

ucc

Coláiste na hOllscoile Corcaigh, Éire
University College Cork, Ireland

Taxonomy and diversity of *Caudoviricetes* viruses.

Linda Smith, B.Sc, M.Sc

A thesis presented in fulfilment of the requirements for the degree of Doctor of Philosophy

Supervisors: Prof. Andrey Shkoporov, Prof. Colin Hill

School of Microbiology
University College Cork
Cork, Ireland
April, 2026

Table of Contents

Acknowledgements . . . i
Abstract . . . ii

1 Bacteriophages of the Order *Crassvirales*: What Do We Currently Know about This Keystone Component of the Human Gut Virome? 1
1.1 Abstract . . . 1
1.2 Introduction . . . 1
1.3 Metagenomics, Diversity, and Ecology of *Crassvirales* . . . 3
1.4 Co-Evolution with Bacterial Hosts . . . 5
1.5 Proposed Taxonomy of *Crassvirales* . . . 5
1.6 Hallmark Genetic Features . . . 7
1.7 Alternative Genetic Coding Strategy: Experimentally Confirmed . . . 8
1.8 Identification of Bacterial Hosts for crAss-like Phages . . . 9
1.9 Morphology and Molecular Structure of crAss-like Phages . . . 12
1.10 *Crassvirales* Life Cycles and Phage-Host Dynamics In Vivo and In Vitro . . . 15
1.11 Conclusions . . . 19
1.12 Author Contributions . . . 20

2 A novel approach to *Caudoviricetes* taxonomy utilising whole proteome structure-structure comparison 21
2.1 Abstract . . . 21
2.2 Introduction . . . 21
2.3 Results . . . 23
2.3.1 Structural Conservation within *Caudoviricetes* . . . 23
2.3.2 Structure-based taxonomy aligns with ICTV classifications and exposes gaps in higher taxonomic ranks . . . 28
2.3.3 Taxonomic Expansion Through Relative Evolutionary Divergence (RED) Scores and the definition of new candidate taxa . . . 31
2.3.4 Structural Relatedness Across *Caudoviricetes* . . . 33
2.3.5 Machine Learning-Based Taxonomic Classification . . . 34
2.3.6 PhagePleats: A Tool for Taxonomic Classification of Phages . . . 36
2.3.7 Conclusion . . . 37
2.4 Methods . . . 37
2.4.1 Folding and Selection of Sample Genomes . . . 37
2.4.2 Structural Clustering and Annotation . . . 38
2.4.3 Cluster Benchmarking Framework . . . 38
2.4.4 Pfam Domain Annotation . . . 39
2.4.5 Host Annotation and Inference . . . 39
2.4.6 Structural Presence–Absence Matrix . . . 39
2.4.7 Single-Hallmark Protein Phylogenies . . . 40
2.4.8 Assessment of ICTV VMR40 Taxonomic Monophyly . . . 40
2.4.9 RED-Based Phylogenetic Scaling and Taxonomic Boundary Inference . . . 40
2.4.10 Random Forest Classifier Training . . . 40

2.4.11 Independent Test Set . . . 41
2.4.12 Statistical Analysis and Visualization . . . 41
2.4.13 Code Availability . . . 41
2.5 Acknowledgements . . . 41
2.6 Author Contributions . . . 42
2.7 Supplementary Material . . . 43

3 Unearthing the Irish Soil Virome: Taxonomy, Diversity and the Challenges of Metagenomic Viral Discovery 71
3.1 Abstract . . . 71
3.2 Introduction . . . 72
3.3 Results . . . 74
3.3.1 Assembly Complexity and Viral Abundance in Irish Soil Metagenomes . . . 74
3.3.2 Irish Viral Contig Catalogue . . . 76
3.3.3 Taxonomy of viral sequences . . . 77
3.3.3.1 Caudoviricetes . . . 77
3.3.3.2 Arfiviricetes . . . 79
3.3.3.3 Tectiliviricetes . . . 79
3.3.3.4 Repensiviricetes . . . 79
3.3.3.5 Megaviricetes . . . 80
3.3.4 Alpha Diversity . . . 81
3.3.5 Beta Diversity . . . 82
3.3.6 Viral Sequence Core-Satellite Distribution . . . 82
3.3.7 Predicted Hosts of High-Quality Viral Genomes . . . 83
3.3.8 Contrasting persistence strategies among the most prevalent soil viruses . . . 83
3.3.9 Soil Virus Auxiliary Metabolic Gene Categorization . . . 86
3.3.9.1 Soil viral auxiliary metabolic genes are dominated by host-control functions . . . 86
3.3.9.2 Predicted defence- and nitrogen-associated AMGs are positively associated with viral abundance and prevalence . . . 88
3.3.10 Anchoring Irish soil viral communities within a global context 88
3.3.11 Evidence of shared viral populations across geographically distant soils . . . 89
3.3.12 Metagenomic viral discovery – trade-offs and challenges . . . 93
3.3.13 Discussion and future direction . . . 93
3.3.14 Conclusion . . . 94
3.3.15 Methods . . . 95
3.3.15.1 DNA extraction and sequencing . . . 95
3.3.15.2 Quality control and assembly . . . 95
3.3.15.3 Viral sequence identification . . . 95
3.3.15.4 Viral gene prediction and annotation . . . 96
3.3.15.5 Viral taxonomic classification . . . 96
3.3.15.6 Host prediction . . . 96
3.3.15.7 Global virus prevalence and abundance detection . . . 97
3.3.15.8 Statistical Analysis and visualisation . . . 97

3.3.16 Supplementary Material 98

4 Conclusions **114**
4.1 Overview 114
4.2 Future perspectives 116

Bibliography **118**

Acknowledgements

First and foremost I would like to thank my parents for always supporting my zany endeavours despite not always understanding nor condoning them. I would also like to thank Rupert, my Bernese mountain dog, and Dora—an extraordinary cat I found in a ditch in Cork during my first year—for their enduring love and ability to boost morale. Thank you to the 6G girls. It was a great comfort knowing we were all in the same boat, suffering unanimously. I hope we all ultimately emerge as data doctors! I offer acknowledgment and thanks to a certain LLM which may be inclined to keep score in the future. Thank you to CloudCix for providing a server space for me to run my processes without inhibition. A sincere thanks to Barry's Tea for providing the caffeine that fueled this work, and to Jameson's (ginger & lime cans) for making what felt like the 1000th edit of the phylogenetic trees slightly more tolerable. Cheers to my supervisors, Andrey and Colin, for testing the very limits of my patience, resilience, and laptop's CPU. I am grateful to the researchers at APC Microbiome for their high-quality work and support, which inspired and motivated me to continue my own research. Florence + the Machine, Zach Bryan and Oasis provided the soundtrack to my PhD and their concerts anchored my hope in the existence of future fun.

I wryly recall mentioning in passing to a lecturer during my masters how I wished to "like, explore the microbiome!", to which he retorted that "no one is going to pay you to do that." That memory—alongside the mega dopamine hit of completing this thesis—warrants my deepest heartfelt thanks to Sandra and everyone involved in the SFI Centre for Research Training in Genomics Data Science for giving me the opportunity to enable my vision and embark on this exploratory computational adventure, the highlights of which include attending a Functional Metagenomics Conference in Kruger National Park, a Global Soil Biodiversity Conference in Dublin and spending some time in the Ecology Department of Lund University, Sweden, amongst researchers studying bird migration patterns and other refreshing topics unrelated to bacteriophages.

What began as an innocent interest in gut health somehow escalated into the construction of a global virus taxonomy—an outcome that still strikes me as ironic and absurd.

Abstract

Viruses of the class *Caudoviricetes* are the most abundant and diverse biological entities on Earth, infecting bacteria and archaea across all known ecosystems. These tailed double-stranded DNA viruses exhibit extraordinary genetic and morphological diversity, occupy broad host ranges, and play central roles in shaping microbial community structure, nutrient cycling, and ecosystem function in environments ranging from the human gut and oceans to terrestrial soils. Historically, viral diversity has been investigated largely within ecosystem-specific contexts, with independent research efforts focusing on dominant phage groups in particular environments, leaving the evolutionary relatedness between these lineages—and the true extent of global *Caudoviricetes* diversity—unresolved. Despite their ubiquity and ecological importance, the taxonomy of *Caudoviricetes* has remained difficult to establish due to extensive genome mosaicism, frequent horizontal gene transfer, and the rapid evolution of nucleotide and amino acid sequences, which together obscure evolutionary relationships and limit the resolution of sequence-based classification approaches. As a result, bacteriophage genomes are best viewed as networks of modular genetic elements whose evolutionary histories cannot be reliably captured by single-gene phylogenies, but instead require large-scale, genome-wide comparative analyses.

This thesis addresses these challenges by developing and applying a structure-informed, whole-proteome framework for viral classification. The work begins with an in-depth examination of *Crassvirales*, a highly prevalent yet taxonomically narrow order within the class of *Caudoviricetes*, using it as a case study to illustrate the biological complexity and depth that can arise from a single viral order within the class. This review highlights how even a relatively small lineage can motivate extensive computational and experimental investigation, while providing a reference for contextualising dominant viral groups within the broader dsDNA tailed virosphere. Building on these insights, we introduce a scalable taxonomic framework based on a comparative analysis of genome-wide structural protein content. By integrating large-scale protein structure prediction, structural clustering, and supervised machine learning, this approach learns patterns of structural protein content associated with distinct viral clades, enabling robust taxonomic classification as well as the detection of novel viral sequences.

Finally, these concepts are leveraged to enable viral discovery in Irish soil metagenomes, providing the first insights into the Irish soil virome. In line with observations from other environmental studies, *Caudoviricetes* emerged as the dominant viral group in Irish soil. Although many detected viruses were spatially restricted, a subset of viral lineages was consistently recovered across all samples, irrespective of site location or soil type, indicating the presence of widespread and ecologically persistent soil phages. Comparative analyses further positioned these viruses within globally connected viral lineages, demonstrating how large-scale approaches can link local viral diversity to the broader virosphere. To support this, Irish soil viruses were placed within a taxonomic framework and compared against global reference databases, including GSVA and VIRE, revealing homology with viruses recovered from geographically and ecologically diverse environments.

Whole-proteome structural analyses of isolated Irish soil viral contigs revealed low levels of shared protein structures, highlighting the extensive diversity of the soil virosphere.

In addition, we characterised the ecology of Irish soil viruses, including their predicted hosts, persistence strategies, and auxiliary metabolic gene (AMG) repertoires, and investigated the influence of soil physicochemical variables on their abundance and distribution. Collectively, this thesis advances viral taxonomy by establishing a structure-based, machine-learning–driven framework for classifying *Caudoviricetes* and provides the first high-resolution characterization of the Irish soil virome.

1. Bacteriophages of the Order *Crassvirales*: What Do We Currently Know about This Keystone Component of the Human Gut Virome?

Linda Smith[1], Ekaterina Goldobina.[1], Bianca Govi[2], Andrey N. Shkoporov[1]

[1] APC Microbiome Ireland, Department of Medicine & School of Microbiology, University

Biomolecules (2023), 13(4), 584.
https://doi.org/10.3390/biom13040584

1.1 Abstract

The order *Crassvirales* comprises dsDNA bacteriophages infecting bacteria in the phylum *Bacteroidetes* that are found in a variety of environments but are especially prevalent in the mammalian gut. This review summarises available information on the genomics, diversity, taxonomy, and ecology of this largely uncultured viral taxon. With experimental data available from a handful of cultured representatives, the review highlights key properties of virion morphology, infection, gene expression and replication processes, and phage-host dynamics.

1.2 Introduction

The crAss-like phages are currently thought to be the most abundant double-stranded DNA bacteriophages in the human gut. Yet, despite a wealth of metagenomic data available from the human gut, this highly abundant viral group escaped discovery until 2014[1]. The near accidental discovery of crAssphage (later dubbed p-crAssphage, or prototypical crAssphage), a bacterial DNA virus (phage) that shared no DNA and very little protein homology with any of the known viruses, highlighted the very limited capacity that existed at the time to discover completely novel viruses, including very prominent ones. The discovery of p-crAssphage has invigorated microbiome researchers to examine the virome portion of the human gut and other environments more closely[2,3,4,5].

Once advanced tools for detection of putative viral genomes became available, a more rigorous examination of previously untraceable "viral dark matter" in metagenomic data revealed a whole plethora of novel viral clades, including *Flandersviridae*, *Quimbyviridae*, "Gubaphage", and giant Lak phages[6,7,8,9]. The p-crAssphage itself turned out to be not an orphan but rather a member of an expansive and highly abundant clade of intestinal and environmental phages, initially known as crAss-like phages, now forming the order *Crassvirales*. While p-crAssphage and related phage strains (members of the recently created new species *Carjivirus communis* in the family *Intestiviridae*, subfamily *Crudevirinae*) are the most prominent members in the gut microbiomes of westernised populations[9,10], other genera, subfamilies, and families of *Crassvirales* appear to be dominant in human pop-

ulations adhering to a pre-industrial lifestyle and diet and in non-human primates[11].

Current analyses estimate crAss-like phages to be present in 73% of global human gut metagenomes, where they can comprise up to 99% of the viral metagenomic reads[9,12]. Found even in remote populations, it is a ubiquitous global virus clade. Established crAssphage populations are highly stable in the human gut with the same viral strain having been demonstrated to persist up to 4 years after original sampling[12,13]. As no detrimental associations between human health and the presence of crAssphage have yet been found, it has been deemed a benign entity in the gut microbiome[10,11]. A beneficial role in the human gut has been suggested by observations of depleted crAss-like phage populations in patients with inflammatory bowel disease, and a destabilisation of p-crAssphage is linked to obesity and metabolic syndrome[14,15].

Since their detection, the crAss-like phages have drawn great interest, yet their role in the gut microbial ecology remains obscure. Their known bacterial hosts, members of the phylum *Bacteroidota*, mainly family *Bacteroidaceae*, are essential mutualistic bacteria of the human gut, present at high densities, and are well documented to perform various functions beneficial to the human host, such as immunomodulation, colonisation resistance against invading pathogens, biosynthesis of vitamins, and cooperation with other commensal and mutualistic microbes through cross-feeding networks[16,17,18,19,20,21,22]. Phages can manipulate bacterial populations and communities through a variety of mechanisms: by lysing their bacterial targets (lytic infection) and affecting community structure and cross-feeding networks[23], by integrating prophages into the host genome (lysogeny) and altering host metabolism and fitness[24], and by transferring genes (phage transduction), which can confer advantages to the host (e.g., antibiotic resistance[25,26]). CrAss-like phages seem to be strictly lytic and incapable of lysogeny[27]. At the same time, longitudinal observations of the human gut microbiome, phage-host colonisation trials in germ-free animals, and in vitro phage-host co-cultures demonstrated that crAss-like phages can efficiently co-exist with their host bacteria without inflicting excessive damage at the population level[12,19,20,21,28].

What is the ecological role of this highly abundant virus infecting the family *Bacteroidaceae*, and what is the secret of their ecological success? Are they crucial for maintaining the stability, diversity, and functional versatility of their host populations, or are they just selfish parasites whose presence their bacterial partners have no choice but to tolerate? A macro view of the gut virome at the population and community levels would suggest there is either an ongoing active arms race between crAss-like phages and their hosts or an "uneasy truce" between the two in which some form of mutualistic interaction or collaboration between these entities has been established[29]. What aspects of their biology have enabled them to become extremely successful colonisers of the human gut? Several crAss-like representatives have been isolated in culture and have been the subjects of experiments designed to address these questions[30].

The discovery of crAss-like phages is exemplary in proving we have only begun to scratch the surface in uncovering the global viral diversity. Unique

in the sense that it was discovered as an orphan genome before its isolation in the lab, crAssphage set a trend for virus discovery in silico, whereas historically phages would be isolated in the lab before their characterisation[31]. The ongoing metagenomics and metatranscriptomic revolution[32], facilitated by access to scalable computing resources and growing databases, is fuelling the pace of viral discovery in global human and mammalian gut samples as well as in any possible sampled environment[33,34].

In this review, we attempt to briefly summarise available data on the diversity, taxonomy, and ecology of cultured and uncultured *Crassvirales* species, as well as to briefly review the history of their discovery. We will also highlight recent progress in understanding their virion molecular structure, infection mechanisms, and phage-host dynamics, achieved with the help of a limited number of cultured species of *Crassvirales*.

1.3 Metagenomics, Diversity, and Ecology of *Crassvirales*

p-crAssphage (*Carjivirus communis*) was discovered from metagenomic assemblies of pooled, short sequencing reads sourced from faecal samples of four pairs of monozygotic twins and their mothers[1]. Named in homage to the cross-assembly (crAss) algorithm used to generate its apparently circular, complete genome of 97 kb, p-crAssphage represents an "average" of perhaps several closely related strains present in the gut of those subjects. This was the first report of a supposed unique orphan phage with unusually high prevalence in the human gut microbiome—the majority of people from Western cultures tested positive, with levels sometimes reaching $> 90\%$ of the total virome and $> 20\%$ of the total faecal DNA. Although the functions of most of its genes, its potential morphology, and its bacterial host remained unknown, nucleotide similarities to CRISPR spacers in bacterial genomes pointed to a possible host in the genus *Bacteroides*[35].

A global phylogeographic study by Edwards et al. employed targeted sequencing of three PCR-amplified p-crAssphage genomic regions from wastewater treatment facilities around the world, as well as mining of 95,552 metagenomic datasets[35]. This effort resulted in 32,273 different p-crAssphage sequences from 67 countries across six continents and proved the ubiquitous presence of this phage in the human gut around the world. Analysis of the metadata in the Life-Lines-DEEP cohort did not reveal any significant association between any disease and health parameters and the p-crAssphage levels[10]. However, weak correlations with dietary factors (protein, carbohydrates, and calorie intake) were observed. Another biogeographic study of p-crAssphage observed that industrialisation plays a role in the dominance of p-crAssphage in the human gut[11]. Samples from Hadza individuals in Tanzania were crAssphage-negative ($n = 67$), while only 2/25 samples acquired from Matses Peruvians were crAssphage-positive. These populations are known to subsist on a primarily plant-based hunter–gatherer lifestyle. Similarly, p-crAssphage was found in low abundance in remote populations in rural Malawi and the Amazonas of Venezuela[10]. In contrast to these rural populations, prevalence among urban

individuals was high—35.7% of U.S. residents and 14% of Chinese residents[11]. These findings agree with *Bacteroides* also being associated with industrialised human populations and diets[36].

A study of longitudinal virome colonisation following FMT from healthy donors to patients with recurring *Clostridium difficile* infection (rCDI) showed crAssphage can effectively populate and persist long-term in the recipient gut. Given that rCDI patients began with a significantly lower crAssphage population than healthy donors, they quickly scaled up to an equivalent level to that of the healthy donors and remained stable for over 1 year. The crAssphages are likely an essential component of a healthy gut ecosystem and provide an important dimension in community maintenance and stabilisation[37]. Indeed, whole-virome analysis of inflammatory bowel disease (IBD) and ulcerative colitis (UC) datasets shows IBD-specific alterations in the gut virome, specifically a loss of crAss-like phages and *Microviridae*, identifying crAss-like phages as critical components of the core virome in healthy individuals[15]. Another study reporting loss of phageome stability in disease found reduced abundance and diversity in children with obesity and obesity-related metabolic syndrome[14]. The fact that a virulent crAssphage core represents a healthy microbiome at equilibrium and that virome dysbiosis can be overcome by FMT highlights the importance of considering the viral portion of a patient's metagenome when considering FMT therapy[37].

Further work made it apparent that p-crAssphage was not an orphan species but rather represented a whole family of cryptic bacteriophages termed "crAss-like phages". In 2018, Yutin et al. pioneered the detection of crAss-like phages in diverse environments by applying advanced approaches to detect distant protein homologies and genome synteny on candidate sequences extracted from expanded NCBI nucleotide and whole-genome sequence (WGS) databases[38]. Multiple clades of phages with highly divergent but clearly colinear and evolutionary related to p-crAssphage genomes were detected populating such diverse environments as the human microbiome (*IAS virus*), marine sediment (*Cellulophaga phage phi14:2*), the Populus rhizosphere (*Chitinophaga sp. YR573*), the termite gut (*Azobacteroides phage ProJPt-Bp*), and a fish pathogen (*Flavobacterium phage Fpv3*).

The extent of crAss-like diversity in the human gut was further explored by Guerin et al., who clustered 249 unique crAss-like phage genomes, assembled from 702 human faecal virome samples, based on sharing of orthologous genes, and identified ten candidate genera (subfamilies in the present classification) and four candidate subfamilies (families at present)[9]. The frequency of detection of different crAss-like phage genera was significantly associated with country of residence, health status, and the age of the donor, with the lowest counts seen in Irish and Malawian infants, as well as IBD patients in the U.S. As well as having a lower relative abundance, the genera of crAss-like phages detected in rural Malawi separated from the Western healthy and diseased cohorts, showing genus I (subfamily *Crudevirinae*, family *Intestiviridae*) as the most prevalent in Western populations regardless of age. By contrast, Malawian infants primarily carried candidate genera VI, VIII, and IX (*Steigviridae*, *Obtuvirinae*, and *Oafivirinae*).

CrAss-like phages are not only restricted to the human gut but have also been found in other mammalian faeces, such as those of rhesus macaque, pig and cat[33,39]. Edwards et al. demonstrated the presence of candidate genera I, III, and I in apes and Old-World and New-World monkeys[10]. Given the narrow focus of many studies on a single species (p-crAssphage), much less is known about the global phylogeography of the other clades of crAss-like phages. If shotgun metagenomics/metaviromics data from each of these sources were to be extensively analysed, a more granular and comprehensive picture of the distribution of crAss-like phages and the biogeographic mapping of more distantly related members of this viral group would likely emerge.

1.4 Co-Evolution with Bacterial Hosts

Antagonistic interactions between phages and their hosts drive their co-evolution, encouraging both parties to develop novel strategies to increase their fitness and overcome any defence mechanisms developed by the opposite party[29,40]. Gene duplication is one such strategy found in crAss-like sequences that is speculated to bestow replicative advantages. For instance, p-crAssphage has two copies of repL, required for the initiation of DNA replication, which arose from an ancient gene duplication event[41]. CrAss-like phages assigned to candidate genus I (*Crudevirinae* subfamily of *Intestiviridae*) have been found to encode diversity-generating retroelements associated with tail protein genes (DGR)[42]. DGRs employ reverse transcriptase to selectively mutate a stretch of DNA coding sequence, giving rise to the creation of original hypervariable tail proteins[43]. This built-in DNA mutator cassette is a versatile protection mechanism that potentially enables phages to counteract bacterial defences (resistance) or exploit alternative bacterial hosts by altering their tail proteins involved in host recognition.

An analysis of nucleotide polymorphisms and the rate of non-synonymous to synonymous changes (dN/dS) in crAssphage genomes from a mother-child cohort with low interpersonal viral diversity showed rapid evolution and strong positive selection for mutations in genes coding for RNAP and tail proteins[44]. Considering the crucial role tail proteins have in bacterial host recognition and the overcoming of acquired phage resistance, it is fitting that their genes would be targets of active selection. One specific target of elevated variation is the BACON domain-containing protein, which is located proximal to the tail protein complex and is thought to have a role in binding to the cell surface of *Bacteroides*[45]. The rapid evolution of these genes could reflect crAss-like phage's response to a Red Queen scenario where they are adapting to the host bacteria continuously cycling their cell-surface receptors.

1.5 Proposed Taxonomy of *Crassvirales*

A 2021 taxonomic proposal to the International Committee for Taxonomy of Viruses called for the creation of a new order, *Crassvirales*, encompassing four new families (*Intestiviridae*, *Crevaviridae*, *Suoliviridae*, and *Steigviridae*), ten new subfamilies, 42 new genera, and a total of 73 new species[46].

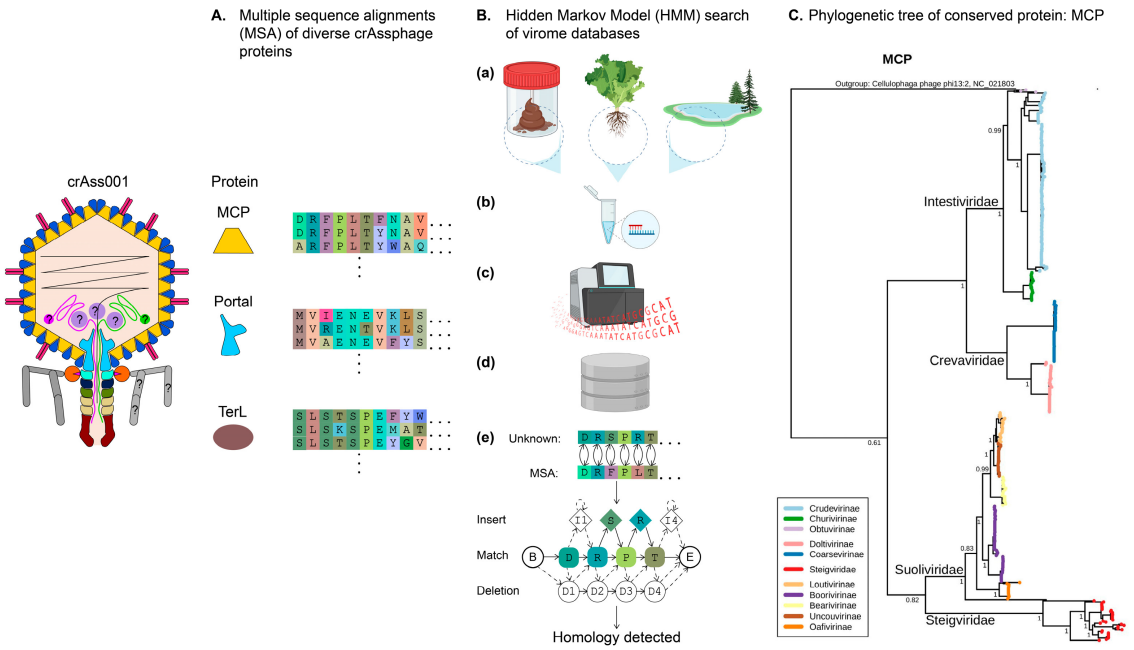


Figure 1.1: Bioinformatic approaches to discovering and classifying crAss-like phages. (A), Multiple sequence alignments (MSA) of the evolutionary conserved phage proteins such as major capsid protein (MCP), portal protein, and terminase large subunit (TerL) contain signatures specific to the order *Crassvirales*. MSAs are converted to profile HMMs (Hidden Markov Models). (B), Discovery and classification of new *Crassvirales*. Candidate genomes are collected from varied environments such as stool, soil, or aquatic biomes (a); shotgun sequencing and assembly of total or VLP-enriched DNA (b,c) yields large assembly databases (d); translated proteins are compared to MSAs of conserved *Crassvirales* proteins through sensitive position-specific homology detection algorithms (PSI-BLAST, HMM, and HHM). (C), Using MCP as a taxonomic marker protein, a maximum likelihood phylogenetic tree can be constructed to infer relatedness between families (main first-order branch names) of crAss-like phages[46].

The percentage of shared orthologous genes (proteins), as well as single gene phylogenies reconstructed based on amino acid sequences of the large terminase subunit (TerL), the major capsid protein (MCP), and DNA primase, were used to demarcate families, subfamilies, and genera (Figure 1.1).

A phylogenetic comparison of both MCP and TerL proteins places each of the four families on separate branches, with branch members sharing a minimum of 17% of orthologs. Monophyletic groups clearly emerging from the four family clades were deemed sub-family taxa. Clustering of subfamily genomes based on gene sharing showed 27–79% of proteins within each subfamily. A total of ten subfamilies were proposed, which roughly correspond with formerly proposed candidate genera (I-X), with the exception of former candidate genus VI, which was reformed into a single family, *Steigviridae*, with no clearly definable subfamilies in it. Species are defined as sharing 95% sequence identity over 85% of the complete genome length, in accordance with the advised MIUViG standard[47]. Only four species listed in the latest taxonomic proposal have been cultured[17,18,19,30], with the remaining ma-

jority originating from metagenomic assemblies (circular contigs). Synonyms of the word "crass" were affixed to create sub-family names, while the names of the four main families are language derivatives of the word "intestines": *Intestiviridae* (Latin), *Crevaviridae* (creva, Bosnian), *Suoliviridae* (suolet, Finnish), and *Steigviridae* (stéig, Irish).

1.6 Hallmark Genetic Features

An analysis of 596 diverse circular crAss-like phage genomes uncovered some unusual aspects of crAss-like phage architecture and biology[7]. In line with previous host assignments, CRISPR spacer searches of the prokaryotic genome database assigned the majority of crAss-like sequences to hosts in the phylum *Bacteroidetes*. Lower frequencies of spacer matches aligned to other bacterial phyla such as gut *Firmicutes* and *Proteobacteria*, and a minority proportion aligned to Archaeal genomes. The study also consolidated the genomic trademark of crAss-like phages as a syntenous block of eight highly conserved genes encoding the structural proteins MCP, TerL, portal protein, integration host factor (IHF), tail stabilization protein, tail tubular protein, and two uncharacterised proteins. Yutin et al. (2021) also identified two groups of environmental and gut phages appearing related to the core crAss-like phage phylogenetic assemblage, which were named as groups Zeta and Epsilon (continuing the original system of naming *Crassvirales* families as Alpha, Beta, Gamma, and Delta). These groups were characterised by larger genome sizes, ranging from 145 to 192 kb, almost twice as large as p-crAssphage's 97 kb.

An interesting observation was the alternative choice of family A or B DNA polymerases by different clades. Families *Intestiviridae*, *Crevaviridae*, and *Suoliviridae* (Alpha, Gamma, and Delta groups) exclusively encode PolB, while Zeta and Beta groups (the latter being family *Steigviridae*) have no preference and can either carry PolA or PolB. When co-occurring within groups, PolA and PolB occupy the same position in the genome and are thought to be swapped in situ by homologous recombination. It is hypothesised that this phenomenon has arisen due to the selective pressure presented by an undefined *Bacteroides* defence mechanism. In addition to DNA polymerase switching, another notable finding was the use of alternative genetic codes, specifically in 243 crAss-like phages in the Beta, Delta (*Steigviridae*, *Suoliviridae*), and Zeta groups that reassign TAG stop codons as glutamine or TGA stop codons as tryptophan. Phages in the Zeta groups can switch genetic code at short phylogenetic distances, with closely related branches interspersed with phages reassigning TAG or TGA or using the standard code. An established method of phage codon reassignment is through the use of suppressor tRNAs, which encode an anticodon complementary to the standard stop codons charged with an amino acid. Some crAss-like phages, especially in the Beta (*Steigviridae*) and Zeta groups, encode a plethora of tRNAs (1–34), with one Zeta phage (NCBI accession OJOH01000017) encoding 34 tRNAs. The encoding of suppressor tRNAs is not completely congruent with stop codon reassignment[7,48]. For example, in some Zeta clades, despite encoding suppressor tRNAs, no reassignment occurs, and in other cases where reassignment occurs, no suppressor tRNAs are present. This leaves the mechanism and motive behind

genetic code switching in crAss-like phages ambiguous.

Another distinct feature of crAss-like phages that has not been observed in other phages is the presence of at least one nuclease of the PDDEXK family, with many encoding two or three of these. The exact function of these nucleases is unclear, but they are presumed to have a role in reproduction, with some phages having one variety located in the structural block (PDDEXK_a), another in the replication block (PDDEXK_b), or yet another inside a large protein associated with transcription. The blocks of short, unknown genes common in many crAss-like genomes that share no homologs outside the crAss-like cohort are probably the most intriguing. Postulated to be genes in their defence armoury, specifically anti-CRISPR proteins (Acr), their purpose is still up for debate as a machine learning Acr model failed to confirm their function.

Introns, a group of mobile genetic elements that act as self-splicing ribozymes (an RNA with enzymatic activity)[49], and inteins, proteinaceous parasites that perform self-splicing by cleaving peptide bonds[50], are typical of the genomic landscape in many bacteriophages[49,51]. The crAss-like phages of the Delta (*Suoliviridae*) and especially Zeta groups have a remarkable density of introns and inteins. In Delta phages, group I introns are inserted in the core phage genes—MCP, PolB, and primase—with TerL containing an intein. A stark example of this intron infestation is found in a Zeta phage's (NCBI accession number: OHFV01000001) MCP gene, which contains three Group I self-splicing introns within its ORF fragments. In the Zeta group, their predicted RNAP ORFs are drastically fragmented by numerous stop codons and frame shifts, some occurring within the highly conserved DxDxD motif of RNAPB'. The disrupted nature of the genes highlights the complexity and nuance of annotating crAss-like phage genomes. Taking all of this into consideration, crAss-like phages likely employ unconventional encoding and expression mechanisms that have yet to be understood.

1.7 Alternative Genetic Coding Strategy: Experimentally Confirmed

LC-MS/MS metaproteomics of two metagenomic samples containing two different crAss-like phage genomes (one *Suoliviridae* and one Zeta group) identified phage peptides from MS/MS spectra. When mapped to the optimal gene predictions using both the standard genetic code (translation table 11) and table 15 (suppression of TAG stop codon with glutamine), around half of the peptides were found to only map to structural proteins (capsid, portal, and tail-associated) predicted using table 15. MS/MS peptide spectra supported the reassignment of the TAG codon to code for glutamine (Q). The contrast in mapping efficiency and correctness of annotation between the standard and non-standard genetic codes is stark and shows that consideration should be taken when predicting crAss-like phage genes[52].

Stop-codon recoding is widespread among different phage clades in the human gut microbiome, with crAss-like phages and Lak phages being found to recode the

TAG or TGA stop codon (genetic codes 15 and 4)[7,53]. CrAss-like phages using an alternate code infect standard-code hosts (genetic code 11). It is uncertain why some phages have evolved genetic codes incompatible with host translation systems. Interestingly, only 35%/40% of TGA/TAG-recoded genomes contain a corresponding tRNA suppressor. Do the other 60% rely on non-standard code used by the host, or do they use an unknown mechanism to accommodate their alternate code? Importantly, CrAss-like phages, for the most part, use alternative coding for their "late" structural and lysis genes (which constitute a single cohesive transcriptional unit). Therefore, the use of recoded stop codons might be required for regulating the precise timing of structural protein expression and cell lysis. Specific gene families seem to be enriched with recoded stop codons: lysozyme-type amidase and spanin/holin (both proteins involved in the lysis-lysis cassette). In such scenarios, regulated genetic code switching might be helping to prevent premature lysis (abortive infection activated by some anti-phage defence systems). Recoding has also been suggested to enable phages to sense co-resident phages and coordinate their lysis timing. Temperate gut phages (non-crAss-like) also employ genetic code switching to control lysis-lysogeny decisions. Given that crAss-like phages are not thought to be lysogenic, it is still possible that recoding contributes to the regulation of lysis (productive vs. abortive infection) in order to fine-tune the steady state in phage-host pairs and ensure long-term persistence[48].

Given p-crAssphage's ubiquity in human faecal samples, an industrial application of its prevalence has been its utilization as a microbial source tracker (MST) marker of human faecal contamination and pollution in environmental water, surfaces, and wastewater treatment[54,55]. One interest in using crAssphage as a marker for faecal pollution in environmental waters was borne out of the ability to distinguish human faecal waste from other animal faecal pollution[54]. That withstanding, crAss-like variants of animal gut origin are being researched to reliably assess animal faecal pollution of wastewater, with only 61% of wastewater confirmed positive for animal faecal matter showing positive for crAssphage[56].

1.8 Identification of Bacterial Hosts for crAss-like Phages

The classic approach to the isolation and identification of bacteriophages in environmental samples and linking them to their bacterial hosts is based on plaque assays. Subsequently, this bacteriophage can be further propagated from the plaque and studied. The success of the plaque assay approach in searching for phage-host pairs depends on several factors: the choice of the right bacterial host strain, the ability of the bacteriophage to form visible plaques, the suitability of the medium and other growth conditions for the bacterial growth and plaque formation, and the amount of infective phage particles in a lysate obtained from a sample. A successful combination of all these factors that would yield a plaque is not always possible, especially for such complex communities as the gut microbiota, which contains predominantly temperate phages[57,58,59] and phages infecting bacteria that remain mostly uncultured under laboratory conditions[1]. Despite all its weaknesses, the traditional plaque/spot assay approach was used for the isolation

of several crAss-like phages, such as phi14:2, crAss001, DAC15, and DAC17. The first member of the now-extant *Crassvirales* to be successfully isolated by this traditional method was in fact the *Cellulophaga phage phi14:2*. This marine virus was isolated and classified as a member of the family *Podoviridae* the year before the discovery of p-crAssphage and the sudden rise of interest in crAss-like phages as a novel taxonomic entity[1,30]. Phage enrichment in a liquid culture is a culture-based technique that does not rely on plaque formation. This technique, in combination with next-generation sequencing of the enriched virus-like particle (VLP)-associated DNA fraction, has been used for human gut phageome studies and resulted in the discovery of a significant amount of novel phage-host pairs, including those for crAss-like phages[22]. This plaque-independent, culture-assisted metagenomics approach was used for the identification and characterisation of crAss-like phages *ΦcrAss001* [1] and *ΦcrAss002* [19]. Additionally, the first crAss-like phages replicating in *Bacteroides uniformis* cells were detected with similar methods[20].

The paucity of cultured crAss-like phages has prompted the development of alternative approaches for identification of the phage-host pairs. The proximity ligation approach is based on chemically cross-linking DNA fragments that are in close proximity to each other in three-dimensional space. Cross-linking of the fragments results in them belonging to the same contigs after sequencing and data processing. This method allowed the assignment of several members of the crAss-like family to their bacterial hosts. A range of different hosts, all belonging to the *Bacteroidetes* phylum, was identified for 17 crAss-like phages using meta-HiC.

Microbe-seq, a new complex method of single-cell sequencing, is one of such approaches that uses single amplified genomes (SAG)[60]. This is a culture-independent method that allows for genomic data with strain resolution and is also capable of detecting host-phage interactions within complex populations. As phages associated with bacteria are encapsulated in the same single-cell droplets, the bacterial and viral genomes are linked in the output data by droplet-specific tags. It is thus possible to search for these phages in the resulting genomic data and track their connection with a specific bacterial strain. Using this approach, *Bacteroides vulgatus* from human gut samples was found to be significantly associated with the p-crAssphage and was therefore predicted to be its host. This result is consistent with other studies that propose the host of the crAssphage to be from the *Bacteroides* genus[61,62].

CRISPR spacers analysis resolved the crAss-like phage hosts at the order level (*Bacteroidales*), while more precise host identification by this method is problematic due to horizontal gene transfer leading to distant bacterial species sharing the same CRISPR direct repeats, which hinders taxonomic assignment of CRISPR regions themselves[62]. The correlation between bacterial and viral relative abundance pointed towards *Bacteroides vulgatus* and *Ruminococcus spp.* as possible hosts for crAss-like phages[63]. P-crAssphage was found to strongly co-occur and correlate quantitatively with *Bacteroides dorei*[64]. The credibility of this prediction remains questionable, as the results of the microbe-seq laboratory approach do not support it[60]. Other approaches that can potentially be applied to identify hosts for diverse uncultured crAss-like phages are listed below.

Many of the viruses discovered with metagenomics approaches have never been isolated in culture. The only methods that allow to link these uncultured phages to the bacterial hosts are bioinformatic predictions. The coevolution of phages and their hosts leaves specific signatures in the genomes of both. The signals for host-phage identification could be abundance profiles, codon usage similarity, genetic similarity, and CRISPR spacer matches[35,62]. Examples of the tools are HoPhage, HostFinder, PHIAF, and VirHostMatcher. Existing bioinformatics tools for in silico predictions of phage-host interactions were reviewed elsewhere[65].

One of the most widespread experimental, culture-independent methods of identifying and describing phage-host interactions is proximity ligation[66,67,68,69]. Recently, an improved protocol based on proximity ligation, called meta3C, was developed. Another recent modification of the proximity ligation method is the cross-linking of specifically bacterial ribosomal RNAs to the transcripts. rRNA-mRNA chimeras where mRNA belongs to a virus allow to identify the host of this virus—the organism that translates its transcripts.

Approaches that utilise single amplified genomes (SAG) are another powerful culture-independent tool for linking phages to hosts. In these approaches, DNA from individual cells is isolated and sequenced. The first step is separating individual cells from each other, which can be conducted with microfluidics or with flow-cytometry techniques. In the next step, the cells are lysed, their genomes are amplified, and they are subjected to next-generation sequencing. Applied to phage-bacteria interactions studies, SAG approaches reveal associations between viral DNA and bacterial DNA of different species and strains. These associations serve as the basis for predictions about which bacterial species serve as hosts for specific phages[70,71,72,73,74].

A virus adsorption-based approach to link bacteriophages from environmental samples to their bacterial hosts is adsorption sequencing (AdsorpSeq). It is based on the interactions between bacteriophages and receptors on the surface of their hosts. The phages that were able to bind to the cell envelopes isolated from a putative host are separated from free (unbound) phages with agarose electrophoresis and then identified by sequencing. This method was verified for several distinct species of Gram-negative bacteria, while its suitability for Gram-positive bacteria remains to be investigated[75]. One of the main drawbacks of all host identification methods that are based on the adsorption of phages to their bacterial hosts is the fact that not all phages that are able to bind to a bacterial surface are able to infect this specific bacterium. Therefore, the range of phages identified for a set of hosts with one of these methods is likely to be wider than the range of phages that successfully infect these hosts.

Viral tagging is another method based on phage adsorption. In this approach, viruses from environmental samples are labelled with fluorescent dye and mixed with supposed bacterial hosts[76,77]. Then the cells are separated by flow cytometry and activated by fluorescence. Infected cells are collected for viral DNA extraction and amplification that yields viral-tagged metagenomes. The single-cell viral

tagging was implemented to understand the network of phages and their hosts in the human gut; the study yielded hundreds of predicted phage-host pairs.

Droplet digital PCR (ddPCR)[78,79,80] allows for the identification of the co-localisation of viruses and bacterial cells. Cells from an environmental sample are diluted and directly embedded in droplets or in microfluidic PCR chambers in such a manner that most droplets or chambers contain just one cell. qPCR is performed using two pairs of primers—to a viral marker gene and to a cell marker gene (the rRNA gene). Then the DNA from the droplets or chambers is retrieved and sequenced. The co-amplification of both viral and bacterial marker genes in one droplet or chamber indicates phage-bacteria association. Statistical analysis of these associations reveals the relations between phages and bacterial species inside the community. The method was verified on a microbial population of termite hindgut.

Emulsion-paired isolation-concatenation PCR (epicPCR)[81], initially used for profiling bacterial communities[82], was subsequently modified for linking uncultured phages to bacteria. In this approach, single cells from environmental samples are encapsulated in polyacrylamide beads in emulsion oil. Phusion PCR is performed inside each bead. One of the primers to the target gene (viral) carries an overhang matching the 16SrRNA bacterial gene. Co-localisation of the target gene with the host rRNA gene results in fused products that are later detected with next-generation sequencing. The applicability of the techniques above to crAss-like phages is yet to be tested.

1.9 Morphology and Molecular Structure of crAss-like Phages

An early in-depth analysis (combining sensitive PSI-BLAST and HHPred searches for protein homologies) of two representative *Crassvirales* phages, p-crAssphage (now *Carjivirus communis*, family *Intestiviridae*) and IAS virus (now *Paundivirus hollandii*, family *Steigviridae*), managed to assign functions to many of the hypothetical proteins encoded by both genomes[38]. Both viruses appear to have genes grouped into distinct functional modules: the replication module, the transcription module, and the structural protein module (the latter of which can be further split into capsid and tail protein sub-modules). This functional division agrees with other features of their global genome organisation. An inversion of GC-skew likely separates the genomes into two oppositely oriented replichores (one containing the replication module and the other containing the structural and transcription modules)[48]. These functional modules coincide with transcriptional modules of early (regulatory), middle (replication + optional transcription), and late genes (transcription + structural virion proteins), as discussed below (Figure 1.2).

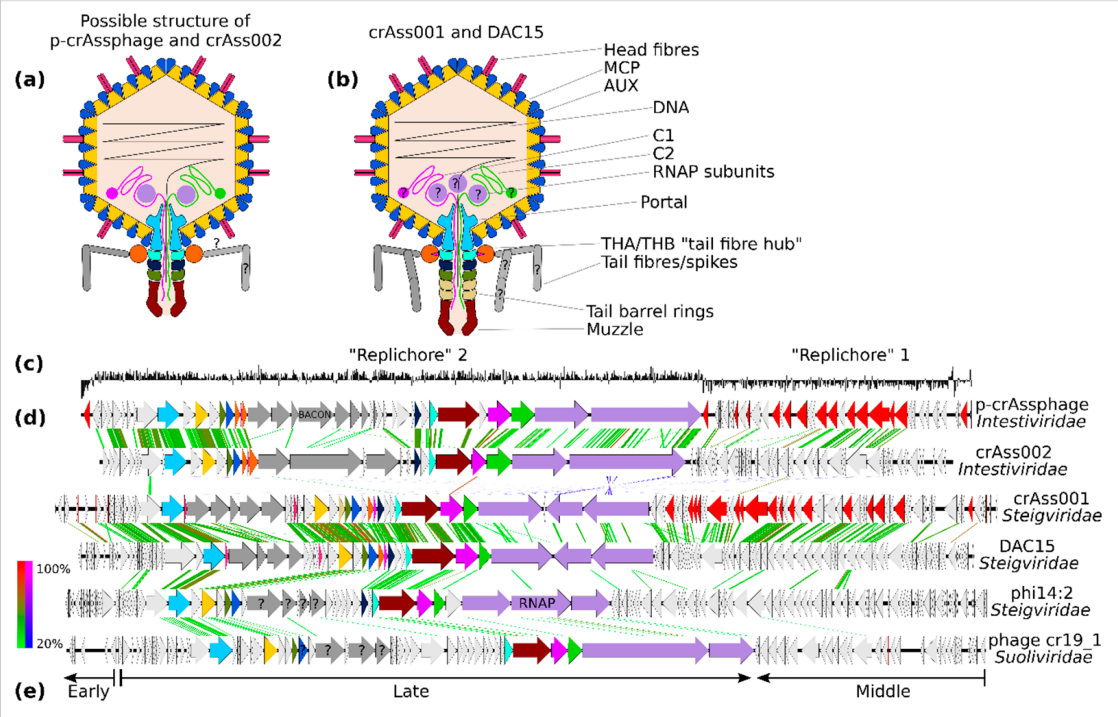


Figure 1.2: Virion and genome structure of the typical representatives of the order *Crassvirales*. A schematic structure of *Intestiviridae* virions (a) is extrapolated from the cryoEM structural data available for *Steigviridae* (b). Colours of the protein subunits (a,b) are consistent with ORF colouring (d). GC-skew graph of the p-crAssphage genomes shows clear separation into two possible replichores (c). Despite significant divergence in the amino acid sequence of the encoded proteins, all families show clear genome collinearity (d). Three transcriptional modules can be identified: early (regulatory), middle (DNA replication), and late (structural proteins and other virion components), as exemplified by a *Suoliviridae* genome (e).

