The assay was repeated in triplicate for each protein.

## Author Contributions

C.A.T, O.W.B, A.N.S and C.H contributed to the conceptual development of this study. B.G created protein constructs. C.A.T carried out protein expression,

pulldown assays, depolymerase assays and esterase assays. O.W.B collected Cryo-EM data. C.A.T and O.W.B processed Cryo-EM data and built and refined atomic models. C.C performed AlphaFold predictions and analysis. C.A.T, O.W.B, C.C, A.A, A.N.S and C.H interpreted the results. C.A.T wrote the manuscript with editorial contributions from O.W.B, A.N.S and C.H.

Supplementary

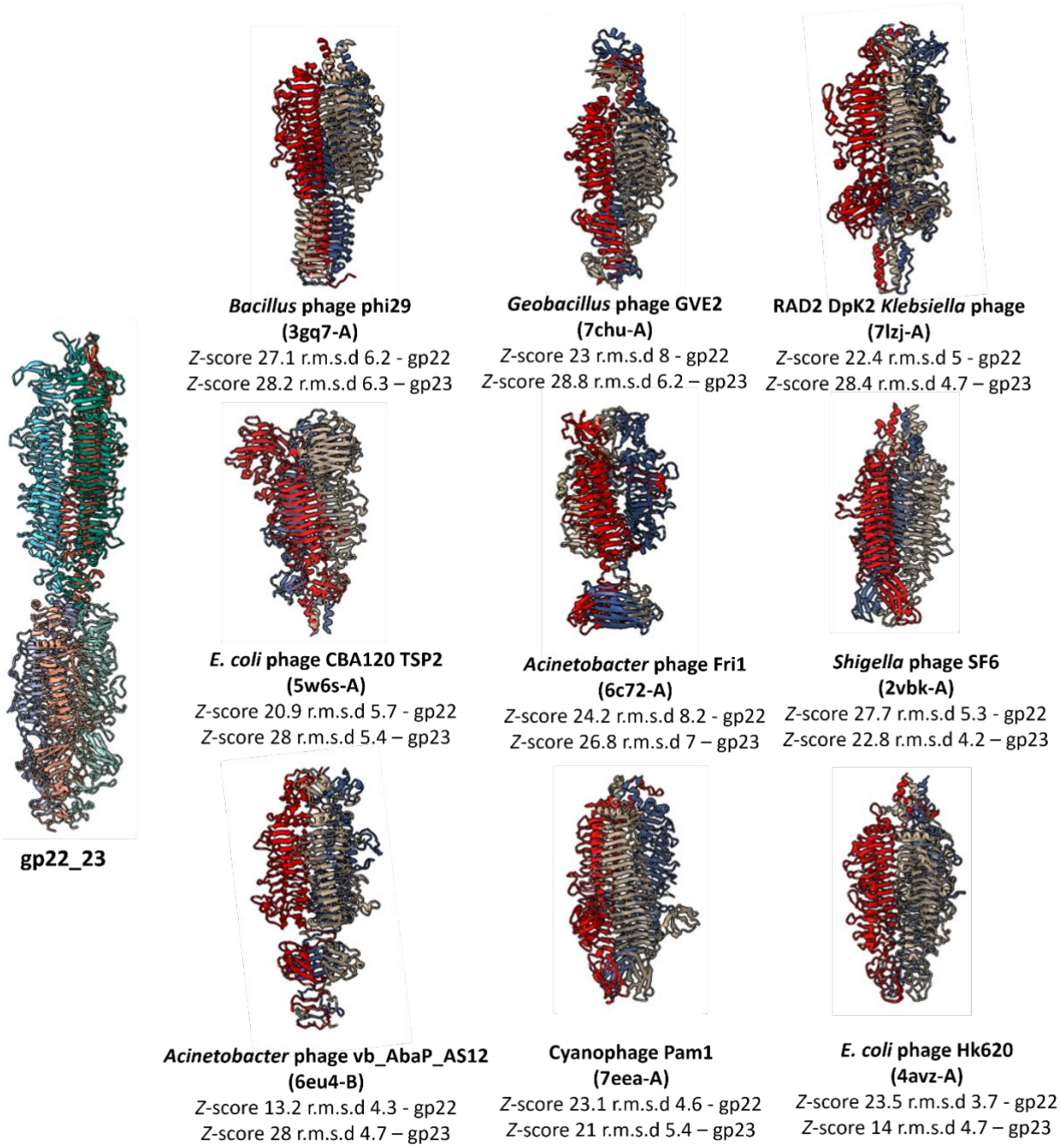



**Figure S1. Gp22_23 TSP has structural homology to a number of other phage TSPs.** All TSP's are orientated from N-terminal to C-terminal after alignment. Structural homology is based on DALI *Z*-scores and r.m.s.d values. Each individual monomer of gp22 is coloured blue, red and green and each monomer of gp23 is coloured purple, orange and light green. Each monomer of the TSP trimers is coloured red, light brown and dark blue.

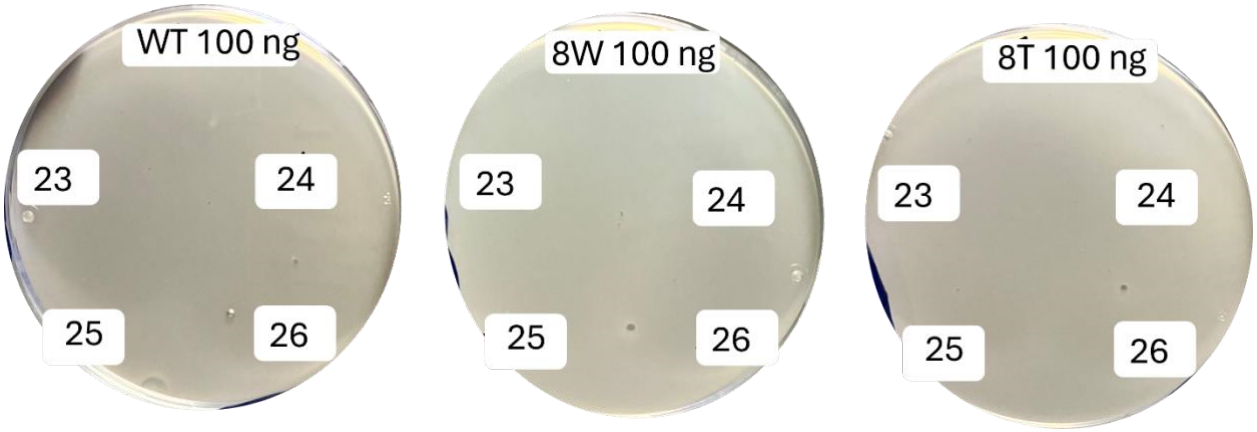



**Figure S2. Gp23, gp24, gp25 and gp26 lack capsule depolymerase activity.** One hundred nanograms of each protein were spot on WT, 8W and 8T. Each spot test was repeated in triplicate.

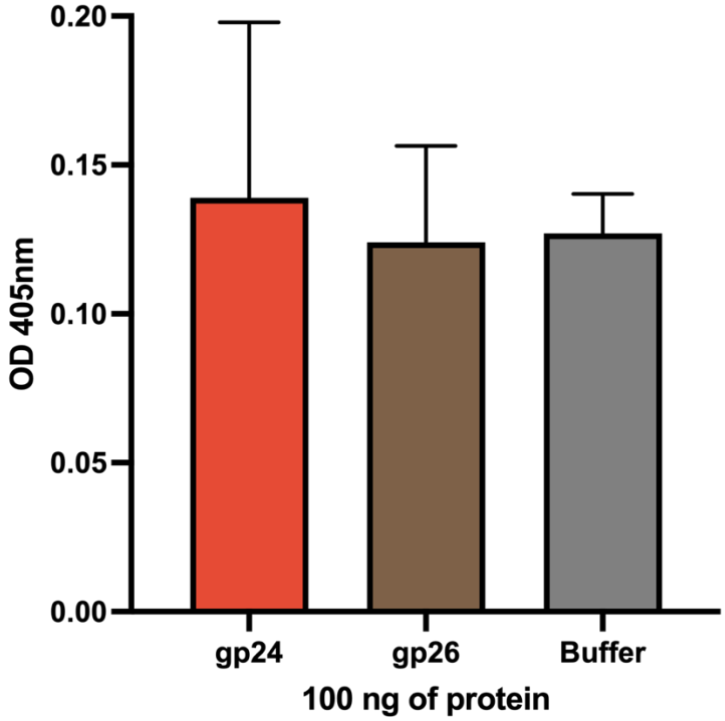



**Figure S3. Gp24 and gp26 have no esterase activity.** One hundred nanograms of each protein was incubated with 4-nitrophenyl acetate for two minutes and activity was measured by absorbance at 405 nm in a microplate reader. Error bars represent the standard deviation of three independent replicates.

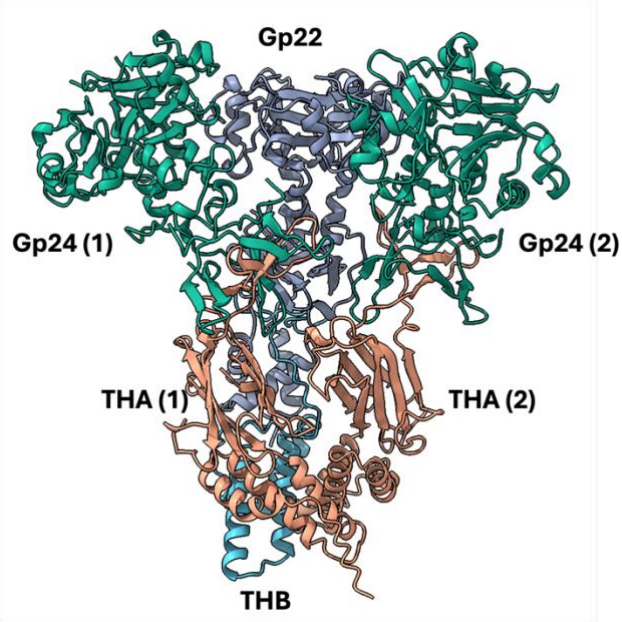



**Figure S4**. **Alternative AlphaFold predicted interactions of gp24 and THA.** Two trimers of gp24 (residues 1-110) are predicted to interact with each monomer of THA. A trimer of gp22 (residues 1130) are also shown interacting with THB monomer.

| | gp22_23 | gp25 |
|---|---|---|
| **Model composition** | | |
| Non-hydrogen atoms | 63,825 | 36,081 |
| Protein residues | 4,134 | 2,343 |
| Ligands | 0 | 0 |
| **R.M.S deviations** | | |
| Bond lengths/ Å | 0.004 | 0.004 |
| Bond angles/ ° | 0.754 | 0.665 |
| **Ramachandran plot** | | |
| Favoured/ % | 87.15 | 92.13 |
| Allowed/ % | 12.63 | 7.57 |
| Outlier/ % | 0.22 | 0.3 |
| **Validation** | | |
| Rotamer outliers/ % | 0.12 | 0.55 |
| Clashscore | 1.79 | 1.83 |
| Molprobity | 1.4 | 1.14 |