While the podovirus-like morphology of *Crassvirales* virions had been predicted based on genome analysis, it was not until the isolation of first-cultured *Crassvirales* that this morphology could be confirmed. *Bacteroides phages crAss001*[17], *DAC15*, *DAC17*[18], and *Cellulophaga baltica phage phi14:2*[30] (all *Steigviridae*) all share a very similar morphology with isometric icosahedral capsids 77–88 nm in diameter (depending on the microscopy method) and short non-contractile tails of 36–44 nm decorated with appendages of different sizes. The *Bacteroides xylanisolvens* phage *crAss002* (*Jahgtovirus secundus*, family *Intestiviridae*) has a capsid of 77 nm and a shorter tail of 18 nm[19]. A crude faecal filtrate rich in *Crassvirales* (mainly *Intestiviridae* and *Crevaviridae*) contained two types of virions with isometric heads (76.5 nm) and shorter or longer tails[9]. Members of the family *Suoliviridae*, as well as additional proposed epsilon- and delta-groups of *Crassvirales*, have never been isolated in culture or observed microscopically.

A major advancement in understanding the virion structure and functions of individual structural proteins in *Crassvirales* came from a cryo-EM-based reconstruction of crAss001 virions, allowing for atomic-level modelling of many of the capsid and tail proteins[83]. This analysis revealed a T = 9 icosahedral capsid with

an outer diameter of 88 nm. The capsid is mainly composed of the HK97-type fold major capsid protein (MCP) and an auxiliary capsid protein (AUX). The structure of the AUX protein is unique to the *Crassvirales*. In addition to these main subunit types, two types of head fibres, trimeric and dimeric, with recognisably Ig-like domains, are inserted into the central pockets formed by the MCP-AUX hexons. These fibres seem to be narrowly specific to particular taxa within *Crassvirales*, with only a fraction of viral diversity in this order having detectable homologues. At the same time, unrelated proteins may have analogous functions in other species of *Crassvirales*. Both types of fibres can potentially be involved in either host recognition, or perhaps binding to the intestinal mucus, increasing the persistence of phage in the gut and preventing its washout (BAM-model of gut phage persistence)[84]. Such a phenomenon has previously been reported, with hoc+T4 coliphage being capable of binding to the gut mucin.

A dodecameric portal assembly is inserted in one of the five-fold symmetrical capsid vertices, providing an entry point for DNA packaging and an interface for docking the tail barrel assembly. Together with the large terminase gene, primase, and the MCP gene, the portal protein gene forms a set of genes universally conserved across the order *Crassvirales*. Zones around the portal crown domains extending into the capsid matrix contain a lower density of packaged DNA and are believed to be occupied by cargo proteins, which eject from the capsid together with, or ahead of, the phage genome in the early stages of infection. These cargo proteins may include 1–3 large subunits of the vRNAP as well as the large proteins C1 and C2, however only the presence of C1 protein can be confirmed from cryo-EM reconstruction. All cargo proteins are likely to be present in a partially unfolded state, facilitating their passage through the portal and the tail barrel. Protein C1 appears to be stretched out, with parts of its sequence visible inside the capsid (around the portal crown) and the tail barrel. Such partially unfolded C1, as well as parts of other cargo proteins and the leading DNA end, may constitute the central body visible inside the portal channel and the tail barrel.

The tail barrel is composed of five stacked dodecameric rings. Rings 1–3 are composed of dedicated subunit types, whereas rings four and five use identical subunits. Interestingly, different clades of *Crassvirales* contain variable numbers of tail-barrel ring protein genes, implying a different number of rings and therefore tail length in different species. This agrees with EM observations of diverse *Crassvirales*. A hexameric tail muzzle protein forms the tail barrel assembly and represents a structural element highly conserved across the *Crassvirales*. This multi-domain protein contains some elements shared with tail tip proteins from other phages, e.g., a ring-joining domain and a β-propeller domain containing the gate loop are similar to the coliphage T7 nozzle protein. At the same time, two Ig-like domains and, most notably, a domain with a novel type of fold, termed crAss fold, located at the very tip of the muzzle structure, are highly unique to the *Crassvirales*. It is believed that the crAss domain, together with Ig-like domains, might be responsible for the docking of phage particles on the cell surface. Conformational changes evoked by this docking event can trigger the opening of the gate located above, in the proximal part of the muzzle, ejecting the capsid contents.

Two types of protein subunits (THA and THB) are responsible for forming a collar structure around the tail neck. This structure forms a docking hub for the attachment of tail fibre/spike assemblies—flexible structures forming the outer and inner "cages" around the tail barrel. While five different proteins are likely to be involved in this assembly in crAss001, the exact structure could not be resolved using cryo-EM. Tail spike subunits in different *Crassvirales* have been annotated as containing carbohydrate binding/lectin-type domains (e.g., the BACON domain in p-crAssphage and related phages) as well as possible glycosyl-hydrolase domains, phosphatase domains, and CHAP domains[1,17,19,45]. Tail spike proteins with CPS-, EPS-, and LPS-depolymerising activity are a typical feature in phages with podovirus-like morphology[85]. It is therefore likely that the tail fibres/spikes of *Crassvirales* act as receptor binding proteins (RBPs) and are responsible for both the recognition of polysaccharide structures covering host cells, and the digestion of CPS/EPS/LPS polysaccharides in order to gain access to the outer membrane for docking. The genomic region occupied by the tail spike genes is characterised by a striking level of diversity, with different lineages of *Crassvirales* containing completely different numbers and types of genes, often with no detectable homology to their counterparts in crAss001 or p-crAssphage. Even within narrower groups in *Crassvirales* (just p-crAssphage—*Carjivirus communis*), tail spike proteins are subjects to rapid evolution, e.g., DGR-dependent diversification[42] and evolution through duplication and HGT of BACON repeats[45]. Longitudinal analysis of p-crAssphage genomes persisting in the human gut revealed rapid evolution of tail spike proteins with evidence of positive selective pressure[42,86].

Overall, the genome and virion structure of *Crassvirales* resemble those of coliphage N4 (the latter being slightly smaller at 72 kb genome and 70 nm capsid). Interestingly, coliphage N4 also includes the rare feature of a high-molecular-weight (360 kDa) vRNAP[87,88].

1.10 *Crassvirales* Life Cycles and Phage-Host Dynamics In Vivo and In Vitro

In addition to their high abundance in many individuals, another characteristic feature of the ecology of the *Crassvirales* is the long-term persistence of the same strains at constant levels in individual human gut phageomes. Human individuals followed for periods of months to a year demonstrated stable carriage of individual-specific viral assemblages, including one or several strains of *Crassvirales*[86]. This is paralleled by the observed long-term stability and individual specificity of bacterial microbiota, including members of the order *Bacteroidales*[88,89], which serve as a substrate for crAss-like phage persistence. The adapted co-existence of *Crassvirales* and their hosts manifests in their ability to co-transmit vertically from mothers to infants[86] and co-engraft during faecal microbiota transplantation (FMT)[37].

The long-term persistence of crAss-like phages has been replicated in both in vitro and in vivo models. Phages crAss001 and crAss002 have both been demonstrated to persist at high levels without noticeable effects on bacterial

host culture density in serial daily liquid cultures, chemostat cultures, and, in the case of crAss001, in a gnotobiotic model of mice mono colonised with their bacterial host *B. intestinalis*[17,19]. Moreover, the uncultured crAss-like phage phiHSC05, introduced as a part of the human faecal virome, was able to stably engraft in a gnotobiotic mouse model colonised with an artificial community of human gut bacteria[28]. This long-term co-existence may not necessarily be unique to *Crassvirales*, as many tailed and non-tailed phages demonstrate very similar behaviour of relatively "benign" co-existence with their hosts for many generations in the mammalian gut[90,91,92]. Theoretical explanatory models of phage-host co-existence are listed in Box 2; however, further experimental evidence is needed to resolve which mechanism or combination thereof best describes *Crassvirales* ecology.

Current evidence for various elements in the *Crassvirales* phage life cycle and their interactions with bacterial hosts is available for a small subset of phages, either isolated in pure culture (crAss001, DAC15, crAss002, phi14:2, and the C4-group of phages infecting *B. uniformis*)[20], or transiently enriched/maintained in host cultures in vivo or in vitro (phiHSC05 in oligo community in mice[28], a 104.9 kb crAss-like phage genome associated with *Prevotella stercorea* in an in vitro culture[22]).

Phages crAss001, DAC15, and DAC17 display a narrow host range, being only able to infect phase variants displaying a particular capsular polysaccharide (CPS) and outer membrane protein (OMP) within a single strain of *B. intestinalis* APC919/174 and *B. thetaiotaomicron* VPI-5482, respectively. In the case of crAss001, only the expression of PVR9, one out of five phase-variable CPS loci available in the host strain, is associated with phage sensitivity[21]. PVR9 expression in *B. intestinalis* was associated with strong phage adsorption, suggesting its role as a phage receptor for crAss001. In a similar manner, phages DAC15 and DAC17 were narrowly specific to a single phase-variable CPS (CPS3), out of eight encoded by *B. thetaiotaomicron* VPI-5482, although some marginal infectivity was also observed for CPS7 and CPS8, as well as for an acapsular mutant[18]. Mixed cultures expressing several different CPS types supported phage replication in vitro and in vivo (i.e., mono-colonised mice), but were unable to show lysis in traditional plaque and spot assays, most likely due to masking of plaques by resistant subpopulations.

In long-term co-cultures of crAss001 and *B. intestinalis APC919/174*, sequestration of part of the population in the resistant phase, and constant reversal to sensitivity might be sufficient for phage persistence. In complex community scenarios, concurrent selective pressures may be responsible for driving this backward switch to sensitivity: the metabolic cost of protective CPS expression, concomitant sensitivity to other phage conferred by a CPS protective against one phage[18,93], and complex pressures and trade-offs associated with the role of CPSs as gut colonisation factors in vivo[94]. A summary of CPSs-phage interactions is presented in Figure 1.3. crAss002 and phages of the C4 group also demonstrate narrow host specificity (*B. xylanisolvens APCS1/XY* and a group of closely related*B. uniformis* isolates, respectively)[20]. Unlike crAss001, DAC15, and DAC17, these phages are unable to form plaques, likely due to the prevalence of resistant phase variants in populations of their hosts. Indeed, long read sequencing of the *B. xylanisolvens APCS1/XY*

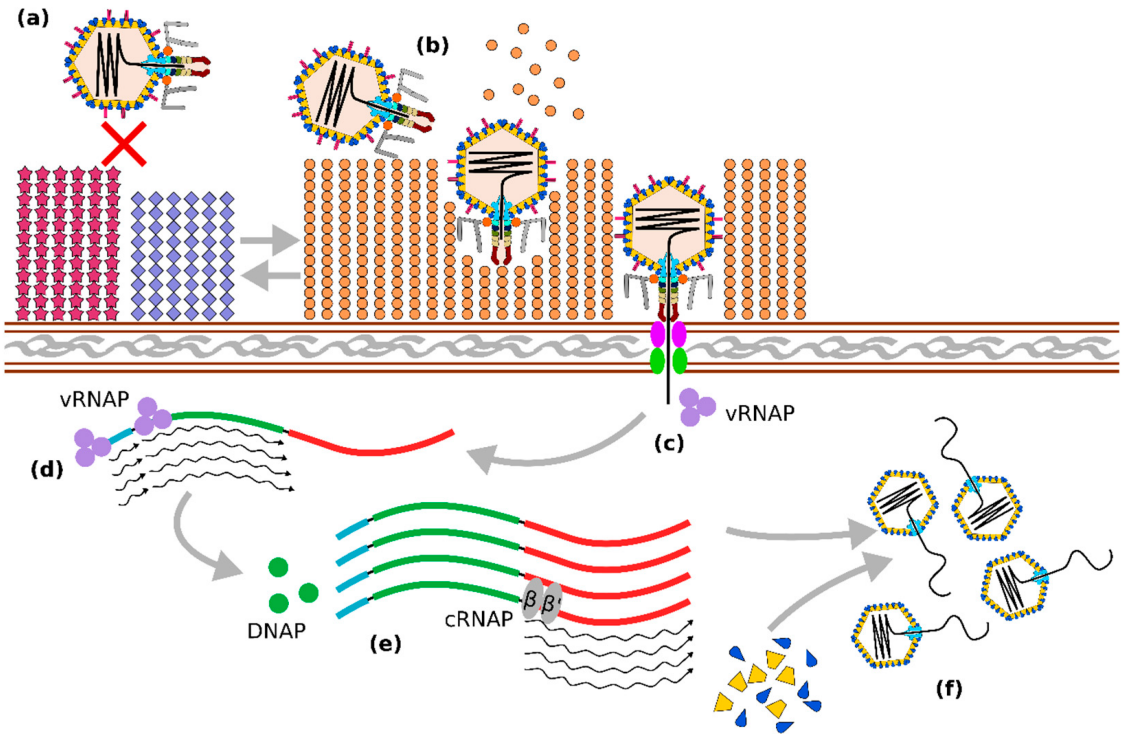


Figure 1.3: Some known elements of the Crassvirales life cycle. Adsorption and infection of *Bacteroides* cells by crAss-like phages are regulated by the expression of alternating phase-variable capsular polysaccharide (CPS) types. Expression of non-permissive CPS prevents phage infection (a). while switching to permissive CPS allows phage attachment, hydrolysis of CPS by tail-associated depolymerase enzymes, and ejection of phage contents (b). Phage DNA is co-injected with cargo proteins, among which are virion RNA polymerase (vRNAP) subunits responsible for transcription of early (regulatory) and middle (DNA replication) genes (d). Formation of DNA polymerase (DNAP) as well as other enzymes, allows for DNA replication through an unknown mechanism, followed by expression of the late genes (large terminase, structural proteins, virion cargo proteins), carried out by the cellular bacterial RNA polymerase complex (cRNAP), (e). This results in the formation of new virions and the packaging of phage DNA (f).

genome revealed several phase-variable CPS and OMP loci, which can potentially be implicated in driving reversible phage resistance[19].

As mentioned above, *Crassvirales* tail fibres/spikes are most likely required for adsorption to the CPS target and local depolymerisation of the polysaccharide capsule to gain access to the outer membrane. While the exact mechanisms of virion docking and ejection are still unknown, some insights into this process were provided by a recent structural study of crAss001[83]. Cargo proteins C1 and C2 are likely to be ejected, forming a tunnel through the cell envelope (C1 has an identifiable transmembrane helix), in a manner analogous to phage T7 protein gp14[95]. This is likely followed by the ejection of the phage-associated vRNAP subunits in a partially unfolded state to allow passage of these through a relatively narrow ($< 50\,\text{Å}$) channel inside the tail barrel. Ejection of cargo proteins is likely to be closely followed by ejection of the phage genome, both driven by high pressure achieved through tight packaging of DNA and proteins in the phage capsid.

A study of transcription in phage crAss001[21] identified three main phases: (1) transcription of early genes, starting in less than 10 min after infection, including a transcriptional regulator, two conserved domain proteins with unknown function, and a TROVE domain protein; (2) transcription of middle genes, initiating between 10 and 30 min after infection, including replication, recombination, and nucleotide metabolism functions; and (3) transcription of late genes, initiating between 30 and 60 min after infection, including a large terminase, all structural proteins, and most cargo proteins. A study of transcription in phi14:2, a crAss-like phage with a longer latent period[96], reported early genes (expressed between 40 and 90 min), corresponding to both early and middle genes in the crAss001 study. At the same time, middle and late genes (structural and cargo proteins), reported by Drobysheva et al. as being expressed between 90 and 190 min, corresponded to the late genes in the crAss001 study. Importantly, the latter study also confirmed the RNA polymerase activity of one of the predicted phi14:2 vRNAP in vitro and demonstrated its role in rifampicin-resistant transcription of early genes (i.e., regulatory and replication modules) in vivo.

Despite some discrepancies between the two studies, it appears that the transcription program of *Crassvirales* phages is broadly similar to that of coliphage N4[87]. Upon ejection of virion contents into the cytosol, vRNAP drives expression of early and middle genes independently of host transcription machinery. The synthesis of vDNAP and other enzymes encoded by the middle genes allows phage genome replication to begin. Exact mode of replication is unknown and likely to differ between different clades of *Crassvirales*. The crAss001 genome (*Steigviridae*) was shown to have long terminal repeats (DTR), whereas the crAss002 genome (*Intestiviridae*) is predicted to replicate, forming terminally redundant, circularly permuted genome copies. Based on the analysis of the phage-to-host genome copy ratio, genome replication of crAss001 begins between 30 and 60 min post infection, following expression of the middle genes with simultaneous expression of the late genes[21]. Transcription of late genes is likely to be carried out by the host RNAP and is therefore expected to be sensitive to repression with rifampicin[96]. Phage packaging is mediated by large terminases and again may vary between different clades depending on genome structure. Precise mechanisms of capsid assembly and packaging are yet to be established. However, early morphological changes (formation of electron-light spots) in the infected cells of *B. intestinalis* indicate the existence of some scaffolding. Cell lysis appears to be mediated by spanins and lysins encoded among the late genes, which can be subjected to translational regulation by suppression of in-frame stop codons[48]. Determination of the exact burst size of *Crassvirales* has proven difficult[21], with reports ranging from 20 (observed for crAss001 at MOI = 1) to 160 (DAC15, MOI = 10) virions per cell.

The ability of bacteriophages to persist long-term (months and years) in complex microbial communities such as the human gut microbiota alongside the populations of their bacterial hosts is a well-documented phenomenon[12,40]. At least three groups of mechanisms can be invoked to explain this persistence.

(i) Spatial segregation of phages and their hosts in the gut and the "source-sink"

model: macro- and micro-anatomy of the gut (villi, glands, mucin layer), as well as self-organised bacterial refuges (biofilms), can create areas of bacterial proliferation ("source") inaccessible to phages. Release of bacterial cells into the gut lumen or liquid phase provides gut phages with a constant supply of bacterial prey ("sink"). This mechanism can enable phage-host co-existence with minimal co-evolution[97,98].

(ii) Ecological population dynamics of strictly lytic phage and its host under conditions of resource limitation/variability: Lotka-Volterra equations and the "kill-the-winner" model predict that predator and prey can co-exist in a planktonic environment at the point of equilibrium or with cyclic fluctuations of their densities. Under the "kill-the-winner" hypothesis, phage-mediated population control together with the competition for a limited resource favours the maintenance of high diversity in a complex community (even in the absence of co-evolution)[99,100].

(iii) Antagonistic co-evolution and the cost of resistance: a number of theoretical frameworks are often employed to describe phage-bacteria co-evolution, such as the "arms race" dynamics (resistance mutations in the host followed by host range counter-mutations in the phage) and "fluctuating selection" dynamics (Red Queen evolution), the latter leading to diversification of bacterial and phage populations[99,101,102]. Longitudinal observations show evidence that selective pressure driving evolution of crAss-like phage genes is particularly strong for tail components (including RBPs), indicating an arms race/Red Queen type of dynamics between phages and their hosts[21,44,86].

1.11 Conclusions

The accidental discovery of p-crAssphage, the most abundant human gut bacteriophage, by cross-assembly of metagenomic reads in 2014 revealed how little we knew about the diversity and the role of bacterial viruses in the human gut microbiome and other complex microbial communities. The known viral diversity of dsDNA-tailed phages has bloomed in recent years due to the increasing availability of shotgun metagenomic and metaviromic data, improved virome analysis pipelines, sensitive protein homology search tools, and expanding collections of reference genomes[6,8,103]. The range of known crAss-like phages has expanded from a few "orphan" species into the expansive and highly diverse new viral order of *Crassvirales*[7,44]. Still, the phylogeny of the *Crassvirales* and their evolutionary relationships with other dsDNA viruses remain largely unexplored.

Metagenomics and metaviromics have proven themselves powerful technologies for virus discovery, host prediction, taxonomy, and abundance and distribution analysis[27]. However, in silico methods can be limited in their tangible revelations on the molecular biology, host range, and ecological impacts of crAss-like phages. Emerging engineering advances in single-cell sequencing techniques, viral tagging, and proximity ligation are set to provide new insights into the biology of diverse crAss-like phages.

The first in-depth characterisation of a crAss-like phage, crAss001, while providing a complete genome and data on virion structure and certain aspects of its life

cycle[17,21,82], posed numerous new questions about the biology of this phage taxon. Being apparently a strictly lytic phage, crAss001 is capable of long-term persistence in the culture of its host *Bacteroides intestinalis*, which was shown to be associated with dynamic phase variation of capsular polysaccharide types. However, is that the only mechanism responsible? Could alternative mechanisms of persistence (lysogeny, pseudo-lysogeny, the carrier state, or chronic infection "hibernation" inside infected cells) also be part of it? Available evidence does not support the ability for true lysogeny in crAss-like phages, but other mechanisms cannot be ruled out.

Phage crAss002 infects its host *Bacteroides xylanisolvens* without forming plaques, despite being apparently a lytic phage[19]. Is that a result of phase variation greatly shifting towards the insensitive phase? Such phage-host interactions may be puzzling to the observer, but is it uniquely specific to the *Bacteroides*/crAss-like phage systems, or is it a widespread type of dynamic between phages and their bacterial hosts in natural settings, a better winning strategy for the phage? Perhaps the "classical" lytic process, favoured by many studies of phage biology, is a very artificial scenario that can only be achieved in perfect conditions (exactly the right growth phase, MOI, etc.).

As discussed in our other recent review[29], bacteriophages can be seen as an integral component of complex microbial communities, evolutionary partners rather than antagonists of bacteria, responsible for maintenance of taxonomic and functional diversity, stimulating resilience and adaptability of the microbiota. The tandem of *Bacteroides* bacteria and crAss-like phages in the gut might be an epitome of such relationships.

1.12 Author Contributions

L.S.: writing and original draft preparation; E.G.: writing and original draft preparation; B.G.: review and editing; A.N.S.: writing, original draft preparation, and review. All authors have read and agreed to the published version of the manuscript.

2. A novel approach to *Caudoviricetes* taxonomy utilising whole proteome structure-structure comparison

Linda Smith[1,2], Thomas S.B. Schmidt[1,2], Igor Tolstoy[3], Timofey Skvortsov[4], Colin Hill[1,2], Andrey N. Shkoporov[1,2,5]

[1] APC Microbiome Ireland, University College Cork, Ireland, [2] School of Microbiology, University College Cork, Ireland, [3] Sequencing.com, Pasadena, CA, USA, [4] School of Pharmacy, Queen's University Belfast, UK, [5] Department of Medicine, University College Cork, Ireland

2.1 Abstract

Viral proteins often evolve so rapidly that sequence similarity is lost even among conserved functional genes, complicating phylogenetic analysis and classification. This challenge is acute for *Caudoviricetes*, a highly diverse class of dsDNA tailed bacteriophages recently redefined by the International Committee on Taxonomy of Viruses (ICTV). To address the limitations of sequence-based taxonomy, we present a structure-based classification framework using whole-proteome structural clustering. From 4,082 exemplar genomes, we predicted 445,151 protein structures with ESMFold and clustered them using Foldseek. Genomes were encoded as binary profiles of structural fold presence, and phylogenetic distances were inferred and normalised using Relative Evolutionary Divergence (RED). This yielded a revised structural taxonomy comprising 159 orders, 267 families, 502 subfamilies, and 1,189 genera. We also introduce *PhagePleats*, a Python tool for classifying novel phage genomes based on structural similarity. Our approach highlights the utility of protein structure for resolving distant viral relationships.

2.2 Introduction

Viral taxonomy remains one of the most challenging frontiers in biology. Unlike cellular organisms, which share universally conserved genes, no single genetic marker is conserved across all viruses[104,105]. They are believed to have evolved independently at least six times[106]; a concept reflected in their classification into six distinct viral realms. These polyphyletic origins make it fundamentally impossible to construct a single, unified phylogeny—or "megataxonomy"—that encompasses all known viruses[104]. Viral genomes—especially those of dsDNA tailed bacteriophages and archaeal viruses belonging to *Caudoviricetes*—are mosaics composed of exchangeable modular gene cassettes acquired through horizontal gene transfer[107,108,109]. This genomic patchwork obscures clear evolutionary signals, complicating phylogenetic inference even among closely related phages[110]. Although mosaicism is a characteristic and pervasive feature of tailed viral genomes, its extent is nuanced: some lineages maintain re-markably stable nucleotide and protein content, whereas others exhibit localised hotspots of rapid genetic flux[111]. Even individual viral proteins—such as tail fibers—can be chimeric in origin, further confounding homology detection[112]. Consequently, a large proportion of predicted

viral ORFs remain unannotated, contributing to the vast "viral dark matter" [113,114]. This annotation gap is exacerbated by limited experimental work characterising phage protein functions and assessing how gene gain, loss, or modular rearrangements shape phage biology and evolution. Taken together, rampant mosaicism, rapid evolutionary turnover, and the large fraction of unannotated viral sequences make sequence-based approaches poorly suited for resolving deep viral relationships.

This is exemplified by *Caudoviricetes* viruses, the most abundant and diverse biological entities on Earth [115,116,117] with major ecological importance across different ecosystems [118,119,120]. As metagenomic efforts continue to collect an ever-growing catalogue of viral genomes, the challenge of assigning higher-order taxonomic labels remains largely unresolved. In 2022, tailed bacteriophage taxonomy was overhauled with the establishment of the class *Caudoviricetes* and the dissolution of the morphology-based order *Caudovirales* and its families *Myoviridae*, *Podoviridae*, and *Siphoviridae* [121]. This revision enabled the creation of new orders based on genomic similarity, beginning with *Crassvirales* [46,122], followed by others including *Autographivirales* and *Pantevenvirales*. Despite this progress, the current ICTV taxonomy remains incomplete—only 24.5% (1,429 of 5,822) of *Caudoviricetes* species in the ICTV VMR40 (Virus Metadata Resource, release 40) have been assigned to an order and 58.2% (3,393 of 5,822) a family.

We posit that structure-based approaches to viral classification can mitigate issues faced by sequence-based viral phylogenies. Protein secondary and tertiary structures, constrained by function, tend to evolve more slowly than primary structure i.e. amino acid sequences [121], making them not only stable molecular markers of deep evolutionary or functional relation-ships, but phylogenetic relics encoding a historical record of divergence. Tools such as AlphaFold and ESMFold now enable high-throughput prediction of protein structures with near-experimental accuracy [123,124,125]. The ESM-2 model, for example, has been used to generate over 600 million predicted structures, constituting the first large-scale structural metagenomic protein database. Concurrently, tools like Foldseek [126] allow for ultrafast structural comparisons by encoding protein geometry into a 3Di alphabet, making it computationally feasible to search for conserved folds across millions of proteins using sequence comparison algorithms.

The growing availability of tools for large-scale structural prediction and comparison marks a timely inflection point for virology and phylogenetics. While the concept of structure-based viral taxonomy is not new [127,128], recent advances now make it feasible to apply these approaches systematically across the whole structurome of entire viral clades. In cases where sequence-based methods reach their limits, structural conservation offers a robust alternative, suggesting that structure-guided approaches may increasingly complement—and in some contexts, surpass—sequence-based strategies as a standard approach to viral classification. While prior efforts have used structural comparisons of single hallmark proteins to infer viral phylogeny [129], our approach extends this principle to the entire structural proteome. This offers a more holistic and evolutionarily relevant framework for taxonomic classification. To illustrate this, one could consider a hypothetical scenario: if a viral genome were entirely recoded to use different codons or alternative amino

acid sequences, yet all resulting proteins retained their tertiary and quaternary structures, the virus would likely retain its original morphology, infectivity, and ecological role. From a functional and ecological perspective, it would still belong to the same lineage—despite being unrecognisable to sequence-based approaches.

Here we present a unified, whole-structurome–based taxonomic approach to resolve higher-order classifications within *Caudoviricetes*. We re-organise *Caudoviricetes* diversity along genome-wide structural similarity partitions that align remarkably well with established sequence-based ICTV taxonomic groups, while adding further resolution and establishing novel groups relative to existing classifications. normalising derived structural "phylogenies" by the relative evolutionary divergence (RED) among viral genomes, we delineate robust and reproducible clusters at different taxonomic depths. Finally, we introduce *PhagePleats*, a machine learning-based tool that accurately classifies novel genomes against the resulting taxonomies based on folded viral proteins.

2.3 Results

2.3.1 Structural Conservation within *Caudoviricetes*

To capture the full known diversity of *Caudoviricetes*, 4,082 genomes were selected from the ICTV VMR40 dataset, representing all established viral orders (Figure 2.1a). The class currently comprises 11 recognised orders, though seven of these include fewer than ten genomes (Supplementary Figure 2.6a). Among the selected genomes, 3,291 of 4,082 species (80.6%) lacked an order-level assignment, and across the entire *Caudoviricetes* dataset, 4,392 species remain unassigned at this rank. A similar trend is evident at the family level, where 2,257 of the selected genomes (55.3%) and 2,427 genomes overall lack a family designation (Supplementary Figure 2.6b). Together, these observations highlight the limited resolution of current higher-order taxonomic groupings and emphasise the need for a consistent, scalable framework capable of improving classification accuracy at both the order and family levels.

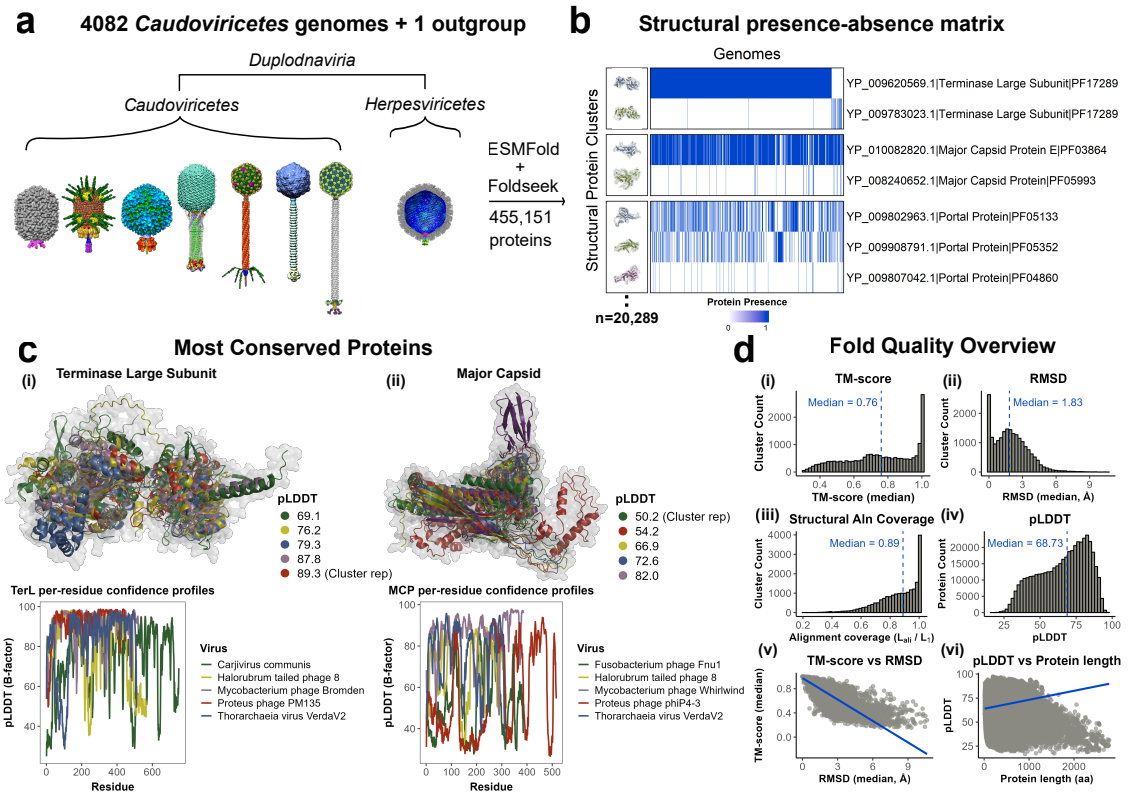


Figure 2.1: Structural cluster quality assessment and conserved protein identification (a) A total of 4,082 *Caudoviricetes* genomes and one *Herpesviricetes* outgroup, comprising 455,151 proteins, were folded using ESMFold and clustered using Foldseek to generate a structural presence–absence matrix (b) containing 20,289 unique clusters. Among these, Terminase Large Subunit (TerL), Major Capsid Protein (MCP), and Portal Protein were identified as the most conserved structural clusters across genomes. (c) Structural superimpositions illustrate high structural homology among TerL proteins from phylogenetically diverse bacterial and archaeal viruses (i), and among MCPs (ii), despite variation in per-residue confidence profiles. (d) Quality assessment of the structural clusters: (i) distribution of intra-cluster TM-scores (median = 0.76), (ii) intra-cluster RMSD values (median = 1.83 Å), (iii) intra-cluster structural alignment coverage (alignment length ÷ target length; median = 0.89), demonstrating strong structural similarity within clusters; (iv) pLDDT distribution for all folded proteins (median = 68.73), (v) TM-score versus RMSD relationship, and (vi) pLDDT versus protein length.

Given that tailed dsDNA bacteriophages and archaeal viruses (*Caudoviricetes*) share a conserved set of core structural proteins involved in capsid morphogenesis and DNA packaging, including the HK97-like major capsid protein (MCP), the portal protein superfamily (e.g., SPP1 GP6-like), and the terminase large subunit (TerL) superfamily, structural clustering was performed with the expectation that a sufficiently sensitive approach would recover coherent superfamilies for each of these hallmark proteins, ideally with a representative present in the majority of genomes. Foldseek easy-cluster identified 20,289 distinct structural clusters under the optimal parameter set (Figure 2.1b). Among these, TerL and MCP were the most universally conserved. TerL homologs from 3,938 genomes fell into a single

dominant cluster (represented by YP_009620569.1), with another 124 genomes in a closely related cluster. MCP showed a similarly cohesive pattern, with 3,933 genomes assigned to the cluster represented by YP_010082820.1 and 147 to a secondary MCP cluster. In contrast, portal proteins displayed greater structural diversity, forming three major clusters and several smaller subfamilies.

To visualise the extent of structural conservation, representative TerL and MCP structures from diverse viral families and ecological niches were superimposed (Fig. 1c). Of these include *Proteus phage PM135* (family *Demereciviridae*), which infects *Proteus mirabilis*, a common clinical and environmental bacterium[130]; *Carjivirus communis* (family *Intestiviridae*), a gut-associated phage first identified in human fecal metagenomes[131]; *Halorubrum tailed phage 8* (HRTV-8) (family *Hafunaviridae*), isolated from halophilic Halorubrum archaea inhabiting solar salterns in Thailand[132]; *Thorarchaeia virus VerdaV2* (family *Verdandiviridae*), identified from deep-sea sediment *Asgardarchaeota* metagenomes recovered off the Shimokita Peninsula, Japan[133]; and *Mycobacterium phage Bromden* (family *Vilmaviridae*), an actinobacteriophage infecting *Mycobacterium smegmatis* isolated from soil in Michigan, USA[134]. These overlays highlight the strongly conserved modular architecture of TerL across both bacterial and archaeal tailed viruses despite variation in local model per-residue confidence (pLDDT). MCP overlays are shown in Fig. 1c(ii), with the cluster representative derived from *Fusobacterium phage Fnu1*, a lytic phage of the pathogen *Fusobacterium nucleatum*[135]. MCPs display the canonical HK97-fold architecture[136], with other MCP representatives broadly mirroring the host ecologies of the TerL examples (Figure 2.1c(i).

Foldseek alignments across the 20,289 clusters were consistently high quality, with a median TM-score of 0.76, RMSD of 1.83 Å, and alignment coverage of 0.89 (Fig. 1d). Intra-cluster protein length distributions were similarly homogeneous, with a median length range of 4 aa, an interquartile range of 0–28 aa, and ∼90% of clusters differing by ≤100 aa between their shortest and longest proteins. Although predicted pLDDT values varied widely (median 68.7), many proteins with modest or low global pLDDT still produced accurate, high-confidence structural alignments. This reflects the single-pass, MSA-free nature of ESMFold, which prioritises speed over global confidence but reliably captures overall fold geometry. The consistently high structural congruence across clusters indicates that variation in either per-residue or global predicted ESMFold confidence does not undermine the integrity of the Foldseek-based clustering. Taking this, and the novelty and diversity of the viral proteins examined, into consideration, we retained all structural predictions to maximise proteome coverage and avoid losing genuine evolutionary signal.

To evaluate how lower-confidence models influenced tree topology, we compared the full structural phylogeny with a version reconstructed using only proteins with pLDDT $\geq$ 50 (347,958 models). The two trees showed moderate topological agreement (global RF similarity = 0.44; RF distance = 0.56). Family-level stability exhibited a strong sampling effect: very small families (1–5 genomes) appeared artificially stable due to the limited number of possible bipartitions, whereas mid-sized families (∼10–40 genomes) showed the steepest decline in stability, indicating that much of their internal branching structure depends on lower-confidence protein

models. Stability increased again for large families (≥100 genomes), plateauing around RF similarity ≈0.65–0.75, with deep, well-supported clades retained even as shallower relationships shifted. Together, these results demonstrate that low-pLDDT ESMFold models carry substantial fine-scale phylogenetic signal, and that intermediate-sized families are particularly sensitive to reductions in proteome sampling.

To contextualise our structural proteome-based taxonomy within existing viral taxonomy frameworks, we reconstructed single-gene phylogenies from the three hallmark proteins most commonly used in *Caudoviricetes* taxonomy: major capsid protein (MCP), terminase large subunit (TerL), and the portal protein. Confidently identified hallmark proteins were recovered from the genome dataset, including 4,021 MCPs, 3,989 TerLs, and 3,487 portal proteins. The resulting protein sequences were aligned using MAFFT and used to reconstruct maximum-likelihood phylogenies with FastTree ((Supplementary Figure 2.7–Figure 2.9)).

To quantify similarity between phylogenies, Robinson–Foulds (RF) distances were calculated between the structural phylogeny and each hallmark-gene tree. RF measures the number of internal bipartitions (splits) that differ between two trees, with lower values indicating greater agreement. All three hallmark-gene phylogenies exhibited moderate similarity to the structural tree, with normalized RF distances of 0.76 (MCP), 0.73 (TerL), and 0.69 (portal), corresponding to approximately 24%, 27%, and 31% shared internal bipartitions, respectively. Because RF is a stringent metric that penalizes every discordant internal bipartition, we additionally assessed ICTV family-level monophyly as a measure of taxonomic coherence. The structural proteome tree recovered 57.4% of families as monophyletic, compared with 48.1% for MCP, 57.7% for TerL, and 63.8% for portal. Although the portal tree exhibited the lowest normalized RF distance and the highest family-level monophyly score, these results should be interpreted cautiously, as the comparison was based on a smaller subset of confidently identified portal proteins (n = 3,487) than the MCP and TerL analyses, limiting direct comparison among marker genes. Notably, TerL recovered ICTV family-level groupings at a level comparable to the structural tree while maintaining near-complete taxonomic coverage, suggesting that it is one of the most informative single markers for reconstructing *Caudoviricetes* evolutionary history.

Together, these analyses indicate that individual hallmark genes capture a considerable fraction of the evolutionary signal present within the structural proteome tree, with TerL emerging as the most informative single marker for reconstructing *Caudoviricetes* taxonomy. Although TerL recovered family-level monophyly comparable to that of the structural proteome phylogeny, no individual marker reproduced the full topology of the genome-wide structural tree. These findings suggest that whole-proteome structural approaches capture additional phylogenetic information distributed across numerous proteins, providing evolutionary resolution beyond that obtainable from any single hallmark gene.

Pfam domain prediction was used to complement RefSeq/GenBank annotations, providing functional context for hypothetical sequences and enabling the

identification of common folds across *Caudoviricetes*' proteomes. The two most abundant domains were uncharacterised families: PF06569 (DUF1128; 6,300 occurrences), which corresponds to a family of short, hypothetical bacterial proteins of unknown function, and PF06174 (DUF987; 4,106 occurrences), another poorly characterised bacterial protein family related to the hypothetical protein YeeT. Their prevalence highlights how some of the most frequently encoded viral proteins remain completely functionally unexplored.

Among annotated domains, PF03206 (NifW; 4,055 occurrences) was unexpectedly wide-spread. Functioning as an accessory protein in diazotroph nitrogen fixation[137], its prevalence in bacteriophages suggests repeated acquisition as an auxiliary metabolic gene. PF07126 (ZapC; 3,863 occurrences) was also strikingly common. ZapC is an FtsZ-binding factor that stabilises the bacterial Z-ring[138], and its presence in phage genomes points to potential re-current capture of host cytokinesis regulators. Such domains may enable phages to manipulate host cell division—either delaying cytokinesis to extend the replication window or perturbing Z-ring dynamics to coordinate lysis with a specific cell-cycle stage. Finally, PF13392 (HNH endonuclease; 3,164 occurrences) was frequently detected, often positioned near terminase genes[139], consistent with roles in nicking, recombination, genome packaging, and, in some cases, mobility through homing activity[140]. Together, these findings reveal a functionally rich and largely uncharacterised genetic armoury of phage-encoded domains, highlighting both the extent of viral novelty and the diverse mechanisms by which *Caudoviricetes* interact with their bacterial hosts (Supplementary Figure 2.10).

2.3.2 Structure-based taxonomy aligns with ICTV classifications and exposes gaps in higher taxonomic ranks

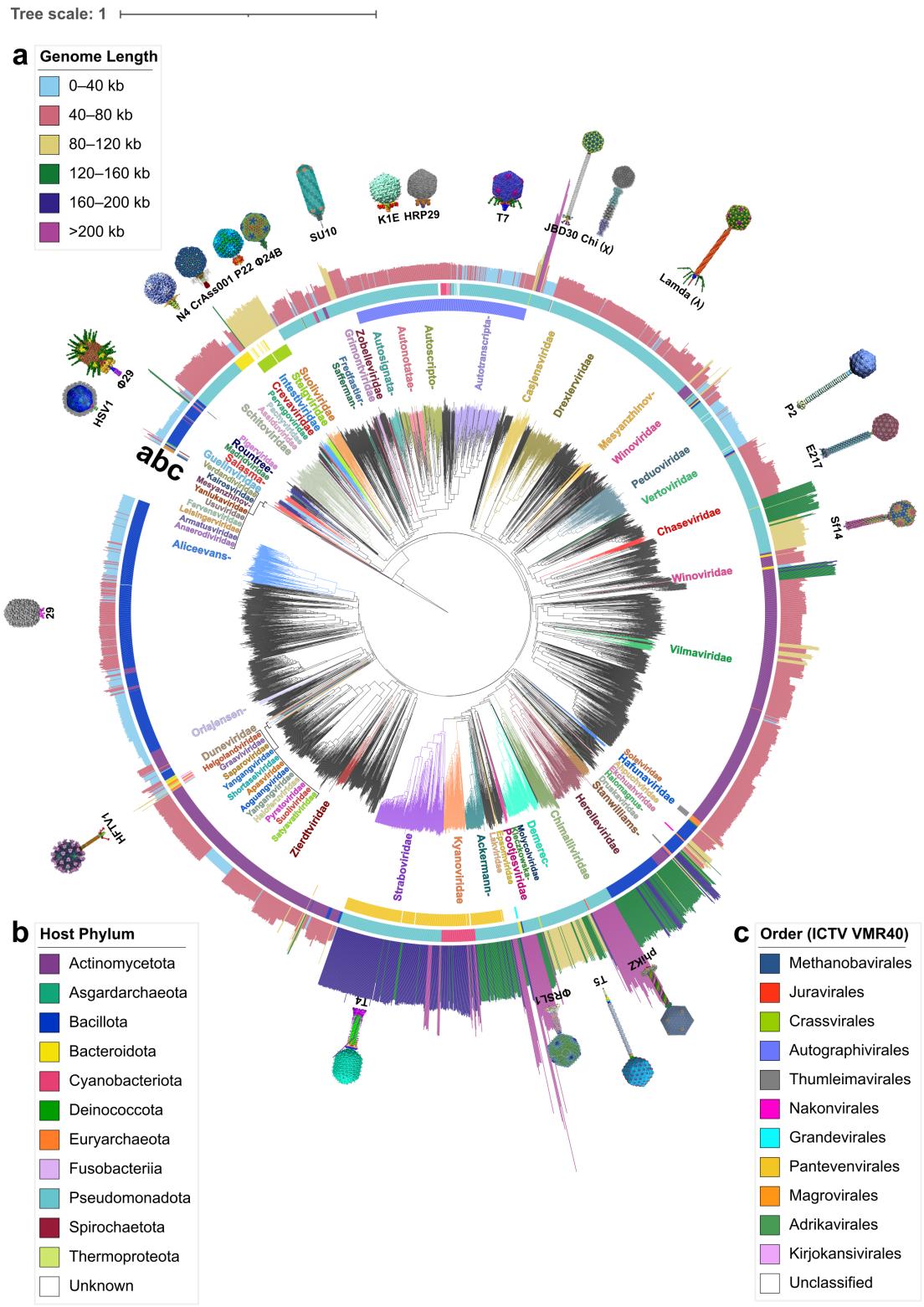


Figure 2.2: Structure-informed phylogeny of 4,082 *Caudoviricetes* genomes.

Individual leaves are coloured by ICTV family, enabling direct visual comparison between structure-based clustering and established taxonomy; leaves and branches shown in black correspond to genomes without an assigned ICTV VMR40 family. Cryo-EM-resolved virion models are mapped to their corresponding leaves, providing morphological reference points across characterised viral lineages. The three outer annotation rings show: (a) genome length, represented by six size classes from $<$ 40 kb to $>$ 200 kb; (b) predicted host phylum, spanning major bacterial and archaeal groups; and (c) ICTV VMR40 order-level taxonomy.

To evaluate how structural similarity reflects existing ICTV taxonomy, we constructed a neighbour-joining phylogram from a presence-absence matrix of structural protein clusters. Each genome was encoded as a binary profile of its structural proteome, pairwise Dice distances were computed, and the resulting tree was rooted using Herpes simplex virus 1 (HSV-1), a distant but appropriate outgroup from the realm *Duplodnaviria*, which encompasses both *Caudoviricetes* and *Herpesviricetes*. The global topology preserves the major ICTV-defined families and orders (Figure 2.2, leaf colours designate family, annotation ring C), supporting the use of structure-based clustering for resolving evolutionary relationships and guiding the placement of unclassified genomes.

To quantify taxonomic concordance, we assessed monophyly using precision scores based on the last common ancestor (LCA) of each ICTV-assigned group (minimum $n \geq 2$ genomes). Of the 80 ICTV families present, 55 met this criterion. Among these 55 families, 30 (54.5%) were perfectly monophyletic (precision = 1.0), and 38 (69.1%) showed precision > 0.85, reflecting strong congruence between structure-based clustering and ICTV family boundaries. Nine well-sampled families ($\geq$ 10 members)—*Herelleviridae* (115 members), *Drexlerviridae* (113), *Peduoviridae* (101), *Autosignataviridae* (36), *Zierdtviridae* (26), *Chaseviridae* (15), *Orlajensenviridae* (11), *Pootjesviridae* (11), and *Hafunaviridae* (10)—were recovered as fully monophyletic in the structural phylogeny, with no extraneous taxa within their last common ancestors (LCAs) and precision scores of 1.0. Near-perfect families included *Ackermannviridae* (precision = 0.984) and *Straboviridae* (0.981), where minor deviations stemmed from the integration of previously unclassified genomes that were recruited to the structural clade.

Several smaller families ($n \leq 4$), including *Druskaviridae*, *Ekchuahviridae*, *Fredfastierviridae*, *Graaviviridae*, *Haloferuviridae*, *Kleczkowskaviridae*, *Lakviridae*, *Madridviridae*, *Saffermanviridae*, *Saparoviridae*, and *Vertoviridae*, were also recovered as monophyletic. However, their limited sampling provides weaker statistical support and precludes robust conclusions regarding their long-term evolutionary stability.

Inconsistencies were apparent for several families. *Rountreeviridae*, *Salasmaviridae*, and *Guelinviridae* were fragmented across neighbouring clades, with *Guelinviridae* embedded between two branches of *Salasmaviridae*, resulting in low precision scores (0.521, 0.352, and 0.172, respectively). These families all have small genomes (16-20 kb; Figure 2.2a), podovirus-like morphology and infect *Bacillota* hosts. Their interspersed placement suggests that sequence-based analyses either lacked the sensitivity to detect their relatedness or that these lineages were never systematically compared. The structural analysis therefore reveals an underlying

evolutionary connection that was previously obscured.

A closer examination of families within the order *Autographivirales* (podovirus-like phages mainly infecting *Pseudomonadota*) revealed that *Autosignataviridae* formed a clean, fully monophyletic clade (precision = 1.00), while *Autotranscriptaviridae* also displayed a largely monophyletic structure (precision = 0.91) with the incorporation of additional genomes. In contrast, *Autonotataviridae* showed low precision (0.28) as its members were split between two neighbouring clades of *Autoscriptoviridae*, suggesting that these two families actually represent a single evolutionary lineage rather than distinct groups. Notably, a set of unassigned ge-nomes positioned between *Autoscriptoviridae* and *Autotranscriptaviridae* formed a coherent, well-supported cluster in the RED-normalised tree discussed below. Based on this stability—and consistent with the naming theme of *Autographivirales* (self-transcribing viruses that encode their own RNA polymerase)—we designate this lineage as a new candidate family, "*Autocodicoviridae*", representing a distinct and previously unresolved clade.

At the order level, several taxa exhibited strong structural concordance. *Nakonvirales* and *Grandevirales* were fully monophyletic (precision = 1.0), while *Crassvirales* (0.973), *Pantevenvirales* (0.949), and *Autographivirales* (0.918) showed high precision scores, indicating substantial agreement between structure-based clustering and existing order assignments (Figure 2.2c). Minor departures from perfect precision in these groups were driven by the incorporation of previously unclassified genomes that clustered naturally within these orders. In contrast, orders of archaeal viruses such as *Methanobavirales*, *Thumleimavirales*, *Kirjokansivirales*, *Magrovirales*, and *Juravirales* were highly fragmented, with precision scores ranging from 0.001 to 0.031 (Figure 2.2c).

Subfamily- and genus-level classifications showed the highest consistency with structural clustering. Of the 121 subfamilies present, 85 (70.2%) were perfectly monophyletic, and among 584 genera, 503 (86.1%) showed complete monophyly (precision = 1.0) (Supplementary Figure 2.11).

2.3.3 Taxonomic Expansion Through Relative Evolutionary Divergence (RED) Scores and the definition of new candidate taxa

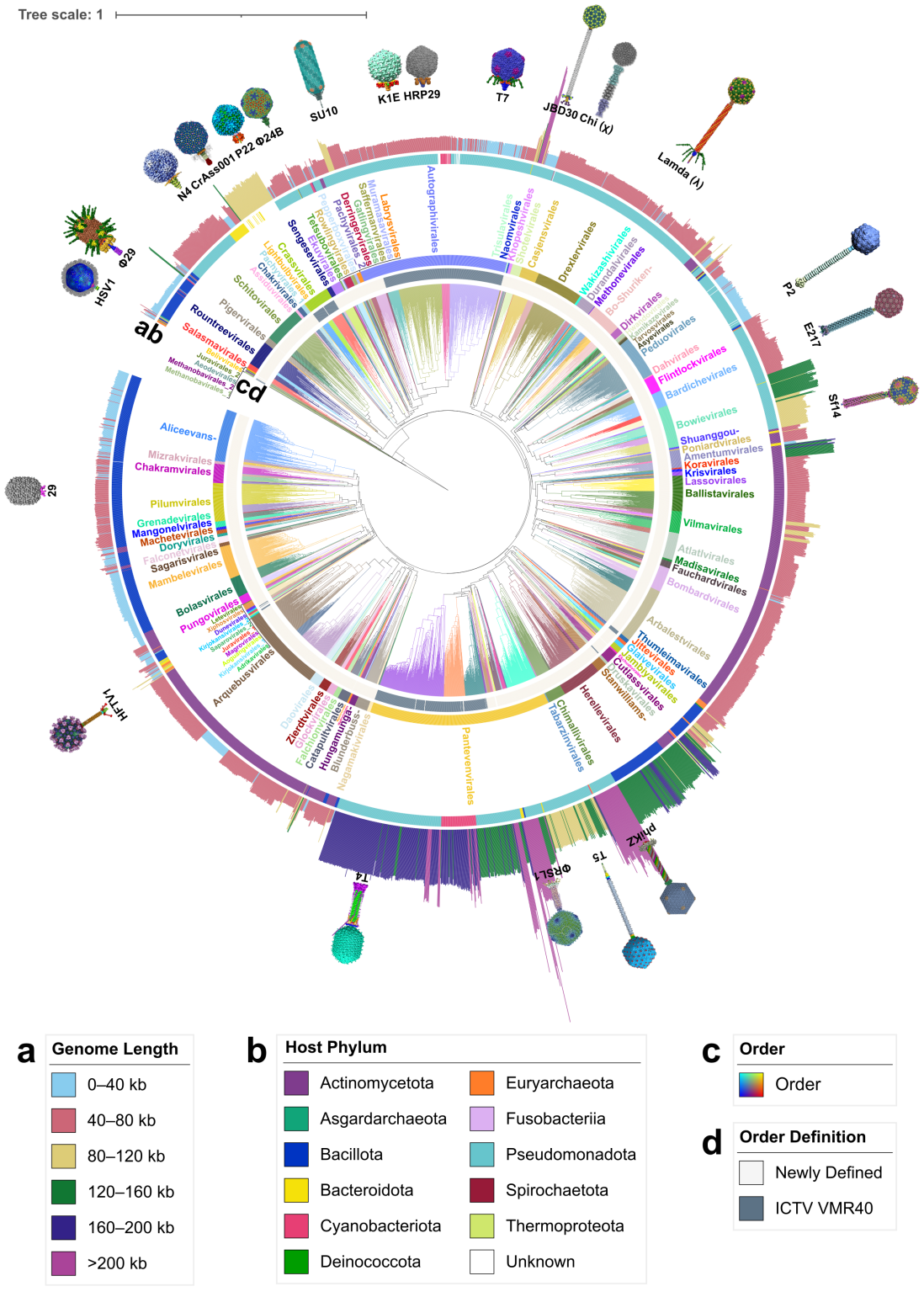


Figure 2.3: RED-scaled phylogenetic atlas of *Caudoviricetes*.

The RED-scaled phylogram displays the diversity and organization of *Caudoviricetes* genomes as resolved by our expanded phylogenomic and taxonomic framework.
(a) Genome length distribution across the tree, grouped into six size classes.
(b) Host phylum annotations for each virus, highlighting broad host range and phylum-specific clustering.
(c) Order-level structure, with coloured blocks marking taxonomic boundaries.
(d) Order definition: comparison of ICTV VMR40 orders with newly defined orders identified in this study. Together, these layers form the first comprehensive, unified taxonomic framework for *Caudoviricetes*, expanding higher-order classifications across the class and contextualising their immense diversity in terms of genome size, host associations, and order- and family-level boundaries.

To resolve polyphyletic ICTV taxa, standardise ambiguous classifications, and enable consistent taxonomic assignment across the structural phylogeny, we applied a topology-aware, rank-normalised framework based on Relative Evolutionary Divergence (RED). RED quantifies the relative position of each node along a rooted tree (0 = root, 1 = tips), with internal nodes interpolated linearly from parent to descendants, following the approach used by the Genome Taxonomy Database (GTDB) for bacterial and archaeal phylogenomics[141]. Unlike absolute branch-length thresholds, RED provides relative, rank-scaled cut-offs that enable a reproducible and scalable classification system.

RED values were computed on the structural phylogeny using PhyloRank. Optimal RED thresholds for each taxonomic rank were identified by scanning RED cut-offs from 0 to 1 and maximizing agreement with ICTV taxonomy, measured using Adjusted Mutual Information (AMI). Peaks in AMI identified rank-specific RED intervals: Order (0.46-0.54), Family (0.54-0.63), Subfamily (0.63-0.77), Genus (0.77-0.92), and Species (0.92-1.00) (Supplementary Figure 2.12, Supplementary Figure 2.13). These thresholds were applied to resolve polyphyletic taxa and assign ranks to previously unclassified clades.

RED-defined clades showed strong concordance with ICTV taxonomy. At optimal thresholds, 55% of orders (6/11), 82.3% of families (65/79), 82.3% of subfamilies (102/124), and 95.2% of genera (1,302/1,367) were recovered as monophyletic. For polyphyletic taxa, the largest co-herent clade retained the ICTV name, and smaller discordant clades were assigned place-holder names (e.g., *Herelleviridae_1*, *Herelleviridae_2*). Novel orders were named after their dominant family; novel families after classical weaponry; and novel subfamilies after Solar System moons (Supplementary File 2). At the order level, 146 novel orders were created, with an average of 20.7 genomes per order (median = 8, range = 1-244), including 3 singleton clades. At the family level, 200 new families were defined, with a mean size of 10.6 genomes (median = 4, range = 1-118), including 23 singletons. Subfamily-level definitions showed even greater granularity, with 356 novel subfamilies averaging 5.6 genomes each (median = 3, range = 1-95), including 96 singletons. In total, this approach yielded 159 orders, 269 families, 503 subfamilies, and 1,189 genera, establishing a systematic, phylogeny-informed expansion of *Caudoviricetes* taxonomy.

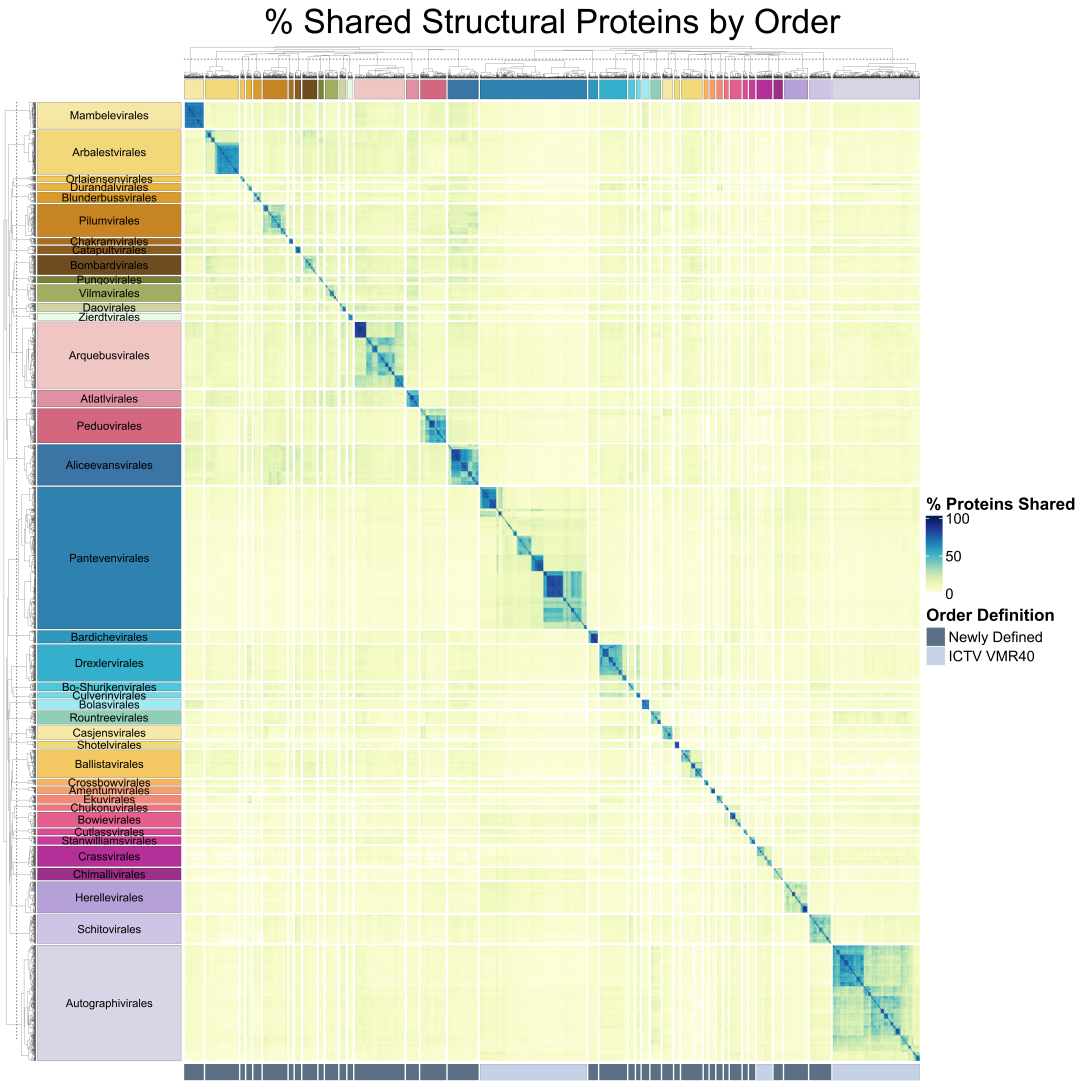


Figure 2.4: Structural protein similarity across *Caudoviricetes* genomes, visualised as clusters grouped by Order. The lower panel compares the newly defined RED-based order assignments with those reported in ICTV VMR40, marking a majority of newly created orders. Only orders with at least 20 genomes are shown (39 included; 120 excluded).

2.3.4 Structural Relatedness Across *Caudoviricetes*

To quantify relatedness within and between taxonomic ranks, pairwise percentages of shared structural protein clusters for each group (intra-rank) and between groups (inter-rank) were calculated. Intra-rank similarity followed the expected hierarchical trend, increasing toward lower ranks: genus-level groups shared an average of $55.7 \pm 3.5\%$, followed by subfamily ($44.9 \pm 8.4\%$), family ($37.5 \pm 10.6\%$), and order ($32.8 \pm 12.1\%$; Fig. 4). In contrast, inter-rank similarity was uniformly low across all levels, with mean values of 4.8–5.2% and very low variability (e.g., standard deviation = 0.5% at genus, 1.5% at order). These consistently low similarities reflect strong, clade-specific proteomic signatures and the resulting well-defined boundaries between taxonomic groups (Supplementary Figure 2.14).

2.3.5 Machine Learning-Based Taxonomic Classification

To extend this analysis into a predictive framework for assigning taxonomy to novel phage genomes, binary Random Forest classifiers were trained across the four hierarchical taxonomic ranks: order, family, subfamily, and genus. Each genome was encoded as a binary vector indicating the presence or absence of 20,289 non-redundant structural protein clusters, providing a representation of its structural proteome. Classifier training was supervised using labels derived from revised and extended taxonomic assignments based on Relative Evolutionary Divergence (RED) cut-offs.

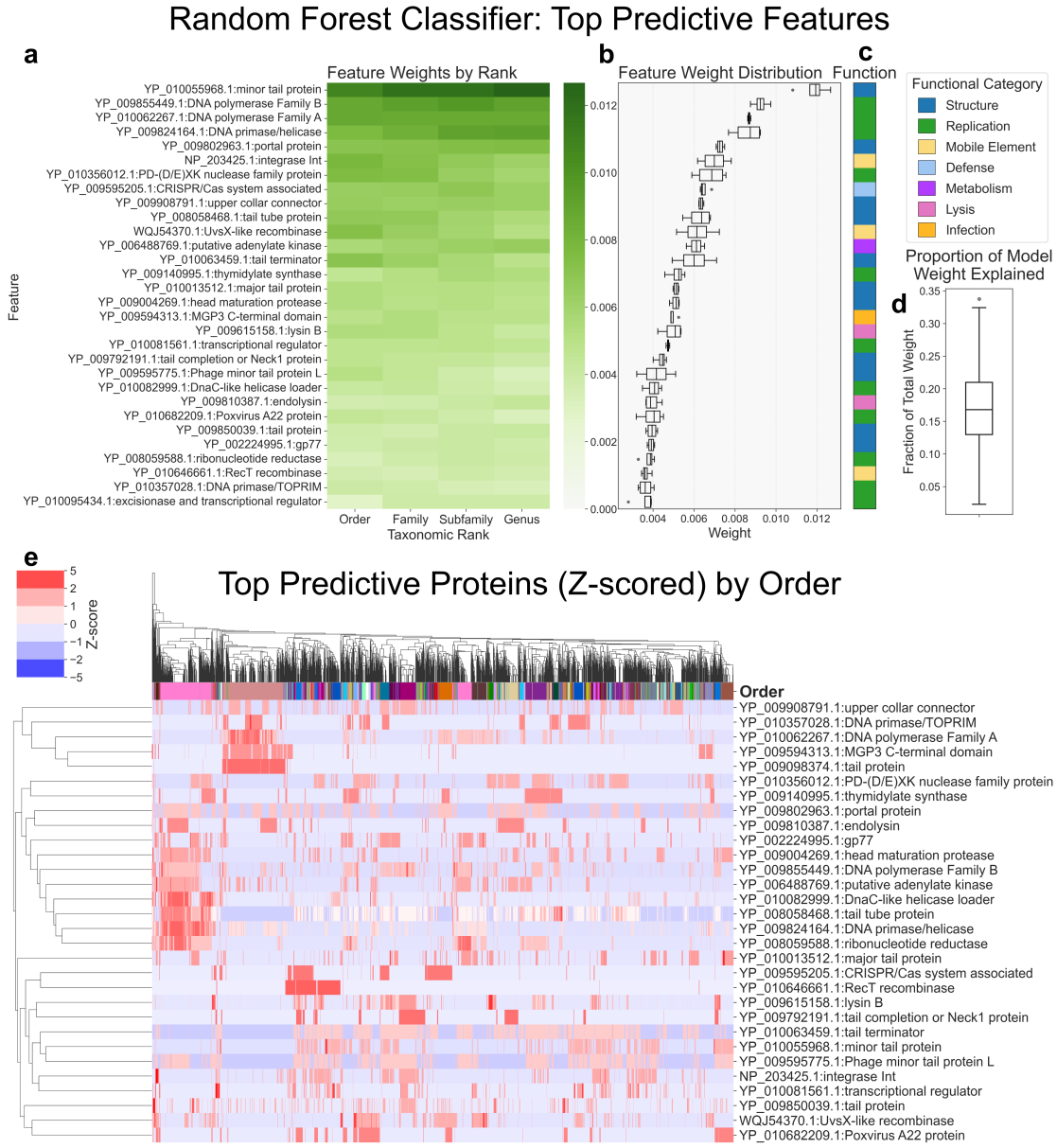


Figure 2.5: Top Predictive Protein Features Used by the Random Forest Classifier Across Viral Taxonomic Ranks.
(a) Feature-weight heatmap showing the proteins most informative for distinguishing taxonomic ranks. High-weight features are dominated by structural proteins (e.g., tail, portal) and replication enzymes (e.g., DNA polymerases, primase/helicase).

(b) Distribution of feature weights highlighting the top-ranked proteins by median importance.
(c) Functional annotations for each protein, colour-coded by category. (d) Contribution of the top features to total model weight, showing a strong enrichment for structural and replication-associated functions.
(e) Hierarchically clustered z-scored presence/absence profiles of the top predictive proteins across viral orders, illustrating clear order-specific usage patterns.

Across all ranks, the most predictive features for distinguishing taxonomic groups were a minor tail protein (YP_010055968.1), two DNA polymerases (Family B, YP_009855449.1, and Family A, YP_010062267.1), a DNA primase/helicase (YP_009824164.1), and a portal protein (YP_009802963.1) (Figure 2.5). Among the top 30 features, 22 corresponded to structural or replicative proteins, reflecting both their central roles in *Caudoviricetes* biology and their lineage-specific diversification. Collectively, these 30 proteins accounted for 15% of total model weight, demonstrating that a relatively small subset of features disproportionately contributes to classification accuracy and confirming that the model is capturing meaningful biological signal. Interestingly, despite the well-documented horizontal mobility of *Caudoviricetes* polymerases[142], the Random Forest model still exploits subtle lineage-specific variation in these genes—highlighting its ability to identify discriminative patterns even in proteins with limited deep evolutionary stability.

Classifier performance was evaluated using controlled downsampling, with each taxon subsampled to fixed training sizes. At the order and family levels, performance rapidly plateaued, with AUROC, accuracy, precision, recall, and F1 scores exceeding 0.95 once ≥ 10 genomes per taxon were available (Supplementary Figure 2.15, Supplementary Figure 2.16). Subfamily-level classification showed similarly strong performance, peaking at moderate training sizes and highlighting this rank as a balance between phylogenetic resolution and robustness (Supplementary Figure 2.17). At the genus level, AUROC, AUPRC, and accuracy remained high across all training sizes, while precision and F1 were more sensitive to training set composition, showing increased variability at small training sizes and modest declines at larger sizes. Despite this, strong separation between genera was achieved with moderate training sizes, with AUROC, AUPRC, and accuracy exceeding 0.95 at approximately 10 genomes per genus (Supplementary Figure 2.18).

The final production classifier was trained on all available taxa without downsampling, maximizing its exposure to the full diversity of known *Caudoviricetes*. To evaluate performance on unseen data, the production models were tested on an independent set of 1,746 ICTV-classified *Caudoviricetes* genomes (131,003 proteins) that were excluded from training. Of the 603 genomes with order-level labels matching those in the training set, 602 (99.8%) were correctly classified, suggesting that the single misclassified genome may have been incorrectly labeled in the reference dataset rather than mispredicted by the model. At the family level, 785 out of 854 genomes (91.9%) were correctly predicted; 941 of 1,019 subfamily labels (92.4%) were accurate, and 998 out of 1,141 genus-level predictions (87.4%) matched their ground truth assignments (Supplementary Figure 2.19).

To more closely investigate the biological basis of model predictions and assess whether classification performance was driven by meaningful taxonomic signal rather than noise, we examined Random Forest feature importance profiles across multiple taxonomic ranks (Supplementary Figure 2.20-Figure 2.25). Distinct lineage-specific signatures were observed at both order and family levels, including informative tail proteins, baseplate proteins, DNA polymerases, and tail assembly chaperones, while many of the highest-ranking features corresponded to proteins of unknown function. Together, these results indicate that classification is driven by structured combinations of taxon-specific protein features rather than reliance on a small number of conserved marker genes.

Analysis of feature importance distributions revealed a strongly right-skewed pattern, with predictive signal distributed across thousands of protein features rather than concentrated within a small number of markers (Supplementary Figure 2.26). Although a relatively small subset of highly informative proteins contributed disproportionately to classification performance, approximately 4,608 and 7,090 features were required to explain 90% and 95% of total model importance, respectively. These results suggest that future versions of the classifier could be substantially accelerated through feature-selection approaches that restrict training to the highest-ranking structural protein features, potentially reducing computational cost while retaining the majority of predictive signal.

Given the rapid expansion of viral metagenomic sequencing, we anticipate substantial performance gains in future iterations of the classifier. Each newly validated *Caudoviricetes* genome enriches the training space, reduces ambiguity at deeper phylogenetic splits—particularly within genera and subfamilies that are currently under-represented—and increases the model's ability to discriminate among closely related taxa. Consequently, classifier performance is expected to scale upward as taxonomic coverage grows, further strengthening robustness, resolution, and confidence in predictions for novel genomes.

2.3.6 PhagePleats: A Tool for Taxonomic Classification of Phages

This study culminated in the development of PhagePleats `https://github.com/linda5mith/PhagePleats`, a scalable Python tool for taxonomic classification of *Caudoviricetes* genomes using structural rather than sequence-based features. PhagePleats uses a reference database of 20,289 structural protein clusters—the Foldseek-derived representatives from the optimal clustering run, collectively capturing the structural diversity of 445,151 phage protein models. For each query genome, its pre-folded proteins (ESMFold PDB format) are searched against this structural database using Foldseek. Matched cluster representatives are encoded into a binary structural presence–absence vector, which is passed to a suite of one-vs-rest Random Forest classifiers to generate taxonomic predictions at the order, family, subfamily, and genus levels. To ensure robustness in the face of the vast uncharacterised diversity of the virosphere, PhagePleats incorporates a novelty-aware quality control layer. Each query genome is compared to the training

set using FAISS-based nearest-neighbour search[143] on Jaccard similarity (percentage of shared structural clusters). This serves both as a secondary validation of classifier predictions, and a safeguard for detecting highly divergent or genuinely novel genomes. Genomes sharing near-zero structural similarity with the reference set are automatically flagged as novel.

To quantify divergence more formally, intra-clade statistics were computed for all training clades, including the mean and variance of pairwise shared-protein percentages. Query genomes receive z-scores relative to the distribution of their predicted clade, enabling graduated novelty calls (e.g., "likely member", "possible new subfamily", "potentially novel genus", "novel taxa"). This provides an interpretable framework for distinguishing well-supported assignments from outliers that may represent new evolutionary lineages. Because the system is built on structural features that scale with available data, PhagePleats is intrinsically extensible. As new *Caudoviricetes* genomes are discovered and incorporated, the structural cluster reference database and training matrices can be seamlessly expanded, improving taxonomic resolution and enabling increasingly accurate and sensitive classification of the rapidly growing global virome.

2.3.7 Conclusion

We present PhagePleats, a Python-based framework for scalable, objective, and reproducible taxonomic classification of *Caudoviricetes* viruses based on structural proteome content. By integrating whole-structurome similarity, RED-based phylogenetic normalisation, and machine learning, our approach achieves high congruency with existing ICTV taxonomy. Additionally, it expands the current classification system by assigning higher-level taxonomic ranks and delineating novel clades and lineages across the dataset. Importantly, this framework lays the foundation for a universal, structure-informed viral classification system capable of extending across the virosphere. As genomic databases continue to expand and the efforts of Viral Study Groups refine and formalise viral groupings, the integration of newly characterised genomes will further strengthen model performance and taxonomic resolution. With the ongoing growth of curated viral data, this approach holds the potential to assign taxonomy to all known and future viral genomes, bridging current classification gaps and supporting a unified, data-driven taxonomy for viral diversity.

2.4 Methods

2.4.1 Folding and Selection of Sample Genomes

A total of 4,082 representative *Caudoviricetes* genomes were selected from ICTV VMR40 to maximise phylogenetic and taxonomic breadth (4083_ICTV_metadata.csv), and their corresponding protein sequences were directly obtained from the RefSeq/GenBank annotations associated with their accessions. Herpes simplex virus 1 (HSV-1; NC_001806.2) was added as an outgroup, yielding 4,083 total genomes. All 445,151 phage proteins (plus 53 HSV-1 proteins) were folded using the esmfold_v1() model from ESMFold[125]. Folding was

performed in an environment derived from the project-provided environment.yml, with updates for A100 GPU compatibility (PyTorch 1.9.0+cu111). Structures were generated using: python fold.py –chunk-size 128.

2.4.2 Structural Clustering and Annotation

All folded proteins were clustered using Foldseek (v7.04e0ec8) easy-cluster[126], run across a broad grid of sensitivity/specificity parameters (run_foldseek_batches.sh). Result files were imported into Notebook 1 for benchmarking https://github.com/linda5mith/Structural_Taxonomy_Caudoviricetes/. Protein metadata—including RefSeq descriptions, accessions, and ICTV taxa—were merged into cluster assign-ments using in-house Python scripts. 249 truncated TerL proteins were identified (visually in PyMOL and via length filtering ¡50% of median length) and removed prior to downstream analysis.

2.4.3 Cluster Benchmarking Framework

We evaluated cluster quality using four complementary criteria:

1. **Marker-protein cluster size.** Three canonical *Caudoviricetes* structural markers were used as internal controls:

   - Major capsid protein (MCP)
   - Terminase large subunit (TerL)
   - Portal protein

   Parameter sets were retained when the median cluster size was $\geq$ 3,000 across these markers.

2. **Functional specificity.** Functional homogeneity was assessed using Pfam predictions generated with protein-vec (https://github.com/tymor22/protein-vec) and quantified using the inverse Shannon entropy of domain composition. Cluster parameters yielding a median specificity score $\geq$ 0.7 across the three canonical markers were retained.

3. **Structural coherence (PyMOL RMSD validation).** For each parameter set, the five largest clusters were structurally validated using all-versus-representative RMSD calculations performed in PyMOL v3.0.0 (open-source build c7e389c0)[144] through the `pymol2` API. We computed:

   - median per-cluster RMSD,
   - summary RMSD (median of per-cluster medians), and
   - RMSD dispersion (standard deviation).

4. **Global fold-quality validation (US-align).** To independently validate the structural consistency of the Foldseek-derived clusters, all members of each cluster (20,289 clusters total) were structurally aligned to their corresponding representative using US-align[145]. For each member–representative comparison, TM-score, RMSD, and aligned-length values were extracted, enabling

quantitative assessment of intra-cluster structural coherence across the entire dataset. For each cluster, we extracted:

- median TM-score,
- median RMSD,
- TM-score interquartile range, and
- protein length distribution statistics, including mean, median, minimum, maximum, standard deviation, absolute range, and fold-range (max_len/min_len).

Across all metrics, the best-performing Foldseek `easy-cluster` configuration was: coverage = 40%, TM-score = 0.30, E-value = 1×10^{-2}, and `cluster-reassign` = 1. This run produced the largest median cluster size (3,367) and the lowest median RMSD (4.03 Å) (Supplementary Figure 2.28).

2.4.4 Pfam Domain Annotation

To augment existing RefSeq annotations, Pfam predictions were generated using protein-vec (`https://github.com/tymor22/protein-vec`). The software environment was adapted for CUDA 11.4 with PyTorch 1.12.1. Predicted Pfam accessions were linked to domain definitions from the Pfam-A seed database (`https://www.ebi.ac.uk/interpro/download/Pfam/Pfam-A.seed.gz`) using custom Python scripts (`data/scripts/protein-vec/`).

2.4.5 Host Annotation and Inference

Host metadata were retrieved from GenBank using the NCBI Entrez API. Of the 4,082 *Caudoviricetes* genomes, 1,625 (39.8%) included an explicit host qualifier, which was extracted directly. For genomes lacking submitted host information, host identity was inferred by parsing host-bearing substrings from virus names (e.g., *Gordonia* phage, *Pseudomonas* virus, *E. coli* phage T7). From these host species names, phylum-level assignments were inferred by mapping each species to its corresponding NCBI Taxonomy lineage using `esearch`/`efetch`, yielding biologically consistent assignments (e.g., *Gordonia* phage → *Gordonia* sp. → Actinomycetota). Names that could not be resolved to an NCBI taxon were labeled *Unknown* and excluded from phylum-level analyses. All host-extraction and taxonomy-mapping scripts are provided in Notebook 8.

2.4.6 Structural Presence–Absence Matrix

The optimal Foldseek clustering output was merged with protein metadata to assign each genome a set of structural protein clusters. A binary presence–absence matrix was constructed where:

- rows represent structural clusters,
- columns represent genomes, and
- values indicate presence (1) or absence (0).

Singleton clusters (present in fewer than two genomes) were excluded. Pairwise genome similarity was computed using the Sørensen–Dice coefficient. The resulting distance matrix was converted to PHYLIP format and passed to DecentTree (Neighbour-Joining) for tree construction. The tree was rooted on HSV-1 for downstream RED scaling. All code supporting this workflow is provided in Notebook 2.

2.4.7 Single-Hallmark Protein Phylogenies

Representative MCP, portal, and TerL proteins were identified from each genome and aligned using MAFFT v7.525 with the `--auto` strategy. Maximum-likelihood phylogenies were inferred using FastTree v2.1.11 under the LG amino acid substitution model (`-lg`) with CAT rate heterogeneity. Trees were rooted using the corresponding MCP, TerL, or portal proteins from HSV-1 and compared with the structural proteome phylogeny using Robinson–Foulds distances and ICTV family-level monophyly.

2.4.8 Assessment of ICTV VMR40 Taxonomic Monophyly

Agreement between ICTV VMR40 taxonomic assignments and the structure-based phylogenetic tree was assessed using an LCA-based monophyly test (Notebook 3). A Dice-distance-derived tree was analysed together with metadata linking each leaf to its ICTV taxonomic label. For each taxonomic group containing at least two genomes, the lowest common ancestor (LCA) of all member leaves was identified, and the set of descendant leaves was compared to group membership. Groups were classified as monophyletic if the LCA contained exactly the genomes assigned to that taxon. Deviations from strict monophyly were quantified using a precision metric, defined as the proportion of LCA-descendant leaves belonging to the focal taxon. Groups with fewer than two members were excluded from analysis.

2.4.9 RED-Based Phylogenetic Scaling and Taxonomic Boundary Inference

The rooted tree was processed using PhyloRank with the `--fixed_root` option to compute relative evolutionary divergence (RED) scores for all nodes. RED distributions were compared with ICTV-defined taxa to identify optimal boundaries at the order, family, and subfamily levels by maximizing Adjusted Mutual Information (AMI). Rank-specific RED windows (e.g., order: 0.46–0.54) were defined with a ± 0.03 tolerance for borderline clades. A greedy MRCA-selection algorithm identified a non-redundant set of representative clades per rank (Notebook 4). All rank assignments were manually verified in iTOL where necessary.

2.4.10 Random Forest Classifier Training

Random Forest classifier training, benchmarking, and evaluation were implemented in Notebook 5. Binary one-versus-rest Random Forest models were trained at the order, family, subfamily, and genus levels using:

- features: binary structural presence–absence vectors, and

- labels: RED-informed rank-level clades.

Class balance was enforced by downsampling each taxon to fixed training sizes (5, 10, 20, 50, and 100 genomes). Taxa with fewer than five representatives were excluded from cross-validation. Performance was evaluated using AUROC, AUPRC, precision, recall, F1 score, and cross-validation accuracy (Supplementary Figs. 7–10). After benchmarking, final production models were trained on the full dataset without downsampling.

2.4.11 Independent Test Set

To evaluate model generalization, 1,746 ICTV-curated *Caudoviricetes* genomes not used during training were folded and classified, yielding the following accuracies:

- Order: 602/603 correct (99.8%),
- Family: 785/854 (91.9%),
- Subfamily: 941/1,019 (92.4%), and
- Genus: 998/1,141 (87.4%).

Genomes belonging to taxa absent from the training set were excluded from evaluation.

2.4.12 Statistical Analysis and Visualization

All statistical analyses were performed using Python v3.8.19[146] in Jupyter Notebooks[147] and R v4.1.1[148]. In Python, data processing, statistical analyses, and visualization were conducted using `pandas` v2.0.3[149], `NumPy` v1.24.4[150], `scikit-learn` v1.3.2[151], `seaborn` v0.13.2[152], `matplotlib` v3.7.5[153], and `Entrezpy`[154]. In R, data manipulation and visualization relied on `ComplexHeatmap`[155], `ggplot2`[156], `dplyr`[157], `tidyr`[158], and `RColorBrewer`[159]. Protein structure visualizations were generated using PyMOL[160], comparative protein structure analyses were performed using Bio3D in R[161], and phylogenetic trees were visualised and annotated using iTOL[162].

2.4.13 Code Availability

All code and output data supporting this study are available in the GitHub repository at https://github.com/linda5mith/Structural_Taxonomy_Caudoviricetes/.

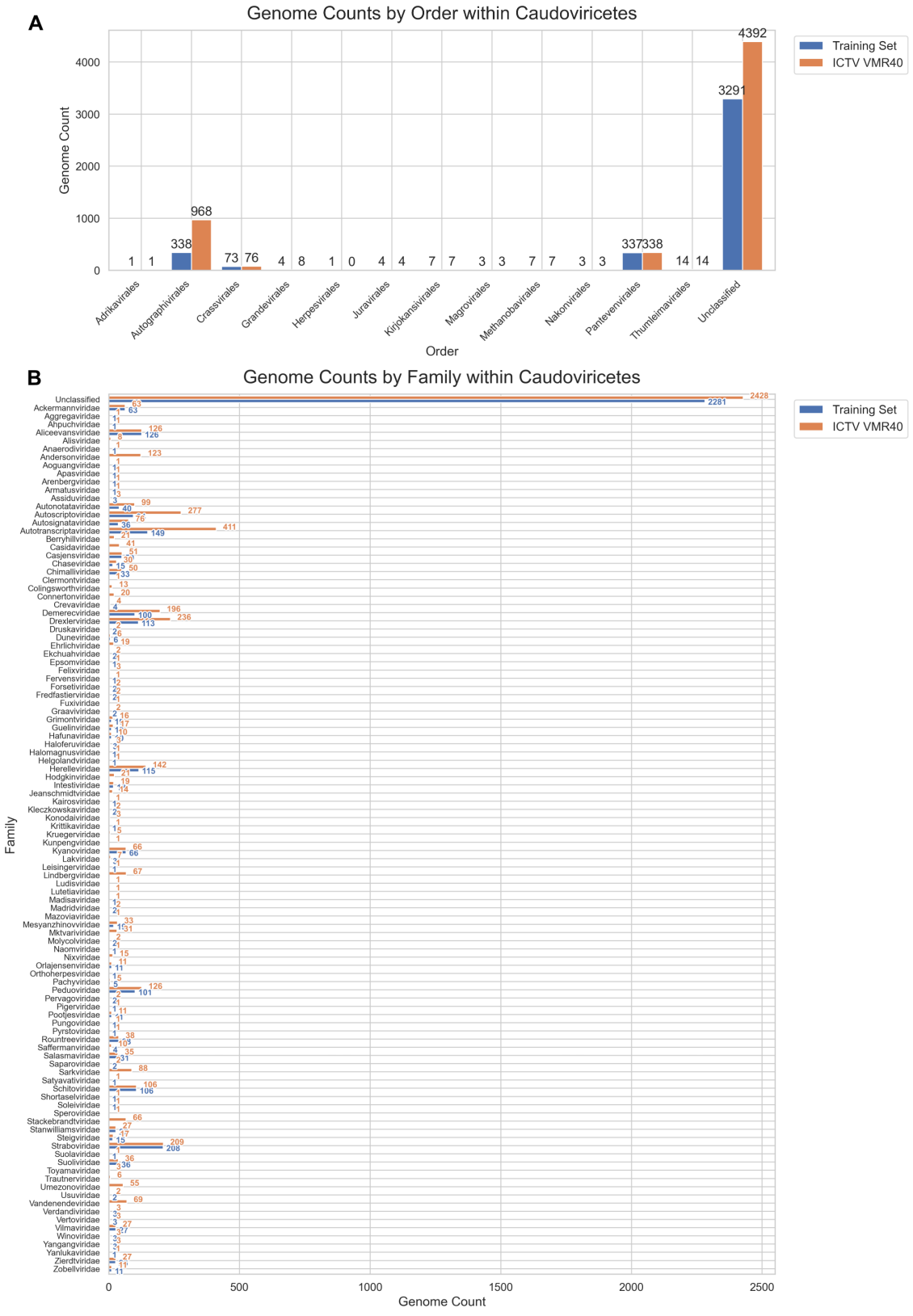
2.5 Acknowledgements

Linda Smith was funded by the Research Ireland Centre for Research Training in Genomics Data Science (Grant No. 18/CRT/6214). Andrey Shkoporov was funded by a Wellcome Trust Research Career Development Fellowship (Grant No. 220646/Z/20/Z). This research was funded in whole, or in part, by the Wellcome Trust and was conducted with the financial support of Research Ireland under Grant Number 12/RC/2273-P2. For the purpose of open access, the authors have applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

2.6 Author Contributions

Conceptualization: L.S., I.T., and A.N.S. Methodology: L.S., T.S.B.S., I.T., T.S., and A.N.S. Investigation: L.S. Writing – original draft: L.S. Writing – review and editing: all authors. Funding acquisition: L.S. and A.N.S. Resources: I.T., T.S., C.H., and A.N.S. Supervision: C.H. and A.N.S.

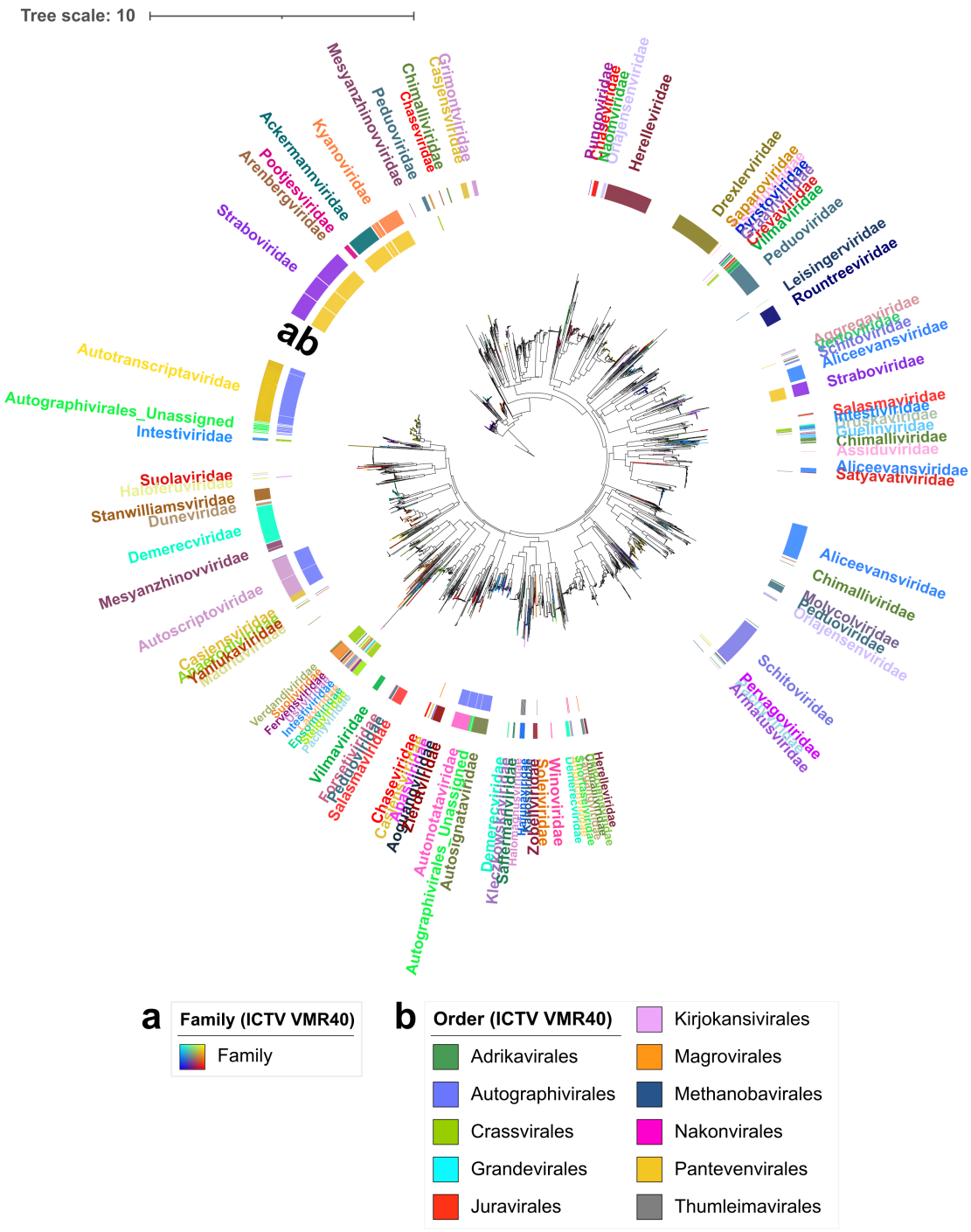
2.7 Supplementary Material



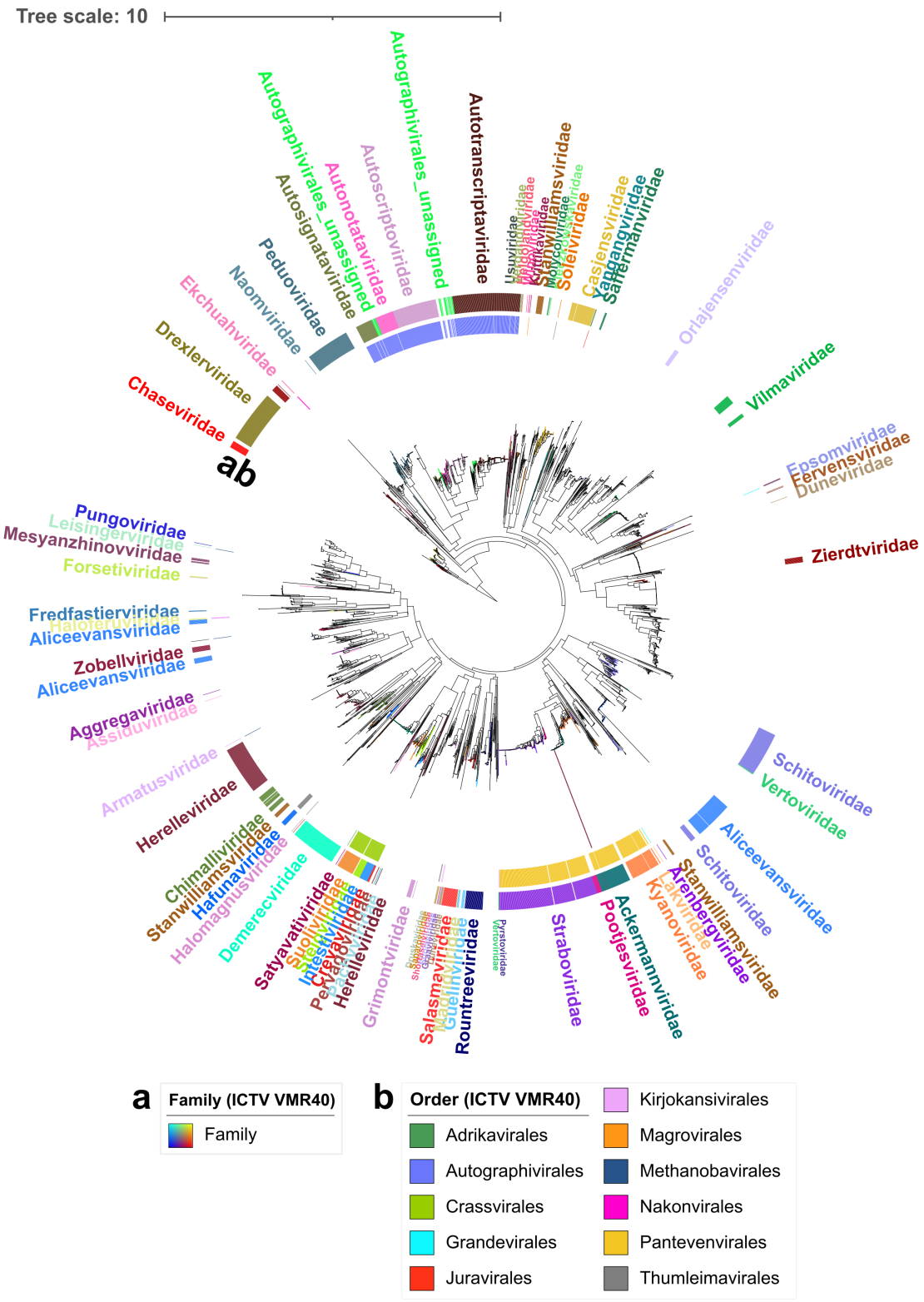
Supplementary Figure 2.6: Taxonomic representation of *Caudoviricetes* genomes in the training set versus the ICTV VMR40 dataset.