**Table S1. Model composition and refinement statistics of gp22_23 complex and gp25.** Statistics were obtained after real space refinement in phenix.

| GENES | PRIMERS | LENGTH |
|---|---|---|
| **22_23** | CrAss001_gp22_F1 ATGTTTTTTACACAAGAAGATTATAGAAAAATAG<br>CrAss001_gp23_R1 GTCTTCATATCTTCCCTCTTGAC | 4736 |
| **23_24** | CrAss001_gp23_F1 ATGGCAATACAAGATAAATTTAATAAGCC<br>CrAss001_gp24_R1 AATCTTTCTTAGTCTTGTATCAGTTTC | 3240 |
| **24_25** | CrAss001_gp24_F1 ATGAAAGATATACAACAACTAATTAAGAAG<br>CrAss001_gp25_R1 GTCATTTCTTTGTATTACTCTATCTGCC | 3438 |
| **25_26** | CrAss001_gp25_F1 ATGGATAGAATACAATTAAGAAGAGAC<br>CrAss001_gp26_R1 TACTTCTTCATATCTTCTTCCCTTA | 5009 |

**Table S2. Primer sequences for consecutive gene pairs amplification**

## References


1. Braga, L. P. P. *et al.* Impact of phages on soil bacterial communities and nitrogen availability under different assembly scenarios. *Microbiome* **8**, (2020).
2. Brown, T. L., Charity, O. J. & Adriaenssens, E. M. Ecological and functional roles of bacteriophages in contrasting environments: marine, terrestrial and human gut. *Curr Opin Microbiol* **70**, 102229 (2022).
3. Guerin, E. & Hill, C. Shining Light on Human Gut Bacteriophages. *Front Cell Infect Microbiol* **10**, (2020).
4. Borodovich, T., Shkoporov, A. N., Ross, R. P. & Hill, C. Phage-mediated horizontal gene transfer and its implications for the human gut microbiome. *Gastroenterol Rep (Oxf)* **10**, (2022).
5. Benler, S. *et al.* Thousands of previously unknown phages discovered in whole community human gut metagenomes. *Microbiome* **9**, (2021).
6. Koonin, E. V. & Yutin, N. The crAss-like Phage Group: How Metagenomics Reshaped the Human Virome. *Trends Microbiol* **28**, 349–359 (2020).
7. Turner, D. *et al.* Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ICTV bacterial viruses subcommittee. *Arch Virol* **168**, (2023).

8. Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D. & Lawley, T. D. Massive expansion of human gut bacteriophage diversity. *Cell* **184**, (2021).

9. Yutin, N. *et al.* Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features. *Nat Commun* **12**, (2021).

10. Honap, T. P. *et al.* Biogeographic study of human gut-associated crAssphage suggests impacts from industrialization and recent expansion. *PLoS One* **15**, (2020).

11. Edwards, R. A. *et al.* Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. *Nat Microbiol* **4**, (2019).

12. Dutilh, B. E. *et al.* A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat Commun* **5**, (2014).

13. Guerin, E. *et al.* Isolation and characterisation of ΦcrAss002, a crAss-like phage from the human gut that infects Bacteroides xylanisolvens. *Microbiome* **9**, (2021).

14. Papudeshi, B. *et al.* Host interactions of novel Crassvirales species belonging to multiple families infecting bacterial host, Bacteroides cellulosilyticus WH2. *Microb Genom* **9**, (2023).

15. Hryckowian, A. J. *et al.* Bacteroides thetaiotaomicron-Infecting Bacteriophage Isolates Inform Sequence-Based Host Range Predictions. *Cell Host Microbe* **28**, (2020).

16. Ramos-Barbero, M. D. *et al.* Characterization of crAss-like phage isolates highlights Crassvirales genetic heterogeneity and worldwide distribution. *Nat Commun* **14**, (2023).

17. Shkoporov, A. N. *et al.* ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. *Nat Commun* **9**, (2018).

18. González-García, V. A. *et al.* Conformational Changes Leading to T7 DNA Delivery upon Interaction with the Bacterial Receptor. *Journal of Biological Chemistry* **290**, 10038–10044 (2015).

19. Ye, T.-J. *et al. Klebsiella pneumoniae* K2 capsular polysaccharide degradation by a bacteriophage depolymerase does not require trimer formation. *mBio* **15**, (2024).

20. Xia, G. *et al.* Wall Teichoic Acid-Dependent Adsorption of Staphylococcal Siphovirus and Myovirus. *J Bacteriol* **193**, 4006–4009 (2011).

21. Nobrega, F. L. *et al.* Targeting mechanisms of tailed bacteriophages. *Nat Rev Microbiol* **16**, 760–773 (2018).

22. Dunne, M., Prokhorov, N. S., Loessner, M. J. & Leiman, P. G. Reprogramming bacteriophage host range: design principles and strategies for engineering receptor binding proteins. *Curr Opin Biotechnol* **68**, 272–281 (2021).

23. Cuervo, A. *et al.* Structural Characterization of the Bacteriophage T7 Tail Machinery. *Journal of Biological Chemistry* **288**, 26290–26299 (2013).

24. Steinbacher, S. *et al.* Phage P22 tailspike protein: crystal structure of the head-binding domain at 2.3 Å, fully refined structure of the endorhamnosidase at 1.56 Å resolution, and the molecular basis of O-antigen recognition and cleavage. *J Mol Biol* **267**, 865–880 (1997).

25. Zheng, J. *et al.* Asymmetric Structure of Podophage GP4 Reveals a Novel Architecture of Three Types of Tail Fibers. *J Mol Biol* **435**, 168258 (2023).

26. Šiborová, M. *et al.* Tail proteins of phage SU10 reorganize into the nozzle for genome delivery. *Nat Commun* **13**, 5622 (2022).

27. Bayfield, O. W. *et al.* Structural atlas of a human gut crassvirus. *Nature* **617**, (2023).

28. Knecht, L. E., Veljkovic, M. & Fieseler, L. Diversity and Function of Phage Encoded Depolymerases. *Front Microbiol* **10**, (2020).

29. Dunstan, R. A. *et al.* Mechanistic Insights into the Capsule-Targeting Depolymerase from a Klebsiella pneumoniae Bacteriophage. *Microbiol Spectr* **9**, (2021).

30. Shkoporov, A. N. *et al.* Long-term persistence of crAss-like phage crAss001 is associated with phase variation in Bacteroides intestinalis. *BMC Biol* **19**, (2021).

31. Cortés-Martín, A. *et al.* Adaptations in gut Bacteroidales facilitate stable coexistence with their lytic bacteriophages. Preprint at https://doi.org/10.1101/2024.11.17.624012 (2024).

32. Smith, N. L. *et al.* Structure of a group A streptococcal phage-encoded virulence factor reveals a catalytically active triple-stranded β-helix. *Proceedings of the National Academy of Sciences* **102**, 17652–17657 (2005).

33. Cuskin, F. *et al.* How nature can exploit nonspecific catalytic and carbohydrate binding modules to create enzymatic specificity. *Proceedings of the National Academy of Sciences* **109**, 20889–20894 (2012).

34. Anderson, A. C., Stangherlin, S., Pimentel, K. N., Weadge, J. T. & Clarke, A. J. The SGNH hydrolase family: a template for carbohydrate diversity. *Glycobiology* (2022) doi:10.1093/glycob/cwac045.

35. Penttinen, L. *et al.* Elucidating Sequence and Structural Determinants of Carbohydrate Esterases for Complete Deacetylation of Substituted Xylans. *Molecules* **27**, 2655 (2022).

36. Subramanian, S., Bergland Drarvik, S. M., Tinney, K. R., Doore, S. M. & Parent, K. N. Moo19 and B2: Structures of *Schitoviridae* podophages with *T* = 9 geometry and tailspikes with esterase activity. *Sci Adv* **10**, (2024).

37. Prokhorov, N. S. *et al.* Function of bacteriophage G7C esterase tailspike in host cell adsorption. *Mol Microbiol* **105**, 385–398 (2017).

38. North, O. I. *et al.* Phage tail fibre assembly proteins employ a modular structure to drive the correct folding of diverse fibres. *Nat Microbiol* **4**, 1645–1653 (2019).

39. Sciara, G. *et al.* Structure of lactococcal phage p2 baseplate and its mechanism of activation. *Proceedings of the National Academy of Sciences* **107**, 6852–6857 (2010).

40. Ji, W. *et al.* A novel endolysin disrupts *Streptococcus suis* with high efficiency. *FEMS Microbiol Lett* **362**, fnv205 (2015).

41. Wu, L. & Davies, G. J. Structure of the GH9 glucosidase/glucosaminidase from *Vibrio cholerae*. *Acta Crystallogr F Struct Biol Commun* **74**, 512–523 (2018).

42. Cheng, Y., Han, J., Song, M., Zhang, S. & Cao, Q. Serine peptidase Vpr forms enzymatically active fibrils outside Bacillus bacteria revealed by cryo-EM. *Nat Commun* **14**, 7503 (2023).

43. Clark, L. C. *et al.* Staphylococcal Periscope proteins Aap, SasG, and Pls project noncanonical legume-like lectin adhesin domains from the bacterial surface. *Journal of Biological Chemistry* **299**, 102936 (2023).

44. Zhang, L. *et al.* Structural and functional characterization of the deep-sea thermophilic bacteriophage GVE2 tailspike protein. *Int J Biol Macromol* **164**, 4415–4422 (2020).

45. Ouyang, R., Ongenae, V., Muok, A., Claessen, D. & Briegel, A. Phage fibers and spikes: a nanoscale Swiss army knife for host infection. *Curr Opin Microbiol* **77**, 102429 (2024).

46. Xiang, Y. *et al.* Crystallographic Insights into the Autocatalytic Assembly Mechanism of a Bacteriophage Tail Spike. *Mol Cell* **34**, 375–386 (2009).

47. Gage, M. J. & Robinson, A. S. C-terminal hydrophobic interactions play a critical role in oligomeric assembly of the P22 tailspike trimer. *Protein Science* **12**, 2732–2747 (2003).

48. Müller, J. J. *et al.* An Intersubunit Active Site between Supercoiled Parallel β Helices in the Trimeric Tailspike Endorhamnosidase of Shigella flexneri Phage Sf6. *Structure* **16**, 766–775 (2008).

49. Subramanian, S., Bergland Drarvik, S. M., Tinney, K. R. & Parent, K. N. Cryo-EM structure of a Shigella podophage reveals a hybrid tail and novel decoration proteins. *Structure* **32**, 24-34.e4 (2024).

50. Porter, N. T. *et al.* Phase-variable capsular polysaccharides and lipoproteins modify bacteriophage susceptibility in Bacteroides thetaiotaomicron. *Nat Microbiol* **5**, 1170–1181 (2020).

51. Guerin, E. *et al.* Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. *Cell Host Microbe* **24**, 653-664.e6 (2018).