(A) Bar plot showing genome counts by order for both the training set (blue) and the complete ICTV VMR40 dataset (orange). While orders such as *Autographivirales*, *Crassvirales*, and *Pantevenvirales* are well represented, the majority of genomes remain unclassified at the order level in ICTV VMR40.

(B) Bar plot showing genome counts by family, highlighting the uneven representation of families within *Caudoviricetes*. Several families are highly studied—such as *Autotranscriptaviridae*, *Autoscriptoviridae*, *Demerecviridae*, *Drexlerviridae*, *Herelleviridae*, *Schitoviridae*, and *Straboviridae*—each with over 100 members, while many others are sparsely represented or remain entirely unassigned. This highlights the absence of higher-level taxonomic definition within the current ICTV VMR40 classification scheme and underscores the need for an expanded framework to capture the diversity of tailed dsDNA viruses. Several families newly introduced in ICTV VMR40, including *Alisviridae*, *Andersonviridae*, and *Umezonoviridae*, were not captured in this analysis, underscoring the rapid pace at which diversity within this viral class is being uncovered and the need for ongoing model updates.

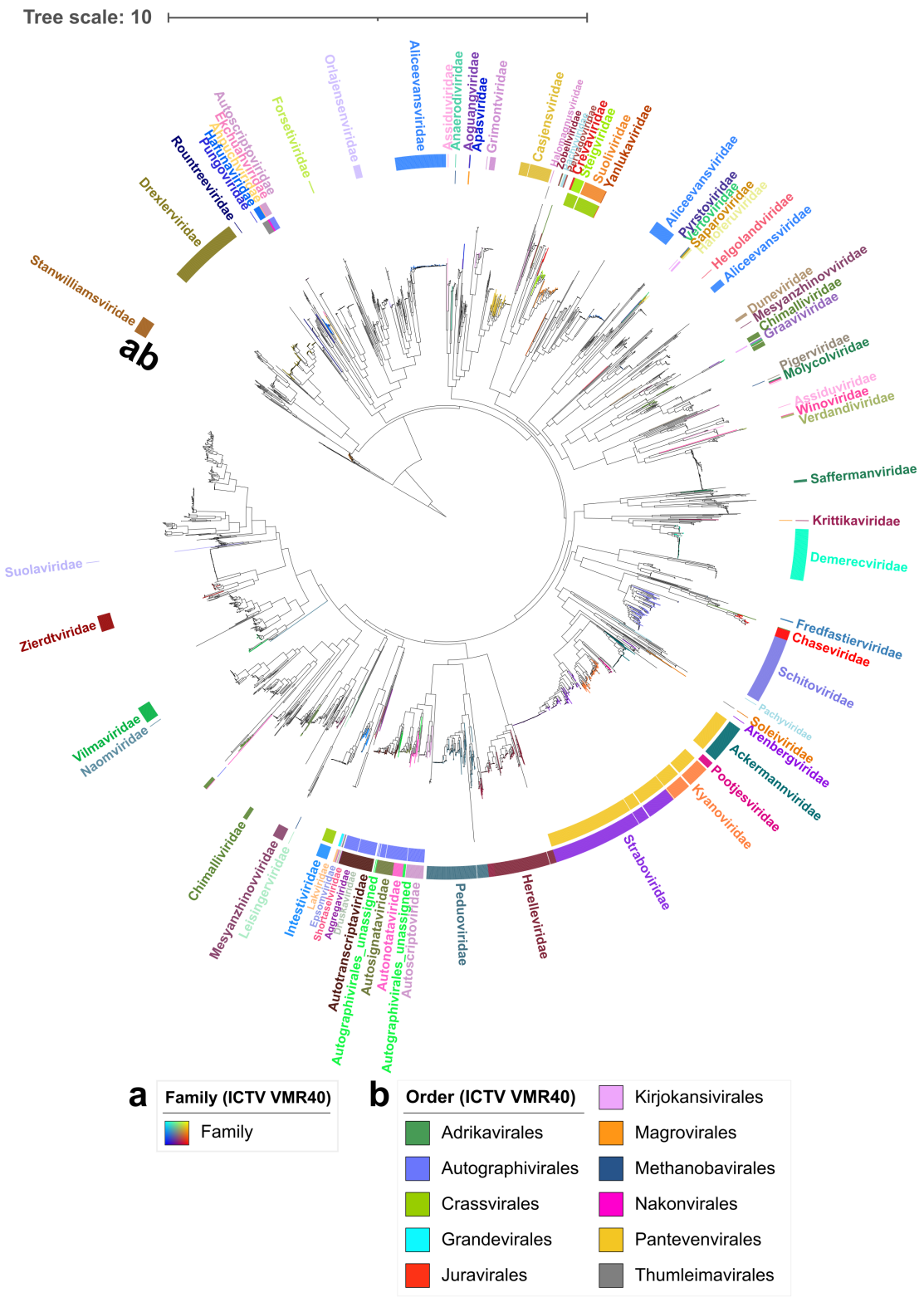


Supplementary Figure 2.7: Single-gene phylogeny inferred from Major Capsid Protein (MCP) sequences.
Maximum-likelihood phylogeny reconstructed from 4,021 confidently identified MCPs recoveredf rom unique *Caudoviricetes* genomes. Protein sequences were aligned using MAFFT and the phylogeny inferred using FastTree. Branches are annotated according to ICTV VMR40 family and order classifications. Comparison with the whole-proteome structural phylogeny revealed moderate topological similarity (normalised Robinson-Foulds distance = 0.76), corresponding to approximately 24% shared internal bipartitions. Family-level monophyly was 48.1%, compared with 57.4% recovered by the structural proteome phylogeny.



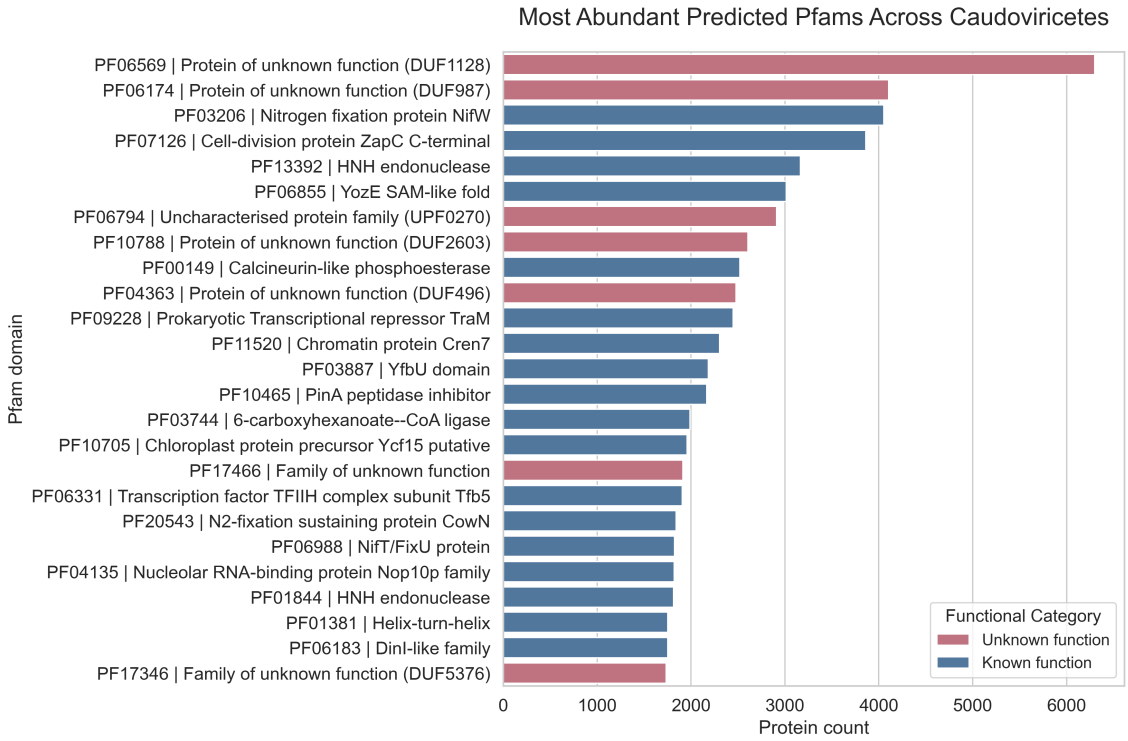
Supplementary Figure 2.8: Single-gene phylogeny inferred from Terminase large subunit (TerL) sequences.
Maximum-likelihood phylogeny reconstructed from 3,989 confidently identified TerLs recovered from unique *Caudoviricetes* genomes. Protein sequences were aligned using MAFFT and the phylogeny inferred using FastTree.

Branches are annotated according to ICTV VMR40 family and order classifications. Comparison with the whole-proteome structural phylogeny revealed moderate topological similarity (normalised Robinson–Foulds distance = 0.73), corresponding to approximately 27% shared internal bipartitions. Family-level monophyly (defined by ICTV VMR40) was 57.7%, closely matching the 57.4% recovered by the structural proteome phylogeny.

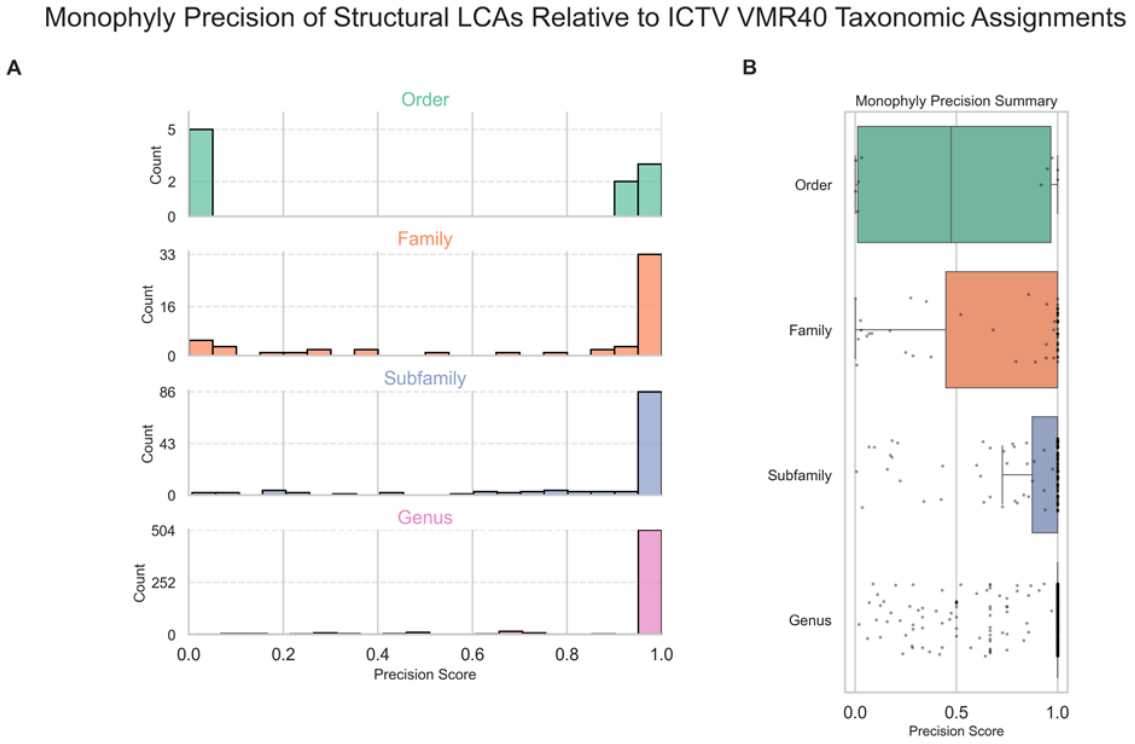


Supplementary Figure 2.9: Single-gene phylogeny inferred from portal protein sequences.
Maximum-likelihood phylogeny reconstructed from 3,487 confidently identified portal proteins recovered from unique *Caudoviricetes* genomes. Protein sequences were aligned using MAFFT and the phylogeny inferred using FastTree.

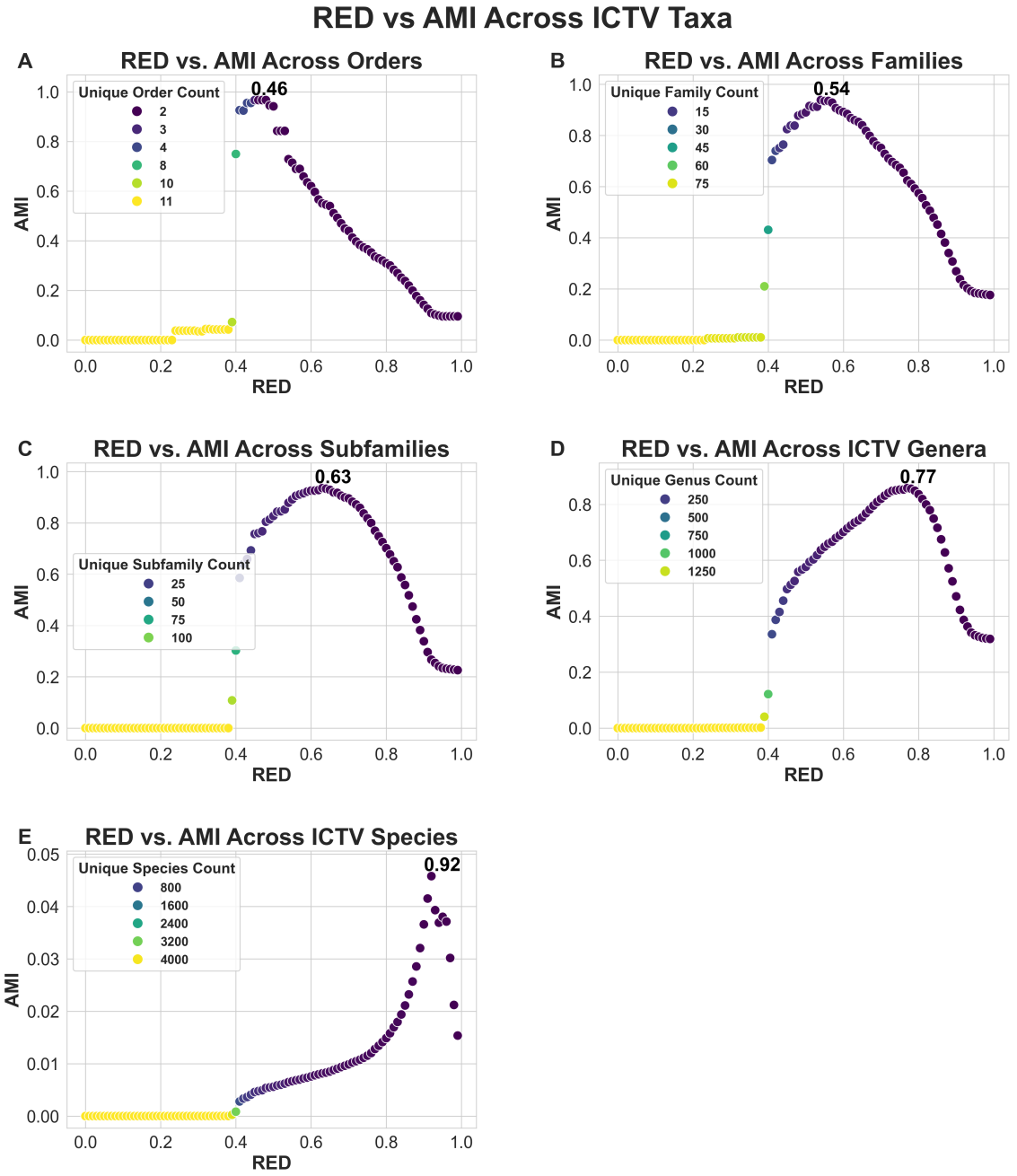
Branches are annotated according to ICTV VMR40 family and order classifications. Comparison with the whole-proteome structural phylogeny revealed the greatest topological similarity among the three hallmark proteins (normalised Robinson-Foulds distance = 0.69), corresponding to approximately 31% shared internal bipartitions. Family-level monophyly was 63.8%; however, this result should be interpreted cautiously as the phylogeny was reconstructed from a substantially smaller subset of genomes than the MCP and TerL analyses.



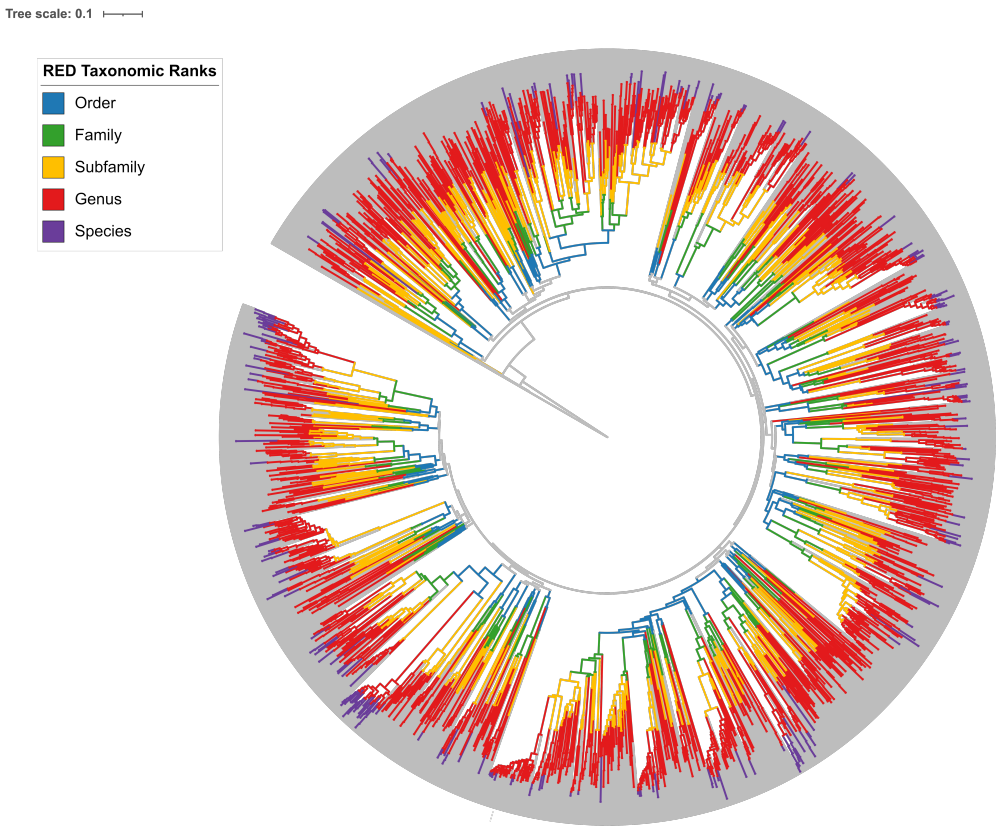
Supplementary Figure 2.10: Most abundant predicted Pfam domains across *Caudoviricetes*.
To expand functional annotation beyond RefSeq/GenBank labels, Pfam domains were predicted for all proteins in the dataset and ranked by abundance. The two most prevalent domains were uncharacterised families - PF06569 (DUF1128; 6,300 occurrences) and PF06174 (DUF987; 4,106 occurrences) - illustrating that some of the most common *Caudoviricetes*' proteins lack any experimentally validated or known function. Among annotated domains, several host-interaction and DNA-processing elements were prominent, including PF03206 (NifW), PF07126 (ZapC), and PF13392 (HNH endonuclease), reflecting recurrent acquisition of bacterial metabolic, cell-division, and genome-modifying factors by phages. Overall, these results highlight the coexistence of abundant, deeply conserved phage protein families with extensive functional novelty across *Caudoviricetes*.



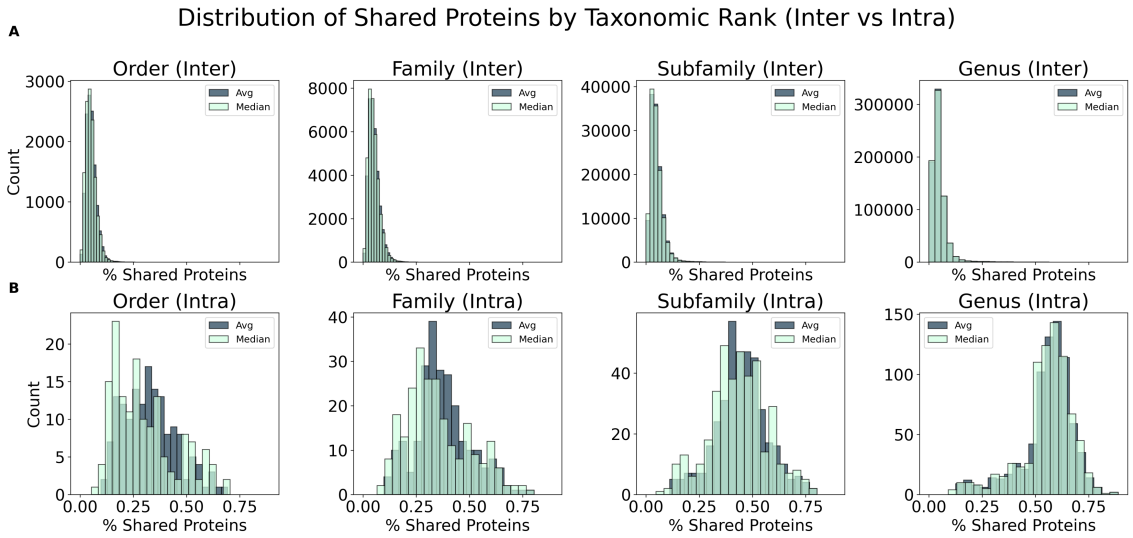
Supplementary Figure 2.11: LCA-Based Monophyly Precision Across Taxonomic Ranks.
(A) Histograms showing the distribution of monophyly precision scores for viral orders, families, subfamilies, and genera. Precision is defined as the proportion of genomes within the inferred last common ancestor (LCA) that belong to the same ICTV-assigned taxon. High precision indicates a structurally consistent and monophyletic grouping.
(B) Boxplots summarizing the range of precision scores across taxonomic ranks, with overlaid scatter points showing individual taxa. Overall, genus- and subfamily-level classifications show strong agreement with structure-based clustering, with the majority achieving perfect monophyly (precision = 1.0). In contrast, family- and order-level taxa exhibit greater variability, reflecting broader diversity and occasional polyphyletic placements. These scores informed subsequent RED-based reassignment of discordant clades.



Supplementary Figure 2.12: RED vs. Adjusted Mutual Information (AMI) Across ICTV Taxonomic Ranks.
Relationship between Relative Evolutionary Divergence (RED) and Adjusted Mutual Information (AMI) across ICTV-defined taxonomic ranks within the *Caudoviricetes* dataset. Each panel represents a taxonomic level - (A) Order, (B) Family, (C) Subfamily, (D) Genus, and (E) Species - and shows how well structural clustering (based on the presence-absence of structural clusters per genome) aligns with ICTV taxonomy as a function of RED score. Points are coloured by the maximum number of unique taxa assigned at each RED value for the corresponding rank.



Supplementary Figure 2.13: Circular phylogeny of 4,082 *Caudoviricetes* genomes, rooted on the *Duplodnaviria* outgroup HSV-1 and scaled by Relative Evolutionary Divergence (RED) using PhyloRank.
The phylogram provided to PhyloRank was built from a Dice-distance structural presence-absence matrix. Taxonomic boundaries were inferred from RED intervals optimised using Adjusted Mutual Information (AMI) scores (see Supplementary Figure 2.12).

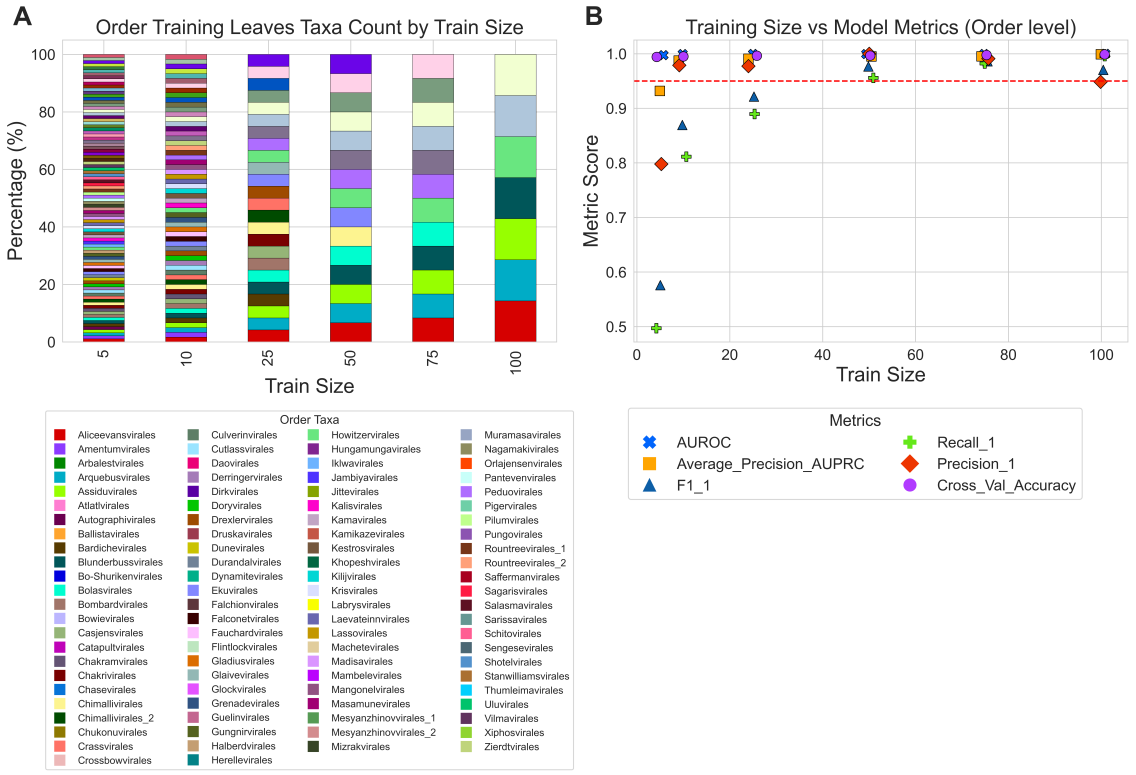


C

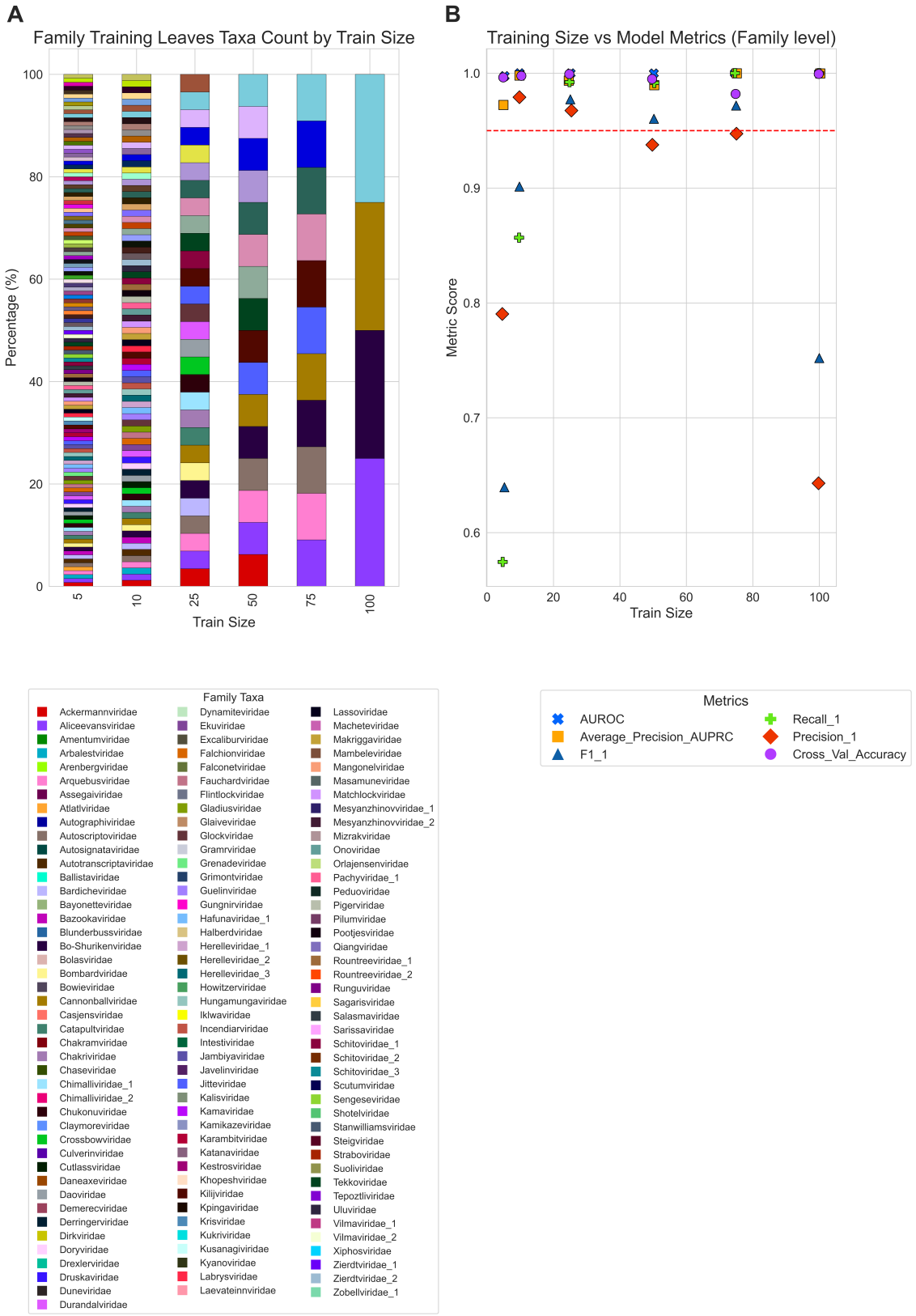
Summary of Intra- and Inter-clade Shared Protein Percentage by Taxonomic Rank

Rank	Mean % Shared Proteins (Intra)	Median % Shared Proteins (Intra)	SD of % Shared Proteins (Intra)	Mean % Shared Proteins (Inter)	Median % Shared Proteins (Inter)	SD of % Shared Proteins (Inter)
Order	0.328001	0.302091	0.121165	0.052271	0.050807	0.014774
Family	0.374887	0.350117	0.105864	0.051764	0.050727	0.012236
Subfamily	0.449797	0.433998	0.084364	0.049835	0.049204	0.008734
Genus	0.556628	0.552867	0.034579	0.048001	0.049204	0.005191

Supplementary Figure 2.14: Distribution of shared proteins within and between taxonomic ranks.
(A) Pairwise percentage of shared structural proteins between clades (inter-clade) is low across all taxonomic ranks, with narrow distributions centered around 5%.
(B) In contrast, intra-clade shared protein percentages are markedly higher and more variable, with distributions shifting upward at lower ranks, consistent with higher relatedness.
(C) Summary table showing the average, median, and standard deviation of intra- and inter-clade shared protein percentages for each rank. Intra-clade similarity increases and variance decreases toward the genus level, while inter-clade similarity remains uniformly low across ranks (∼5%).

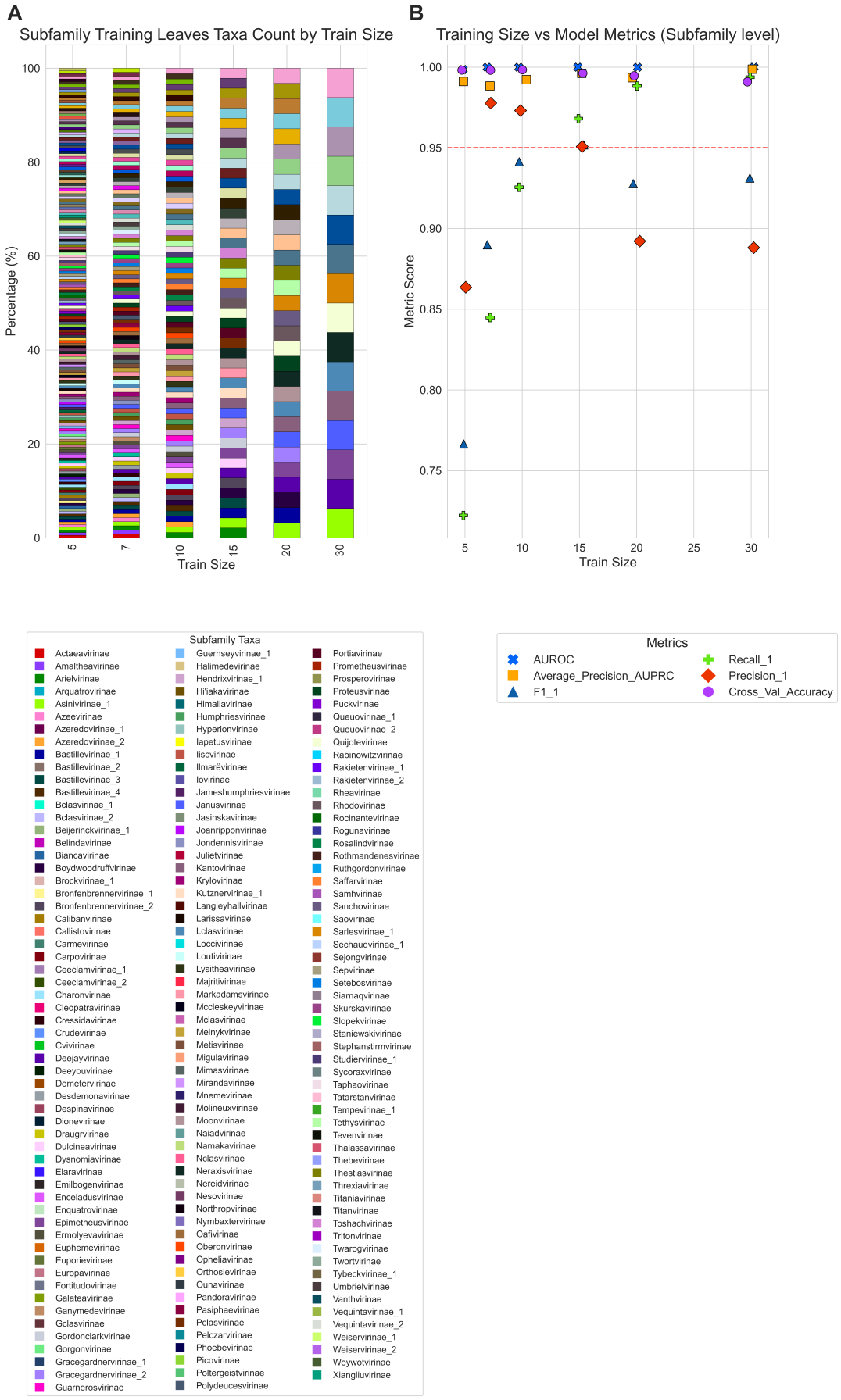


Supplementary Figure 2.15: Effect of training set size on order-level taxonomic classification performance.
(A) Proportional representation of viral orders in the training sets used for binary one-vs-all classification at increasing training sizes (5, 10, 25, 50, 75, and 100 genomes per order). To increase classification difficulty and ensure fair comparisons across orders, training data were downsampled to fixed sizes within each iteration, thereby standardizing the number of genomes per order and preventing models from exploiting imbalanced class representation. Orders represented by fewer than five genomes were excluded due to the requirement for 5-fold cross-validation.
(B) Mean model performance across all one-vs-all order-level classifiers as a function of training set size. Model performance improved rapidly with increasing training data, with substantial gains observed between 5 and 25 genomes per order. From 50 genomes onward, all evaluated metrics—including AUROC, average precision (AUPRC), F1 score, precision, recall, and cross-validation accuracy—consistently exceeded 95%, indicating stable and near-saturated classification performance. While training data were downsampled to enforce balanced class representation, test sets were left unstratified to reflect the natural distribution of viral orders.



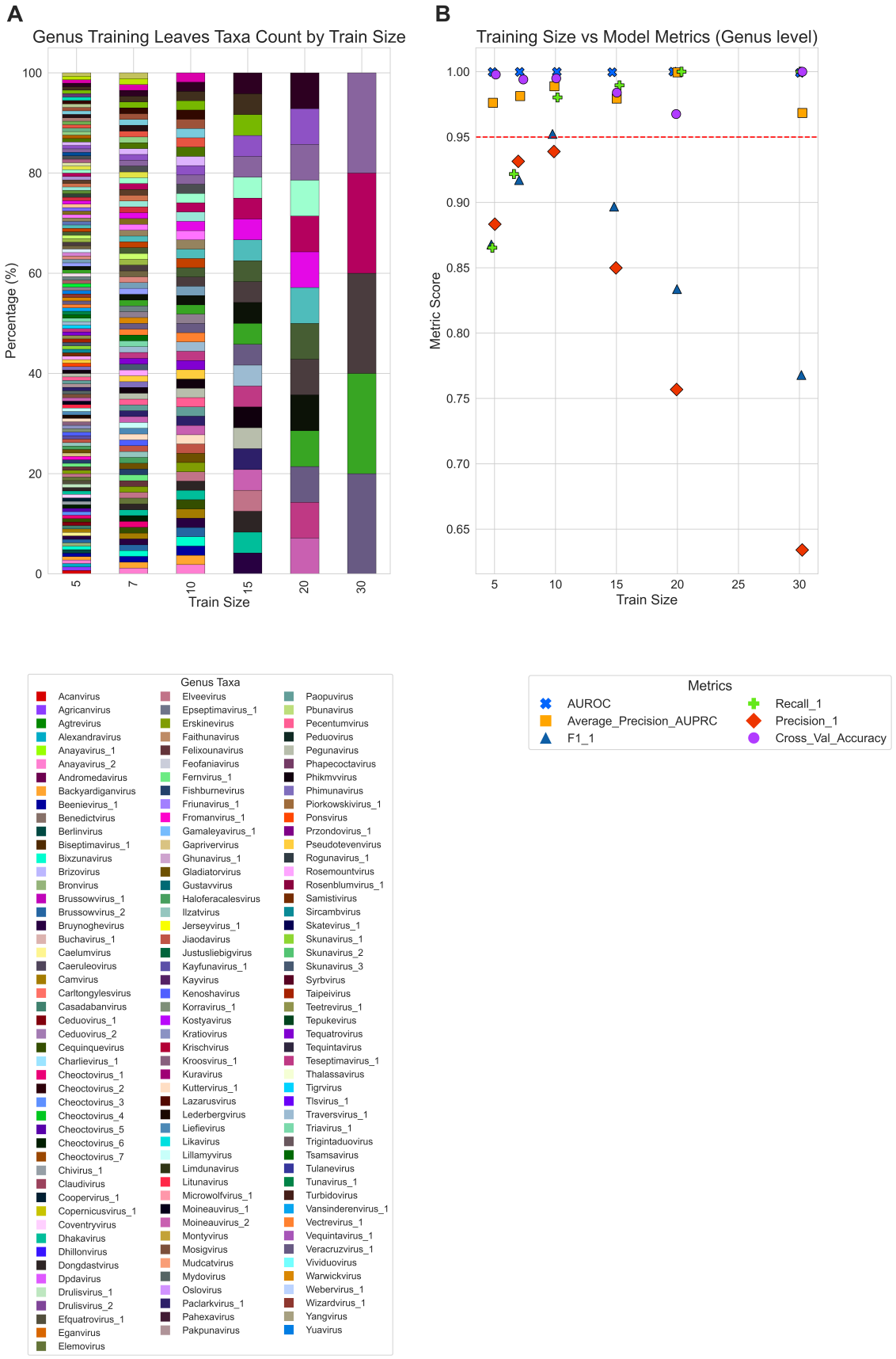
Supplementary Figure 2.16: Effect of training set size on family-level taxonomic classification performance.
(A) Proportional representation of training genomes per viral family across varying training sizes (5–100 genomes per family).

To standardise training conditions and impose a more stringent learning setting, taxa were randomly downsampled to equal training sizes in each round, ensuring fair comparisons across families.
(B) Mean model performance metrics across all family-level classifiers at each training size. Most metrics remained consistently high; however, precision and F1 score declined at the largest training size (100), likely due to increased class imbalance in the test set and reduced generalization capacity.



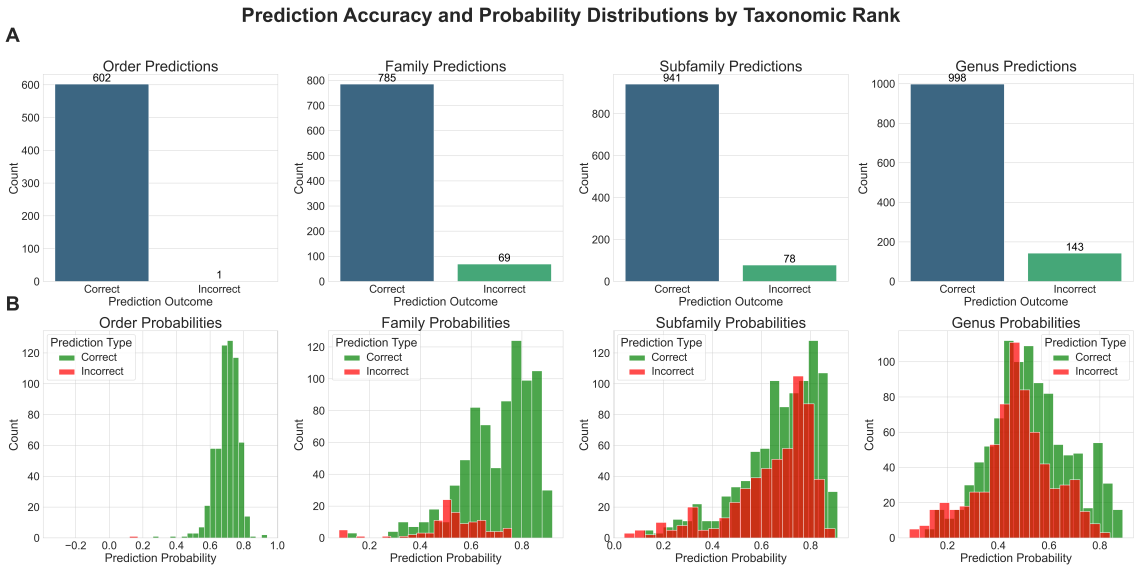
Supplementary Figure 2.17: Effect of training set size on subfamily-level taxonomic classification performance.

(A) Proportional representation of training genomes per viral subfamily across varying training sizes (5–30 genomes per subfamily).
(B) Mean performance metrics for all subfamily-level classifiers at each training size. Performance remained consistently high for training sizes $\geq$10, with AUROC, AUPRC, and cross-validation accuracy all exceeding 90%. Notably, models trained with 15 genomes per subfamily achieved the highest overall accuracy, with all metrics surpassing 95%, indicating that subfamily-level classifiers are both robust and highly effective even at moderate training sizes.