52. Brown, B. P. *et al.* crAssphage genomes identified in fecal samples of an adult and infants with evidence of positive genomic selective pressure within tail protein genes. *Virus Res* **292**, 198219 (2021).

53. Scheres, S. H. W. RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *J Struct Biol* **180**, (2012).

54. Rohou, A. & Grigorieff, N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J Struct Biol* **192**, (2015).

55. Jamali, K. *et al.* Automated model building and protein identification in cryo-EM maps. *Nature* **628**, 450–457 (2024).

56. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**, 486–501 (2010).

57. Adams, P. D. *et al.* PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* **66**, 213–221 (2010).

58. Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Science* **30**, 70–82 (2021).

59. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

60. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

61. Holm, L. Dali server: structural unification of protein families. *Nucleic Acids Res* **50**, W210–W215 (2022).

# Chapter 6: Capsid structure of *Crassvirales* phage DAC15

Ciara A Tobin, Oliver W Bayfield, Andrey N Shkoporov and Colin Hill

## Abstract

Phages of the *Crassvirales* order are highly abundant in the human gut, with mechanisms such as phase variation allowing them to persist with their *Bacteroidetes* hosts. The cryogenic electron microscopy-based reconstruction of *Kehishuvirus primarius* (CrAss001) provided structural and functional annotation to the majority of the proteins that constitute the virion. The capsid was determined to have a HK97-fold and a triangulation number of nine. In addition to the major capsid protein (MCP), each structural unit had nine copies of an auxiliary protein and both the size and fold of this protein had not been observed in the capsids of virions previously. Furthermore, two copies of a head fibre dimer (HFD) and one copy of a head fibre trimer (HFT) were associated with each structural unit. In this study, the capsid of a *Crassvirales* phage within the same family (*Steigviridae*) was resolved to 2.95 Å. The capsid of *Wulfhauvirus bangladeshii* (DAC15) is 86 nm in diameter and also has a triangulation number of nine. Both the MCP and auxiliary protein of DAC15 have strong structural matches to their counterparts in CrAss001, with the only discernible difference being an insertion of nine residues to the N-arm of the MCP of CrAss001. The N-terminal domains of both the HFD and HFT were structurally similar to those in CrAss001 and they bind in the pocket of each hexon created by the interactions of the MCP and auxiliary protein. The C-terminal domains of both the HFD and HFT were predicted by AlphaFold. The HFD has an Ig-like fold with both monomers protruding vertically from the capsid. This differs to the HFD of CrAss001, whose Ig-like domains are predicted to angle away from each other. The HFT of DAC15 has a 26-residue loop insertion which results in the HFD being projected several nanometres further from the capsid than the HFT of CrAss001. These head fibre proteins may aid in persistence of *Crassvirales* phages, either through binding to the bacterial host or to surfaces in the human gut. An extensive number of non-covalent interactions (primarily hydrogen bonds and salt bridges) were identified both between MCP chains and between MCP and auxiliary proteins. The conservation of the MCP-auxiliary protomer lends weight to the idea that this is a structural hallmark of *Crassvirales*.

## Introduction

Tailed, double-stranded DNA bacteriophages (phages) belonging to the class *Caudoviricetes* constitute ~96% of all currently known phages[1]. Phage DNA is densely packed in an icosahedral, proteinaceous shell, known as a capsid. The capsid is connected to the tail through a dodecameric portal vertex. The tail is responsible for bacterial host recognition, penetration of the cell membrane and ejection of the phage DNA into the cytoplasm[2]. Stabilisation of the capsid is crucial, both to withstand harsh environments and due to the high internal pressure generated by the densely packed DNA. The capsids of all tailed double-stranded phages are composed of a major capsid protein (MCP) which has a conserved fold, referred to as the HK97 fold[3] (first identified in HK97 phage[4]). There are four highly conserved domains: the N-terminal arm (N-arm), the peripheral domain (P-domain), the axial domain (A-domain) and the extended loop (E-loop). There are additional domains that appear in some capsids, such as the insertion domain[5,6] (I-domain) and the glycine-rich loop[7] (G-loop). The icosahedral capsid is composed of 60 asymmetrical structural units, with the number of copies of the MCP in each structural unit referred to as the triangulation (T) number[8]. This allows for great variation in the size of capsids, to accommodate for different genome lengths of phages. Some phages have additional auxiliary or decoration proteins incorporated into the capsid which can reinforce stability[9,10] or aid in bacterial host cell adhesion[11,12].

Phages of the order *Crassvirales* are highly abundant and prevalent in the human gut, where they infect bacteria in the phylum *Bacteroidetes*[13,14,15,16,17,18]. Despite their high abundance, the first member of this order, *Carjivirus communis* (prototypical or p-crAssphage) was only first identified in silico in 2014. Experiments with the small number of *Crassvirales* phages that have been successfully isolated with their bacterial hosts since then have revealed mechanisms such as phase variation[19,20] and delayed lysis of phage infected bacterial cells[19] that have provided insights into how these phages and their bacterial hosts co-exist. The cryo-electron microscopy-based reconstruction of *Crassvirales* phage *Kehishuvirus primarius*

(CrAss001) allowed the structural and functional assignment to the majority of proteins that constitute the virion[21]. The capsid was determined to have an outer diameter of approximately 88 nm and had a triangulation number of nine. In addition to nine copies of the MCP, each asymmetric unit also had nine copies of an auxiliary protein, two copies of a head fibre dimer (HFD) and one copy of a head fibre trimer (HFT). The auxiliary protein had a unique fold which had not been observed in viral capsids before[21].

In this study, the capsid of a *Crassvirales* phage *Wulfhauvirus bangladeshii* (DAC15) was resolved to 2.95 Å. DAC15 is in the same family as CrAss001 (*Steigviridae*) and infects *Bacteroides thetaiotaomicron*[15], while the host of CrAss001 is *Bacteroides intestinalis*[13]. Like CrAss001, the capsid was determined to have a triangulation number of nine and both the MCP and auxiliary protein were highly structurally similar to those in CrAss001. Extensive interactions, both between MCP chains and between MCP chains and auxiliary chains were identified, which aids capsid stability. The majority of these interactions were mediated by hydrogen bonds while the remaining interactions were formed by salt bridges. The HFD and HFT of DAC15 were structurally similar to those of CrAss001. Differences were observed in positioning of the C-terminal domain of the HFD's and the insertion of a 26 amino acid loop in the HFT of DAC15 that helps protrude it further from the capsid than the HFT of CrAss001. Both the conservation of the auxiliary protein and the head fibre proteins suggest that the capsids of *Crassvirales* phages are highly stable.

## Results

To determine the structure of the DAC15 capsid, 23,662 cryo-EM images of the virion were collected **(Fig 1a)**. Data collection, processing and refinement statistics are shown in **Table S1**. A total of 200,159 particles were selected and the capsid was reconstructed to a resolution of 2.95 Å by imposing icosahedral symmetry. **(Fig 1b)** The capsid has an outer diameter of approximately 86 nm and has a triangulation number of nine. Similar to CrAss001, each asymmetric unit is composed of nine copies of the major capsid protein (*gp091*), nine copies of the

auxiliary protein (*gp088*), two copies of the head fibre dimer protein (*gp094*) and one copy of the head fibre trimer protein (*gp103*).

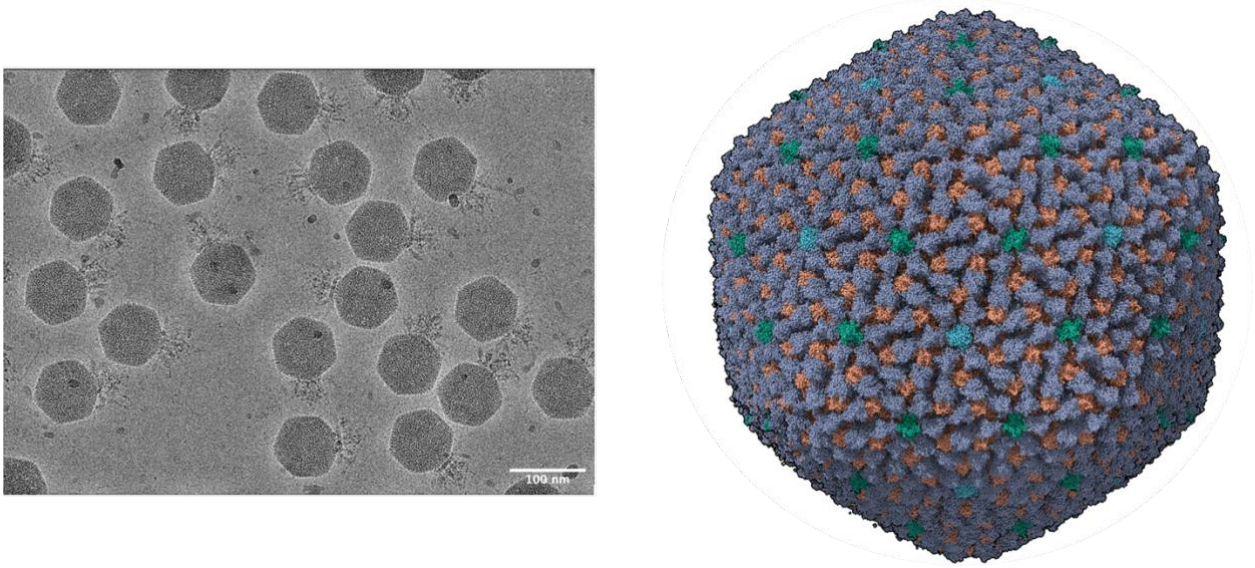



**Figure 1. Overall structure of DAC15 capsid. A)** Cryo-electron micrograph of DAC15 virions, n = 23,662 micrographs. **B)** Molecular surface of capsid reconstruction. The major capsid protein is coloured orange, auxiliary protein purple, head fibre dimer (residues 1-50) green and head fibre trimer (residues 1-39) blue.

## Structure and interactions of MCP

The major capsid protein is composed of five domains: An N-arm domain (residues 2-49), a P-domain (residues 50-60, 192-289 and 431-497), an E-loop (residues 61-76 and 168-191), an I-domain (residues 77-167) and an A-domain (residues 290-430, **Fig 2a**). Each MCP makes contact with six other surrounding MCP's (termed chain B-chain G), primarily by hydrogen bonding and a number of salt bridges adding extra stability (**Fig2b**). The length of the N-arm allows it to extend towards three surrounding MCP's, interacting with the P-domain and A-domain of chain B, the I-domain and E-loop of chain C and the P-domain of chain D, forming 23 hydrogen bonds and five salt bridges (**Fig 2c**). The P-domain is composed of a central α-helix surrounded by a glycine rich loop (G-loop), a three stranded anti-parallel β-sheet and a β-hairpin, referred to as the P-loop. This is a central point of connection with five other MCP's. The G-loop, two of the three β-strands and the P-loop collectively form 16 hydrogen bonds and two salt bridges with the I-domain and E-loop of chain

B. The loops that extend from the β-sheet form five hydrogen bonds and one salt bridge with the P-domain of chain C. The third β-strand forms six hydrogen bonds with the P-domain of chain E. A further three hydrogen bonds facilitate interactions with the N-arms of chains D and F (**Fig 2d**). The A-domain is composed of two triple-stranded β-sheets, two α-helices that are perpendicular to each other, and a β-hairpin that extends away from the domain (A-loop). The A-loop and the loop connecting the two β-sheets facilitate eleven hydrogen bonds and two salt bridges with the A-domain of chain B, as well as two hydrogen bonds and one salt bridge with the E-loop domain and three hydrogen bonds and one salt bridge with the I-domain. Two of the three β-strands closest to the A-loop are involved in 16 hydrogen bonds and one salt bridge with the A-domain of chain G. Additionally, one of the α-helices forms three hydrogen bonds with the N-arm of chain F (**Fig 2e**). The E-loop is an extended β-hairpin which connects to the I-domain. This extension allows this domain to make contact with three surrounding MCP's. The E-loop of chain A predominantly interacts with the P-domain and A-domain of chain G, through 12 hydrogen bonds. Five hydrogen bonds are formed with the N-arm of chain E and a further three hydrogen bonds with the N-arm of chain F (**Fig 2f**). The I-domain is composed of a six-stranded β-barrel and primarily makes contact with the A-domain and P-domain of chain G through seven hydrogen bonds, while a single hydrogen bond and salt bridge connects the I-domain to the N-arm of chain E (**Fig 2g**). These extensive interactions identified between MCPs likely confer stability to the capsid.

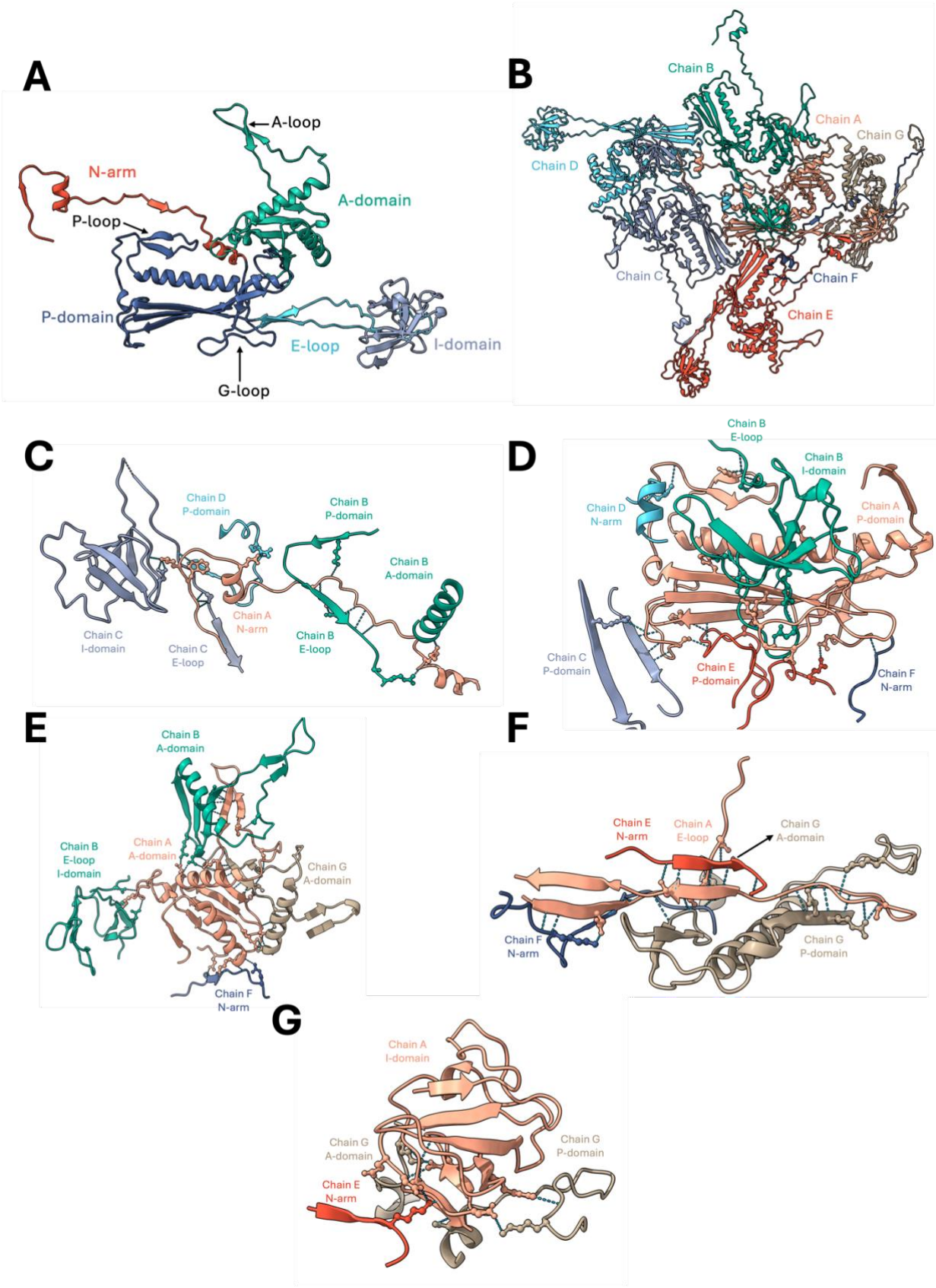



**Figure 2. Structure of MCP and interactions with surrounding chains. A)** Each MCP is composed of five domains: the N-arm (residues 2-49), the P-domain (residues 50-60, 192-289 and 431-497), the

E-loop (residues 61-76 and 168-191), the I-domain (residues 77-167) and the A-domain (residues 290-430). **B)** Each MCP chain, represented by chain A in this figure, interacts with six surrounding MCP's (termed chain B-chain G). **C)** The N-arm of chain A interacts with three MCP's, chains B-D. The N-arm interacts with the P-domain, A-domain and E-loop of chain B, the I-domain and E-loop of chain C and the P-domain of chain D. **D)** The P-domain of chain A interacts with five surrounding MCP chains. The P-domain interacts with the I-domain and E-loop of chain B, the P-domains of chain C and E and the N-arms of chain D and F. **E)** The A-domain of chain A makes contact with three surrounding MCP's. The A-domain interacts with the A-domain, E-loop and I-domain of chain B, the N-arm of chain F and the A-domain of chain G. **F)** The E-loop of chain A makes contact with three surrounding MCP chains. It interacts with the N-arms of chain E and F, and the A-domain and P-domain of chain G. **G).** The I-domain of chain A makes contact with the N-arm of chain E and the P-domain and A-domain of chain G. Bonds between chains are represented by blue dotted lines.

## Structure and interactions of auxiliary protein

The auxiliary protein is composed of two domains, the FD domain (residues 76-236) and the PH domain (residues 1-75 and 237-329, **Fig 3a**). The FD domain is composed of two anti-parallel β-sheets with four strands each and an α-helix. It connects to the A-domain of the MCP through five hydrogen bonds. The PH domain consists of two anti-parallel β-sheets, one with three strands and one with four, as well as two α-helices. The PH domain interacts with the I-domain of the MCP through three hydrogen bonds (**Fig 3b)**. Additionally, each MCP makes contact with four surrounding auxiliary proteins. The P-domain, primarily through the G-loop and the E-domain forms six hydrogen bonds and one salt bridge with the PH domain of a second auxiliary protein. The N-arm forms five hydrogen bonds primarily with the PH domain of a third auxiliary protein. The I-domain forms two hydrogen bonds with the PH domain of a fourth auxiliary protein. The N-arm and A-domain also make contact with a fifth auxiliary protein through two bonds with the FD domain (one hydrogen bond and one salt bridge, **Fig 3c**). Additional interactions with surrounding auxiliary proteins is also likely to aid in stability of the capsid.

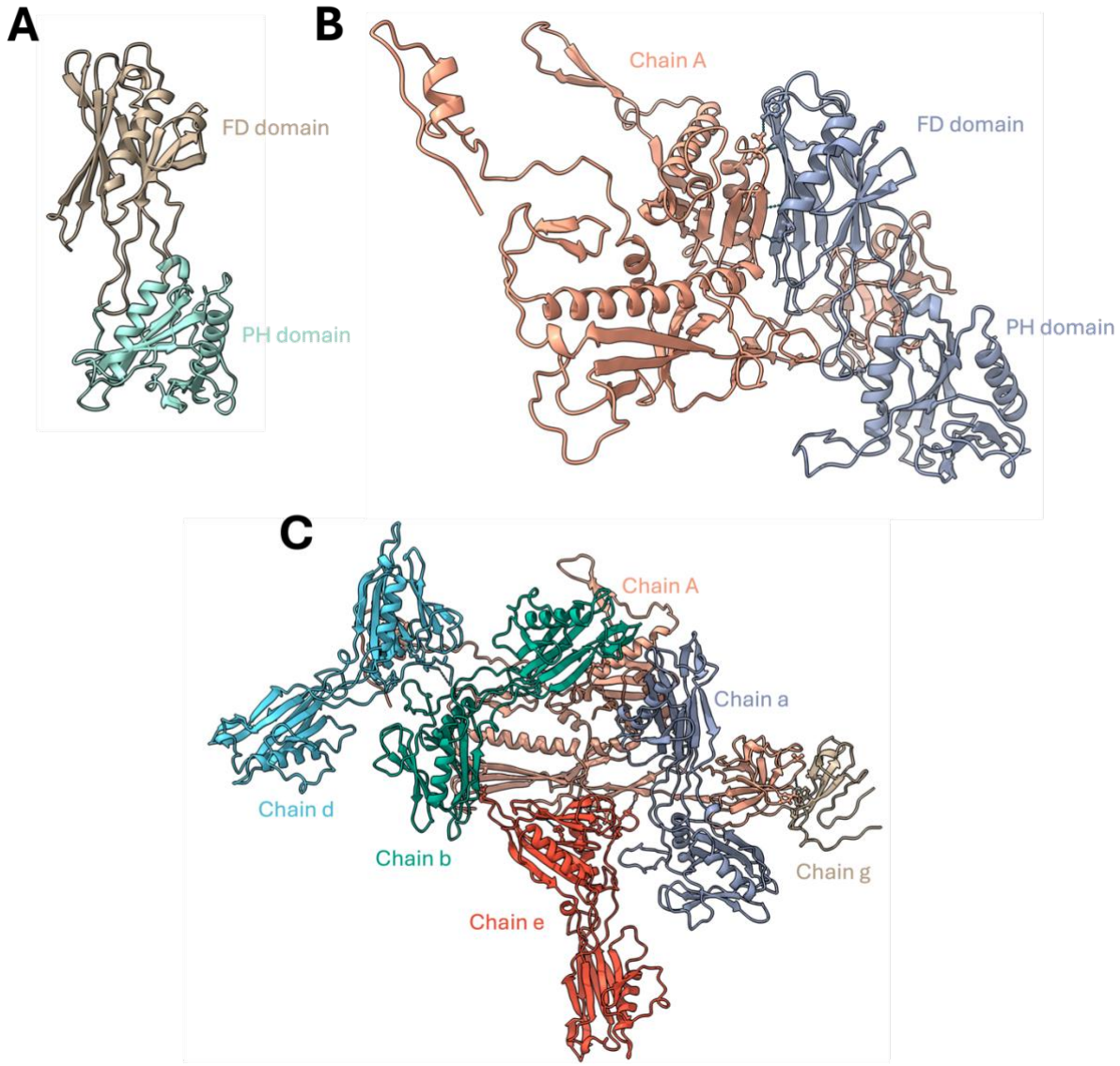



**Figure 3. Structure of auxiliary protein and interactions with MCP. A)** The auxiliary protein is composed of two domains: the FD domain (residues 76-236) and the PH domain (residues 1-75 and 237-329). **B)** Each MCP docks an auxiliary protein through the A-domain and I-domain. **C)** Each MCP interacts with an additional four surrounding auxiliary proteins. Bonds are shown as blue dotted lines.