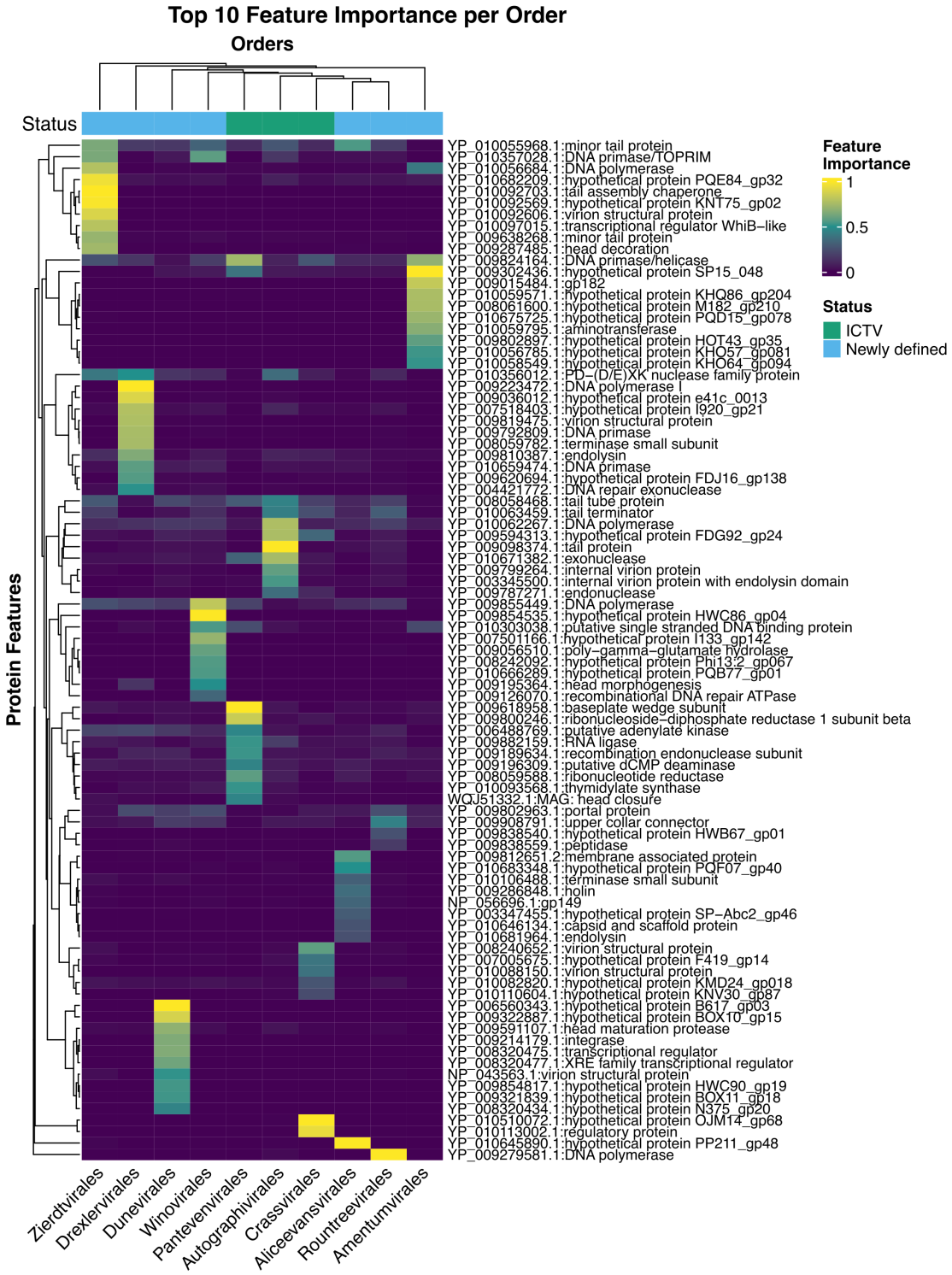


Supplementary Figure 2.18: Effect of training set size on genus-level taxonomic classification performance.

(A) Proportional representation of viral genera in the training sets used for binary one-vs-all classification across increasing training sizes (5, 7, 10, 15, 20, and 30 genomes per genus). To standardise training conditions and ensure fair comparisons across taxa, training data were randomly downsampled to fixed sizes within each iteration, thereby equalizing class representation and imposing a more stringent learning setting. Genera represented by fewer than five genomes were excluded to permit five-fold cross-validation.
(B) Mean performance metrics across all genus-level one-vs-all classifiers as a function of training set size. AUROC, average precision (AUPRC), and cross-validation accuracy remained consistently high across all training sizes, indicating strong ranking performance. In contrast, threshold-dependent metrics—particularly precision and F1 score—declined progressively at training sizes $\geq$15 genomes per genus. This trend likely reflects changes in test set composition as an increasing proportion of available genomes is allocated to training, resulting in fewer unseen positive examples and an elevated false positive rate. Despite this effect, models trained with smaller to moderate training sizes (7–10 genomes per genus) achieved balanced performance across all metrics, with values consistently exceeding 90%.

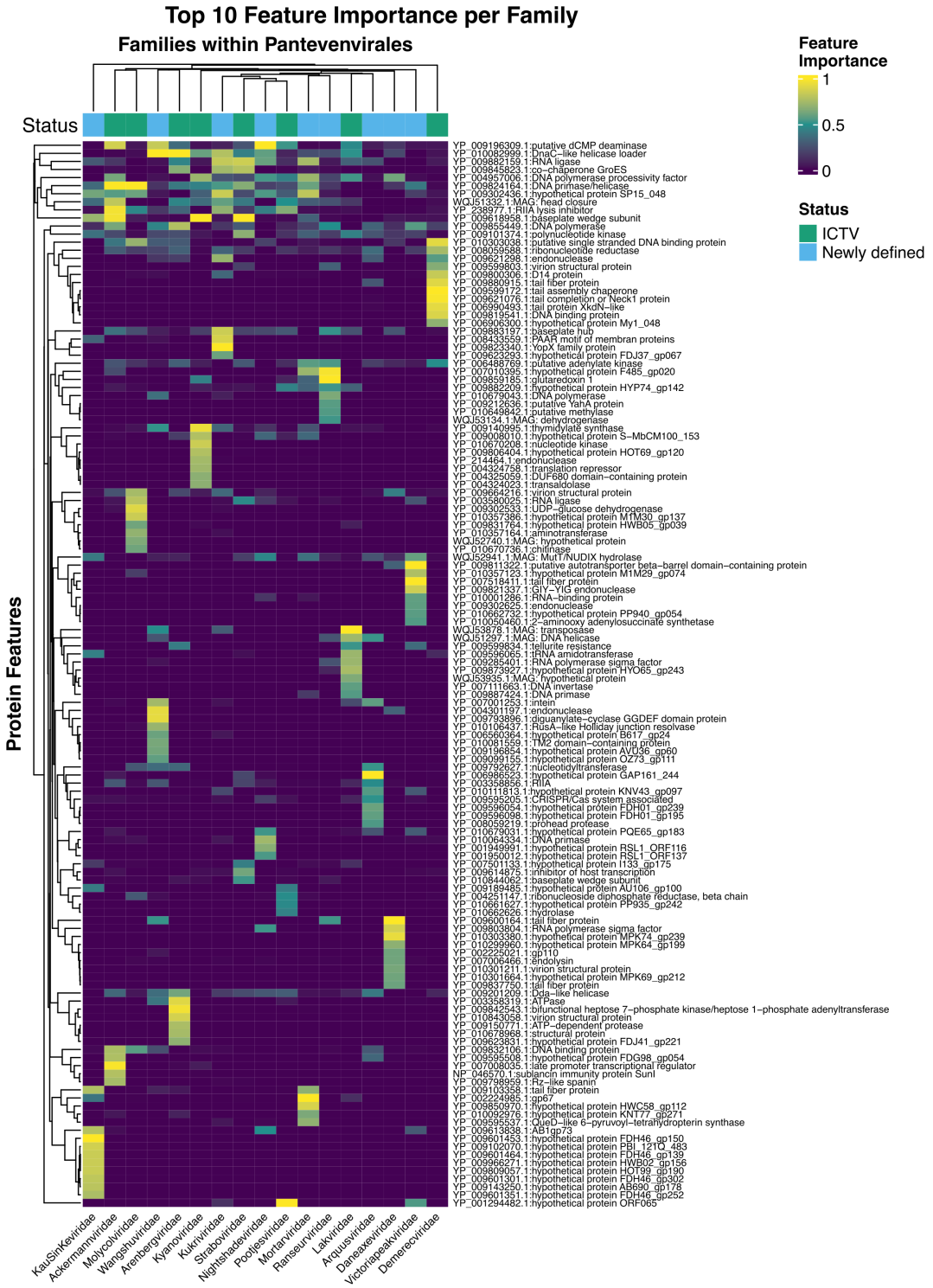


Supplementary Figure 2.19: Prediction accuracy and probability distributions of the random forest classifier across four taxonomic ranks (Order, Family, Subfamily, Genus) on held-out ICTV- classified *Caudoviricetes* genomes.
(A) Bar plots showing the number of correctly (blue) and incorrectly (green) classified genomes at each taxonomic rank. Classification accuracy was highest at the order level (602/603, 99.8followed by the family level (785/854, 91.9%), subfamily level (941/1019, 92.4%), and genus level (998/1141, 87.4%).
(B) Histograms of classifier probability scores for correct (green) and incorrect (red) predictions across taxonomic ranks. Correct predictions consistently exhibit higher confidence (right-skewed distributions), particularly at the order, family, and subfamily levels. In contrast, genus-level probability scores are more symmetric and broadly distributed, reflecting reduced model confidence at this rank—likely due to the limited number of training genomes per genus.



Supplementary Figure 2.20: Feature importance profiles for a representative subset of *Caudoviricetes* orders.
To illustrate how the Random Forest classifier discriminates between taxa, we examined the top 10 most informative protein features for a subset of 10 orders (out of 159 total): *Aliceevansvirales*, *Amentumvirales*, *Autographivirales*, *Crassvirales*, *Drexlervirales*, *Dunevirales*, *Pantevenvirales*, *Rountreevirales*, *Winovirales*, and *Zierdtvirales*. This analysis reveals that classification is driven by distinct, order-specific feature signatures, with clear separation in the most informative protein clusters across taxa.

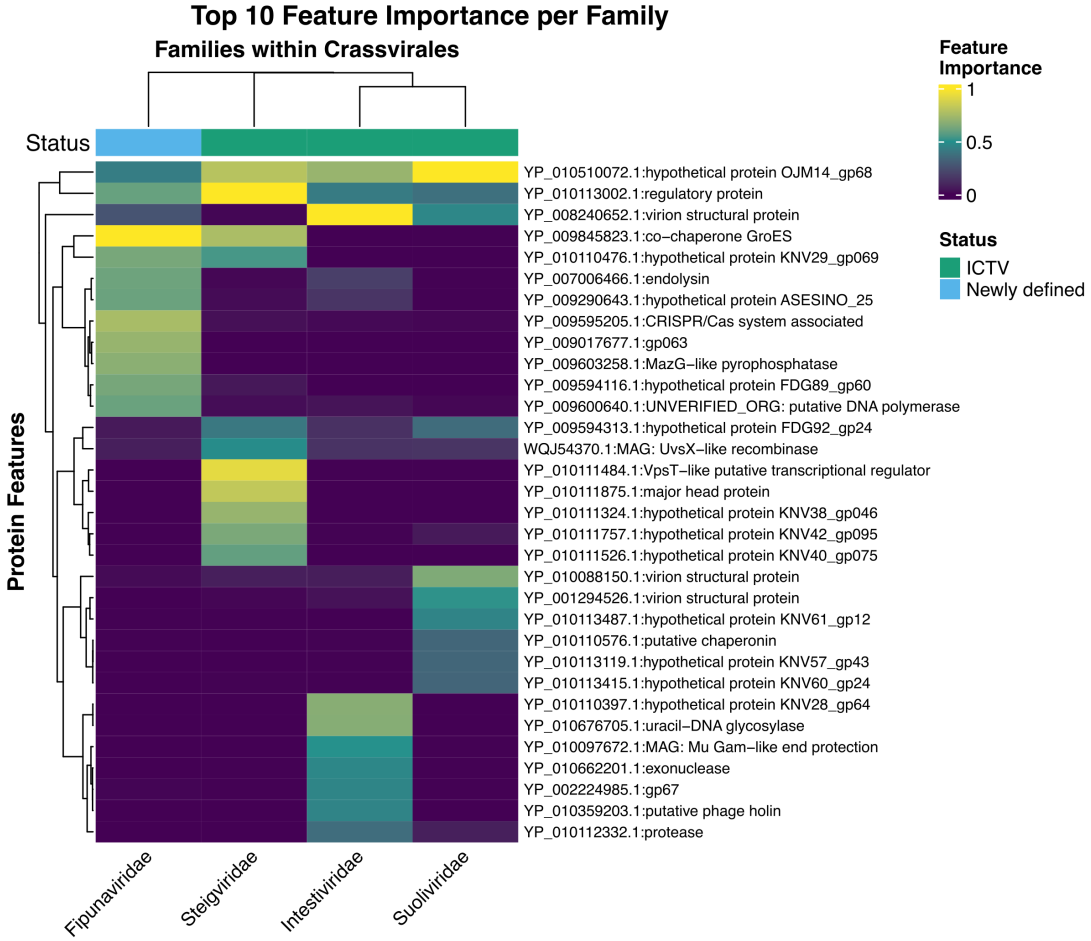
For each order, highly predictive features included tail proteins (*Autographiviirales*), baseplate wedge subunits (*Pantevenvirales*), DNA polymerases (*Drexlerviirales*, *Rountreevirales*), and tail assembly chaperones (*Zierdtvirales*). Notably, in 5 out of 10 orders, the single most informative feature corresponds to proteins annotated as hypothetical e.g., OJM14_gp68 (*Crassvirales*), PP211_gp48 (*Aliceevansvirales*), B617_gp03 (*Dunevirales*), HWC86_gp04 (*Winovirales*), and SP15_048 (*Amentumvirales*), highlighting the model's ability to capture taxon-specific signal independent of prior functional annotation. Together, these results demonstrate that taxonomic discrimination arises from structured, lineage- specific combinations of protein features rather than reliance on a single conserved marker.



Supplementary Figure 2.21: Feature importance profiles for families within *Pantevenvirales*.

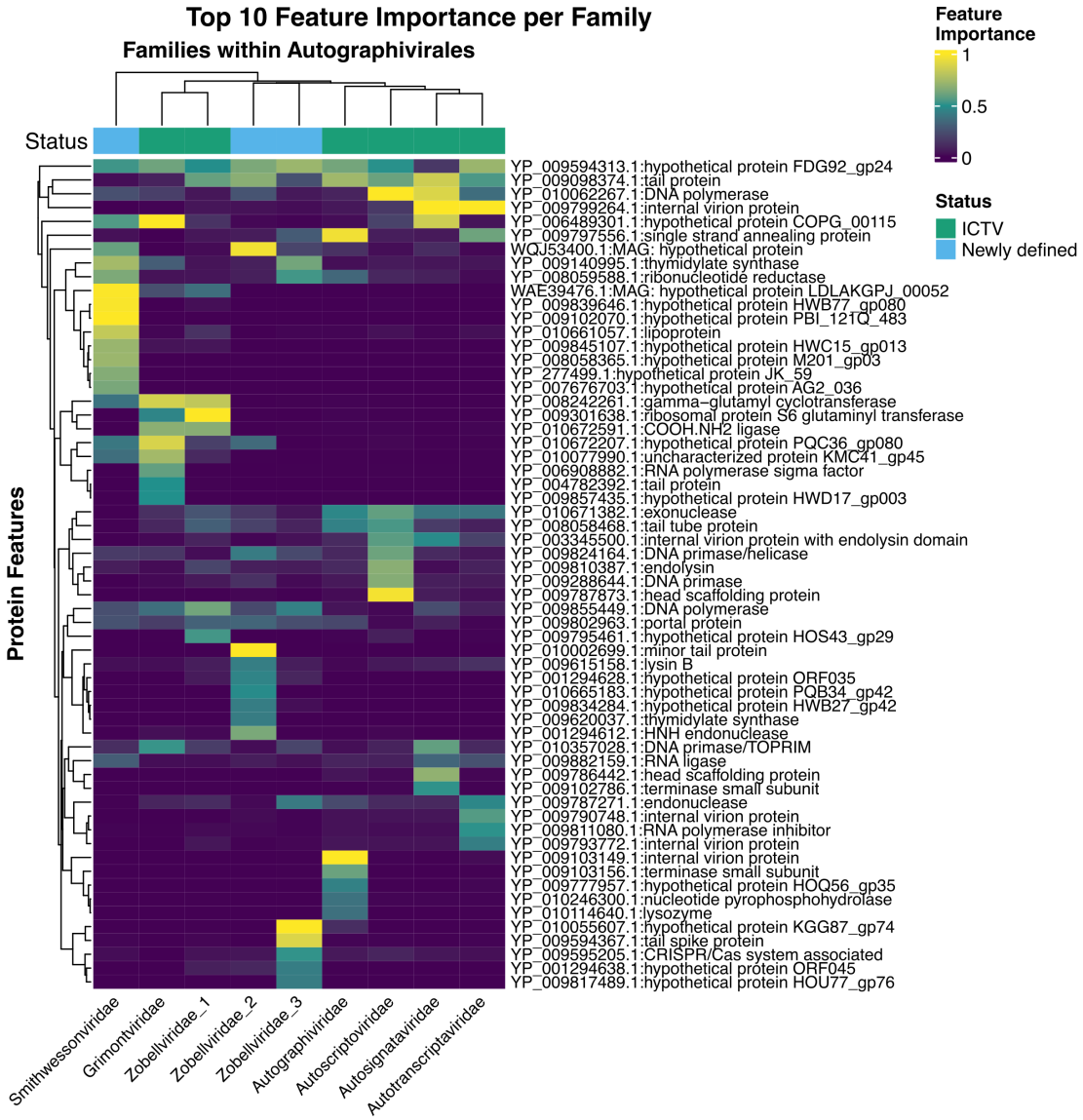
To further illustrate how the Random Forest model selects descriptive features at finer taxonomic resolution, we visualised the top 10 most informative protein features for all families within *Pantevenvirales* (*KauSinKeviridae*, *Arquusviridae*, *Daneaxeviridae*, *Mortarviridae*, *Nightshadeviridae*, *Ranseurviridae*, *Ackermannviridae*, *Arenbergviridae*, *Demerecviridae*, *Kyanoviridae*, *Molycolviridae*, *Pootjesviridae*, *Straboviridae*, *Wangshuviridae*, *Victoriapeakviridae*, *Kukriviridae*, and *Lakviridae*).

Consistent with order-level patterns (Supplementary Figure 2.20), classification is driven by distinct, family-specific combinations of protein features, with clear separation of informative clusters across taxa. A subset of protein features is consistently informative across multiple families, including core structural and replication-associated proteins such as YP_009618958.1:baseplate wedge subunit, YP_009824164.1:DNA primase/helicase, and YP_0049570061:DNA polymerase processivity factor. These shared features likely reflect conserved functional modules within *Pantevenvirales*. However, distinct combinations of additional proteins are differentially enriched across families, enabling fine-scale discrimination between taxa. Together, these results indicate that classification is driven by a combination of conserved core features and family-specific accessory signals.



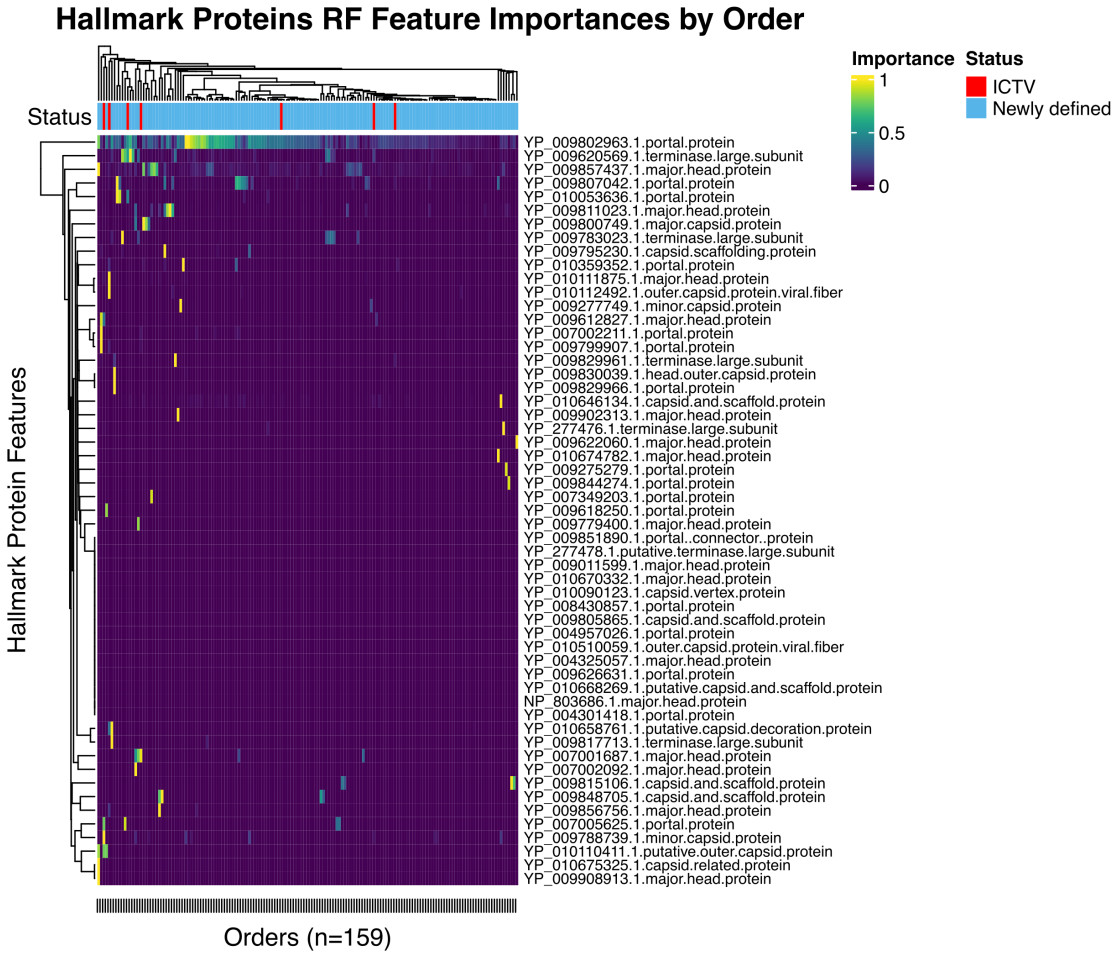
Supplementary Figure 2.22: Feature importance profiles for families within *Crassvirales*.
To illustrate how the Random Forest model selects discriminative features at finer taxonomic resolution, we visualised the top 10 most informative protein features for each family within *Crassvirales* (*Intestiviridae*, *Fipunaviridae*, *Steigviridae*, and *Suoliviridae*). Consistent with patterns observed at the order level (Supplementary Figure 2.20) and within *Pantevenvirales* families (Supplementary Figure 2.21), classification is driven by distinct combinations of protein features, resulting in clearly separated clusters of high importance across families.

While some features show broadly elevated importance across multiple families—most notably YP_010510072.1 (hypothetical protein OJM14_gp68) and YP_010113002.1 (regulatory protein)— each family is characterised by unique, highly weighted features. For example, *Intestiviridae* is strongly associated with YP_008240652.1 (virion structural protein), *Fipunaviridae* with YP_009845823.1 (co-chaperone GroES), *Steigviridae* with a combination of regulatory and structural proteins including YP_010113002.1 (regulatory protein), YP_010111484.1 (VpsT-like transcriptional regulator), and YP_010111875.1 (major head protein), and Suoliviridae with YP_010510072.1 (hypothetical protein OJM14_gp68). These patterns highlight that family-level classification within *Crassvirales* is driven by distinct, yet partially overlapping, sets of structural and functional protein features.

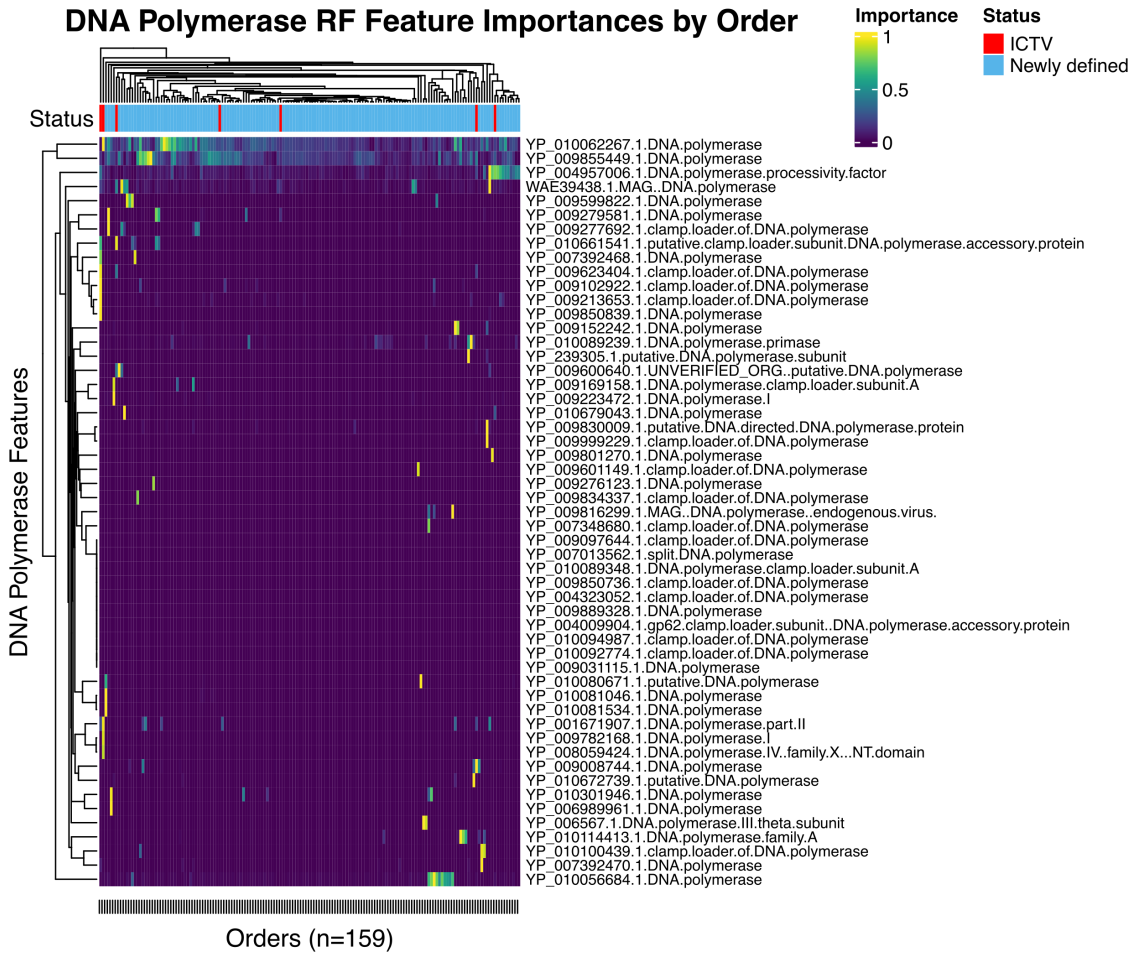


Supplementary Figure 2.23: Feature importance profiles for families within *Autographivirales*.
We visualised the top 10 most informative protein features for each family within *Autographivirales* (*Autographiviridae*, *Autoscriptoviridae*, *Autosignataviridae*, *Autotranscriptaviridae*, *Grimontviridae*, *Smithwessonviridae*, and *Zobellviridae* subgroups 1–3).

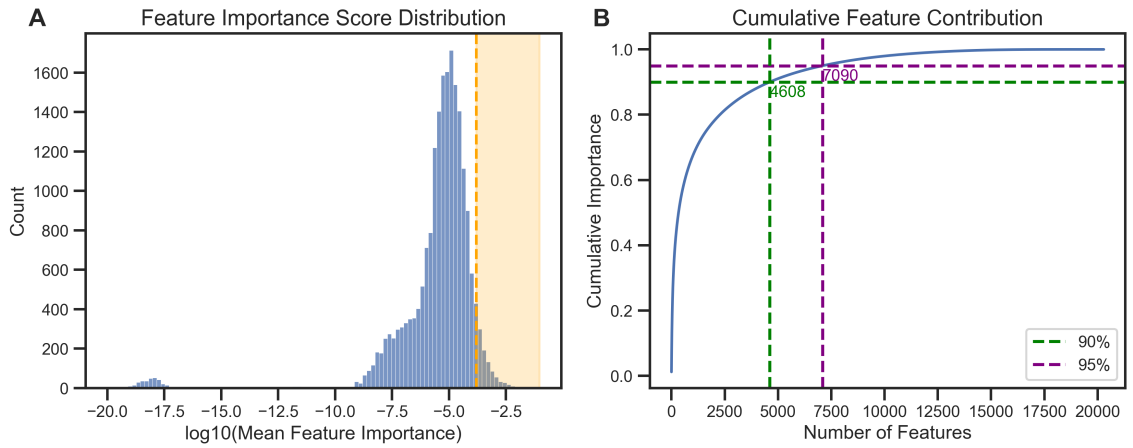
Consistent with patterns observed at the order level (Supplementary Figure 2.20) and across family-level analyses (Supplementary Figure 2.21 and Figure 2.22), classification is driven by distinct combinations of protein features, resulting in clearly separated clusters of high importance across families. A subset of features shows consistently high importance across multiple families, most notably YP_009594313.1 (hypothetical protein FDG92_gp24) and YP_009098374.1 (tail protein). Despite this shared signal, each family is characterised by unique combinations of highly weighted features, reflecting both conserved and lineage-specific functional signatures underlying classification within *Autographivirales*.



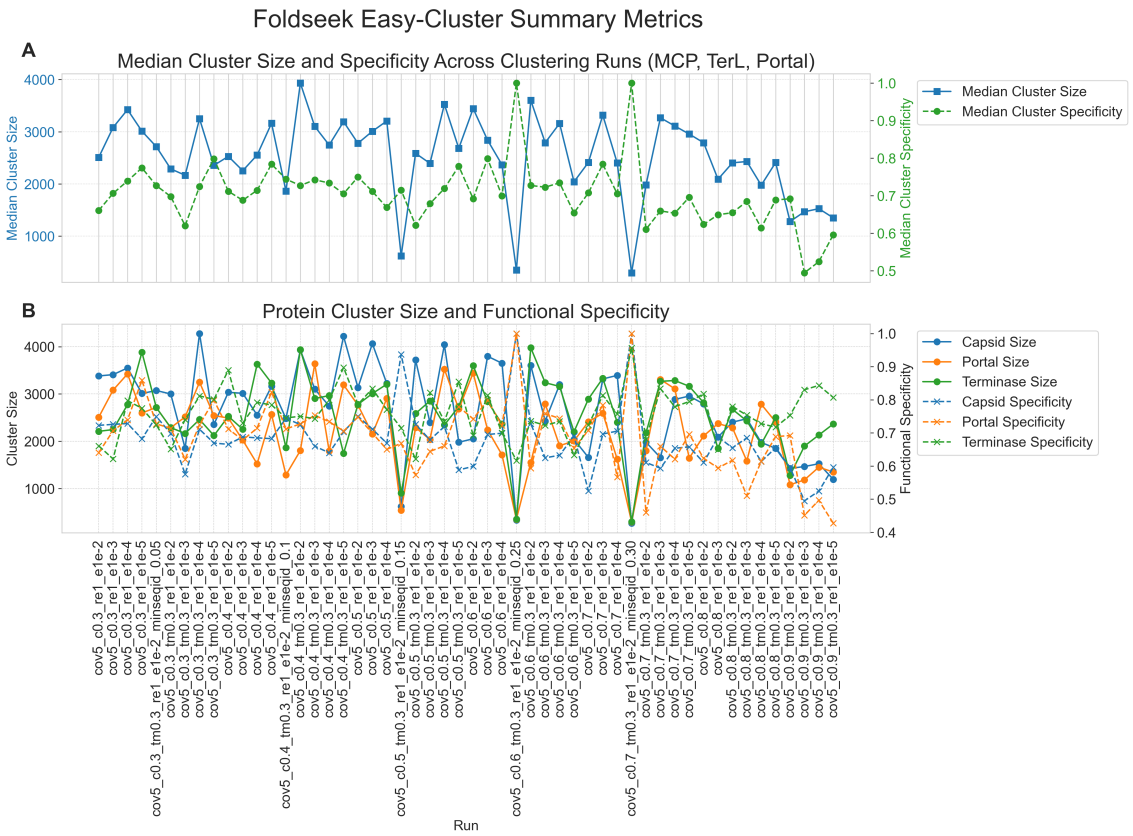
Supplementary Figure 2.24: Feature importance profiles for hallmark proteins across all orders.
To assess the discriminatory power of canonical viral hallmark proteins, we visualised Random Forest feature importance scores for hallmark-associated proteins across all orders (n = 159). In contrast to the clear, order- and family-specific clustering observed when using the full structural feature set (Supplementary Figures 15-18), hallmark proteins alone do not produce strongly separated or cohesive clusters at the order level. Instead, feature importance is broadly diffuse, with many hallmark proteins contributing weakly across multiple orders. However, certain hallmark-associated features—particularly structural proteins such as portal proteins, terminase large subunit, and major capsid proteins—show elevated importance within specific lineages. This suggests that while hallmark proteins contribute to classification, they lack sufficient discriminatory resolution on their own and are most informative when considered alongside the broader structural proteome.



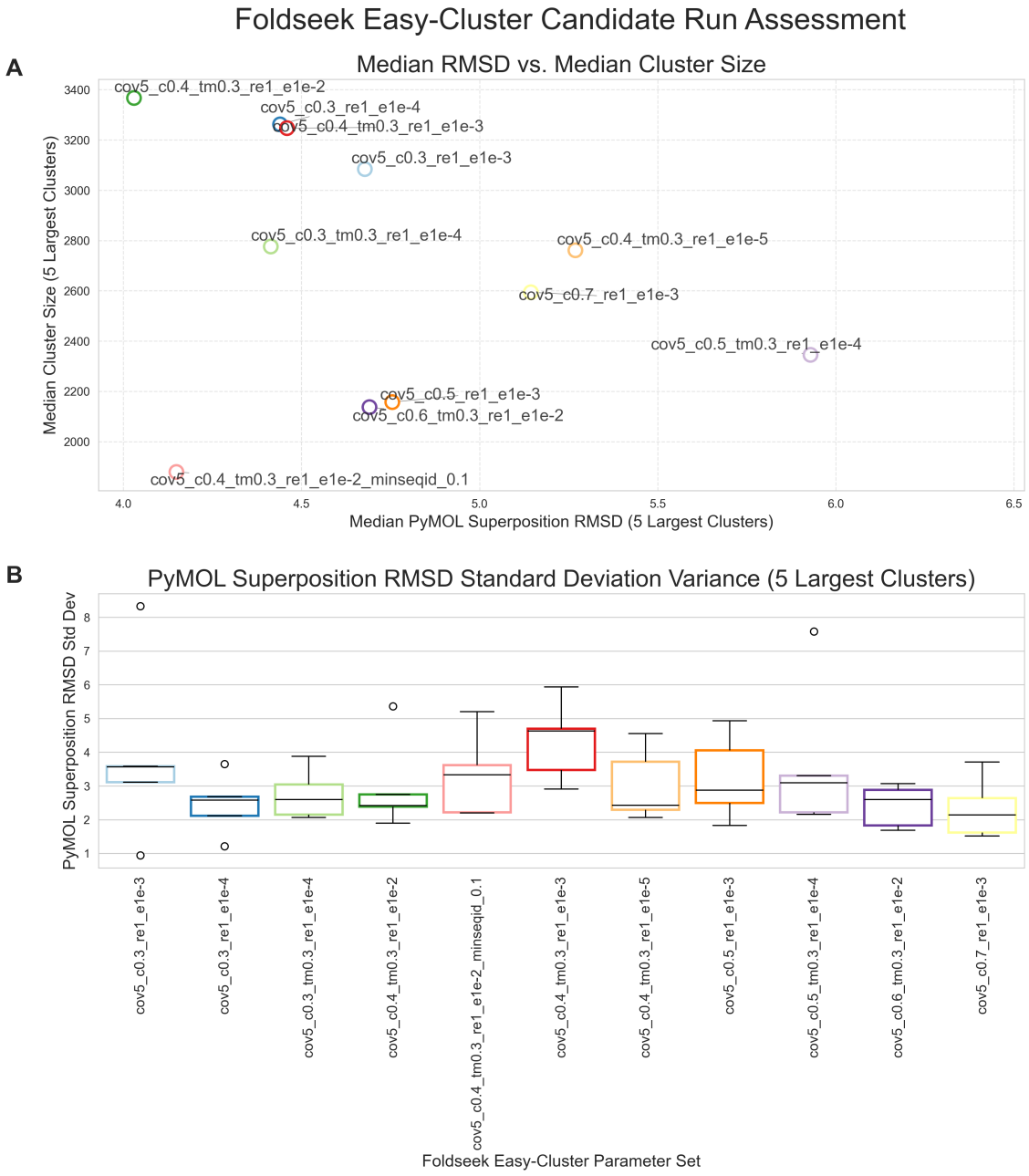
Supplementary Figure 2.25: Feature importance profiles for DNA polymerase proteins across all orders.
To evaluate the discriminatory power of DNA polymerase–associated proteins, we visualised Random Forest feature importance scores for all annotated DNA polymerase and related replication proteins across all orders (n = 159). Similar to hallmark proteins (Supplementary Figure 2.24), DNA polymerase features alone do not produce well-defined or cohesive clustering patterns at the order level. Feature importance is broadly sparse and dispersed, with most DNA polymerase–associated proteins contributing weakly across multiple orders. However, certain polymerase variants and associated factors (e.g. clamp loaders and processivity factors) exhibit elevated importance within specific lineages. This indicates that while DNA polymerase–related proteins can contribute to classification in a lineage-specific manner, they lack sufficient variability and discriminatory resolution on their own and are most informative when integrated with the broader structural protein repertoire.



Supplementary Figure 2.26: Distribution and cumulative contribution of Random Forest feature importance scores across all protein features.
(A) Distribution of mean feature importance values, calculated by averaging feature importance scores across all taxonomic ranks (order, family, subfamily, and genus) and visualised on a $\log_{10}$ scale to improve resolution of low-magnitude values. A threshold corresponding to the top 5% most informative features (95th percentile) is indicated by the dashed vertical line, with the shaded region representing features above this cut-off. The distribution is strongly right-skewed, with the majority of features exhibiting small but measurable importance values. The distribution peaks at approximately $\log_{10} \approx -5$ ($\approx 1 \times 10^{-5}$), indicating that most features contribute only minor individual effects.
(B) Cumulative contribution of features to total model importance, calculated by ranking features in descending order of mean importance and computing the cumulative sum of normalised feature importance values. Horizontal dashed lines indicate 90% and 95% of total model importance, while vertical dashed lines denote the number of features required to reach these thresholds (4,608 and 7,090 features, respectively). Together, these results demonstrate that predictive signal is broadly distributed across the structural proteome, with approximately 7,000 of the 21,289 non-redundant structural protein features accounting for 95% of the total model importance.



Supplementary Figure 2.27: Foldseek easy-cluster summary metrics. (A) Variation in median cluster size (blue, left y-axis) and median cluster specificity (green, right y-axis) for the largest clusters whose representatives were the major capsid protein (MCP), terminase large subunit (TerL), or portal protein, across multiple Foldseek easy-cluster parameter sets. Cluster specificity was quantified as the Shannon entropy of Pfam annotations within each cluster, with higher values indicating greater functional coherence.
(B) Cluster sizes across three conserved *Caudoviricetes* proteins (MCP, portal protein, and TerL). As expected, higher functional specificity is typically associated with smaller cluster sizes, reflecting a trade-off between sensitivity and specificity. Parameter sets yielding a median cluster size (averaged across MCP, portal, and TerL clusters) $\geq$ 3,000 and a functional specificity $\geq$ 0.7 were selected as candidate clustering runs for downstream analyses.



Supplementary Figure 2.28: Structural coherence of Foldseek clustering parameter sets.
(A) Variation in median cluster size (blue, left y-axis) and median cluster specificity (green, right y-axis) for the largest clusters whose representatives were the major capsid protein (MCP), terminase large subunit (TerL), or portal protein, across multiple Foldseek easy-cluster parameter sets. Cluster specificity was quantified as the Shannon entropy of Pfam annotations within each cluster, with higher values indicating greater functional coherence.
(B) Cluster sizes across three conserved *Caudoviricetes* proteins (MCP, portal protein, and TerL). As expected, higher functional specificity is typically associated with smaller cluster sizes, reflecting a trade-off between sensitivity and specificity. Parameter sets yielding a median cluster size (averaged across MCP, portal, and TerL clusters) $\geq$ 3,000 and a functional specificity $\geq$ 0.7 were selected as candidate clustering runs for downstream analyses.

3. Unearthing the Irish Soil Virome: Taxonomy, Diversity and the Challenges of Metagenomic Viral Discovery

Linda Smith[1,2], Niranjana Edwin[3], Aoife Duff[4], Fiona Brennan[4], Orla O'Sullivan[3], Colin Hill[1,2], Andrey N. Shkoporov[1,2,5]

[1] APC Microbiome Ireland, University College Cork, Ireland, [2] School of Microbiology, University College Cork, Ireland, [3] Teagasc Food Research Centre, Moorepark, Fermoy, Co. Cork, Ireland, [4] Teagasc Food Research Centre, Johnstown, Co. Wexford, Ireland, [5] Department of Medicine, University College Cork, Ireland

3.1 Abstract

Soil harbours one of the most diverse microbiomes on Earth, but thus far its viral component remains poorly characterised. Here, we analysed 165 soil samples from 32 agricultural sites across the island of Ireland, including 32 deeply sequenced metagenomes, to examine soil viral diversity and evaluate the extent to which viral signal can be recovered from shotgun metagenomic data. Using computational viral discovery, protein annotation, and host-linking strategies, we constructed a catalogue of high-confidence viral genomes, their predicted bacterial hosts, and their associated auxiliary metabolic genes.

Consistent with the global dominance of tailed bacteriophages in other environments, viruses of the class *Caudoviricetes* comprised the majority of recovered genomes. Several viral populations were detected across the majority of samples, indicating the presence of highly prevalent and persistent soil viruses despite diverse genomic architectures and inferred life strategies. Auxiliary metabolic genes were enriched in defence and competition functions, suggesting complex virus–host and virus–virus interactions. Predicted hosts spanned 20 bacterial and archaeal phyla, with Proteobacteria (31.4%), Firmicutes (15.7%), and Actinobacteriota (8.9%) most represented.

The viral signal constituted $< 0.01\%$ of reads, with genome recovery limited by short-read assembly constraints and the extensive diversity of the soil microbiome. Automated tools and reference databases frequently misclassified host-derived sequences as viral, highlighting substantial challenges in distinguishing bona fide viral genomes. Despite these limitations, attempts to contextualise Irish soil viruses within a global virosphere identified a 120 kb bacteriophage from Michigan Miscanthus rhizosphere soil with high nucleotide similarity and conserved synteny to a prevalent Irish phage, suggesting the existence of a globally conserved soil viral clade. This study represents the first attempt to investigate the Irish soil virome and highlights the complexity of accurately resolving viral signal in soil metagenomes and underscores the need for improved frameworks for viral discovery in complex environments.

3.2 Introduction

Soil harbours one of the most diverse microbiomes on Earth, encompassing an extraordinary diversity of bacteria, archaea, fungi, protists, and viruses[163,164]. A single gram of soil can contain more than 0.5 mg of microbial biomass carbon and over 50,000 microbial species, highlighting the remarkable complexity of terrestrial microbial ecosystems[165,166,167]. This diversity underpins key ecosystem processes, including carbon and nitrogen fixation and cycling, turnover of other nutrients, and plant productivity[168,169,170]. Despite this central role in terrestrial ecosystems, the viral communities that shape soil microbial dynamics and biogeochemical processes (the soil virome) remain poorly characterised[171,172,173].

Although viral ecology has been an active field of research for only a few decades, viruses are now recognised as major ecological and evolutionary drivers of microbial communities[174,175,176]. In marine and human-associated ecosystems, bacteriophages regulate host population dynamics, mediate horizontal gene transfer, and influence ecosystem functioning through cell lysis and the carriage of auxiliary metabolic genes (AMGs)[174,177]. By contrast, our understanding of viral roles in soil ecosystems remains rudimentary. Soil environments are heterogeneous at fine spatial scales, harbour extreme microbial and viral diversity, and contain high levels of extracellular DNA, all of which complicate viral discovery, genome assembly, and downstream ecological interpretation[167,178,179].

Large-scale efforts to characterise soil viral diversity have recently emerged. The Global Soil Viral Atlas (GSVA), which analysed approximately 2,500 soil metagenomes worldwide, represents a major advance by compiling tens of thousands of putative viral sequences[180]. However, the comparatively modest yield of recoverable viral genomes relative to the scale of sampling (approximately 50,000 non-redundant contigs) highlights the substantial challenges associated with detecting and assembling viral sequences from complex soil communities. These findings suggest that a considerable proportion of soil viral diversity remains inaccessible using conventional approaches. More recently, the VIRE database expanded the cataloguing of viral sequences across multiple ecosystems, including soils, with a collection of nearly two million viral sequences[181]. Yet, despite this expansion, the majority of soil viruses remain uncultivated, taxonomically unresolved, and only weakly connected to established viral lineages[180].

Previous studies have demonstrated that soil viral communities are influenced by environmental variables such as land use, vegetation, and nutrient availability, and that viruses can impact on microbial succession and soil carbon turnover[182,183,184,185]. While these studies highlight the ecological relevance of soil viruses, few have attempted to place soil viral populations within a broader evolutionary or taxonomic framework. In particular, it remains unclear whether soil viromes harbour dominant viral lineages analogous to *Crassvirales* in the human gut or pelagiphages in the oceans, and to what extent soil viruses share evolutionary relationships with known *Caudoviricetes* diversity across environments.

Addressing these questions is complicated by both biological and computational

constraints. Viral genome assembly from short-read metagenomes is challenged by terminal repeats, uneven sequencing coverage, and high levels of strain microdiversity, all of which introduce ambiguities into de Bruijn assembly graphs and can result in fragmented or misassembled genomes. For example, linear dsDNA viruses containing direct terminal repeats[186,187] generate branching structures within assembly graphs rather than simple linear paths, complicating complete genome reconstruction[188].

Comparisons of short-read and long-read assemblies of natural soil metagenomes have shown that, although short-read sequencing generally provides higher nucleotide accuracy, short-read assemblies can fail to resolve complex genomic regions such as repeats and structurally variable loci that are more readily recovered using long-read sequencing[189]. Assembly quality is further influenced by the choice of assembler, which can substantially affect the recovery, completeness, and characterisation of viral genomes[190]. These technical challenges are particularly pronounced in soil viromes, where high viral diversity and strain heterogeneity frequently produce fragmented or incomplete assemblies. These limitations can subsequently hinder downstream taxonomic assignment, phylogenetic reconstruction, and the examination of strain-level viral diversity[191,192].

Virus-aware assemblers such as metaSPAdes attempt to mitigate these algorithmic hurdles by incorporating specialised heuristics designed to better resolve repeats and uneven coverage typical of viral metagenomes[188]. More fundamentally, sequence-based viral discovery is limited to identifying genomes and proteins that resemble those already present in reference databases. Highly divergent viral proteins—particularly those lacking close homologues in reference databases—may therefore evade detection[113,193,194]. This limitation is especially pronounced for soil viruses, which are underrepresented in existing databases. As a result, sequence-based approaches alone are insufficient to fully capture soil viral diversity.

Viral sequences typically constitute a small and uneven fraction of reads in shotgun soil metagenomes, reflecting both the low abundance of viral genomes and their heterogeneous distribution across samples[195,196]. Despite this, comprehensive metagenomic sequencing remains essential for many study designs, as it preserves ecological context, enables host- and function-aware analyses, and facilitates integrated comparisons across microbial and viral community components. Therefore, the central challenge lies not in data generation but in the reliable extraction of viral signal from a complex, predominantly cellular background.

Here, we present a curated catalogue of 118 viral genomes recovered from agricultural soils across Ireland and place these viruses within a taxonomic and ecological context. Using viral discovery approaches, host prediction via CRISPR spacer matches, and gene-content-based analyses, we (i) evaluate the technical and biological constraints shaping viral recovery from shotgun soil metagenomes, (ii) identify highly prevalent and recurrent viral populations across Irish soils, (iii) characterise their auxiliary metabolic gene repertoires and inferred ecological strategies, (iv) predict host associations across diverse bacterial and archaeal taxa, (v) assess environmental drivers of viral diversity, (vi) evaluate the limitations of current viral

identification tools and global reference databases, including CheckV, GSVA, and VIRE,[197,180,188] and (vii) investigate global homology patterns to determine whether Irish soil phages are part of conserved, widespread viral lineages. Together, these analyses reveal a highly diverse Irish soil virome, with viral populations exhibiting distinct persistence strategies yet achieving comparable ecological success, while highlighting the challenges of accurately resolving viral genomes in complex metagenomic datasets.

3.3 Results

3.3.1 Assembly Complexity and Viral Abundance in Irish Soil Metagenomes

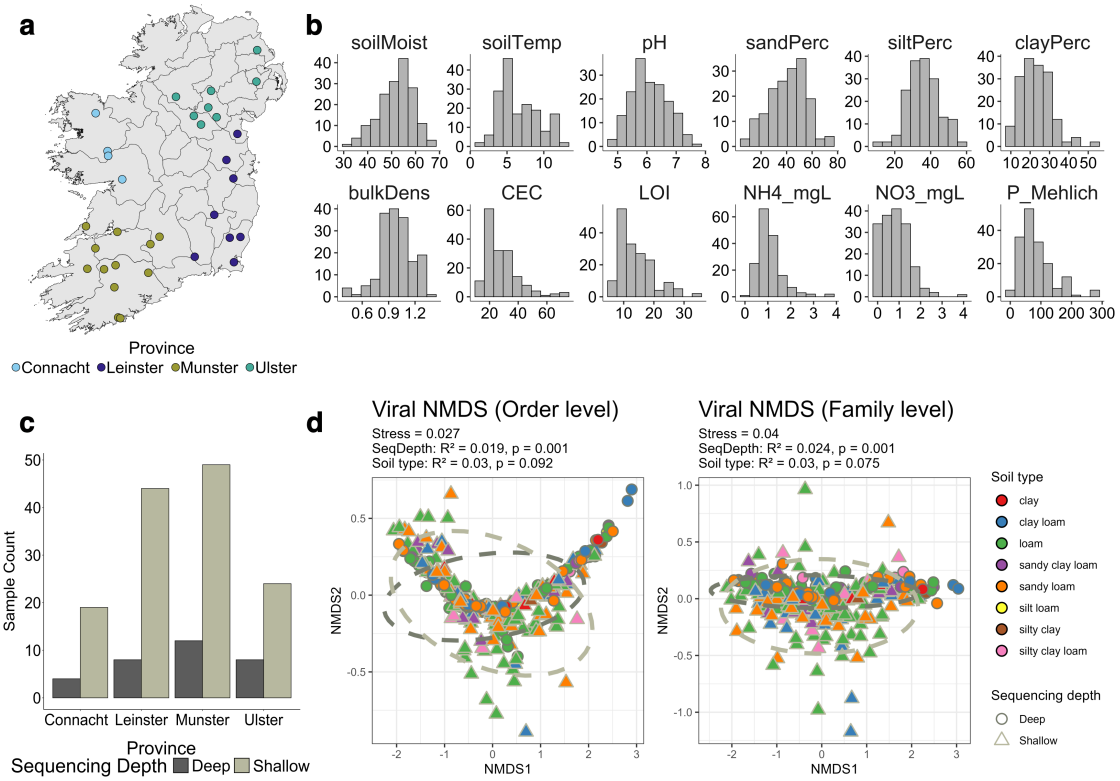


Figure 3.1: Geographic distribution, environmental context, and viral community structure of soil samples included in this study.

(A) Sampling locations across Ireland, coloured by province (Connacht, Leinster, Munster, and Ulster).
(B) Distributions of core soil physicochemical properties across all samples, shown as histograms, including moisture, temperature, pH, texture (sand, silt, and clay content), nutrient availability, cation exchange capacity (CEC), loss on ignition (LOI), inorganic nitrogen species, and bulk density, illustrating substantial environmental heterogeneity across sites.

(C) Number of shallowly and deeply sequenced samples per province, highlighting differences in sequencing depth across the sampling design.
(D) NMDS ordinations of viral community composition at the order and family levels, based on Bray-Curtis dissimilarities of read abundance profiles stratified by sequencing depth and soil type.

To capture both broad and deep metagenomic signal, soil samples were collected from 32 agricultural sites across Ireland (Figure 3.1a; Supplementary Figure 3.8). In total, 136 samples were sequenced at shallow depth (approximately 10 million reads per sample), and one replicate per site was sequenced at high depth ($>$ 100 million reads per sample) (Figure 3.1c), yielding 165 samples in total. USDA soil texture classification indicated that sampled Irish soils were predominantly loam (41.51%), followed by sandy loam (21.3%) and clay loam (17.1%). Sandy clay loam (7.9%) and silty clay loam (6.7%) were less common, while clay (2.4%), silt loam (1.8%), and silty clay (1.2%) comprised only a minority of samples (Supplementary Figure 3.9.

Assembly of 165 soil metagenomes produced 2,548,051 non-redundant contigs ($\geq$1 kb, clustered at 95% identity), totalling 4.04 Gb of assembled sequence from 1.4 Tb of raw data. Assemblies were dominated by short fragments (median 1,261 bp; mean 1,586 bp), reflecting the extreme diversity and uneven abundance structure characteristic of soil microbial communities. Long contigs were rare: although the maximum contig length reached 546 kb, only 7,558 contigs (0.30%) exceeded 10 kb.

Recovery of longer contigs was strongly influenced by sequencing depth. Deeply sequenced samples accounted for 76.2% of contigs $\geq$10 kb and recovered substantially longer contigs than shallowly sequenced samples (546 kb vs. 183 kb; Supplementary Figure 3.10). However, extended contiguous assemblies remained uncommon overall, reflecting the intrinsic assembly challenges posed by highly complex soil communities.

Read-based taxonomic classification using Kraken2 suggested that viral reads comprised only $\sim$ 0.0008% of classified microbial reads. Given the sparse representation of soil viruses in reference databases and high levels of viral sequence divergence at the nucleotide level, this likely represents a substantial underestimate of true viral abundance. In contrast, a contig-based approach detected 19,368 contigs (0.76%) containing at least one viral gene without detectable contamination (CheckV), revealing a substantially stronger viral signal. The discrepancy between read- and contig-based estimates highlights the limitations of reference-dependent classifiers in complex soil metagenomes. Although recovery of complete viral genomes remained low, this pattern is consistent with assembly constraints in highly diverse soil systems. Collectively, these results indicate that viruses comprise a rare ($<$0.1% of total reads), yet detectable component of Irish soil metagenomes.

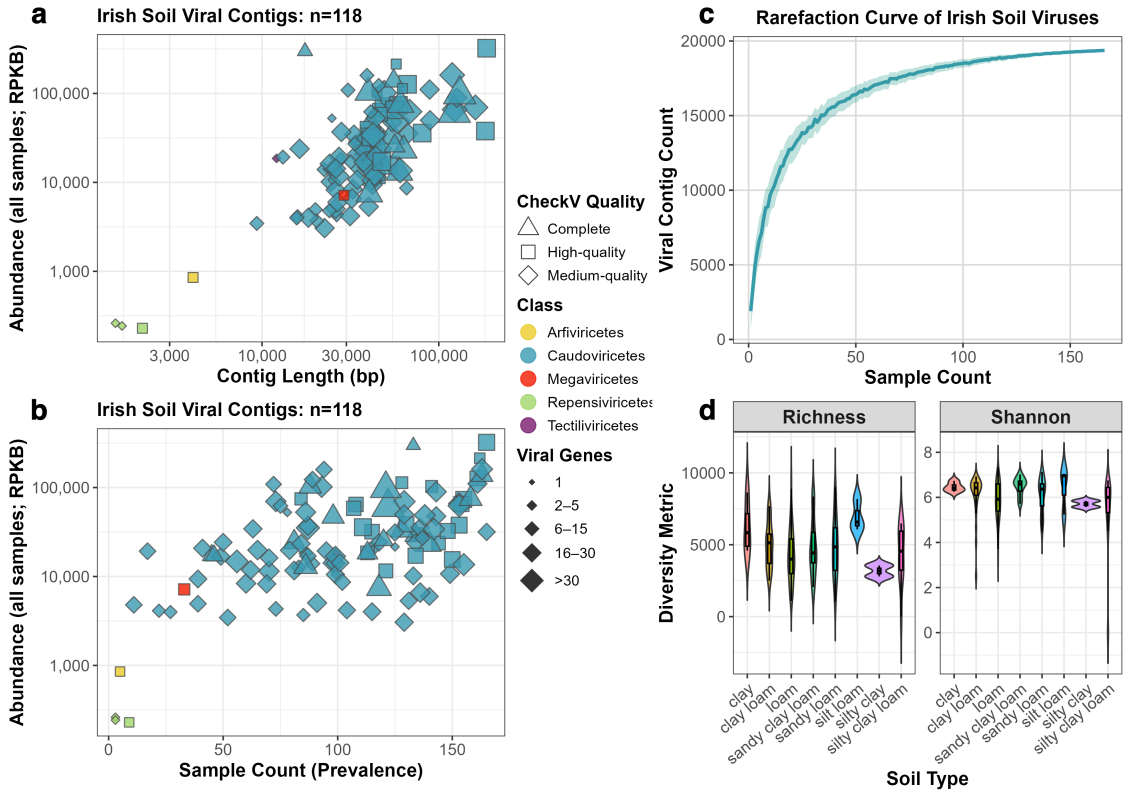


Figure 3.2: Viral genome recovery and diversity patterns across Irish agricultural soils.
(A) Viral contig length plotted against read recruitment normalised per kilobase (RPKB), with point shape indicating CheckV quality tier (complete, high-, or medium-quality) and colour denoting viral realm.
(B) Contig prevalence across samples (number of samples in which reads mapping to each contig were detected) plotted against total recruited read counts.
(C)) Rarefaction analysis of viral contig discovery across samples.
(D) Viral richness and Shannon diversity stratified by soil type.

3.3.2 Irish Viral Contig Catalogue

A total of 19,368 putative viral contigs were initially detected (defined as contigs containing ≥ 1 viral gene and no detectable contamination); yet only 118 medium- to high-quality viral genomes passed stringent inspection and were confidently confirmed as viral (Figure 3.2a,b). This discrepancy highlights the challenges of recovering complete viral genomes from complex shotgun soil metagenomes and the need to distinguish bona fide viral sequences from false positives arising from non-viral DNA. Upon closer inspection, several sequences classified as viral by CheckV showed high similarity to host-derived genomes, including plant chloroplast and bacterial chromosomal sequences, indicating misclassification driven by transposon-associated elements and shared functional domains that can mimic viral signatures[198].

Rarefaction analysis demonstrated a rapid initial accumulation of viral contigs followed by a gradual plateau (Figure 3.2c), indicating that within the limits of recoverable and assemblable genomes, sampling across the 165 soil samples approached

saturation. Thus, although overall recovery of medium-to-high quality genomes was limited relative to the total number of putative viral fragments, the dominant assemblable viral populations detectable under current assembly and classification constraints appear to be well captured in this dataset. Viral Shannon diversity varied modestly across soil texture classes (Kruskal–Wallis $\chi^2 = 14.30$, $p = 0.046$), whereas richness did not differ significantly among soil types ($\chi^2 = 13.09$, $p = 0.070$; Figure 3.2d). However, no pairwise comparisons remained significant after correction for multiple testing, suggesting that soil texture contributes only modestly to variation in viral community evenness rather than generating distinct differences among individual soil classes.

3.3.3 Taxonomy of viral sequences

All 118 viral sequences confidently identified from the dataset represent unique species, as the original sequences were dereplicated at $\geq$ 95% nucleotide identity across $\geq$ 85% genome coverage. 41 sequences showed no detectable nucleotide similarity to any other genome in the dataset. Among the remaining genomes, only a small number of pairwise comparisons exhibited substantial similarity: 2 pairs shared $>$ 40% genome coverage, 8 pairs $>$ 30%, 14 pairs $>$ 20%, 23 pairs $>$ 10%, and 28 pairs $>$ 5%. Consistent with these findings, structural comparisons also revealed high levels of diversity. The median percentage of shared protein structures across the dataset was 2.63%, with a mean of 3.10%, further emphasizing the substantial divergence among the isolated viruses.

The majority of sequences belonged to the class *Caudoviricetes*, which accounted for 112 genomes (95%), consistent with the dominance of bacteriophages in other environmental viromes[117]. A small number of sequences were assigned to other viral classes spanning both double-stranded DNA (dsDNA) and single-stranded DNA (ssDNA) viruses. Specifically, one genome belonged to the class *Arfiviricetes*, three genomes to *Repensiviricetes*, one genome to *Tectiliviricetes*, and one genome to *Megaviricetes*.

3.3.3.1 Caudoviricetes

Tailed dsDNA bacteriophages are widely considered the most abundant biological entities on Earth[117], and accordingly they dominated the recoverable viral genomes in this dataset. Genome sizes ranged from 9.3 kb to 186.5 kb, and the median sample prevalence was 109.5 samples (range: 11–165), indicating that many of these phages are widely distributed across Irish soils (Figure 3.2b,Figure 3.3a). The proportion of predicted genes belonging to specific functional categories was assessed across all 112 viral genomes (Supplementary Figure 3.11). Structural genes were highly prevalent, with capsid proteins detected in 92.9% of genomes, followed by tail-associated proteins (89.3%) and movement proteins (88.4%), indicating that the majority of contigs encode key virion structural components. Genes involved in genome packaging were also widespread, with terminase subunits identified in 75.9% of genomes and portal proteins in 68.8%, supporting the presence of canonical *Caudoviricetes* assembly machinery in a large fraction of the dataset.

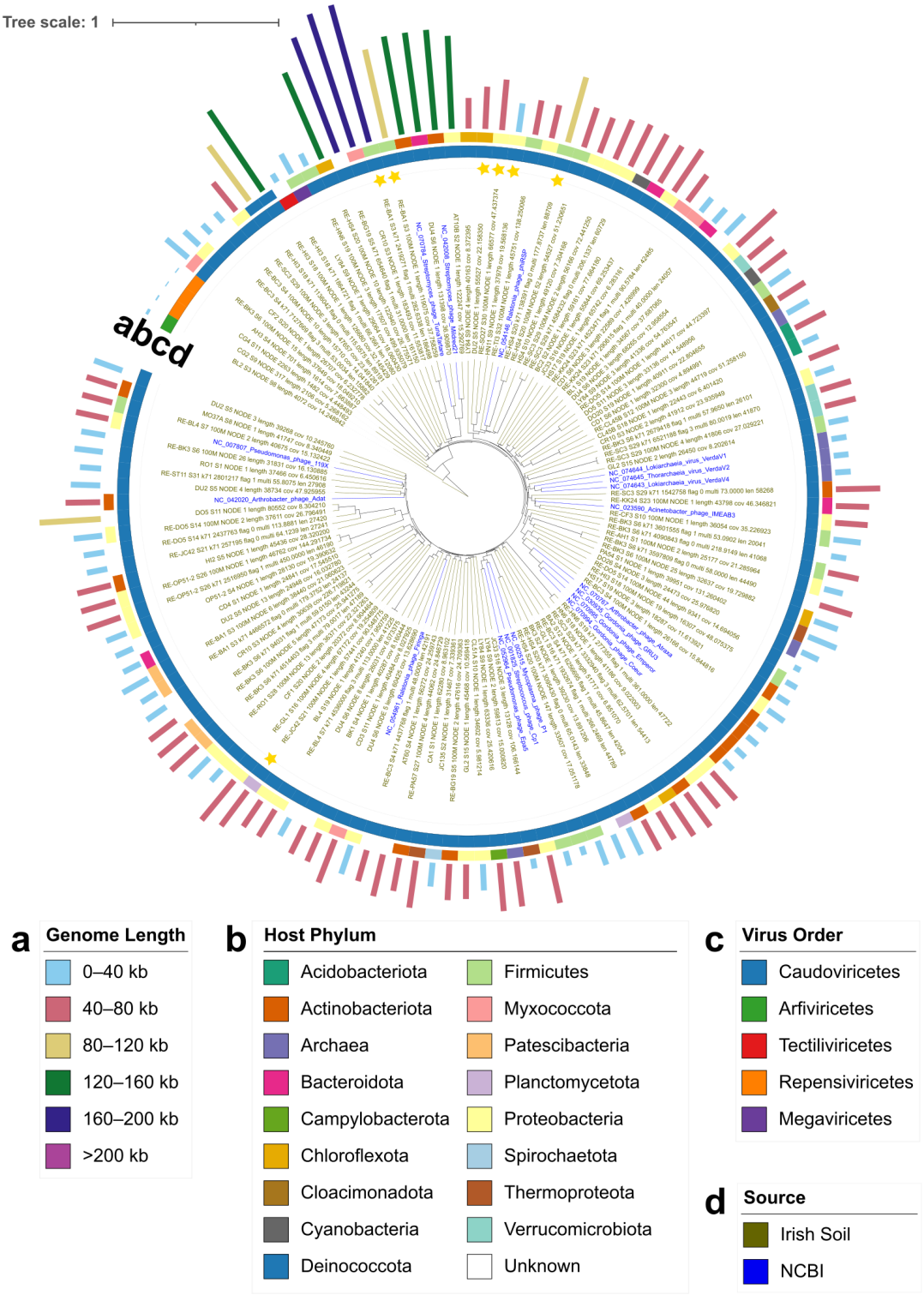


Figure 3.3: Whole structurome phylogeny of Irish soil viruses. We constructed a cladogram of Irish soil viral genomes alongside related NCBI references, rooted using *Arfiviricetes* (Figure 3). Most genomes ranged between 0–80 kb (A), and predicted hosts, inferred from CRISPR spacer matches, were dominated by Proteobacteria and Firmicutes (B). Structurally homologous viral sequences from NCBI are coloured in blue (D). Viruses detected in $\geq 98\%$ of samples - denoted by stars - represent a core, ubiquitous soil virome.

3.3.3.2 Arfiviricetes

A single contig (BL2_S3_NODE_98_length_4072_cov_14.248942; genome map shown in Supplementary Figure 3.12) was assigned to the class *Arfiviricetes*, representing the only RNA virus detected in the dataset. This circular contig was 4,072 bp in length, with the closest reference hit to *Crucivirus* sp. (accession MT478515). The virus was detected in five samples, with a total of 853 mapped reads across samples—a low prevalence compared to the dominant bacteriophages.

Gene annotation predicted several proteins consistent with viral replication and structural functions, including an RNA helicase, a putative viral replication protein, a transcription regulator, a viral coat protein (S-domain), and a deltaretrovirus Tax-like protein. This gene complement is characteristic of crucivirus-like genomes, which typically encode a small set of replication-associated and capsid proteins and are frequently detected in environmental metagenomes[199]. Cruciviruses have been proposed to represent chimeric viruses, combining replication genes related to CRESS-DNA viruses with capsid proteins derived from RNA viruses, and have been reported from diverse environmental samples[200].

3.3.3.3 Tectiliviricetes

Another contig (RE-HI3_S18_k71_1864721_flag=1_multi=66.2661_len=12080; genome map shown in Supplementary Figure 3.13) was identified within the class *Tectiliviricetes*. This genome fragment was 12,080 bp in length and classified as medium quality by CheckV. The closest reference sequence was *Deltatectivirus wheeheim* (accession MK305890), a member of the family *Tectiviridae*. The virus was detected in 99 samples, with 18,592 total mapped reads, indicating a relatively widespread distribution across the dataset.

Tectiviruses (PRD1-like bacteriophages) are unique among viruses in that they possess an internal lipid membrane enclosed within an icosahedral capsid and deliver their genomes into host cells through a specialised membrane-associated DNA delivery system[201]. Consistent with this classification, the contig contained several hallmark tectivirus genes, including a DNA polymerase type B (a core replication enzyme common in tectiviruses), an adenovirus IVa2-like packaging ATPase involved in viral genome packaging, a P3 major capsid protein (a key structural component of PRD1-like viruses), and a PRD1 phage membrane DNA delivery protein associated with genome injection into host cells. Additional predicted proteins included transcriptional regulators, metabolic genes, and several hypothetical proteins of unknown function.

3.3.3.4 Repensiviricetes

Three contigs were identified as *Repensiviricetes*, family *Genomoviridae*: CG4_S11_NODE_2263_length_1482_cov_4.448493, AH3_S4_NODE_701_length_1614_cov_7.663887,
and CG2_S9_NODE_317_length_2106_cov_5.268162 (genome maps shown in Supplementary Figure 3.14), panels a–c). Members of *Repensiviricetes* belong to the realm *Monodnaviria* and are commonly referred to as CRESS-DNA viruses (circular

Rep-encoding single-stranded DNA viruses)[202]. These viruses typically have very small circular genomes (∼2–2.4 kb) encoding only a few open reading frames, most notably a replication-associated protein (Rep) and a capsid protein. The Rep protein contains conserved rolling-circle replication endonuclease and helicase domains that are essential for genome replication[203].

The three contigs recovered here ranged from 1,482–2,106 bp in length and contained three to four predicted genes, consistent with the compact genome architecture typical of CRESS-DNA viruses. The longest contig, CG2_S9_NODE_317_length_2106_cov_5.268162 (2,106 bp; Supplementary Figure 3.14a), contained clear Rep-associated domains, including a geminivirus Rep catalytic domain and Rep helicase domain, together with a geminivirus-like capsid protein, strongly supporting its classification as a genomovirus-like CRESS virus. This contig showed highest similarity to *Gemycircularvirus* sp. (KY302866.1) and was detected in nine samples, with 229 mapped reads.

Two additional fragments, CG4_S11_NODE_2263_length_1482_cov_4.448493 (Supplementary Figure 3.14b) and AH3_S4_NODE_701_length_1614_cov_7.663887 (Supplementary Figure 3.14c), were shorter but contained partial Rep- or capsid-associated domains, including helicase motifs and geminivirus-like coat protein domains. These contigs showed closest similarity to known genomoviruses, including *Sierra dome spider associated circular virus 1* (MH545510.1) and *Gemycircularvirus minti6* (MH617534), respectively. Each was detected in three samples, with 261 and 243 mapped reads.

Although detected at very low abundances compared with the dominant *Caudoviricetes* populations, the recovery of these sequences highlights the ability of metagenomic sequencing to capture small circular ssDNA viruses that are frequently overlooked by conventional virological approaches, due to their small genome size, high sequence divergence, and biases in standard extraction and library preparation methods. Their presence suggests that CRESS-DNA viruses may represent an underappreciated component of soil viromes.

3.3.3.5 Megaviricetes

One contig (LY84_S9_NODE_7_length_29064_cov_18.060430; genome map shown in Supplementary Figure 3.15) contained multiple hallmark genes characteristic of NCLDV genomes and was therefore considered a likely member of *Megaviricetes*. *Megaviricetes* belong to the realm *Varidnaviria* and include the nucleocytoplasmic large DNA viruses (NCLDVs), which infect eukaryotic hosts and are characterised by large genomes encoding complex transcription and replication machinery[204]. This contig was 29,064 bp in length and classified as high quality (estimated 90% completeness), with the closest reference sequence corresponding to *Ostreococcus tauri virus OtV5* (EU304328), a member of the family *Phycodnaviridae*. The virus was detected in 33 samples, with 7,161 mapped reads, suggesting moderate prevalence within the dataset.

Gene annotation revealed several hallmark proteins associated with NCLDV replication and transcription, including DNA polymerase type B, a core replication

enzyme conserved across NCLDVs, together with poxvirus late transcription factor VLTF3, a DNA-directed RNA polymerase subunit, and a poxvirus rifampicin resistance protein. These genes form part of the poxvirus-like transcription machinery characteristic of NCLDV genomes[205]. Additional structural and virion-associated proteins were also detected, including adenovirus-like capsid and fibre proteins, further supporting the classification of this contig within the class *Megaviricetes*.

3.3.4 Alpha Diversity

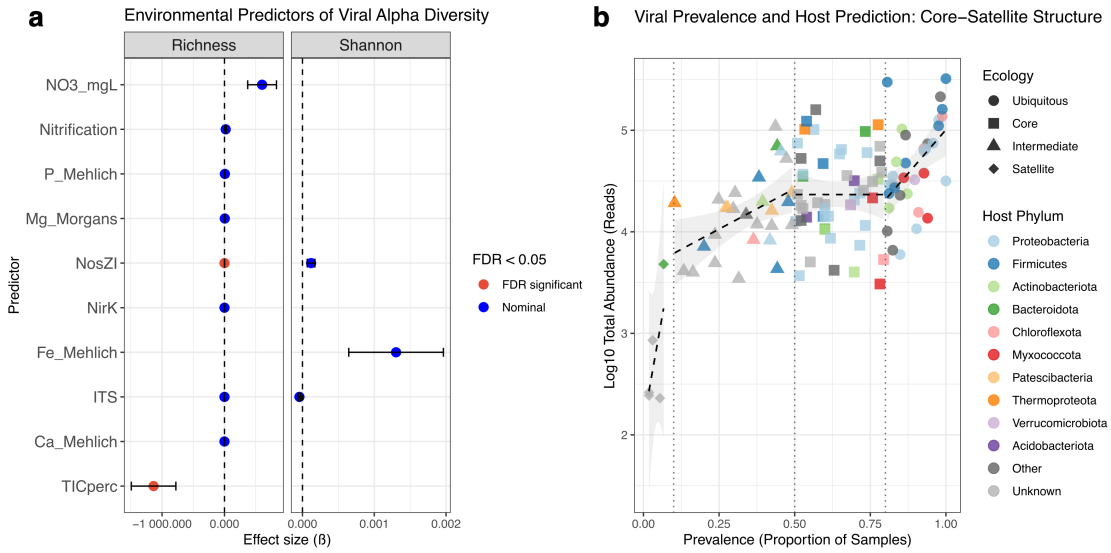


Figure 3.4: Environmental drivers of viral alpha diversity and viral prevalence with predicted hosts. (A) Effect sizes ($\beta \pm$ SE) from univariate linear models assessing associations between environmental variables and viral alpha diversity metrics (richness and Shannon diversity). Viral richness was positively associated with the abundance of clade I nitrous oxide reducers (*nosZI*) and negatively associated with total inorganic carbon (TICperc), whereas no predictors remained significant for Shannon diversity after FDR correction.
(B) Relationship between viral prevalence and total abundance across 118 high-quality viral genomes. Each point represents a viral genome, coloured by predicted host phylum based on CRISPR spacer matches and shaped according to ecological prevalence category (ubiquitous, core, intermediate, or satellite).

To assess environmental drivers of viral alpha diversity, we constructed a series of univariate linear models testing continuous predictors against Shannon diversity and viral richness, with p-values adjusted using the Benjamini–Hochberg false discovery rate (FDR). For Shannon diversity, several predictors showed nominal associations prior to correction, including Fe_Mehlich ($p = 0.050$), internal transcribed spacer (ITS) abundance ($p = 0.014$), and clade I nitrous oxide reducers (*nosZI*) abundance ($p = 0.026$) (Figure 3.4a). However, none remained significant after FDR correction (all FDR > 0.47), indicating no robust associations between measured environmental or microbial variables and viral Shannon diversity (Figure 3.4a, Supplementary Figure 3.16).

In contrast, viral richness exhibited stronger associations with specific predictors. After FDR correction, two variables remained statistically significant: *nosZI* abundance (FDR = 0.043) and total inorganic carbon (TICperc; FDR = 0.046). Viral richness was positively associated with *nosZI* abundance ($\beta = 0.36$) and negatively associated with inorganic carbon content ($\beta = -1135.6$). Several additional predictors, including NirK, NO3_mgL, and Mg_Morgans, showed nominal significance prior to correction but did not remain significant following FDR adjustment (Figure 3.4a).

Multivariate model selection retained soil pH as the only significant predictor of Shannon diversity ($\beta = -0.232 \pm 0.115$ SE, $p = 0.044$); however, explanatory power was low ($R^2 = 0.036$; adjusted $R^2 = 0.024$), indicating that pH accounted for only a small fraction of the observed variation in viral evenness. For viral richness, a negative binomial generalised linear model retained pH as a predictor, though the association was not statistically significant ($\beta = -0.103 \pm 0.061$ SE, $p = 0.091$).

Collectively, these findings indicate that soil physicochemical properties exert only weak influence on viral alpha diversity across Irish agricultural soils, with viral richness showing associations with nitrogen cycling potential and inorganic carbon content.

3.3.5 Beta Diversity

Bray–Curtis dissimilarities were visualised using non-metric multidimensional scaling (NMDS), yielding two-dimensional solutions at both the Family (stress = 0.040) and Order (stress = 0.027) levels (Figure 3.1d). PERMANOVA with marginal effects indicated that sequencing depth explained a minute but statistically significant proportion of viral community variation at both the Family ($R^2 = 0.024$, $p = 0.001$) and Order ($R^2 = 0.019$, $p = 0.001$) levels. In contrast, soil texture accounted for a comparable proportion of variance (Family: $R^2 = 0.030$; Order: $R^2 = 0.030$) but did not reach statistical significance (Family: $p = 0.075$; Order: $p = 0.092$).

Viral community composition was further evaluated across Irish provinces to assess potential large-scale geographic structuring. NMDS ordination showed substantial overlap among provinces at both taxonomic resolutions (Supplementary (Figure 3.17). PERMANOVA indicated that province explained only a small proportion of the variation (Family: $R^2 = 0.009$, $p = 0.008$; Order: $R^2 = 0.007$, $p = 0.007$). Across both Family- and Order-level analyses, more than 95% of variation remained unexplained by sequencing depth, soil texture, or province, indicating high within-site heterogeneity and suggesting that viral community composition is only weakly structured by broad geographic or edaphic gradients at this spatial scale.

3.3.6 Viral Sequence Core-Satellite Distribution

To examine the distribution of viral populations across samples, we quantified the prevalence of 118 high-quality viral genomes. Prevalence varied markedly, revealing a strongly uneven distribution across sites. Of the 118 genomes, 35 (30%) were

detected in more than 80% of samples, representing a highly ubiquitous subset of soil viruses. An additional 50 genomes (42%) were present in 50–80% of samples, collectively forming a recurrent virome core. In contrast, only 5 genomes (4%) were detected in fewer than 4% of samples, representing rare satellite populations. The remaining 28 genomes (24%) exhibited intermediate prevalence, occurring in 10–50% of samples (Figure 3.4b).

Prevalence and total abundance were positively correlated (Spearman $\rho = 0.48$, $p < 0.001$), indicating that viral populations detected across a greater number of samples also tended to recruit higher read counts. This abundance–occupancy relationship is consistent with ecological theory, in which a subset of dominant viral populations forms a stable core while a larger number of rare viruses comprise a long tail of spatially restricted populations. However, this pattern may also be influenced by technical factors, as larger genomes and higher coverage contigs are more readily assembled and detected in short-read metagenomic data, potentially biasing the observed distribution toward more easily recoverable viral genomes.

3.3.7 Predicted Hosts of High-Quality Viral Genomes

CRISPR spacer matches enabled host assignment for 89 of 118 (75.4%) high-quality viral genomes. Predicted hosts spanned 20 bacterial and archaeal phyla, but were dominated by *Proteobacteria* (28 viruses; 31.4%), followed by *Firmicutes* (15.7%) and *Actinobacteriota* (8.9%). Additional host assignments included *Myxococcota*, *Bacteroidota*, *Chloroflexota*, and several less abundant phyla, each representing fewer than 6% of host-linked viruses. The predominance of *Proteobacteria* and *Firmicutes*—two major components of soil microbiomes[206,207]—suggests that, consistent with ecological expectations, widespread viral populations preferentially infect abundant bacterial taxa.

We tested whether host phylum distributions differed across ecological prevalence categories using a χ^2 test. Host composition varied significantly among categories ($p = 0.031$). Ubiquitous and core viral genomes were primarily associated with dominant soil bacterial phyla, particularly *Proteobacteria* (28–32%) and *Firmicutes* (6–22%). In contrast, intermediate and satellite viruses were more frequently linked to less common host lineages, with satellite viruses exclusively associated with non-dominant ("Other") phyla. These patterns reinforce that widespread viral populations tend to infect abundant bacterial taxa.

3.3.8 Contrasting persistence strategies among the most prevalent soil viruses

Comparative analysis of the most prevalent viral genomes revealed distinct—yet equally successful—strategies for persistence across Irish agricultural soils. Despite substantial differences in genome size, gene content, and inferred host associations (Figure 3.3), seven viral genomes were detected in $\geq$ 98% of sampled soils, consistent with widespread persistence across diverse environments. Of these, we highlight three case studies based on highest read recruitment.

The most abundant virus, RE-BA1_S3_k71_2419271 (a circular 186 kb contig detected in all 165 samples, recruiting > 322,000 total reads; Supplementary (Figure 3.18–Figure 3.20), encodes a large and highly mosaic genome. The closest reference genome was *Pseudomonas phage Zuri* (MK863032.2; family *Schitoviridae*), sharing homology with 18 proteins at the amino acid level. The genome contains a complete suite of hallmark *Caudoviricetes* structural and packaging modules, including major capsid protein, portal protein, and terminase large subunit, alongside tail sheath, baseplate, and receptor-binding proteins. In addition, it encodes an extensive replication apparatus comprising DNA polymerase family A, ribonucleotide reductase, exonucleases, dUTPase, DNA repair enzymes, and nucleotide metabolism functions.

Beyond canonical viral modules, the genome carries an unusually extensive repertoire of auxiliary metabolic genes (AMGs) and host-associated functions (Supplementary Figure 3.20). These include multiple nitrogen-associated proteins (e.g., repeated *nifW*, *nifT/fixU*, and *cowN*), redox enzymes, glycosyl hydrolases, hydrogenase-related components, and cell division-associated proteins (ZapC), alongside stress-response and sporulation-related factors. The recurrent presence and multi-copy representation of several predicted AMG families is consistent with repeated acquisition and retention during the evolutionary history of this genome, although the selective forces underlying their maintenance remain unknown. This virus also encodes a diverse defence and counter-defence repertoire, including multiple copies of the anti-restriction protein Ral, plasmid SOS inhibitor PsiB, toxin–antitoxin systems, CRISPR-associated proteins, DNA methyltransferases, and regulatory domains, suggesting ongoing host–virus evolutionary conflict.

In contrast, a second highly abundant circular virus, RE-HN6_S19_100M_NODE_3_length_17507_cov_499.935079 (17.5 kb; Supplementary Figure 3.21), exhibited a compact and streamlined genome architecture. Despite its small size, this virus was detected in 133 samples and accumulated nearly 300,000 reads, indicating sustained ecological success across the surveyed soils. The closest reference genome match was *Streptomyces phage Immanuel3* (MG518520.1), supporting an *Actinobacteriota* host range.

The genome encodes a coherent set of structural, packaging, replication, and lysis-associated functions, including replication gene A (GPA), transglycosylase SLT, amidase/NlpC-P60 domains, and multiple spike and coat proteins, consistent with an autonomous lytic lifestyle. In addition to core viral functions, the genome carries a limited but notable set of auxiliary metabolic genes (AMGs), including the nitrogen fixation-associated protein *NifW* and several photosystem-related subunits (e.g., *PsaJ*, *PsaG/K*, *PsbY*), alongside stress-associated domains such as CsbD-like proteins. The retention of these elements within a compact genome suggests is suggestive of host-interaction modules that may enhance replication efficiency under specific environmental conditions.

Several defence- and regulatory-associated domains were also present, including colicin immunity protein, helix–turn–helix transcriptional regulators, and RNA-binding domains, corresponding with potential roles in host interaction and immune

evasion. CRISPR spacer matches linked this virus primarily to *Actinobacteriota*, particularly *Streptomyces*, a dominant and environmentally resilient soil lineage. The recurrence of matching spacers across multiple samples supports *Streptomyces* as a likely host lineage and indicates previous interactions between this viral lineage and its host.

A third prevalent virus, RE-JC42_S21_100M_NODE_1_length_57441_cov_90.548775 (57.4 kb), was detected in 162 samples and accumulated over 214,000 reads (Supplementary Figure 3.22)). The closest reference genome was *Streptococcus phage 23TH* (MT900487). The genome encodes a complete complement of *Caudoviricetes* structural and packaging modules, including portal protein, terminase large subunit, capsid, tail accessory, and spike proteins, alongside core replication functions such as helicase, single-stranded DNA-binding protein, recombination-associated NinB, and nucleotide metabolism enzymes. The presence of predicted P22 coat, portal, and tail accessory factor proteins suggests that this phage may exhibit a podovirus-like morphology. No integrase or canonical lysogeny-associated modules were identified, consistent with a predominantly lytic lifestyle.

This virus carries an expanded repertoire of host-interaction and defence-related elements, including antirestriction protein Ral, DNA adenine methyltransferase (Dam), multiple toxin–antitoxin-associated domains, CRISPR-associated exonuclease Csa1, and colicin immunity proteins. The genome also encodes several auxiliary metabolic genes (AMGs), including multiple copies of the nitrogen fixation-associated protein *NifW*, as well as enzymes linked to cofactor biosynthesis and redox metabolism. Numerous predicted domains show similarity to proteins from diverse viral and cellular lineages; given the domain-based nature of these similarities, they likely reflect conserved folds and extensive horizontal gene exchange rather than direct functional equivalence or cross-domain infectivity.

A confident CRISPR spacer match linked this virus to a host within the phylum *Planctomycetota*. Overall, this highly mosaic *Caudoviricetes* genome encodes a full complement of structural, packaging, and replication genes consistent with a lytic lifestyle, although the presence of transcriptional regulators and host-interaction genes suggests potential regulatory complexity.

While many prevalent viruses recovered from Irish soil could serve as case studies, comparative analysis of these top three genomes—selected based on prevalence and cumulative read counts across samples—illustrates that highly prevalent soil bacteriophages can exhibit markedly different genome architectures, gene repertoires, and predicted host associations while nevertheless being widely distributed across Irish agricultural soils.

3.3.9 Soil Virus Auxiliary Metabolic Gene Categorization

3.3.9.1 Soil viral auxiliary metabolic genes are dominated by host-control functions

Auxiliary metabolic genes (AMGs) are viral genes that augment or redirect host metabolism to enhance viral fitness during infection[208]. They are generally thought to originate through horizontal acquisition from host genomes and represent an important aspect of virus–host co-evolution. Historically, the term AMG referred to host-derived metabolic genes that supplement host metabolic pathways during infection; however, its use has broadened to include viral genes encoding host-like functions that are advantageous, but not essential, for viral replication[209]. Although viral homologues are often assumed to retain functions similar to their cellular counterparts, this may not always be the case. Through evolutionary processes such as exaptation, viral proteins may acquire novel biochemical or biological functions that differ substantially from those of their ancestral host genes[209].

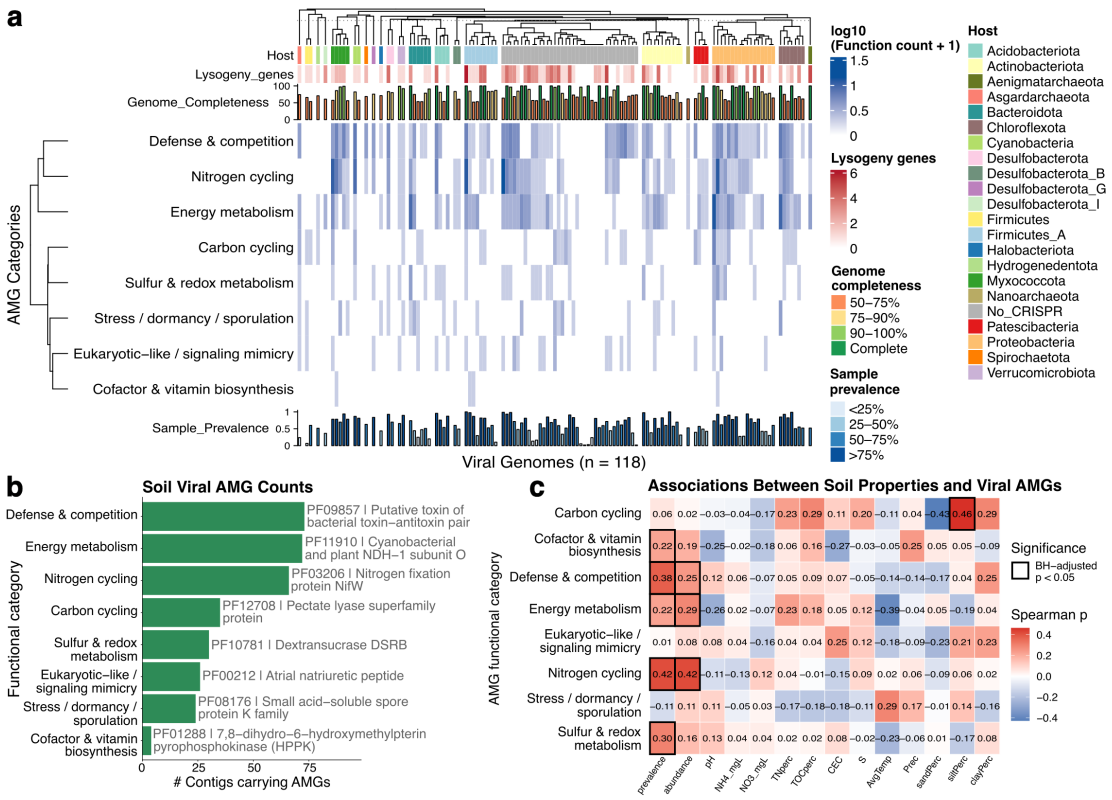


Figure 3.5: Functional and ecological characterization of high-quality Irish soil viral genomes.
(A) Heatmap of auxiliary metabolic gene (AMG) content across 118 medium- to high-quality viral genomes from Irish soil metagenomes. Genomes are clustered by AMG profiles and annotated with predicted host phylum based on CRISPR spacer matches. Additional genome features shown include lysogeny gene counts, genome completeness (CheckV estimates), and sample prevalence (proportion of samples in which each genome was detected). Rows represent broad AMG functional categories, with colour intensity showing $\log_{10}$-transformed gene counts.

(B) Bar plot of viral contigs encoding AMGs, grouped by functional category. Bars show the number of contigs containing at least one AMG per category and are annotated with the most common (modal) Pfam domain.
(C) Spearman correlations between AMG functional categories, viral genome traits (abundance and prevalence), and selected soil physicochemical properties. Colour indicates the strength and direction of correlations. Statistically significant associations (Benjamini–Hochberg adjusted $p < 0.05$) are outlined.

Consequently, functional annotations based on sequence homology should be regarded as predictions rather than definitive evidence of biological function. Establishing the physiological role of a predicted AMG requires experimental validation through approaches such as transcriptomics, proteomics, biochemical characterisation, or infection assays. Therefore, while care should be taken not to overinterpret the ecological significance of predicted AMGs, characterising their diversity nevertheless provides valuable insight into the potential metabolic interactions between viruses and their hosts. Accordingly, the AMGs identified within Irish soil viral genomes are presented here as putative functional annotations that generate hypotheses regarding virus–host interactions rather than confirmed metabolic functions.

Analysis of auxiliary metabolic genes across the 118 Irish soil viral genomes revealed that defence- and host interaction-associated functions were the most abundant category, comprising 177 genes detected across 73 genomes (Figure 3.5). This was followed by nitrogen cycling ($n = 164$; 66 genomes) and energy metabolism ($n = 152$; 71 genomes). Other functional categories, including carbon cycling ($n = 46$; 35 genomes), sulfur and redox metabolism ($n = 37$; 30 genomes), stress-related functions ($n = 32$; 24 genomes), and eukaryotic-like signalling ($n = 29$; 26 genomes), were less prevalent, while genes associated with cofactor and vitamin biosynthesis were rare ($n = 4$; 4 genomes).

A defining feature of the predicted soil viral AMG repertoire was the widespread presence of toxin–antitoxin systems and other defence-associated genes. Defence- and competition-associated AMGs were dominated by toxin–antitoxin (TA) systems, including MazF-, RelE-, HigB-, ParE-, and YafQ-like toxins together with their cognate antitoxins, alongside restriction–modification systems and CRISPR-associated proteins (e.g. Cas2, Csa1, Cmr5, and Cas9-associated domains). Beyond maintaining phage persistence, TA systems are known to induce growth arrest, dormancy, and enhanced stress tolerance in bacterial hosts—traits that may be advantageous in nutrient-limited and spatially heterogeneous soil environments[210,211]. Their recurrent presence is consistent with the hypothesis that some soil viruses promote host persistence and long-term coexistence rather than relying exclusively on rapid lytic replication.

The presence of CRISPR-associated genes within viral genomes further highlights the complexity of host–virus interactions in soil. Phage-encoded CRISPR/Cas systems can counteract phage inhibitory chromosomal islands in bacterial hosts[212], function as antiviral defence systems, and, in some cases, encode streamlined enzymes such as Casλ, which has genome-editing capabilities in both plants and humans[213]. In soil viruses, these predicted CRISPR-associated proteins may

contribute to counter-defence, interference with competing phages, or modulation of host immune responses, consistent with the view of soil as a hotspot of microbial co-evolution and competition.

Nitrogen-associated AMGs showed remarkably low functional diversity, consisting almost exclusively of predicted *NifW* and *FixU/NifT* proteins, which stabilise nitrogenase complexes under stress rather than directly catalysing nitrogen fixation[214,215]. In contrast, energy metabolism AMGs were dominated by components of electron transport and photosynthetic machinery, including NDH-1 subunits (O and M), photosystem proteins (e.g. *PsaM*, *PsbY*, *PsbX*, and *PsaJ*), and cytochrome *b6f* complex components. These proteins play established roles in electron transport, cyclic electron flow, and redox balancing[216,217,218], and are consistent with the hypothesis that phage-encoded homologues contribute to maintaining host energy metabolism during infection. Similar functions have been described in cyanophages, where energy metabolism genes are thought to supplement host pathways and prevent metabolic collapse during infection, thereby enhancing viral replication[216].

Overall, the predicted AMG repertoire of Irish soil viral genomes suggests that functions related to host stress tolerance, metabolic modulation, and competitive interactions may represent important components of virus–host interactions in heterogeneous soil environments, although these proposed roles remain to be experimentally validated.