## Structural and sequence similarity to CrAss001 MCP and auxiliary protein

The MCP has 77% protein sequence similarity with the MCP of CrAss001 (**Fig 4a**) and has a strong structural match (DALI *Z*-score 37.6, r.m.s.d 1.1, 7qoh-A, **Fig 4b**). The main difference between the two is that CrAss001 has an extended N-arm, with an extra nine residues (residues 1-9), which adds an extra loop to the N-terminal end. Additionally, DAC15 inserts an extra two residues into a loop connecting a β-

sheet and an α-helix in the A-domain (residues 299 and 300). Outside of *Crassvirales*, the MCP is structurally most similar to the MCP of podophage VP1 which infects *Vibrio cholerae*[22] (DALI *Z*-score 15.5, r.m.s.d 4.1, 9in6M). The auxiliary protein has 66% protein sequence similarity (**Fig 4c**) and a strong structural match with the auxiliary protein of CrAss001 (DALI *Z*-score 38.6, r.m.s.d 0.9, 7qoh-a, **Fig 4d**). The lowest sequence similarity is across residues 162-177. An insertion of three residues in this region in CrAss001 increases the size of the loop between two β-strands. Additionally, the auxiliary protein has structural similarity to the tail sheath protein of myophage A-1(L), which infects cyanobacteria *Anabaena* sp. PCC 7120[23] (DALI *Z*-score 9.4, r.m.s.d 3.5, 8kee-R). Overall, both the MCP and auxiliary protein of DAC15 have strong structural matches to their counterparts in CrAss001.

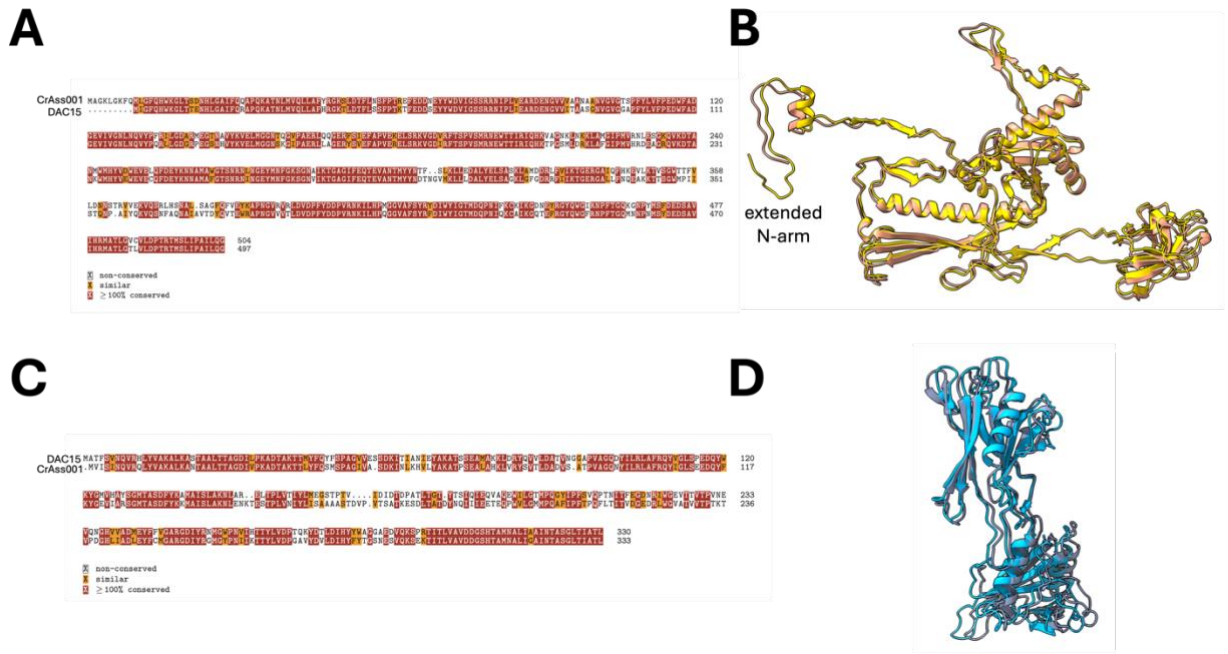



**Figure 4. Amino acid sequence and structural similarity between the MCP and auxiliary protein of DAC15 and CrAss001. A)** The MCP's share 77% sequence similarity at the amino acid level. Amino acids that are 100% conserved are coloured in red. Amino acids that are substituted with an amino acid with similar properties are coloured yellow. Where there is no match, the amino acids are left uncoloured. **B)** Structural comparison of the MCP of DAC15 (orange) and the MCP of CrAss001 (yellow). **C)** The auxiliary proteins of DAC15 and CrAss001 share 66% similarity at the amino acid level. The same colour code is used as in **A**. **D)** Structural comparison of the auxiliary protein from DAC15 (purple) and CrAss001 (blue).

Structure of HFD and comparison to CrAss001 HFD

Residues 1-50 (out of a total of 121) of the HFD could be modelled (**Fig 5a**). Each monomer is composed of a four-stranded β-sheet, with each strand connected to the next by an extended loop (**Fig 5b**). The strongest DALI match was to the N-terminal domain of a tailspike protein from a *Geobacillus* phage (DALI *Z*-score 3.2, r.m.s.d 2.4, 6e1r-E.) Each HFD is at the centre of six auxiliary proteins and the FD domain of each auxiliary protein docks one of the monomers through the formation of one or two hydrogen bonds (**Fig 5c**). AlphaFold prediction of residues 51-121 shows that each monomer forms a β-barrel that extends outwards from the capsid (**Fig 5d**). The HFD of DAC15 and CrAs001 share 61% amino acid homology and have low structural similarity (DALI *Z*-score 2.7, r.m.s.d 1.9, 7qof-l, **Fig 5e**). Comparisons of the AlphaFold predicted portion of the HFD to the AlphaFold predicted CrAss001 HFD also displays structural similarity (DALI *Z*-score 12.7, r.m.s.d 2.6). However, the angles that these β-barrel domains protrude from the capsid are predicted to be different. In DAC15, this dimer points vertically away from the capsid while in CrAss001, the two β-barrel domains are predicted to angle away from each other.

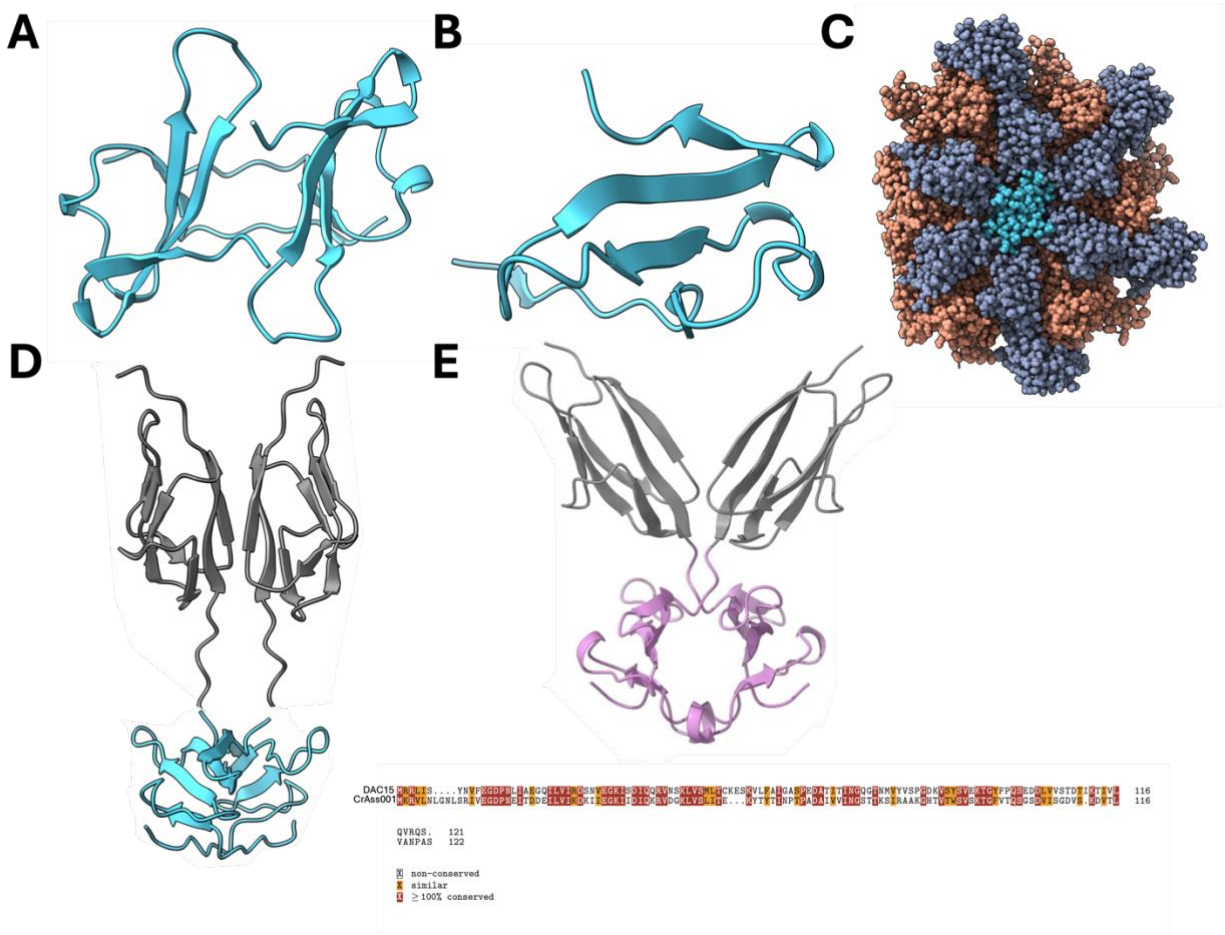



**Figure 5. Structure of HFD and comparison to the HFD of CrAss001. A)** Top view of HFD (residues 150). **B)** Side view of head fibre monomer. **C)** Each HFD is surrounded by six auxiliary proteins and are docked through hydrogen bonds with the FD domain. **D)** Atomic model of HFD (residues 1-50, blue) combined with AlphaFold model (residues 51-121, grey). **E)** Top: atomic model of CrAss001 HFD (residues 1-58, pink) combined with AlphaFold model (residues 59-122). Bottom: The HFD of DAC15 shares 61% amino acid sequence identity. Residues are coloured accordingly to levels of conservation.

## Structure of HFT and comparison to CrAss001 HFT

For the HFT, the first 39 residues could be modelled (**Fig 6a**). Each monomer is composed of a β-sheet with four strands and a loop that extends to the neighbouring monomer to add a fifth strand to the β-sheet (**Fig 6b**). Like the HFD, the HFT is bound, through hydrogen bonds, in the pocket created with six auxiliary proteins and surrounding MCP's (**Fig 6c**). There is 36% protein sequence homology between the HFT of DAC15 and the HFT of CrAss001, as well as structural similarity (DALI *Z*-score 6.6, r.m.s.d 1.2). AlphaFold modelling of residues 40-124 predicts that

each monomer is a long loop of 60 residues that extends approximately 10 nm away from the capsid surface, which is capped by an α-helix (**Fig 6d**). Residues 68-124 have structural similarity to the AlphaFold prediction of the HFT of CrAss001 (DALI *Z*-score 8.3, r.m.s.d 2.1, **Fig 6e**). DAC15 has an extra 27 residues, 26 of which are inserted at residues 42-67 and this is responsible for extending the HFT approximately 4 nm further away from the capsid surface than the HFT of CrAss001.

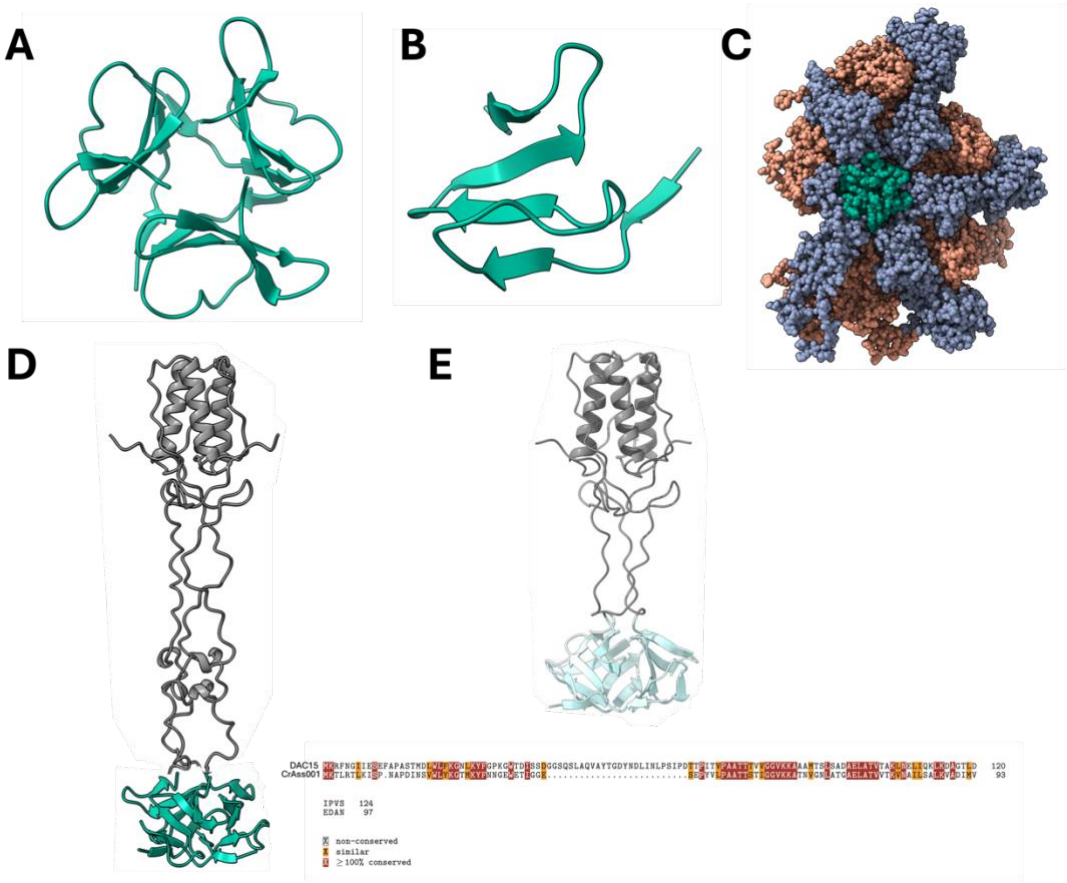



**Figure 6. Structure of HFT and comparison to the HFT of CrAss001. A)** Top view of HFT (residues 139). **B)** Side view of head fibre monomer. **C)** Each HFT is surrounded by six auxiliary proteins and are docked through hydrogen bonds with the FD domain. **D)** Atomic model of HFT (residues 1-39, green) combined with AlphaFold model (residues 40-124, grey). **E)** Top: atomic model of CrAss001 HFT (residues 1-38, pink) combined with AlphaFold model (residues 39-97). Bottom: The HFT of DAC15 shares 61% amino acid sequence identity. Residues are coloured accordingly to levels of conservation.

## Discussion

The cryo-EM reconstruction of CrAss001 was a major development in both our functional and structural understanding of *Crassvirales* phages[21]. In addition to the MCP, CrAss001 was found to have an auxiliary protein with a structure unique to *Crassvirales* phages, as well as dimeric and trimeric head fibre proteins that fit in the pockets created by MCP-auxiliary hexons[21]. Sequence analysis revealed the conservation of the MCP and auxiliary proteins across *Crassvirales* families and outgroups, while homologues of the head fibre proteins were less conserved. To build on this understanding, the capsid of *Crassvirales* phage DAC15 belonging to the same family as CrAss001, was resolved to 2.95 Å. This resolution was sufficient to build full atomic models of the MCP and auxiliary protein, as well as the N-terminal domains of the HFD (residues 1-50) and the HFT (residues 1-39). Both the MCP and auxiliary protein had strong structural matches to those in CrAss001, confirming the conservation of these proteins within the same family. The most discernable difference between the two MCP's was the insertion of an extra nine residues at the N-terminal end of the N-arm in CrAss001. Extension of the N-arm can facilitate interactions with surrounding MCP's and enhance stability[3]. However, we did not identify any additional interactions mediated by these first nine residues in CrAss001. At the time of publication, the auxiliary protein of CrAss001 was noted to have no structural homologues to any viral proteins that were currently in the PDB database. Since then, the structure of myophage A-1(L), which infects cyanobacteria *Anabaena* sp. PCC 7120, has been made available[23]. Interestingly, the auxiliary proteins of DAC15 and CrAss001 have structural similarity to the tail sheath in A-1(L). In A-1(L), subunits of the sheath cover the tail tube, with interactions between the two stabilising the subunits[23].

Structurally characterised decoration or cement proteins generally have one of five main folding motifs, referred to as β-tulip, β-tadpole, OB-fold, Ig-like, and a knotted α-helical fold[12]. Their oligomerisation states range from monomers to trimers, with trimers being the most common. Their most common function is for stability of the capsid[24,25,26], but they have also been observed to facilitate bacterial host

adhesion[11,27]. Low structural similarity was observed between the N-terminal domains of the HFD of DAC15 and CrAss001, which both bind in the hexons created by the MCP's and auxiliary proteins. The AlphaFold predictions of the C-terminal domains are also structurally similar, with the monomers of both having an Ig-like fold. However, these domains are predicted to protrude from the capsid surface at different angles. In DAC15, the Ig-like folds extend vertically, while in CrAss001, the Ig-like domains are at an angle to each other. A greater deal of structural similarity was observed between both the N-terminal and AlphaFold predicted C-terminal domains of HFT in DAC15 and CrAss001. The greatest difference between the two is a 26-residue insertion loop in the HFT of DAC15 which propels the HFT several nanometres further from the capsid than the HFT in CrAss001. The protrusion of these fibres from the capsid would suggest that they have binding capabilities, possibly to the bacterial host or to surfaces in the human gut. The conservation of the predicted C-terminal domains would suggest that the fibres in DAC15 and CrAss001 bind to similar domains. Ig-like domains on phage capsids have previously been shown to adhere to mucous, which has been proposed as a method of phage persistence in the human gut[28].

In addition to decoration or cement proteins, phages use several mechanisms to confer stability to the capsid. MCP's can connect through covalent or non-covalent bonds, such as the covalent isopeptide bond first observed in HK97, which crosslinks a conserved asparagine in the P-domain of one MCP to a conserved lysine in the E-loop of a different MCP[29]. Other phages have been identified that rely solely on intracapsomeric interactions for capsid stability[30,31]. The identification of extensive non-covalent bonds both between MCP's and between MCP's and auxiliary proteins suggest that the capsid of DAC15 is highly stable. Additionally, the conservation of the MCP and auxiliary protein subunit in DAC15 adds weight to the idea that this is a structural hallmark of *Crassvirales* phages.

## Methods

### Virus propagation and purification

A mutant of *Bacteroides thetaiotaomicron* VPI-5482 Δtdk that expresses one capsular polysaccharide only ($CPS3^+$)[32] was used for propagation of DAC15. An overnight of *B. theta* CPS3 was grown in Fastidious Anaerobe Broth (FAB) anaerobically and at 37°C. The bacterial overnight was subcultured 1:20 and infected with DAC15 when the $OD_{600}$ had reached approximately 0.2, at a multiplicity of infection (MOI) of one. A total of two litres of culture was produced, which was centrifuged at 4,500*g* for 15 minutes and filtered through 0.45 µm polyethersulfone (PES) filters. Filtered lysate was concentrated by polyethylene glycol precipitation, through the addition of PEG-8000 at 10% w/v and 0.5M NaCl, and incubated overnight at 4°C. The solution was centrifuged at 5,000*g* at 4°C for 40 minutes and the pellet resuspended in a minimum volume of SM buffer (100 mM NaCl, 8.5 mM $MgSO_4$ and 50 mM Tris-HCl, pH 7.5). The solution was applied to the top of a caesium chloride step gradient comprising of solutions of densities ~1.31 and ~1.55 g $ml^{-1}$ in 20 mM Tris-HCl pH 7.5, 10 mM $MgSO_4$. Gradients were centrifuged at 28,000 rpm in an SW28 rotor at 4 °C for 3 h. The top band, containing mature virions, was extracted by needle side-puncture. Samples were dialysed against buffer containing 20 mM Tris-HCl pH 7.5, 150 mM NaCl, 10 mM $MgSO_4$ overnight at 4 °C.

### Electron microscopy

Data was collected on a Glacios TEM instrument operating at 300kV. Two datasets were recorded and processed together in RELION 5.0[33]. Motion correction of movies was performed using RELION and contrast transfer functions were estimated using CTFFIND-4.1[34], using dose weighting. Initial 2D classifications resulted in six good classes, composed of 168,211 particles. These particles were

extracted and used for 3D refinement, imposing I3 symmetry. This was followed by CTF-refinement of per-particle defocus values and astigmatism, as well as magnification anisotropy estimation.