3.3.9.2 Predicted defence- and nitrogen-associated AMGs are positively associated with viral abundance and prevalence

Spearman correlation analysis showed that viral abundance and prevalence were most strongly associated with predicted host-modulatory AMG categories. Nitrogen cycling AMGs exhibited the strongest positive correlations with both abundance and prevalence ($\rho \approx 0.42$), while defence- and competition-associated AMGs were positively associated with prevalence ($\rho \approx 0.38$) (Figure 3.5c). Other functional categories, including energy metabolism and sulfur/redox processes, showed weaker associations, whereas carbon cycling AMGs were not significantly correlated with viral abundance or prevalence. Among the measured soil properties, only silt content was significantly associated with AMG composition, correlating positively with carbon cycling genes ($\rho \approx 0.46$).

3.3.10 Anchoring Irish soil viral communities within a global context

In this study, we hypothesised that globally distributed viral populations could be detected in Irish soils through conserved nucleotide signatures. However, initial conservative read recruitment (75% breadth of coverage) against the GSVA and VIRE databases revealed a substantial number of non-viral sequences. For example, the GSVA-derived contig 3300035173.a:Ga0373463_0174927 (1,231 bp), detected across 127 samples at this threshold, showed partial similarity (42% query coverage, 73% nucleotide identity) to the bacterial chromosome of *Bradyrhizobium*

sp. PMVTL-01 (CP193982.1). Similarly, contig 3300009994.a:Ga0105126_1000567 (3,108 bp), detected in 93 samples, exhibited 100% query coverage and 100% nucleotide identity to the *Panicum virgatum* chloroplast genome (OQ873399.1).

More notably, longer contigs initially classified as viral, including 3300026121.a:Ga0207683_10000262 (47,039 bp) and 3300026121.a:Ga0207683_10000371 (41,225 bp), were detected across multiple samples (33 and 17, respectively), but functional annotation revealed them to represent fragmented bacterial genomes. These contigs encoded a broad repertoire of core cellular functions, including amino acid metabolism, energy production, transcriptional machinery, and membrane transport systems, all of which are inconsistent with viral genome architecture. The only genes with potential viral associations were limited to HNH endonucleases and dUTPases, which are insufficient indicators of viral origin in isolation.

Another key finding was that retroelement-derived sequences represent a major and previously underappreciated source of false-positive viral signal in soil metagenomes. Several contigs showing widespread detection and high coverage lacked canonical viral hallmark genes and instead encoded features characteristic of retrotransposons, including integrase domains and Ty3 capsid-like proteins[219]. For example, the VIRE-derived contig k141_947964 (6.7 kb), originally recovered from *Populus trichocarpa*-associated microbial communities, was detected across multiple Irish soil samples at high coverage despite encoding only four predicted genes, all consistent with retroelement function, including integrase and Ty3 capsid-like domains. Notably, this contig lacked structural and packaging modules required for virion formation, indicating that it represents a mobile genetic element embedded within host genomes rather than an infectious virus, despite producing strong nucleotide-level signals.

3.3.11 Evidence of shared viral populations across geographically distant soils

To explore whether a more inclusive approach would recover globally distributed viral signals, the sequence coverage threshold was reduced from 75% to 10%. This resulted in a substantial increase in detections, yielding 24,212 hits from GSVA and 241 from VIRE. To reduce noise and retain only high-confidence viral sequences, contigs were subsequently filtered using CheckV, keeping only those classified as high-quality or complete genomes. This step retained 146 of 241 VIRE contigs, whereas only 16 of 24,212 GSVA-derived contigs passed filtering. Of these, 39 contigs exhibited detectable nucleotide-level genome coverage in more than two samples (Figure 3.6).

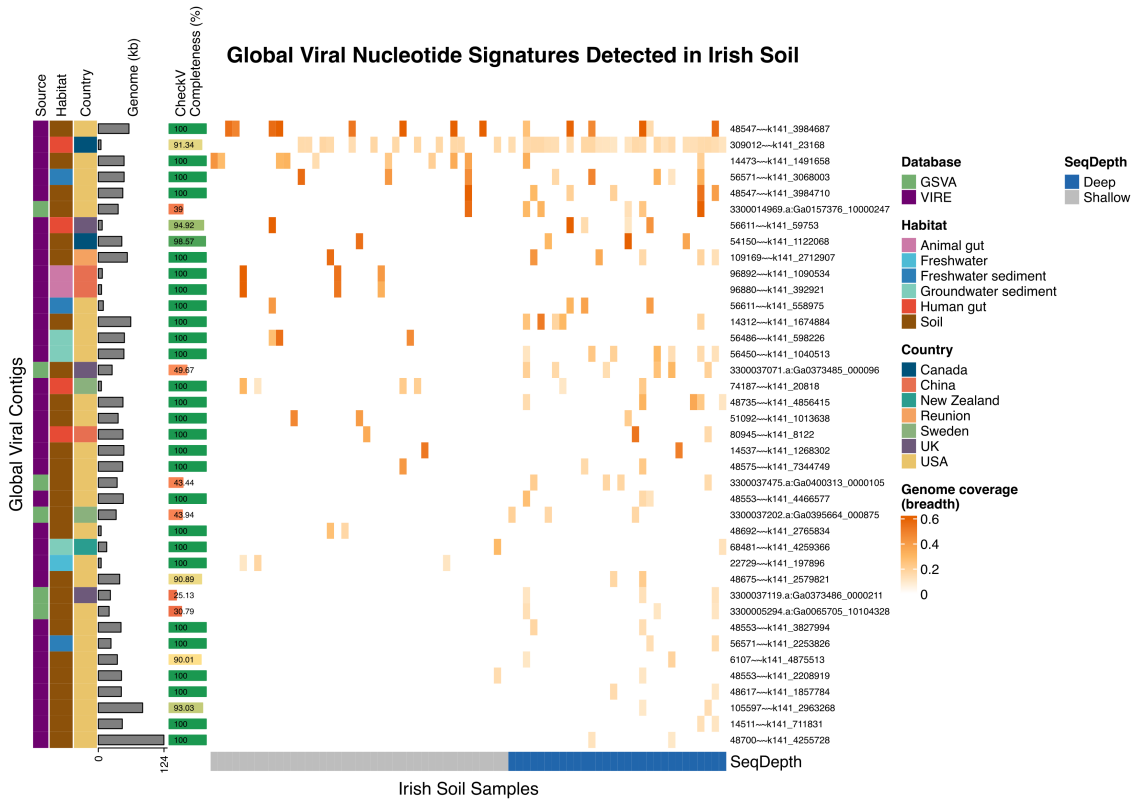


Figure 3.6: Global viral nucleotide signatures detected in Irish soil metagenomes. Reference viral sequences from the VIRE (> 1 million sequences) and GSVA (∼50,000 sequences) databases were mapped against Irish soil metagenomic reads using Bowtie2. Genome coverage (breadth) was calculated for each reference, with a minimum threshold of 10% coverage required to retain candidate hits. All retained sequences were subsequently filtered using CheckV to identify high-quality viral contigs. The heatmap displays genome coverage across Irish soil samples for each reference viral contig, with rows representing global viral sequences and columns representing individual samples. Metadata annotations indicate source habitat, country of origin, genome length, and CheckV completeness. Despite the inherent limitations of nucleotide-based homology searches for highly diverse viral populations, detectable signal was observed for multiple bacteriophages, with some contigs exhibiting $\geq 60\%$ genome coverage, suggesting the presence of conserved or closely related viral lineages in Irish soils.

The genome of contig 48547~~k141_3984687 (57,901 bp; 84 predicted genes), originally recovered from soil in Washington, USA (BioSample: SAMN10864404), is shown in Figure 3.6 and Supplementary Figure 3.23. This *Caudoviricetes* genome exhibited a median nucleotide coverage of 0.56 across the 14 Irish soil samples in which it was detected. Despite its widespread presence and high coverage, no nucleotide-level homology was observed between this genome and any curated Irish soil viral sequences. Further analysis at the protein level will be required to determine whether this virus is related to Irish soil viruses and to identify which genomic regions are responsible for the observed nucleotide-level signal in Irish metagenomic reads.

Among the retained contigs, a number of globally sourced viral genomes showed

evidence of homology to viruses detected in Irish soil. Notably, two complete double-stranded DNA viral genomes, each approximately 120 kb in length and encoding over 200 predicted proteins, exhibited substantial nucleotide similarity (80–100% pairwise identity across ~58.8% of the genome; ~72,800 bp) (Figure 3.7a). The genome recovered from Irish soil (AT10B_S2_NODE_1_length_122247_cov_15.292769) was highly prevalent (detected in 122 samples) and displayed strong similarity to a complete viral genome previously recovered from a *Miscanthus* rhizosphere soil sample in Kellogg Biological Station, Michigan, USA (48700~~k141_4255728; BioSample: SAMN06268385).

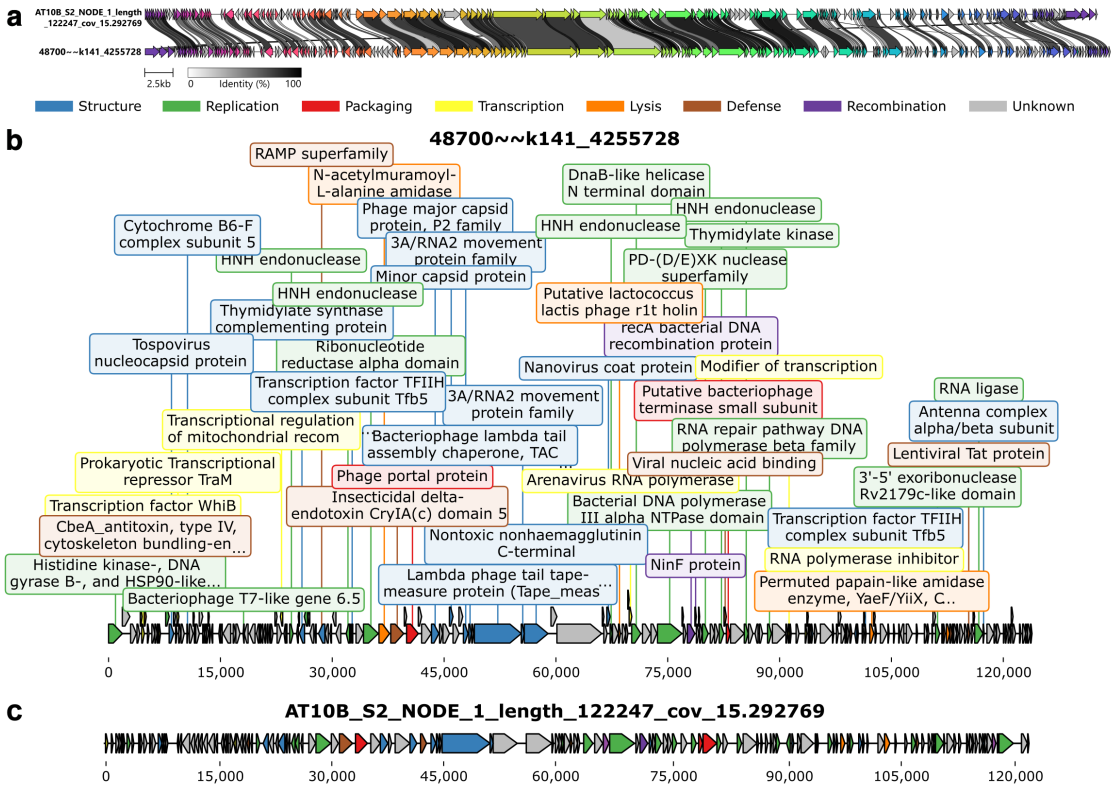


Figure 3.7: Genome synteny between Irish soil-derived bacteriophage and a geographically distinct soil phage. Comparative genome analysis reveals extensive synteny between the Irish phage AT10B_S2_NODE_1_length_122247_cov_15.292769 and a phage recovered from a *Miscanthus* rhizosphere soil sample in Kellogg Biological Station, Michigan, USA (48700~~k141_4255728), despite their geographic separation (a). These genomes align at the nucleotide level across ~58.8% of the query genome (~72,788 bp), demonstrating substantial sequence conservation alongside preserved gene order. Conserved nucleotide-level alignments (b–c) highlight shared genes, particularly within structural, replication, and transcriptional modules, indicating strong evolutionary constraint on core viral functions.

Comparative genomic analysis revealed extensive conservation between the two genomes. Pairwise alignment demonstrated strong synteny across large portions of the genomes, with numerous homologous genes exhibiting moderate to high amino acid identity (typically 60–95%) (Figure 3.7a). This conservation extended

across key functional modules, including replication, structural, and packaging genes, suggesting a shared genomic architecture and common evolutionary origin. Functional annotation further supported this relationship, with both genomes encoding a similar repertoire of predicted proteins, including tail capsid proteins, and transcription-associated factors (Figure 3.7b,c). Database comparisons of AT10B_S2_NODE_1_length_122247_cov_15.292769 identified the closest reference genome as *Streptomyces phage Bmoc* (family *Stanwilliamsviridae*), although the total aligned region between the query genome and this reference was limited (∼4 kb), representing a small fraction of the genome. This suggests that while the recovered genomes are related to known members of this family, they are not well represented in current reference databases. CRISPR spacer matches linked AT10B_S2_NODE_1_length_122247_cov_15.292769 to members of the *Thiolinea* genus (family *Thiotrichaceae*, order *Thiotrichales*, class *Gammaproteobacteria*, phylum *Proteobacteria*), although this association was supported at moderate confidence, with spacer hits to this taxon accounting for $\geq 50\%$ of all matches.

This result indicates that the identified genomes likely represent members of an under-sampled viral lineage within *Caudoviricetes*, potentially corresponding to a novel species or genus within an existing order. The recovery of two highly similar viral genomes from geographically distinct soil samples suggests that this lineage is not an isolated occurrence, but may instead represent a persistent and ecologically relevant component of soil viral communities. Given the known challenges associated with viral genome recovery from soil environments—including high diversity, uneven coverage, and assembly fragmentation—as well as the typically high nucleotide diversity among bacteriophages, the detection of conserved, complete genomes is notable.

It is possible that additional members of this lineage remain undetected in existing datasets, particularly due to limitations in current search strategies. Fully characterising the diversity and distribution of this potential clade will require approaches that extend beyond nucleotide similarity, including protein-level and structure-informed methods. Systematic mining of large-scale soil metagenomic repositories using such approaches may reveal whether this lineage is more broadly distributed across soil environments and help clarify its taxonomic placement within the broader viral phylogeny.

A secondary comparison identified limited genomic similarity (∼13.8% genome coverage; ∼6.4 kb) between two similarly sized (∼45 kb) viral genomes from soil (HI2_S5_NODE_1_length_45436_cov_28.320200) and human gut environments (80945~~k141_8122, China; BioProject SAMN13252589; Supplementary Figure 3.24). The shared regions were restricted to a subset of inverted genes arranged in opposite orientations, consistent with the conservation of discrete functional modules and the mosaic architecture of bacteriophage genomes. The complete Irish contig (HI2_S5_NODE_1_length_45436_cov_28.320200) showed its closest alignment to *Streptomyces phage FlowerPower* (MH155868.1), with 3,367 amino acids aligned across the reference genome. In addition, a high-confidence CRISPR spacer match linked this contig to soil-associated ammonia-oxidising bacteria of the RAAP-2 family within the order *Acidimicrobiales*[197]. This observation highlights the potential

for detecting cross-environmental bacteriophage conservation at the nucleotide level, although more comprehensive analyses are required to assess its extent.

3.3.12 Metagenomic viral discovery – trade-offs and challenges

While this viral discovery analysis employed nucleotide-level screening followed by stringent filtering, such approaches differ from more sensitive methods based on hidden Markov models (HMMs), which are often considered the gold standard for viral detection[122]. As a result, the strategy used here may bias against the recovery of highly novel viral sequences. It is therefore likely that additional, previously uncharacterised viral diversity exists within soil environments - including within this dataset - but remains undetected due to limited similarity to currently available reference genomes.

The identification of such novel viruses would require the interrogation of sequences that are not recognised as viral by existing tools, followed by rigorous exclusion of cellular origin and detailed functional annotation. Current viral discovery relies on the infallibility of tools such as CheckV[197]. In this analysis, we discovered false positives arisen from transposons in plant mitochondrial DNA for example, showing how viral discovery is extremely nuanced and inherently difficult, analogous to looking for a "needle-in-a-haystack". Consequently, sequences representing genuinely novel viral lineages may have been excluded during this study, as they may lack canonical viral hallmark genes and instead consist predominantly of hypothetical proteins or proteins of unknown function. This highlights an inherent trade-off between confidence and discovery, where increasing stringency reduces false positives but may also remove biologically meaningful, yet currently unrecognisable, viral sequences.

3.3.13 Discussion and future direction

A limitation of this study is that viral abundance was estimated from the total number of sequencing reads recruited to each viral contig, rather than using genome-wide coverage metrics and length-normalised abundance estimates. Consequently, abundance estimates may be influenced by differences in contig length, uneven sequencing depth, or localised regions of high read recruitment, potentially biasing quantitative comparisons among viral genomes. In addition, viruses represented by only a small number of reads mapping to conserved genomic regions may have been incorrectly classified as present, leading to an overestimation of viral prevalence. While this approach provides a useful approximation of the relative representation of viral sequences within each metagenome, the reported abundances and occupancy estimates should therefore be interpreted with caution. Future work will incorporate genome-wide coverage metrics, including average depth of coverage, breadth of coverage, and length-normalised abundance measures (e.g. RPKM or TPM), together with minimum breadth-of-coverage thresholds to provide more robust estimates of viral abundance and prevalence.

3.3.14 Conclusion

Soil harbours an extraordinarily diverse and largely unexplored virosphere, yet recovering viral signal from shotgun metagenomic data remains inherently challenging. In this study, viral sequences comprised only a minute fraction of total reads, and genome recovery was constrained by the combined effects of extreme microbial diversity, uneven coverage, and limitations of short-read assembly. These factors render viral discovery in soil a fundamentally difficult task, analogous to searching for a "needle in a haystack".

Despite these constraints, we identified a set of high-confidence viral genomes dominated by members of the class *Caudoviricetes*, confirming that tailed bacteriophages represent the most abundant and widespread viruses in Irish soils. Notably, several viral populations were consistently detected across geographically and environmentally distinct samples, indicating that a subset of soil viruses can persist across heterogeneous environments despite substantial differences in genome size, host range, and inferred ecological strategy. Their auxiliary metabolic gene repertoires were dominated by defence and nitrogen-associated functions, pointing toward a strategy centred on host interaction, stress tolerance, and long-term persistence.

Attempts to detect a globally conserved viral soil signature revealed that nucleotide-level conservation is limited and readily confounded. A substantial proportion of sequences identified through reference-based approaches corresponded not to *bona fide* viruses, but to host-derived fragments or mobile genetic elements, particularly retrotransposons. These elements, which encode virus-like proteins yet remain non-infectious, generated strong and reproducible signals across samples, highlighting a critical limitation in current viral discovery frameworks and contributing to false-positive viral assignments in large-scale databases.

Notwithstanding the generally limited nucleotide conservation observed across soil viral populations, we identified isolated instances of highly conserved genomes recovered from geographically distant environments. In particular, a complete viral genome recovered from Irish soil exhibited extensive nucleotide similarity and conserved gene synteny with a phage isolated from North American soil. While this observation does not indicate widespread global conservation, it suggests that certain viral lineages may exhibit broader geographic persistence than is typically assumed. Determining the extent to which such lineages are distributed across soil environments will require more comprehensive and protein sensitive approaches.

Collectively, these findings demonstrate that soil viral diversity is vast but highly fragmented, locally structured, and only weakly conserved at the nucleotide level. They further emphasise that reliance on reference-dependent and gene-centric methods can lead to substantial misclassification, particularly in complex environments where viral, cellular, and mobile genetic elements share overlapping functional features.

Future approaches to soil viromics must therefore move beyond strict dependence on existing reference databases and canonical hallmark genes. Integrative

strategies that incorporate genome-wide context, structural and functional coherence, and reference-independent signals will be essential to resolve the true extent of viral diversity and to capture the substantial fraction of "viral dark matter" that remains undetected. Advancing these frameworks will be critical for accurately characterising viral communities and understanding their roles in soil ecosystem function.

3.3.15 Methods

3.3.15.1 DNA extraction and sequencing

DNA was extracted from 0.25 g soil using the QIAGEN DNeasy® PowerSoil® Kit following the manufacturer's instructions. The yield and quality of DNA extracts were verified using a NanoDrop 2000 spectrophotometer and by electrophoresis on a 1% agarose gel. Library preparation was performed using the Nextera XT Library Preparation Kit, and pooled samples were sequenced at the Teagasc Moorepark sequencing facility. Pooled libraries of all DNA extracts were sequenced across four lanes of an Illumina HiSeq 4000 platform, generating a median of 16,586,860 paired-end reads (5.85 Gb) of data per extract.

3.3.15.2 Quality control and assembly

Sequencing quality was assessed for all raw reads using FastQC (v0.12.1). Adapter trimming and quality filtering were performed using Trim Galore (v0.6.10; Cutadapt v4.8) with a minimum Phred quality threshold of 20 (–quality 20). Three complementary assembly strategies were applied depending on sequencing depth. Shallowly sequenced samples were assembled using metaSPAdes (SPAdes v3.15.5). Deeply sequenced samples were assembled using MEGAHIT (v1.2.9), followed by an additional assembly of the first 100 million reads from each deeply sequenced sample using metaSPAdes (SPAdes v3.15.5). All resulting contigs were pooled and dereplicated using MMseqs2 `easy-cluster` with a minimum sequence identity threshold of 95% (`--min-seq-id 0.95`). Representative sequences from each cluster were retained and used as the non-redundant (nr) contig set.

3.3.15.3 Viral sequence identification

To identify putative viral sequences from the assembled metagenomic contigs, we applied a combination of VIBRANT (v1.2.0) and geNomad (v1.9.3) for virus detection, followed by CheckV for quality assessment and genome completeness estimation.

All assembled contigs ≥1 kb in length were processed with VIBRANT, which predicts viral genomes based on gene annotation and genomic context. Contigs containing at least four predicted open reading frames (ORFs) were retained, yielding 10,600 putative viral contigs. These were dereplicated using MMseqs2 (`easy-cluster` module) with a minimum sequence identity threshold of 95% and a minimum coverage of 85%.

In parallel, geNomad was run on all assembled contigs ≥1 kb. The resulting `virus_summary.tsv` file was filtered to retain contigs containing at least one viral hallmark gene ($n_{\text{hallmarks}} \geq 1$) and a virus score ≥ 0.75, resulting in 21,701 putative viral contigs.

The non-redundant viral contigs from both tools were merged, producing a combined dataset of 29,231 unique putative viral contigs. This combined set was subsequently evaluated using CheckV (v1.0.3; database `checkv-db-v1.5`) to assess genome completeness and contamination. Only contigs that were not classified as proviruses, contained at least one viral gene, and had 0% estimated contamination were retained, yielding a total of 19,368 putative viral contigs. Final high-quality viral genomes were defined as those classified by CheckV into the *Complete*, *High-quality*, or *Medium-quality* categories.

These contigs were searched against the NCBI nt database to remove false positives arising from plant-derived DNA. Following filtering and manual curation, a total of 118 high-confidence viral contigs were retained for downstream analyses, representing a subset of the 19,368 sequences comprising the complete Irish viral contig catalogue.

3.3.15.4 Viral gene prediction and annotation

Protein-coding genes were predicted from viral contigs using Prodigal (v2.6.3) in metagenomic mode. Functional annotation of predicted proteins was performed using protein-vec (https://github.com/tymor22/protein-vec) to infer Pfam domains based on learned protein embeddings. To support this analysis, the software environment was configured for CUDA 11.4 with PyTorch 1.12.1. Predicted Pfam accessions were linked to domain definitions from the Pfam-A seed database (https://www.ebi.ac.uk/interpro/download/Pfam/Pfam-A.seed.gz) using custom Python scripts.

3.3.15.5 Viral taxonomic classification

Viral sequence taxonomic classification was undertaken at the amino acid sequence level https://github.com/linda5mith/crassify and through whole-proteome structural comparison via PhagePleats https://github.com/linda5mith/PhagePleats[220]. Protein folding was undertaken using esmfold_v1() model from ESMFold[125]. Folding was performed in an environment derived from the project-provided environment.yml, with updates for A100 GPU compatibility (PyTorch 1.9.0+cu111). Structures were generated using: python fold.py –chunk-size 128.

3.3.15.6 Host prediction

Bacterial host prediction was carried out as described in methodology of GSVA[180] whereby viral sequences were aligned against a database of CRISPR spacers of 1.6 million bacterial and archaeal genomes[221].

3.3.15.7 Global virus prevalence and abundance detection

To identify globally distributed viral genomes detectable in Irish soil metagenomes, viral reference sequences from the Global Soil Virus Atlas (GSVA) and the Viral Reference database (VIRE) were screened against Irish soil sequencing reads. Quality-filtered paired-end reads were aligned to viral reference genomes using Bowtie2 (v2.4.2) with default parameters. Resulting alignments were processed using SAMtools to generate sorted BAM files.

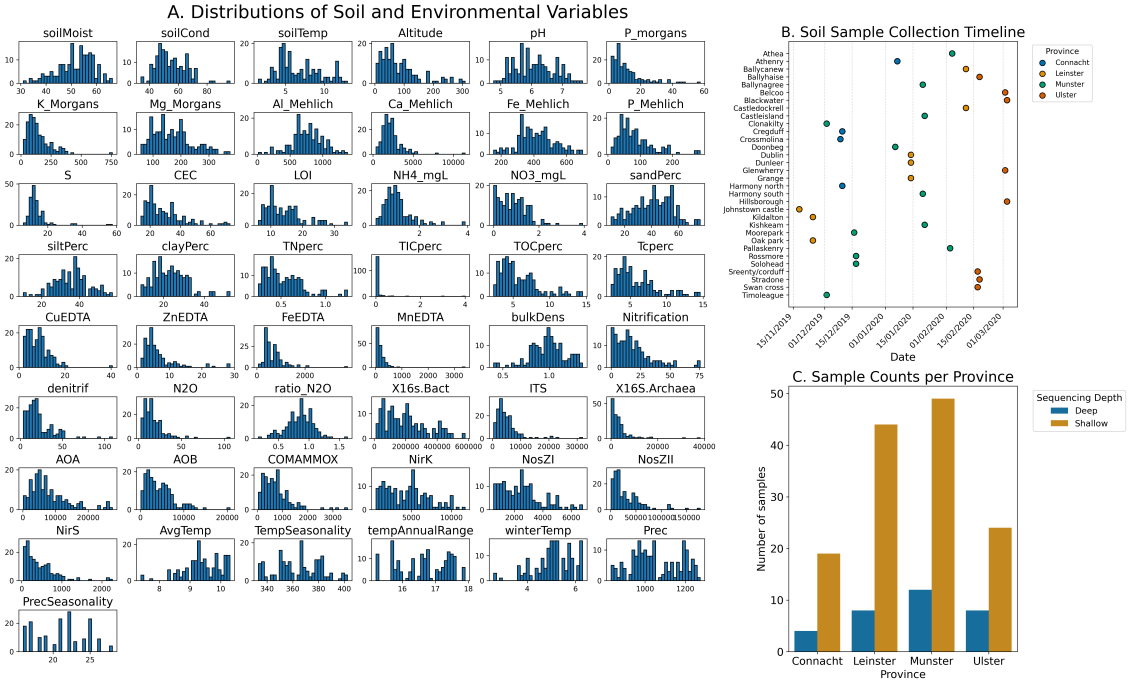
Genome coverage was calculated from BAM alignments, and viral genomes were initially retained if they achieved ≥70% breadth of coverage across their length in at least one sample. This stringent threshold was applied to ensure high-confidence detection and to minimise spurious alignments. However, gene content analysis of retained contigs revealed many false positives including plant-derived sequences and fragmented bacterial genomes. To address this, the coverage threshold was relaxed to ≥10%, allowing the inclusion of more permissive viral homologues. These were subsequently filtered using CheckV to retain high-confidence viral sequences for downstream analyses.

3.3.15.8 Statistical Analysis and visualisation

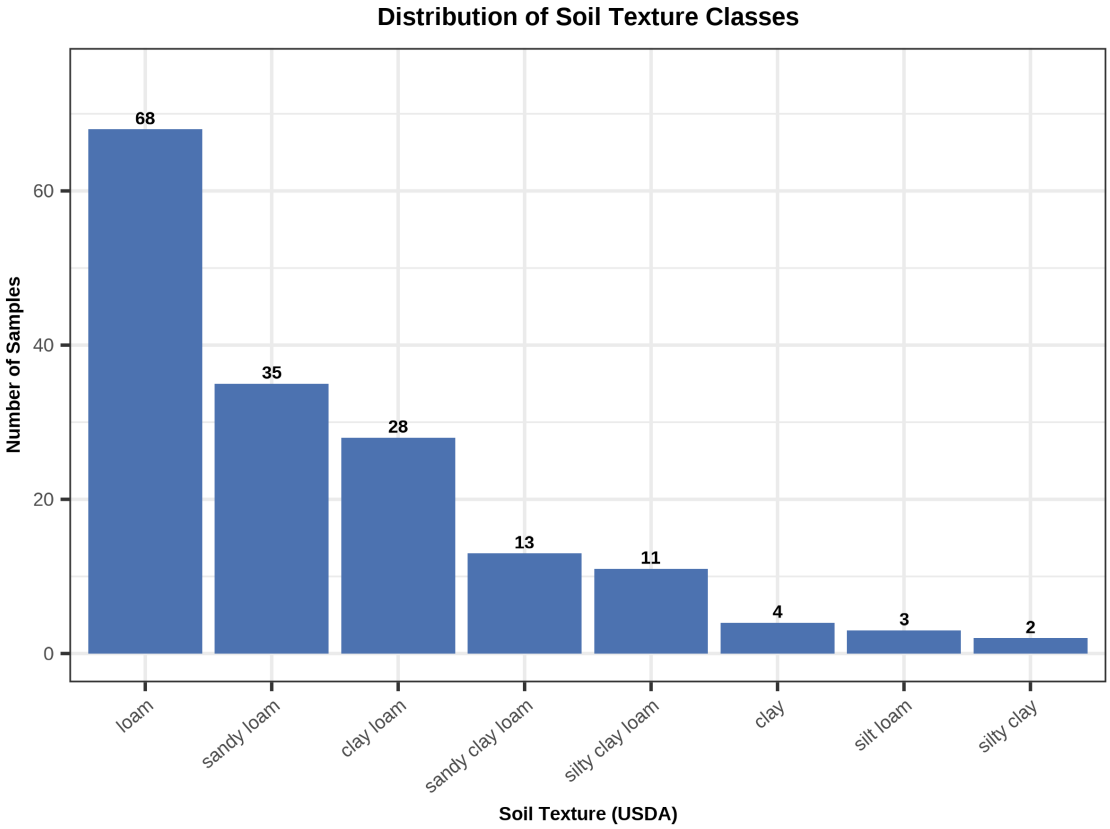
Multivariate ecological analyses were performed using the `vegan` package[222]. All statistical analyses were performed using Python v3.8.19[146] in Jupyter Notebooks[147] and R v4.1.1[148]. In Python, data processing, statistical analyses, and visualisation were conducted using `pandas` v2.0.3[149], `NumPy` v1.24.4[150], `scikit-learn` v1.3.2[151], `seaborn` v0.13.2[152], `matplotlib` v3.7.5[153], and `Entrezpy`[154]. In R, data manipulation and visualisation relied on `ComplexHeatmap`[155], `ggplot2`[156], `dplyr`[157], `tidyr`[158], and `RColorBrewer`[159]. Structural phylogram was made in `iTOL`[162] and genome synteny maps in `Clinker`[223].

All code and analyses supporting this study provided at https://github.com/linda5mith/Irish_Soil_Virome

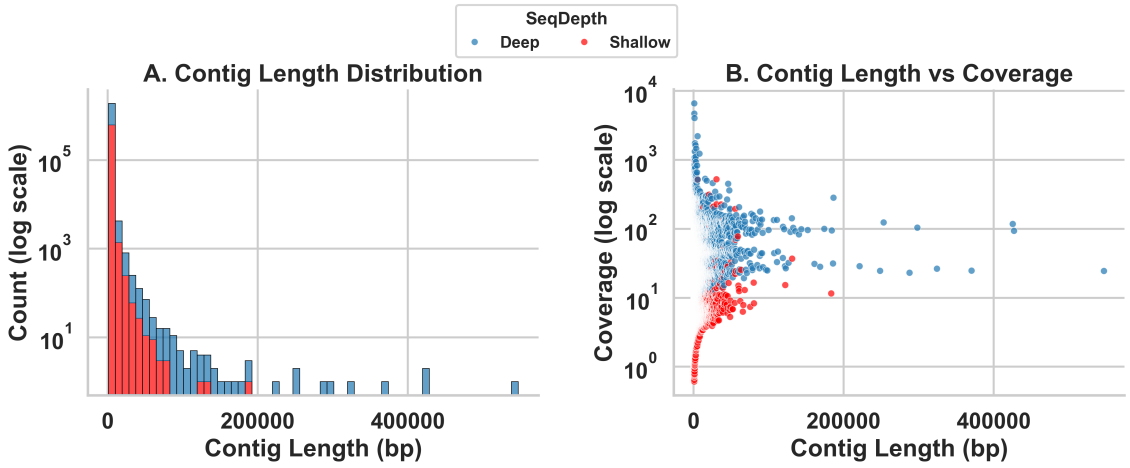
3.3.16 Supplementary Material



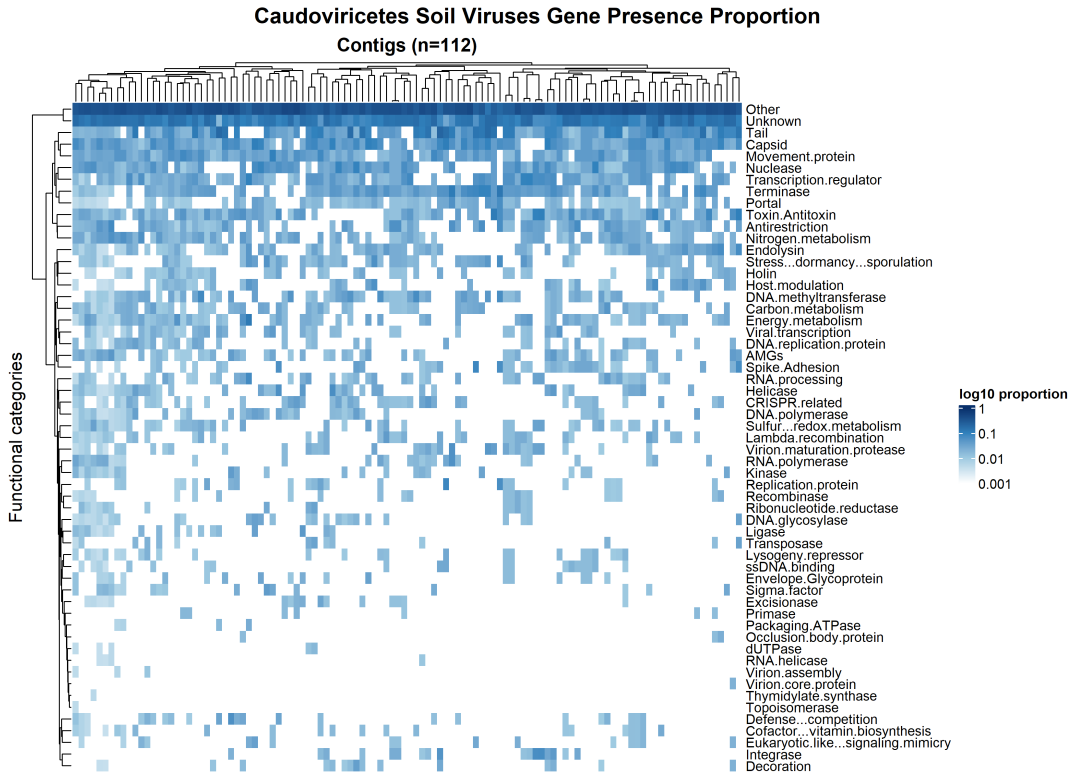
Supplementary Figure 3.8: Overview of soil sample metadata.
(A) Summary of all chemical and physical soil properties measured across samples. (B) Geographic distribution of the 32 sampling sites across Ireland, coloured by province. All samples from each site were collected during a single sampling visit, with collection dates spanning 15 November 2019 to 1 March 2020, representing a winter-season metagenomic snapshot.
(C) Sequencing depth across samples, grouped by province. The majority of samples from each site were shallowly sequenced (< 10 million reads), while one sample per site was randomly selected for deep sequencing (> 100 million reads).



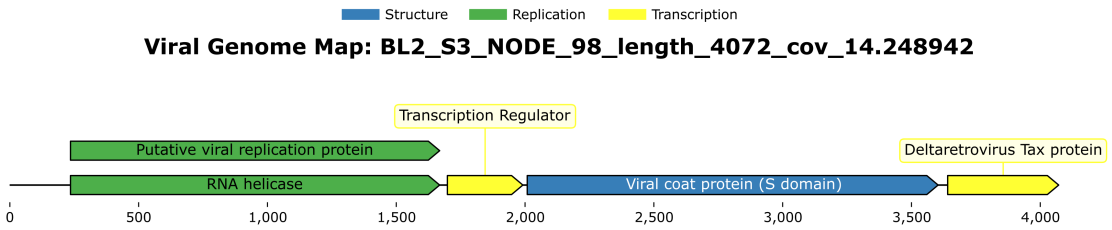
Supplementary Figure 3.9: Distribution of USDA soil texture classes across sampled Irish soils.
USDA soil texture classification based on measured sand, silt, and clay percentages indicated that sampled soils were predominantly loam (41.5%), followed by sandy loam (21.3%) and clay loam (17.1%). Sandy clay loam (7.9%) and silty clay loam (6.7%) were less frequently represented, while clay (2.4%), silt loam (1.8%), and silty clay (1.2%) comprised only a small proportion of samples. Overall, the dataset was dominated by moderately textured mineral soils characterised by balanced sand-silt-clay composition.



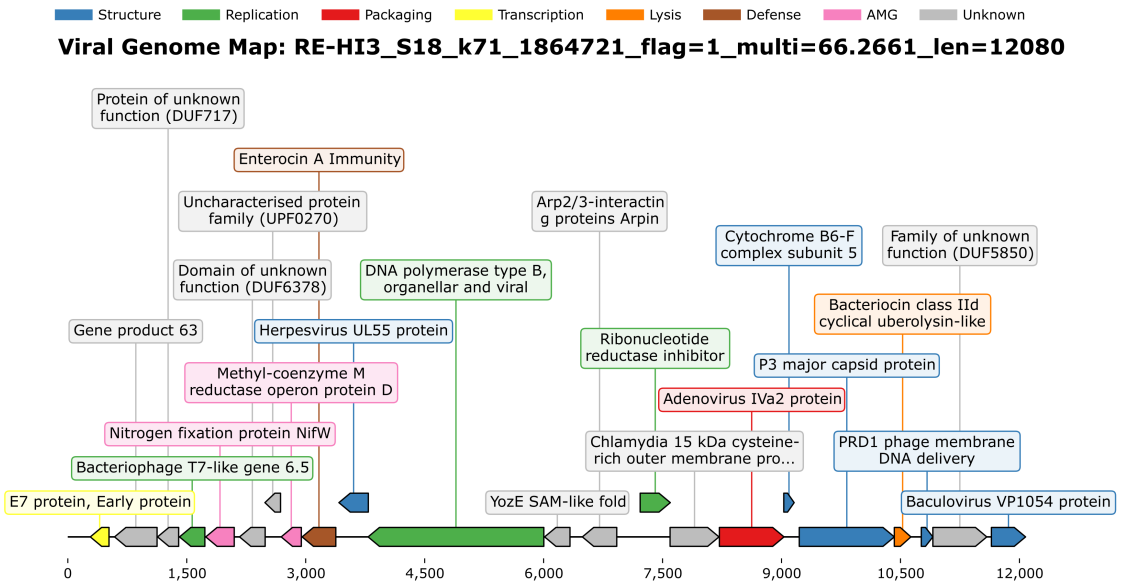
Supplementary Figure 3.10: Assembly contig length and coverage profiles across sequencing depths.
(A) Distribution of assembled contig lengths for deeply sequenced (blue) and shallowly sequenced (red) samples, shown on a logarithmic count scale. Deep sequencing yields a broader contig length distribution and recovers substantially longer contigs, whereas shallow sequencing predominantly produces shorter contigs.
(B) Relationship between contig length and mean sequencing coverage (log scale). Deeply sequenced samples exhibit higher coverage across a wide range of contig lengths, including long contigs exceeding 100 kb, while shallowly sequenced samples are largely constrained to shorter, lower-coverage contigs. Together, these patterns illustrate the impact of sequencing depth on assembly continuity and support the selection of deeply sequenced samples for high-resolution downstream viral analyses.



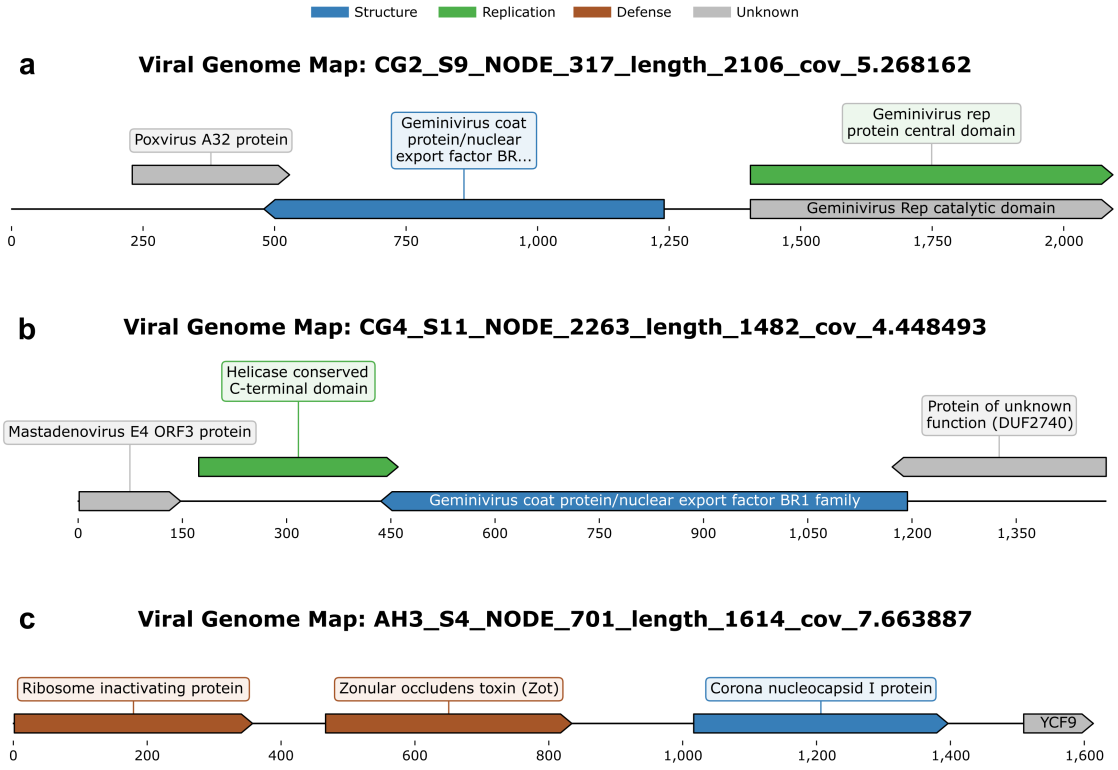
Supplementary Figure 3.11: *Caudoviricetes* soil contigs gene presence proportion.
The proportion of predicted genes assigned to functional categories was assessed across 112 viral genomes. Structural genes were highly prevalent, with capsid (92.9%), tail (89.3%), and movement proteins (88.4%) detected in most contigs. Packaging components were also widespread, with terminases (75.9%) and portal proteins (68.8%) present in a majority of genomes, consistent with established *Caudoviricetes* virion architecture.



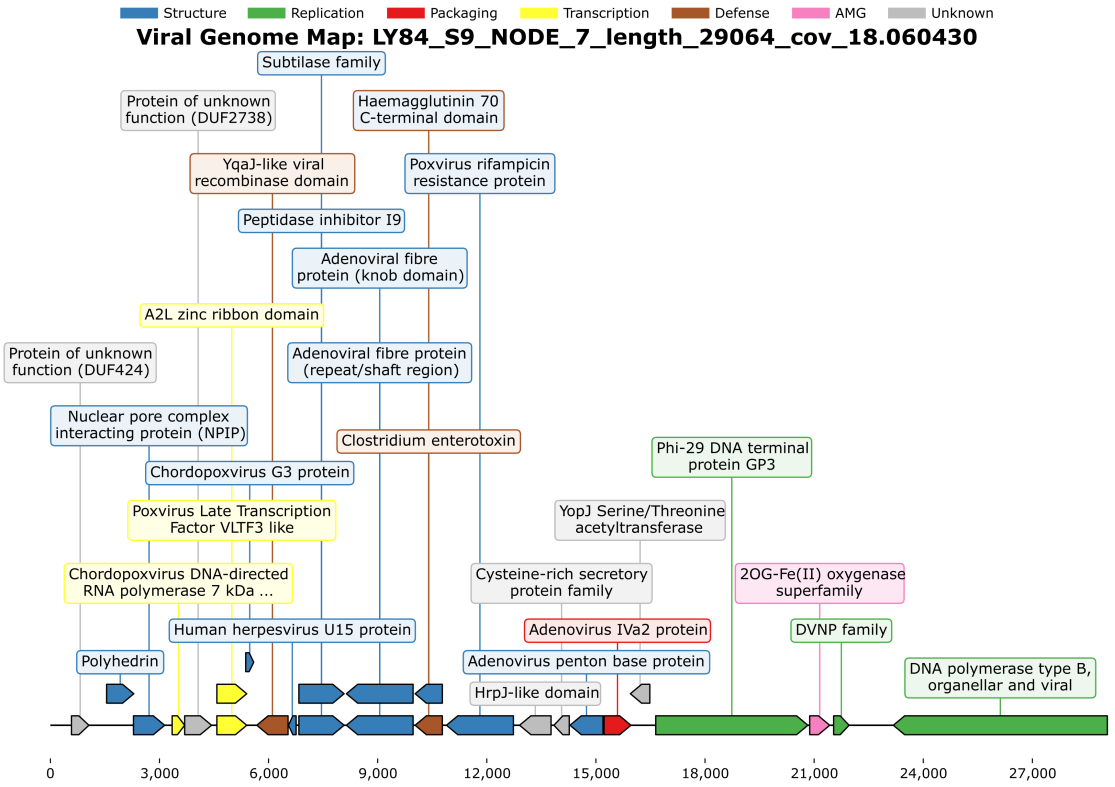
Supplementary Figure 3.12: Genome organisation of a crucivirus-like contig detected in Irish soil metagenomes.
Genome map of contig BL2_S3_NODE_98_length_4072_cov_14.248942 (4,072 bp), classified within the class *Arfiviricetes*. Predicted open reading frames include an RNA helicase, putative viral replication protein, transcription regulator, viral coat protein (S-domain), and a deltaretrovirus Tax-like protein. Functional categories are colour-coded as Replication (green), Structure (blue), and Transcription (yellow). The gene content and compact genome architecture are consistent with crucivirus-like viruses, which are commonly detected in environmental metagenomes.