Atomic modelling and AlphaFold predictions

Initial atomic models were built using ModelAngelo[35], which were refined in Coot[36] and then realspace refined using Phenix[37]. Structures of unmodelled regions of HFD (residues 51-121) and HFT (residues 40-124) were predicted using AlphaFold3[38]. Structures were submitted to the DALI server[39] to search for structurally similar proteins. Images were generated in ChimeraX[40] while amino acid sequence alignments were visualised using msa[41].

## Author contributions

CAT, OWB, ANS and CH contributed to the conceptual development of this study. CAT cultivated and purified phage. OWB collected cryo-EM data. CAT and OWB processed cryo-EM data. CAT built and refined atomic models and analysed structural data. CAT wrote the manuscript with editorial contributions from ANS and CH.

Supplementary

| Data collection | | | |
|---|---|---|---|
| Dataset (optics group) | 1 | 2 | |
| Voltage/kV | 300 | 300 | |
| No. of micrographs | 14,996 | 8,666 | |
| Pixel size/Å | 1.35 | 1.35 | |
| **Data processing and refinement** | | | |
| Symmetry imposed | I | | |
| Pixel size/Å | 1.35 | | |
| No.particles | 200,159 | | |
| Map resolution, FSC0.143/Å | 2.95 | | |
| **Model composition** | **MCP-AUX protomer** | **HFD** | **HFT** |
| Non-hydrogen atoms | 12,795 | 1616 | 1890 |
| Protein residues | 825 | 100 | 117 |
| Ligands | 0 | 0 | 0 |
| **R.M.S deviations** | | | |
| Bond lengths/ Å | 0.004 | 0.003 | 0.004 |
| Bond angles/ ° | 0.692 | 0.672 | 0.699 |
| **Ramachandran plot** | | | |
| Favoured/ % | 94.03 | 96.88 | 97.3 |
| Allowed/ % | 5.6 | 3.12 | 2.7 |
| Outlier/ % | 0.37 | 0 | 0 |
| **Validation** | | | |
| Rotamer outliers/ % | 0.87 | 2.17 | 1.01 |
| Clashscore | 0.94 | 2.48 | 0 |
| Molprobity | 0.78 | 1.29 | 0.5 |

**Table S1. Data collection, processing and refinement statistics of capsid.** Refinement statistics were obtained after real space refinement in phenix.

## References


1. Fokine, A. & Rossmann, M. G. Molecular architecture of tailed double-stranded DNA phages. *Bacteriophage* **4**, e28281 (2014).

2. Nobrega, F. L. *et al.* Targeting mechanisms of tailed bacteriophages. *Nat Rev Microbiol* **16**, 760–773 (2018).

3. Duda, R. L. & Teschke, C. M. The amazing HK97 fold: versatile results of modest differences. *Curr Opin Virol* **36**, 9–16 (2019).

4. Helgstrand, C. *et al.* The Refined Structure of a Protein Catenane: The HK97 Bacteriophage Capsid at 3.44 Å Resolution. *J Mol Biol* **334**, 885–899 (2003).

5. Fokine, A. *et al.* Structural and functional similarities between the capsid proteins of bacteriophages T4 and HK97 point to a common ancestry. *Proceedings of the National Academy of Sciences* **102**, 7163–7168 (2005).

6. Rizzo, A. A. *et al.* Multiple Functional Roles of the Accessory I-Domain of Bacteriophage P22 Coat Protein Revealed by NMR Structure and CryoEM Modeling. *Structure* **22**, 830–841 (2014).

7. Tso, D., Hendrix, R. W. & Duda, R. L. Transient Contacts on the Exterior of the HK97 Procapsid That Are Essential for Capsid Assembly. *J Mol Biol* **426**, 2112–2129 (2014).

8. Louten, J. Virus Structure and Classification. in *Essential Human Virology* 19–29 (Elsevier, 2016). doi:10.1016/B978-0-12-800947-5.00002-8.

9. Lander, G. C. *et al.* Bacteriophage Lambda Stabilization by Auxiliary Protein gpD: Timing, Location, and Mechanism of Attachment Determined by Cryo-EM. *Structure* **16**, 1399–1406 (2008).

10. Qin, L., Fokine, A., O'Donnell, E., Rao, V. B. & Rossmann, M. G. Structure of the Small Outer Capsid Protein, Soc: A Clamp for Stabilizing Capsids of T4-like Phages. *J Mol Biol* **395**, 728–741 (2010).

11. Xiang, Y. & Rossmann, M. G. Structure of bacteriophage $\phi$ 29 head fibers has a supercoiled triple repeating helix-turn-helix motif. *Proceedings of the National Academy of Sciences* **108**, 4806–4810 (2011).

12. Dedeo, C. L., Teschke, C. M. & Alexandrescu, A. T. Keeping It Together: Structures, Functions, and Applications of Viral Decoration Proteins. *Viruses* **12**, 1163 (2020).

13. Shkoporov, A. N. *et al.* ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. *Nat Commun* **9**, (2018).

14. Guerin, E. *et al.* Isolation and characterisation of ΦcrAss002, a crAss-like phage from the human gut that infects Bacteroides xylanisolvens. *Microbiome* **9**, (2021).

15. Hryckowian, A. J. *et al.* Bacteroides thetaiotaomicron-Infecting Bacteriophage Isolates Inform Sequence-Based Host Range Predictions. *Cell Host Microbe* **28**, (2020).

16. Papudeshi, B. *et al.* Host interactions of novel Crassvirales species belonging to multiple families infecting bacterial host, Bacteroides cellulosilyticus WH2. *Microb Genom* **9**, (2023).

17. Ramos-Barbero, M. D. *et al.* Characterization of crAss-like phage isolates highlights Crassvirales genetic heterogeneity and worldwide distribution. *Nat Commun* **14**, (2023).

18. Koonin, E. V. & Yutin, N. The crAss-like Phage Group: How Metagenomics Reshaped the Human Virome. *Trends Microbiol* **28**, 349–359 (2020).

19. Shkoporov, A. N. *et al.* Long-term persistence of crAss-like phage crAss001 is associated with phase variation in Bacteroides intestinalis. *BMC Biol* **19**, (2021).

20. Cortés-Martín, A. *et al.* Adaptations in gut Bacteroidales facilitate stable coexistence with their lytic bacteriophages. Preprint at https://doi.org/10.1101/2024.11.17.624012 (2024).

21. Bayfield, O. W. *et al.* Structural atlas of a human gut crassvirus. *Nature* **617**, (2023).

22. Pang, H. *et al.* Three-dimensional structures of Vibrio cholerae typing podophage VP1 in two states. *Structure* **32**, 2364-2374.e2 (2024).

23. Yu, R.-C. *et al.* Structure of the intact tail machine of Anabaena myophage A-1(L). *Nat Commun* **15**, 2654 (2024).

24. Forrer, P., Chang, C., Ott, D., Wlodawer, A. & Plückthun, A. Kinetic Stability and Crystal Structure of the Viral Capsid Protein SHP. *J Mol Biol* **344**, 179–193 (2004).

25. Qin, L., Fokine, A., O'Donnell, E., Rao, V. B. & Rossmann, M. G. Structure of the Small Outer Capsid Protein, Soc: A Clamp for Stabilizing Capsids of T4-like Phages. *J Mol Biol* **395**, 728–741 (2010).

26. Subramanian, S., Bergland Drarvik, S. M., Tinney, K. R. & Parent, K. N. Cryo-EM structure of a Shigella podophage reveals a hybrid tail and novel decoration proteins. *Structure* **32**, 24-34.e4 (2024).

27. Sathaliyawala, T. *et al.* Functional analysis of the highly antigenic outer capsid protein, Hoc, a virus decoration protein from T4-like bacteriophages. *Mol Microbiol* **77**, 444–455 (2010).

28. Barr, J. J. *et al.* Bacteriophage adhering to mucus provide a non–host-derived immunity. *Proceedings of the National Academy of Sciences* **110**, 10771–10776 (2013).

29. Wikoff, W. R. *et al.* Topologically Linked Protein Rings in the Bacteriophage HK97 Capsid. *Science (1979)* **289**, 2129–2133 (2000).

30. Hryc, C. F. *et al.* Accurate model annotation of a near-atomic resolution cryo-EM map. *Proceedings of the National Academy of Sciences* **114**, 3103–3108 (2017).

31. Guo, F. *et al.* Capsid expansion mechanism of bacteriophage T7 revealed by multistate atomic models derived from cryo-EM reconstructions. *Proceedings of the National Academy of Sciences* **111**, (2014).

32. Hickey, C. A. *et al.* Colitogenic Bacteroides thetaiotaomicron antigens access host immune cells in a sulfatase-dependent manner via outer membrane vesicles. *Cell Host Microbe* **17**, 672–680 (2015).

33. Scheres, S. H. W. RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *J Struct Biol* **180**, (2012).

34. Rohou, A. & Grigorieff, N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J Struct Biol* **192**, (2015).

35. Jamali, K. *et al.* Automated model building and protein identification in cryo-EM maps. *Nature* **628**, 450–457 (2024).

36. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**, 486–501 (2010).

37. Adams, P. D. *et al.* PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* **66**, 213–221 (2010).

38. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

39. Holm, L. Dali server: structural unification of protein families. *Nucleic Acids Res* **50**, W210–W215 (2022).

40. Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Science* **30**, 70–82 (2021).

41. Bodenhofer, U., Bonatesta, E., Horejš-Kainrath, C. & Hochreiter, S. msa: an R package for multiple sequence alignment. *Bioinformatics* **31**, 3997–3999 (2015).

# Thesis summary and future work

Phages of the human gut are highly abundant, with estimates ranging from $10^9$ - $10^{10}$ VLP's per gram of faeces[1]. This sheer abundance makes it possible for phages to exert considerable influence over their microbial environment, through interactions with their bacterial hosts. However, despite this abundance, phages remain understudied compared to their bacterial counterparts. There are a number of issues that contribute to this. Isolating phage in pure culture first requires the successful cultivation of its bacterial host, many of which are strict anaerobes and are therefore difficult to culture. Additionally, many phages are strain specific. The lack of a universal genetic marker, the mosaic nature of phage genomes and rapid mutation rates results in phages often being highly divergent from phages in reference sequence databases. In metagenomic studies, this can often result in a large proportion of phages remaining unclassified, precluding any meaningful analysis. This issue was highlighted in chapter two, where approximately 53% of gut VC's identified in this study could not be classified. This problem was further exacerbated in the oral virome, where the percentage of unclassified viruses was approximately 82%. Additionally, while the movement away from morphology-based classification and towards genomic classification is necessary, it has left a large number of phages floating as unclassified *Caudoviricetes*. This will no longer be an issue in the future as new taxonomic families based on accurate evolutionary relationships are formed, but it hampers meaningful based family comparisons in the interim.

The discovery that the virome and bacteriome from both the gut and oral cavity of patients with acute COVID-19 was largely similar to those in recovery states was interesting. Alterations to the gut bacteriome of acute patients have been identified in numerous studies, characterised by reduced diversity, loss of beneficial bacteria such as SCFA producers and an enrichment of opportunistic pathogens[2,3,4]. Additionally, a number of studies have reported that this perturbation to gut bacteriome persists many months post infection and has been linked to the development of Long Covid[5,6,]. In the context of existing literature, similar gut

viromes and bacteriomes in acute and convalescent patients suggest that alterations to both due to COVID infection is also persisting over time. However, this study was hindered by the lack of available healthy controls, which would be required to conclude if the microbiome of these patients was altered due to COVID infection.

Four chapters of this thesis are concentrated on the most abundant phages of the human gut, the *Crassvirales*. They infect bacteria of the phylum *Bacteroidetes*, mainly of the family *Bacteroidaceae*. These bacteria account for a large proportion of the gut bacteriome and have been documented to perform multiple beneficial roles in the gut, including protection from pathogens through colonisation resistance and biosynthesis of vitamins[7]. These chapters delve into the structure and function of two *Crassvirales* phages, CrAss001 and DAC15, building on the framework of earlier studies of these two phages. CRISPRi was used to target nine predicted essential and non-essential genes in DAC15. Minimal reductions in EOP and phage titres was observed for all nine genes, which led to investigating the DAC15 genome for potential defence strategies to CRISPR. Through a combination of bioinformatics and structural homology, five Acrs were predicted in the genome of DAC15. This predictive strategy was extended to CrAss001, CrAss002 and p-crAssphage, identifying 16 more putative Acrs across the three genomes. It had previously been hypothesised that *Crassvirales* phages encode Acrs but machine learning methods at the time failed to identify any in *Crassvirales* genomes[8]. The prediction of these Acrs highlights the importance of characterising the large number of genes of unknown function in phage genomes. Future work should be focused on experimental verification of these Acrs. Despite a minimal reduction in EOP and phage titre, CRISPRi targeting of TerL resulted in an enrichment of empty phage capsids. This production of empty phage capsids could be used in future studies focused on the mechanism of DNA packaging in *Crassvirales* phages.

Characterisation of structural genes of DAC15 was further built upon in chapter six, with the reconstruction of the capsid of DAC15 to 2.95 Å. This resolution allowed for complete atomic models of the MCP and auxiliary protein to be built, as well as partial models for the HFD and HFT, which were combined with AlphaFold

predictions for the C-terminal domain. Both the MCP and auxiliary proteins had strong structural matches to their counterparts in CrAss001. Additionally, numerous non-covalent interactions (primarily hydrogen bonds and salt bridges) were identified between MCP subunits and between MCP and auxiliary proteins, which are likely to make the capsid of DAC15 highly stable. Some structural differences between the HFD and HFT of DAC15 and CrAss001 were observed. The C-terminal Ig-like domains of the HFD of DAC15 extend vertically from the capsid, facing each other, while these same domains point away from each other in CrAss001. Furthermore, the HFT of DAC15 has a 26 residue insert that allows it to extend several nanometres further from the capsid than the HFT of CrAss001. The structural similarity of these two structures across DAC15 and CrAss001, as well as their protrusion from the capsid, suggests that these proteins may bind to their bacterial host, or aid in persistence through binding to structures in the gut, such as glycan[9]. Future studies should focus on structural models of different families of *Crassvirales*, as sequence homology suggests that some structural proteins are less conserved in some families, such as the auxiliary protein in the family *Crevaviridae*[10]. However, this will first require isolation of representatives of the *Crevaviridae* and *Suoliviridae* families. Currently, there are three *Crassvirales* phages from the *Intestiviridae* family isolated in pure culture. Previous attempts by us to get a high enough titre of one member of the *Intestiviridae* family, CrAss002, for cryo-EM analysis failed. However, the recent isolation of two more *Intestiviridae* phages that have biological characteristics more similar to *Steigviridae* phages[11], may be successful candidates for future structural studies.

Despite the importance of TF's and TSP's in determining host specificity, the structures of these proteins are often not resolved in cryo-EM reconstructions of phages. This is due to the flexibility of these TF's and TSP's compared to the main phage structure, which results in them being averaged out during image processing. To address this issue in CrAss001, we used a combination of structural, biochemical and bioinformatic techniques to determine the structure and function of three TSP's. The inner TSP assembly is composed of two proteins, gp22 and gp23, whose structure was determined to 3.4 Å. Gp22_23 was demonstrated to act as a

depolymerase, forming zones of clearing on sensitive cultures of *B. intestinalis* that predominately expresses CPS PVR9. Gp22_23 TSP was unable to form zones of clearing on a second insensitive culture of *B. intestinalis* that primarily expresses CPS PVR7, suggesting that PVR9 is the receptor for gp22_23 TSP. Gp25, which makes up part of the outer TSP assembly, was resolved to 2.6 Å and was demonstrated to exhibit esterase activity, likely via its SGNH domain. The structures of the remaining two proteins that make up the outer TSP assembly, gp24 and gp26, were predicted by AlphaFold. Based on these structures, gp24 is likely to act as a linker protein for gp25 and gp26. It's C-terminal domain was predicted by AlphaFold to interact with the THA dimer that encircles the tail barrel. Gp26 appears to be a third TSP, with an enzymatic domain that may act on peptidoglycan.

This structural and biological work allowed us to create a schematic of infection for CrAss001. Gp22_23 binds to the first receptor, which is likely to be CPS PVR9. This exposes the second receptor (possibly LPS) for gp25 to bind and hydrolyse. These combined actions allow the tail to dock to the outer membrane, where cargo proteins can be ejected from the tail to form a transmembrane channel into the cytoplasm of *B. intestinalis*. The function of gp26 in this process is currently undetermined. To degrade peptidoglycan, gp26 would have to transverse the outer membrane. Additionally, when gp26 was incubated with CPS from WT *B. intestinalis,* it was able to produce some breakdown products but produced no zones of clearing on overlays of *B. intestinalis* expressing PVR7 or PVR9. Gp26 may have the ability to degrade alternative CPS of *B. intestinalis* (possibly PVR8, PVR11 or PVR12) as an additional mechanism of phage persistence. Future studies should focus on the role of gp26, as well as confirm binding of gp22_23 to PVR9 and gp25 to LPS.

The development of AI tools such as AlphaFold have revolutionised the field of structural biology, providing the structure of a protein in minutes, which can sometimes take up to years to solve experimentally. However, like any AI tool, AlphaFold has its limitations, and caution should be advised on basing entire structures on AlphaFold modelling alone. Instead, its strength lies in its assistance

to structural methods such as X-ray crystallography and cryo-EM, providing the starting structure to these methods, from which atomic models can be built.

Overall, the work in this thesis advances our understanding of several aspects of the gut phageome, primarily through structural and biological work focused on *Crassvirales* phages. Predictions of Acrs in the genomes of *Crassvirales* highlights the complexity of these phage host interactions, while structures of capsids and TSP's provide further insights into the mechanisms of persistence of these phages with their hosts. This builds on over a decade of work focused on *Crassvirales* phages, further establishing their importance in the human gut.

## References


1. Hoyles, L. *et al.* Characterization of virus-like particles associated with the human faecal and caecal microbiota. *Res Microbiol* 165, 803–812 (2014).
2. Zhang, F. *et al.* Gut microbiota in COVID-19: key microbial changes, potential mechanisms and clinical applications. *Nat Rev Gastroenterol Hepatol* 20, (2023).
3. Zuo, T. *et al.* Alterations in Gut Microbiota of Patients With COVID-19 During Time of Hospitalization. *Gastroenterology* 159, (2020).
4. Zhou, T. *et al.* SARS-CoV-2 triggered oxidative stress and abnormal energy metabolism in gut microbiota. *MedComm (Beijing)* 3, (2022).
5. Chen, Y. *et al.* Six-month follow-up of gut microbiota richness in patients with COVID-19. *Gut* 71, 222–225 (2022).
6. Liu, Q. *et al.* Gut microbiota dynamics in a prospective cohort of patients with post Acute COVID-19 syndrome. *Gut* 71, (2022).
7. Zafar, H. & Saier, M. H. Gut *Bacteroides* species in health and disease. *Gut Microbes* 13, (2021).
8. Yutin, N. *et al.* Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features. *Nat Commun* 12, (2021).
9. Barr, J. J. *et al.* Bacteriophage adhering to mucus provide a non–host-derived immunity. *Proceedings of the National Academy of Sciences* 110, 10771–10776 (2013).
10. Bayfield, O. W. *et al.* Structural atlas of a human gut crassvirus. *Nature* 617, (2023).
11. Papudeshi, B. *et al.* Host interactions of novel Crassvirales species belonging to multiple families infecting bacterial host, Bacteroides cellulosilyticus WH2. *Microb Genom* 9, (2023).

# Acknowledgements

Firstly, I would like to thank my supervisors Professor Colin Hill and Professor Andrey Shkoporov for the opportunity to carry out this PhD in your labs. Thank you for the discussions, guidance, advice and support over the last four and a half years. A special thanks to Andrey for taking the time to help me get started in the lab and teaching me invaluable research skills when I first started.

I would like to extend my gratitude to Oli Bayfield, who has been a great mentor to me during this PhD. I am truly grateful for the time you spent teaching me cryo-EM data processing and atomic model building over zoom calls and for all the structural advice you gave me.

Thank you to everyone in the phage lab, both past and present. I have been lucky to work with such a great and supportive group of people. There are a number of people I would like to thank in particular. Fabian, thank you for sharing cat photos that have always cheered me up. Sam, thank you for the lab chats and weekend company. Matthew, thanks for always being interested in my work and for keeping me on my toes with all your questions. Fergal, thanks for the chats and for being a great desk neighbour. Katie, thanks for your visits to the office corner, your chats have always been a welcome distraction. Dave, thanks for your enthusiasm any time you saw a structure on my screen. Lauren, thank you for being both so thoughtful and hilarious. I always enjoyed your visits to the office corner. To the knights of the BSI, thank you for your friendship. Our movie and board games nights were always a welcome break from science.

Hiba, thank you for always being in my corner and for being such a wonderful friend. Thank you for listening to every panicked thought or rant I had and for keeping me consistently fueled with Venchi chocolate. Ellen, thank you for being such a thoughtful and caring friend. You've always given me the best advice and your treats and little notes of encouragement have always cheered me up. I am so

grateful that doing this PhD allowed me to become friends with two such amazing people.

To my oldest and dearest friends, Juana, Aisling and Diana. Your friendship has always given me a break from all things science and I'm so grateful for your unwavering support and understanding.

Finally, to my parents. Your encouragement and support have been invaluable to me. Dad, thank you for the countless lifts and for always believing I'm an expert (no matter the science discipline). Mom, thank you for always being there and ready to listen to whatever story or rant I have.

I am truly grateful for all the advice and support I received during this time.