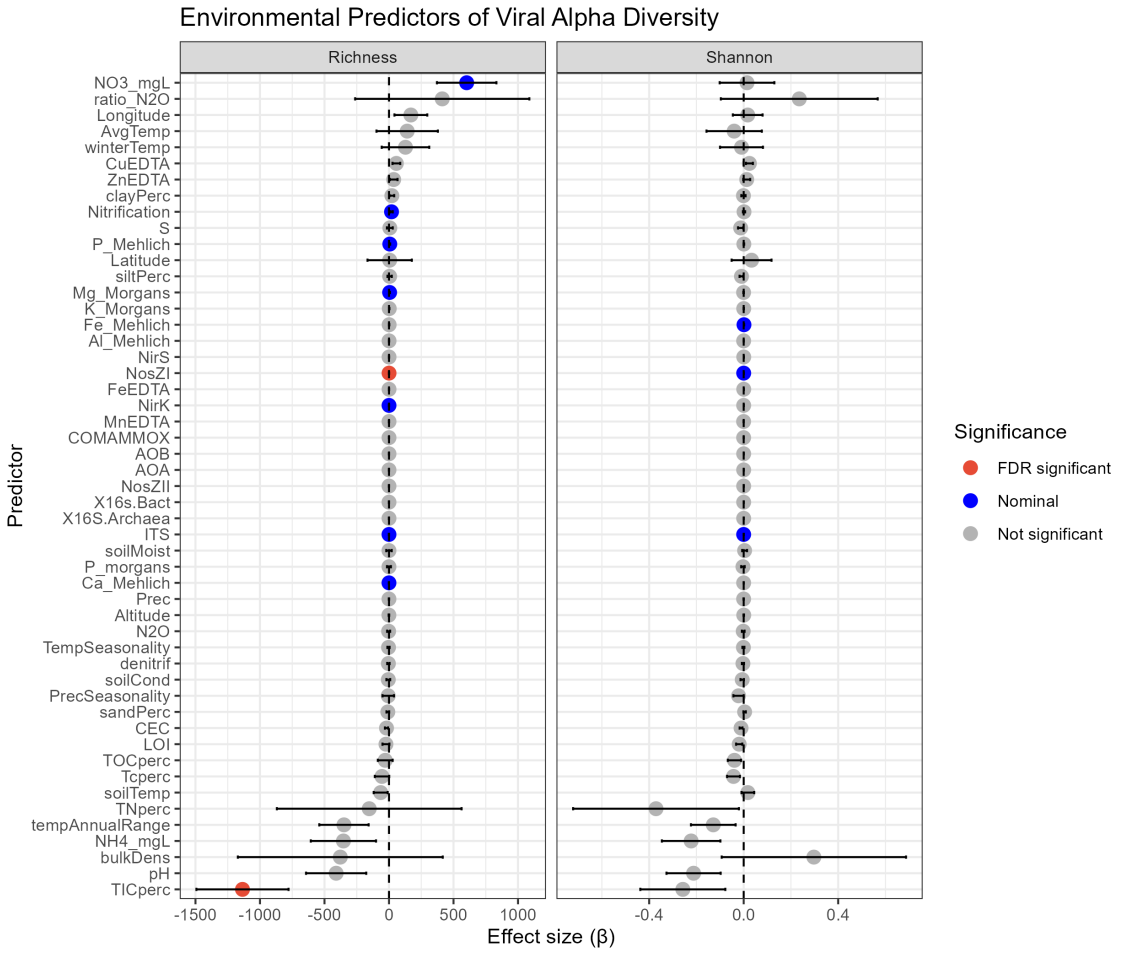
Supplementary Figure 3.13: Genome organisation of a *Tectiliviricetes* contig recovered from Irish soil metagenomes.
Genome map of contig RE-HI3_S18_k71_1864721_flag=1_multi=66.2661_len=12080 (12080 bp) assigned to the class *Tectiliviricetes*. Predicted open reading frames include several hallmark tectivirus proteins, including DNA polymerase type B, Adenovirus IVa2-like packaging ATPase, P3 major capsid protein, and a PRD1 phage membrane DNA delivery protein, which together represent core components of the replication, virion assembly, and genome delivery machinery characteristic of PRD1-like bacteriophages. Additional genes include transcription-associated proteins, auxiliary metabolic genes such as nitrogen fixation protein NifW, and several proteins of unknown function.



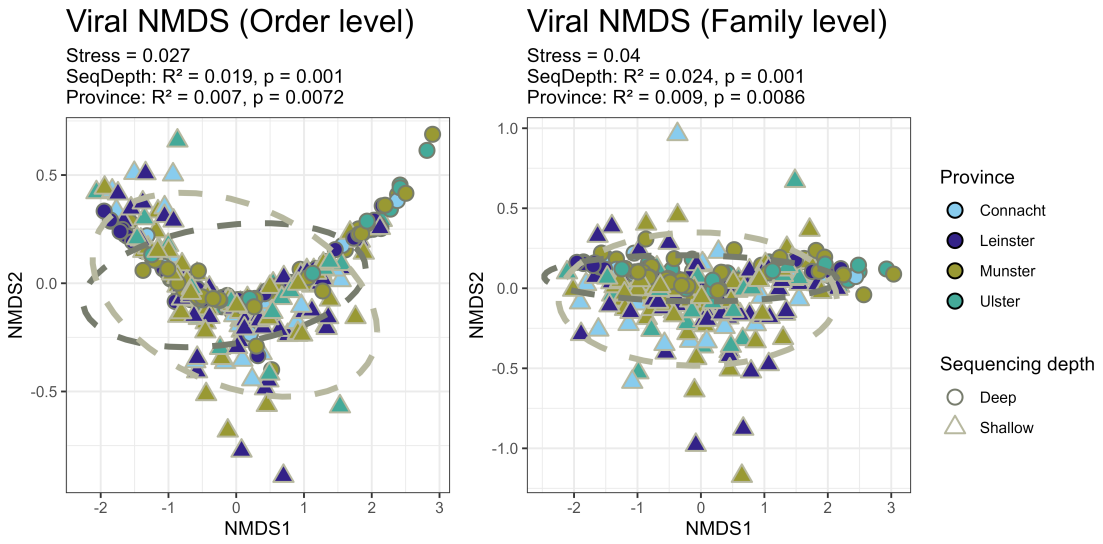
Supplementary Figure 3.14: Genome maps of *Repensiviricetes* (*Genomoviridae*) contigs recovered from Irish soil metagenomes. Predicted gene organisation of three circular CRESS-DNA virus contigs identified in the dataset: (a) CG2_S9_NODE_317_length_2106_cov_5.268162, (b) CG4_S11_NODE_2263_length_1482_cov_4.448493, and (c) AH3_S4_NODE_701_length_1614_cov_7.663887. The longest contig (a) contains canonical CRESS-DNA virus features including a replication-associated protein (Rep) with conserved catalytic and helicase domains and a geminivirus-like capsid protein, consistent with genomovirus genome architecture. The shorter contigs (b-c) contain partial replication or structural protein domains alongside hypothetical proteins of unknown function, reflecting the compact and variable gene content typical of small circular ssDNA viruses.



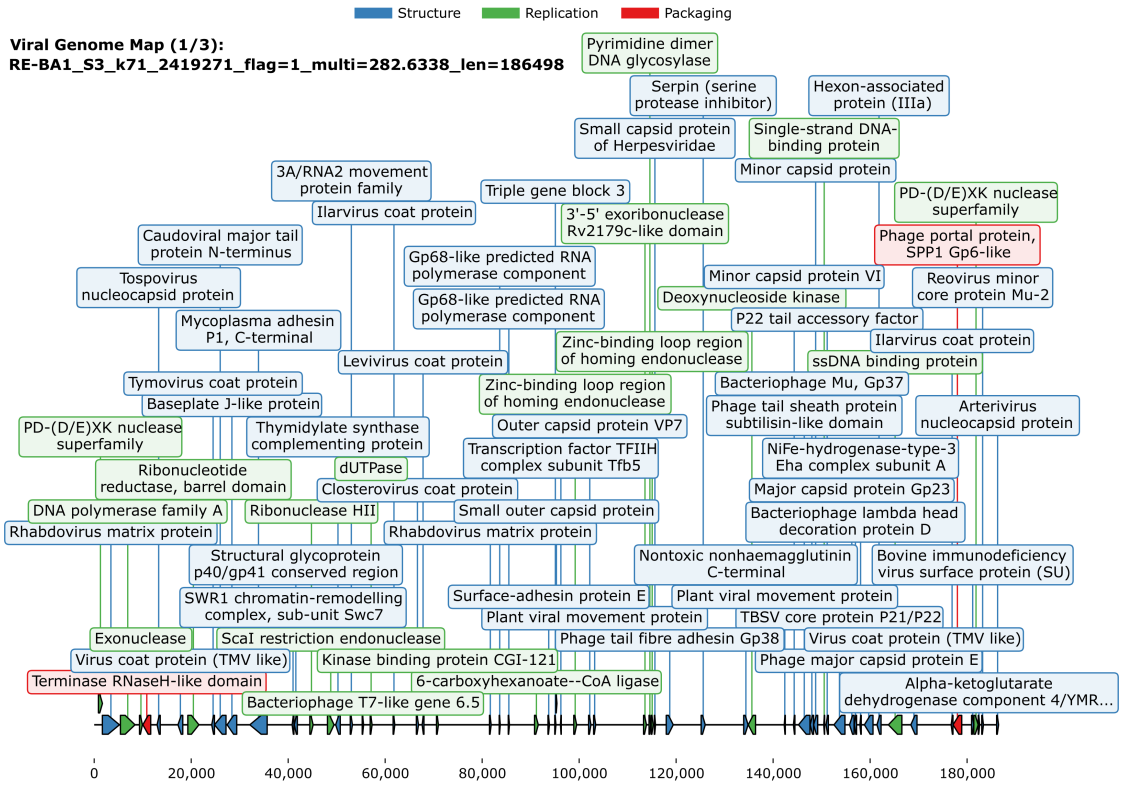
Supplementary Figure 3.15: Genome map of a putative *Megaviricetes* (NCLDV) contig recovered from Irish soil metagenomes.
Predicted gene organisation of contig LY84_S9_NODE_7_length_29064_cov_18.060430 (29,064 bp). Functional annotations were assigned based on Pfam domain predictions and grouped into structural, replication, transcription, packaging, defense, auxiliary metabolic, and unknown functional categories. The contig contains several hallmark genes associated with nucleocytoplasmic large DNA viruses (NCLDVs), including DNA polymerase type B, poxvirus late transcription factor VLTF3, and a DNA-directed RNA polymerase-associated protein, which together form part of the conserved poxvirus-like transcription machinery characteristic of NCLDV genomes. Additional virion-associated structural proteins were also detected, including adenovirus-like fibre and penton base proteins, consistent with capsid and host attachment functions observed in members of the realm Varidnaviria. Other predicted proteins include enzymatic and host-interaction domains such as subtilase family proteases, zinc-ribbon transcription-associated domains, and 2-oxoglutarate–Fe(II) oxygenases, alongside several proteins of unknown function. The presence of these conserved replication and structural genes supports the assignment of this contig to the class *Megaviricetes*.



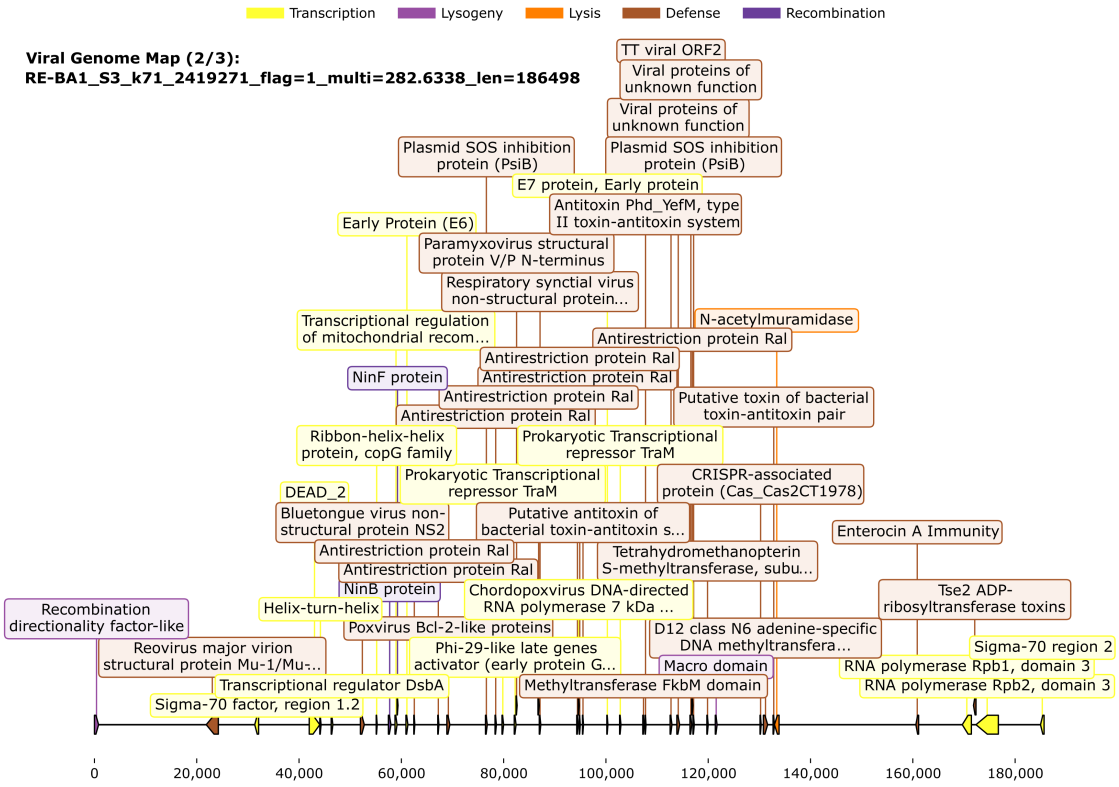
Supplementary Figure 3.16: Environmental predictors of viral alpha diversity.
Univariate linear models were used to test associations between environmental or microbial predictors and viral Shannon diversity and richness. p-values were adjusted using the Benjamini–Hochberg false discovery rate (FDR). Several predictors showed nominal associations with Shannon diversity prior to correction (e.g. Fe_Mehlich, ITS abundance, *nosZI* abundance), but none remained significant after FDR correction, indicating no robust environmental predictors of viral Shannon diversity. In contrast, viral richness showed significant associations with *nosZI* abundance (positive; $\beta = 0.36$, FDR $= 0.043$) and total inorganic carbon (TICperc; negative; $\beta = -1135.6$, FDR $= 0.046$). Additional predictors (e.g. NirK, NO3_mgL, Mg_Morgans) showed nominal significance prior to correction but were not significant after FDR adjustment.



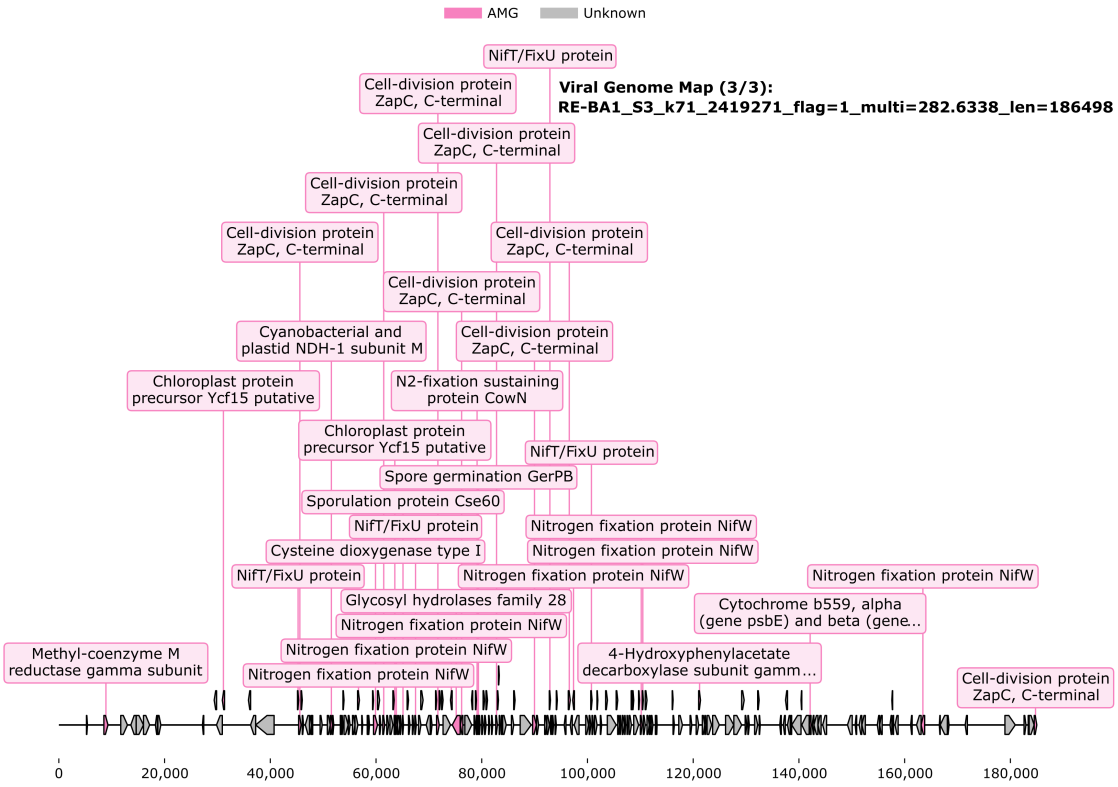
Supplementary Figure 3.17: Viral community composition across Irish provinces.
Non-metric multidimensional scaling (NMDS) ordination of viral communities at the Order (left) and Family (right) levels based on Bray–Curtis dissimilarities. Points are coloured by province and ellipsed by sequencing depth (deep vs. shallow). Viral communities showed substantial overlap across provinces at both taxonomic resolutions, indicating limited large-scale geographic structuring. PERMANOVA revealed that province explained only a small proportion of the variation (Order: $R^2 = 0.007$, $p = 0.0072$; Family: $R^2 = 0.009$, $p = 0.0086$), whereas sequencing depth accounted for a slightly larger but still modest fraction of variation (Order: $R^2 = 0.019$; Family: $R^2 = 0.024$).



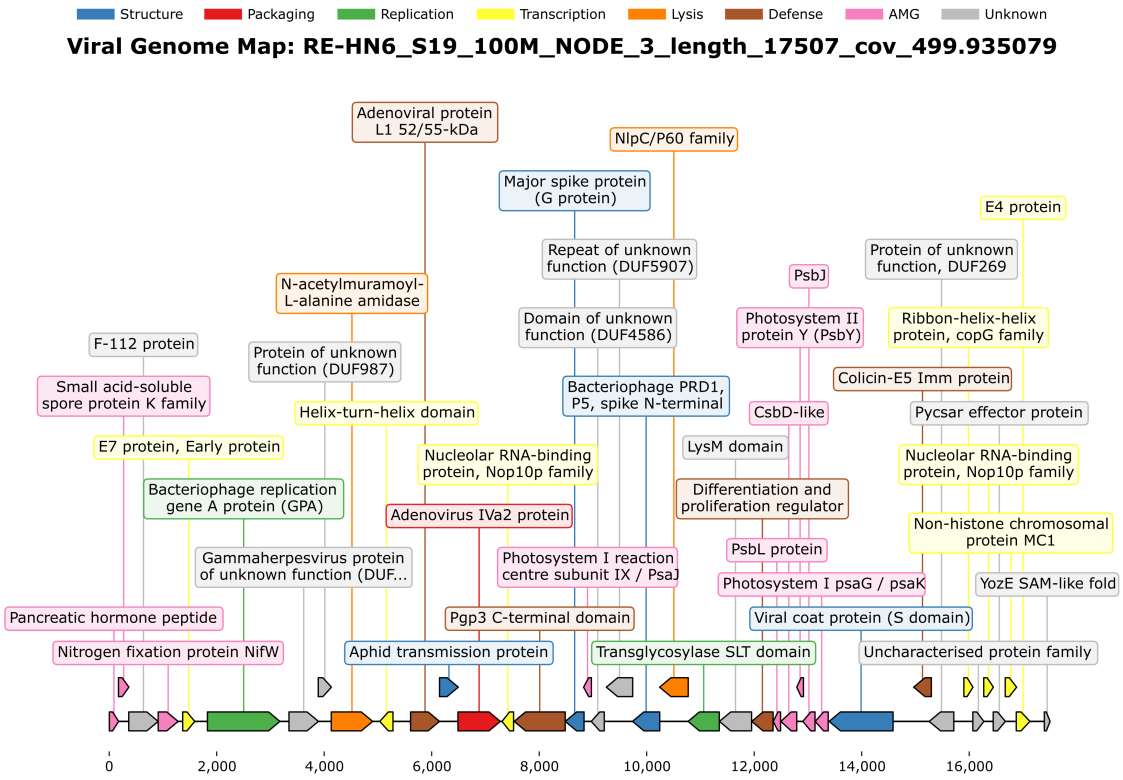
Supplementary Figure 3.18: Genome map (1/3) of the most abundant bacteriophage in Irish soil samples.
The most prevalent virus, RE-BA1_S3_k71_2419271 (186 kb circular contig), encodes a complete suite of hallmark *Caudoviricetes* genes, including the terminase large subunit, major capsid protein, and portal protein. In addition to core structural components, the genome contains numerous tail-associated and accessory proteins, as well as genes involved in replication, transcription, and recombination, reflecting a large, functionally diverse and modular *Caudoviricetes* genome.



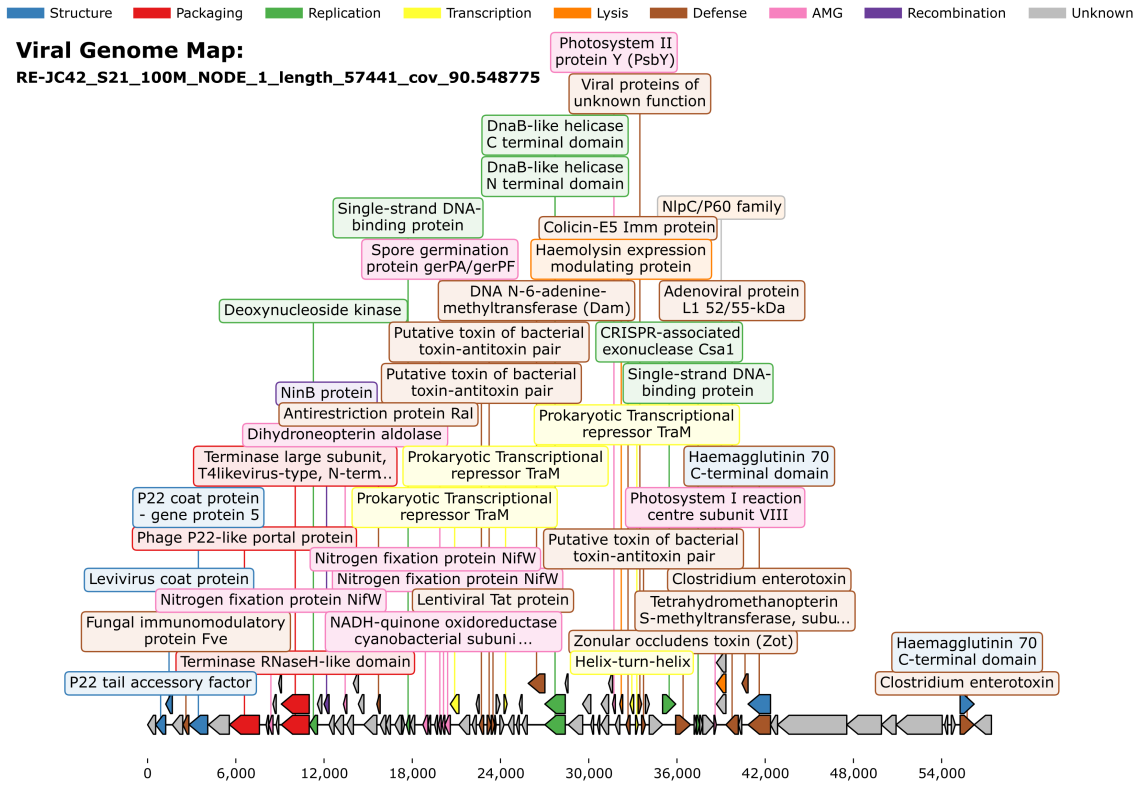
Supplementary Figure 3.19: Genome map (2/3) of the most abundant bacteriophage in Irish soil samples.
This region of RE-BA1_S3_k71_2419271 highlights genes associated with transcription, lysogeny, lysis, defense, and recombination. The genome encodes multiple transcriptional regulators (including sigma factors and RNA polymerase-associated domains), alongside recombination and integration-related proteins, indicating capacity for genome rearrangement and potential lysogenic lifestyle. A diverse array of host interaction and defense-associated genes is also present, including antirestriction proteins, toxin-antitoxin systems, and CRISPR-associated elements, suggesting mechanisms for host manipulation and evasion of bacterial immunity.



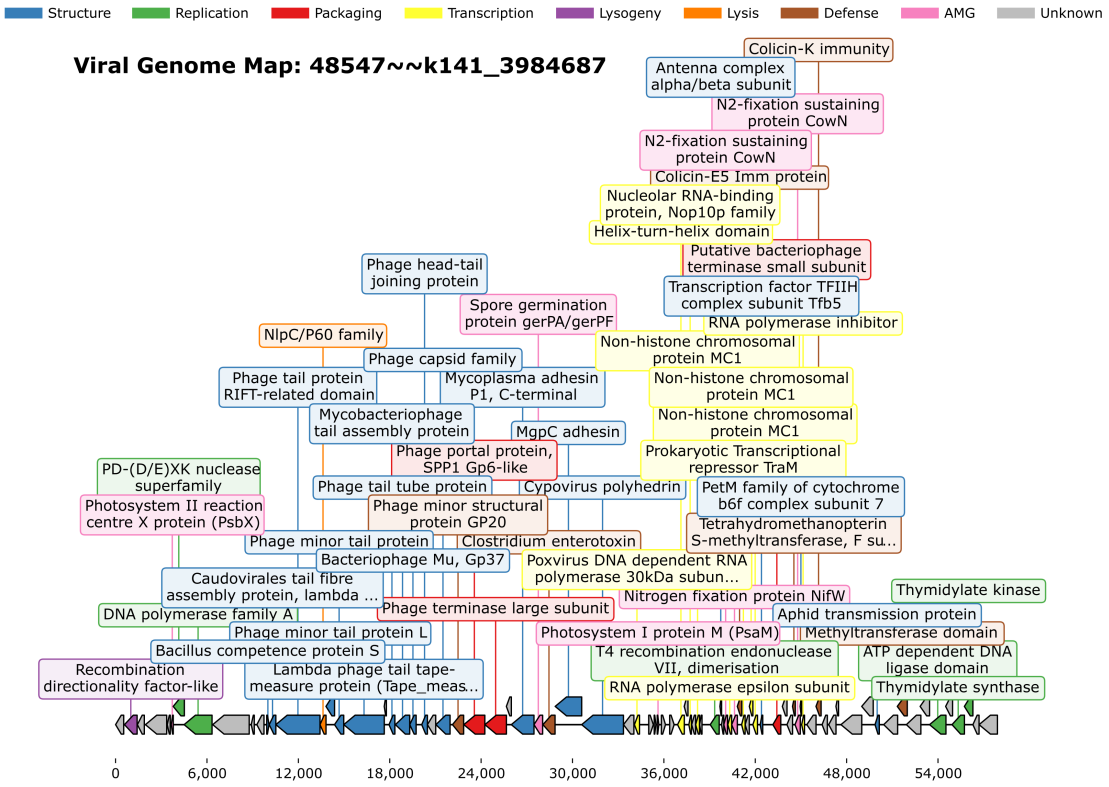
Supplementary Figure 3.20: Genome map (3/3) of the most abundant bacteriophage in Irish soil samples.
This region of RE-BA1_S3_k71_2419271 highlights auxiliary metabolic genes (AMGs) and a large proportion of uncharacterised proteins. Several genes associated with host metabolic processes are present, including those linked to nitrogen metabolism (e.g., Nif proteins), redox functions, and cellular processes, suggesting potential roles in modulating host physiology during infection. Notably, a substantial fraction of the genome encodes proteins of unknown function (unlabelled), underscoring the extensive functional novelty and unexplored genetic diversity within soil *Caudoviricetes*.



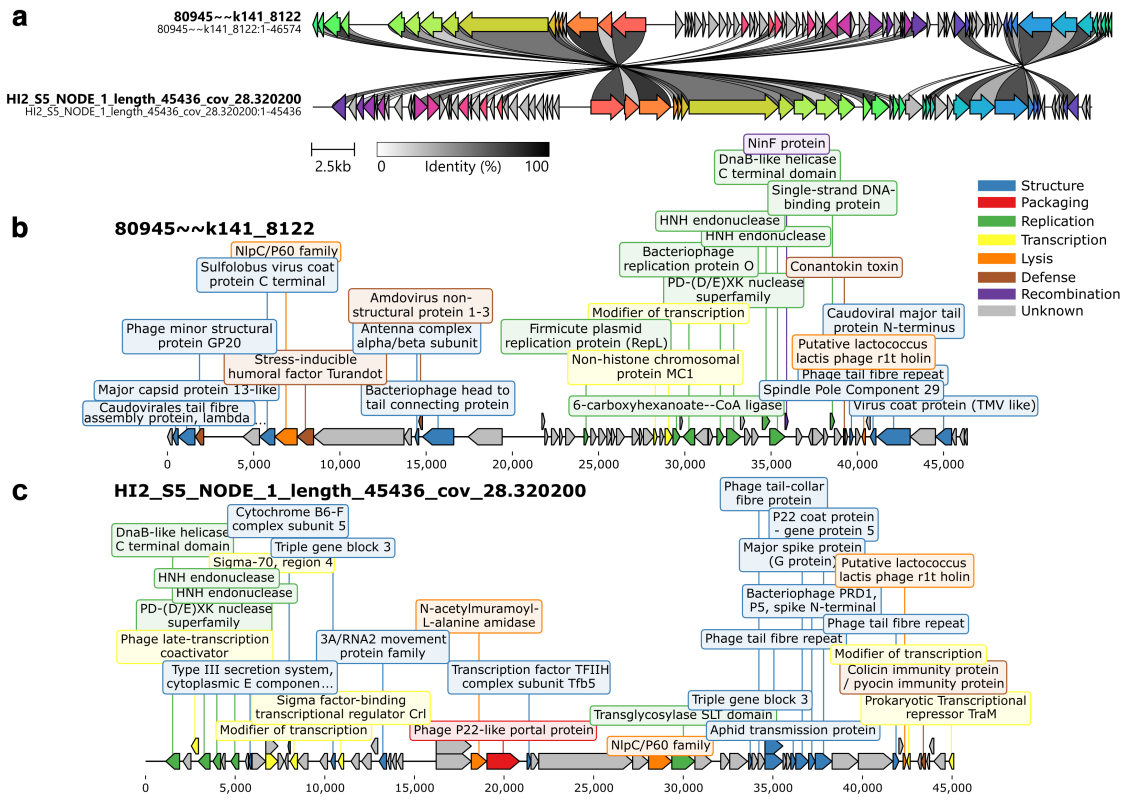
Supplementary Figure 3.21: Genome map of a highly abundant and prevalent bacteriophage in Irish soil samples.
The genome of RE-HN6_S19_100M_NODE_3_length_17507_cov_499.935079 (17.5 kb circular contig) exhibits a compact and streamlined architecture, encoding core structural, packaging, replication, and lysis-associated genes. Key features include replication gene A (GPA), major spike and coat proteins, and cell wall degradation enzymes (amidase/NlpC-P60 and transglycosylase domains), consistent with a lytic lifestyle. In addition to these core functions, the genome contains a limited set of auxiliary metabolic and host-associated genes, including nitrogen- and photosystem-related proteins, alongside regulatory and defense-associated domains.



Supplementary Figure 3.22: Genome map of a highly abundant and prevalent viral MAG in Irish soil samples.
The genome of RE-JC42_S21_100M_NODE_1_length_57441_cov_90.548775 (57.4 kb) exhibits a modular and mosaic architecture typical of *Caudoviricetes*, encoding core structural and packaging proteins (capsid, portal, terminase, tail-associated and spike proteins) alongside replication and recombination functions. The genome also contains an expanded repertoire of host-interaction and defense-related genes, including anti-restriction proteins, DNA methyltransferases, toxin–antitoxin systems, and regulatory domains. Several auxiliary metabolic genes are present, including nitrogen-associated and redox-related proteins, although some predicted functions likely reflect conserved domains rather than direct metabolic roles. A substantial proportion of genes remain uncharacterised, highlighting the functional novelty of soil viral genomes.



Supplementary Figure 3.23: Genome map of a VIRE-derived bacteriophage detected at high nucleotide coverage across multiple Irish soil samples.
The genome of contig 48547~~k141_3984687 (57,901 bp; 84 predicted genes), originally recovered from soil in Washington, USA (BioSample: SAMN10864404), is shown. This *Caudoviricetes* genome exhibited a median nucleotide coverage of 0.56% across the 14 Irish soil samples in which it was detected. Despite its widespread presence and high coverage, no nucleotide-level homology was observed between this genome and any curated Irish soil viral sequences.



Supplementary Figure 3.24: Partial genomic similarity between Irish soil- and human gut-associated bacteriophages reflects conservation of discrete functional modules.

A nucleotide comparison identified limited genomic similarity (∼13.8% genome coverage; ∼6.4 kb) between two similarly sized (∼45 kb) viral genomes originating from soil and human gut environments (80945˜˜k141_8122, China; BioProject SAMN13252589). The conserved regions are arranged in opposite orientations between the two genomes, indicating inversion of shared genomic modules. Two large genes of unknown function, located at approximately 10 kb and 22 kb, respectively, are conserved, as well as phage tail fibre repeats around the ∼42 kb region of the genome. The complete Irish contig (HI2_S5_NODE_1_length_45436_cov_28.320200) showed its closest alignment to *Streptomyces phage FlowerPower* (MH155868.1), with 3,367 amino acids aligned across the reference genome. In addition, a high-confidence CRISPR spacer match linked this contig to soil-associated ammonia-oxidising bacteria of the RAAP-2 family within the order *Acidimicrobiales*.

4. Conclusions

4.1 Overview

Bacteriophages are the most abundant biological entities in the biosphere, outnumbering their cellular hosts by at least an order of magnitude and exerting profound influence on microbial evolution, ecosystem function, and global biogeochemical cycles. This thesis begins with a review of *Crassvirales*, an order of dsDNA bacteriophages that represents a striking example of ecological dominance and global prevalence. Despite their extraordinary abundance in human gut microbiomes, prototypical crAssphage (p-crAssphage) remained undetected until 2013, largely due to limitations in available bioinformatic tools and reference databases. Its discovery through cross-assembly of metagenomic data catalysed a new wave of research into gut viral ecology, including investigations into phage–host interactions, population stability, and potential links to human health.

Within the broader landscape of global bacteriophage—and viral—diversity, *Crassvirales* constitute only a small clade. Placing such dominant lineages into a coherent evolutionary and taxonomic context remains inherently challenging. Viral taxonomy is complicated by the fundamental properties of viral genomes: extreme mosaicism, chimeric gene architectures, rapid evolutionary rates, pervasive horizontal gene transfer, and frequent homologous recombination. Together, these processes obscure phylogenetic signal and blur taxonomic boundaries between viral lineages. Historically, viral relatedness has been inferred using single marker genes such as the major capsid protein, terminase large subunit, or portal protein. While informative, these approaches are limited in resolution and poorly suited to resolving deep evolutionary relationships.

Recent advances in machine learning and structural biology have transformed this research space. Large language models such as ESMFold now enable high-throughput protein structure prediction for hundreds of thousands of proteins, while tools such as Foldseek convert atomic structural coordinates into linear alphabets that allow ultrafast structural comparison and clustering. As a result, analyses that once required computationally intensive all-versus-all protein sequence searches can now be performed rapidly at the highest biologically meaningful resolution—the level of protein structure.

We capitalised on this rapidly evolving bioinformatic landscape by integrating state-of-the-art structure prediction and comparison tools to construct a revised taxonomy of *Caudoviricetes*. Notably, the resulting structure-based classifications showed strong concordance with existing official ICTV taxonomy, with substantial agreement at multiple taxonomic ranks. From the outset, it was evident that higher-order classifications were urgently needed: only a small fraction of known *Caudoviricetes* genomes currently possess order- or family-level assignments, leaving much of their diversity taxonomically unresolved.

A striking outcome of this analysis was the identification of deeply conserved structural features across bacteriophages isolated from diverse environments world-

wide. Nearly all genomes shared a conserved virion assembly core, centred on the terminase large subunit and, to a lesser extent, the major capsid protein, underscoring the evolutionary stability of fundamental packaging and capsid architectures despite extensive genomic divergence elsewhere. At the same time, analysis of a global dataset of 4,082 *Caudoviricetes* genomes encompassing 451,151 predicted proteins revealed that the most prevalent protein folds correspond to domains of unknown function, including DUF1128 (PF06569) and DUF9867 (PF06174). The dominance of these uncharacterised folds highlights the extent of functional novelty within dsDNA bacteriophages and suggests the existence of yet-unknown mechanisms governing phage–host interactions. These findings emphasise the need for targeted experimental characterisation of structurally conserved but functionally enigmatic viral proteins.

This structure-based framework resulted in the delineation of 159 orders, 267 families, 502 subfamilies, and 1,189 genera within *Caudoviricetes*, highlighting the remarkable breadth of tailed phage diversity captured by whole-structurome comparison. Quantitative analysis of proteomic relatedness revealed that inter-order similarity is extremely low, with genomes from different orders sharing on average ∼5% of their structural protein clusters. Despite this deep divergence, the conserved virion assembly core is retained across lineages, with the terminase large subunit (TerL) functioning as a universal molecular motor from which diverse capsid and accessory structural proteins are assembled.

These results emphasise the extraordinary combinatorial diversity of phage proteomes: while the core packaging machinery remains conserved, the surrounding structural components vary extensively, effectively forming a modular repertoire of interchangeable "building blocks" that can be recombined to generate distinct virion architectures. This architectural flexibility provides a mechanistic basis for the vast structural and functional diversity observed across *Caudoviricetes* despite minimal shared proteomic content at higher taxonomic ranks.

Despite the inherent challenges of studying viruses, interest in uncovering new viral species, elucidating their ecological roles, and assembling large collections of viral sequences from diverse environments is accelerating rapidly. This momentum is exemplified by recent large-scale resources such as VIRE (∼1.7 million sequences) and the Global Soil Viral Atlas (GSVA; ∼50,000 sequences), released in 2025. Together, these efforts highlight both the scale of viral diversity being uncovered and the urgency of developing robust frameworks for organising it.

We consider this an especially exciting time for viral taxonomy. Without diminishing the substantial experimental, computational, and conceptual challenges involved in viral discovery and classification—nor the considerable efforts of the research communities undertaking this work—we believe that the taxonomy landscape will naturally shift to incorporate high resolution bioinformatic methods. Increasing access to high-performance computing infrastructure, GPUs, and advanced protein folding models is enabling analytical approaches that were previously infeasible. In this context, whole-structurome comparative strategies are poised to become a foundational component not only of viral taxonomy, but

also of comparative genomics more broadly, offering a scalable means of bridging taxonomic gaps between highly divergent viral genomes where sequence-based approaches alone often fail.

To demonstrate the broader applicability of these concepts, they were extended to the analysis of the Irish soil virome using high-quality shotgun metagenomic data. Despite the inherent challenges of soil viromics, including extreme community complexity and low viral signal-to-noise ratios, this approach recovered a small but quantifiable fraction of viral genomes. Analysis of these genomes revealed recurring and prevalent viral populations across diverse soils, the majority of which belonged to *Caudoviricetes*, consistent with their well-established global dominance, enabling inferences about phage architecture, genome size variation, and auxiliary metabolic gene content.

Notably, Irish soil phages exhibited substantial diversity in genome organisation and metabolic potential while maintaining consistent ecological prevalence and abundance across samples. The prominence of defence- and persistence-associated auxiliary metabolic genes suggests that soil phages may favour long-term coexistence and host survival over aggressive lytic replication, reflecting adaptation to spatially structured and resource-variable environments. Comparative analyses further revealed that these soil viruses are related to viral genomes identified in soils from Michigan, USA, as well as human gut viromes from China, indicating the presence of globally distributed viral lineages and suggesting that conserved genomic features may reflect functionally important and ecologically advantageous traits. Collectively, these findings reinforce a model of viral evolution in which genomes are not fixed entities but dynamic, mosaic assemblies of modular genetic elements that are exchanged, recombined, and reshaped across environments. The recurrence of related viral lineages across geographically and ecologically distinct systems further suggests that these modular components form part of a globally interconnected virosphere, rather than isolated, environment-specific populations.

Taken together, this thesis supports a unifying view of viral diversity in which deeply conserved structural features provide a stable backbone for viral classification, ecological inference, and evolutionary interpretation. Whole-structurome comparison offers a powerful means of resolving viral relationships across ecosystems, identifying the closest known relatives of novel viruses, and bridging the gap between environmental sequences and experimentally characterised phages. As such approaches mature, they hold promise not only for streamlining viral taxonomy, but for revealing the true connectedness of the global viral protein network and, ultimately, ultimately, uncovering viral dark matter and advancing our understanding of viral diversity.

4.2 Future perspectives

Viral taxonomy is entering a transformative phase. Despite the enormous diversity of the virosphere, only ∼18,000 viral species have been formally classified by the International Committee on Taxonomy of Viruses (ICTV; VMR_MLS40), highlighting the vast scale of uncharacterized viral diversity. Structural proteome-

based approaches, such as those implemented in PhagePleats, offer one of the highest-resolution frameworks currently available for viral classification, enabling comparisons that extend beyond sequence similarity into conserved structural features.

The performance and scope of the PhagePleats framework are expected to improve substantially as more curated viral genomes become available through the ICTV and associated study groups. Expanding structural coverage across underrepresented families, and incorporating newly discovered and classified taxa, will enhance both classification accuracy and predictive power. In future iterations, integrating alternative machine learning approaches—such as gradient boosting methods or neural networks—may further refine classification performance and enable more nuanced taxonomic predictions.

Beyond bacteriophages, this framework has the potential to scale across the entire virosphere, encompassing all viral realms and ultimately contributing to a unified, structure-informed system for viral classification. The same principles may also prove applicable to other mobile genetic elements, including plasmids and viroids. At present, widespread adoption is likely hindered by the computational demands of large-scale protein structure prediction, which requires significant GPU resources and specialised computational environments.

In parallel, the broader field of viral discovery continues to face fundamental challenges. Existing databases, including large environmental viral catalogues, likely contain substantial numbers of false-positive and misclassified sequences. Although tools such as CheckV are invaluable for assessing viral genome quality and completeness, they are inherently limited by reliance on known viral signatures and therefore cannot fully capture the extent of viral "dark matter". Likewise, state-of-the-art classifiers such as geNomad and VirSorter have transformed viral discovery from metagenomic data, yet they ultimately assign sequences to predefined categories of viral, chromosomal, or plasmid origin. A substantial fraction of environmental contigs remain unclassified or are assigned with low confidence, raising the possibility that they represent previously unrecognised viruses or entirely novel classes of mobile genetic elements. Systematically investigating this unexplored sequence space will be essential for achieving a more complete understanding of the virosphere.

The continued expansion of global viral datasets presents an opportunity to explore large-scale ecological and evolutionary patterns, particularly in understudied environments such as soils. Realizing this potential will require coordinated efforts to curate high-confidence viral sequence databases, predict and cluster their proteins into orthologous groups, and develop both sequence- and structure-based reference databases. Such efforts would enable deeper insights into viral biogeography, cross-environmental diversity, and the evolutionary relationships that shape the global virosphere. Ultimately, advancing toward a unified, high-resolution viral taxonomy will depend on the integration of improved data quality, scalable computational methods, and increasingly sophisticated predictive models. These efforts will reveal previously hidden dimensions of viral diversity and ecology.

Bibliography

[1] Dutilh, B. E.; Cassman, N.; McNair, K.; Sanchez, S. E.; Silva, G. G. Z.; Boling, L.; Barr, J. J.; Speth, D. R.; Seguritan, V.; Aziz, R. K., *et al.* A Highly Abundant Bacteriophage Discovered in the Unknown Sequences of Human Faecal Metagenomes. *Nature Communications* **2014**, *5*, 4498.

[2] Pratama, A. A.; van Elsas, J. D. The "Neglected" Soil Virome—Potential Role and Impact. *Trends in Microbiology* **2018**, *26*, 649–662.

[3] Coutinho, F. H.; Silveira, C. B.; Gregoracci, G. B.; Thompson, C. C.; Edwards, R. A.; Brussaard, C. P. D.; Dutilh, B. E.; Thompson, F. L. Marine Viruses Discovered via Metagenomics Shed Light on Viral Strategies throughout the Oceans. *Nature Communications* **2017**, *8*, 15955.

[4] Garmaeva, S.; Sinha, T.; Kurilshikov, A.; Fu, J.; Wijmenga, C.; Zhernakova, A. Studying the Gut Virome in the Metagenomic Era: Challenges and Perspectives. *BMC Biology* **2019**, *17*, 84.

[5] Sutton, T. D. S.; Hill, C. Gut Bacteriophage: Current Understanding and Challenges. *Frontiers in Endocrinology* **2019**, *10*, 784.

[6] Benler, S.; Yutin, N.; Antipov, D.; Rayko, M.; Shmakov, S.; Gussow, A. B.; Pevzner, P.; Koonin, E. V. Thousands of Previously Unknown Phages Discovered in Whole-Community Human Gut Metagenomes. *Microbiome* **2021**, *9*, 78.

[7] Yutin, N.; Benler, S.; Shmakov, S. A.; Wolf, Y. I.; Tolstoy, I.; Rayko, M.; Antipov, D.; Pevzner, P. A.; Koonin, E. V. Analysis of Metagenome-Assembled Viral Genomes from the Human Gut Reveals Diverse Putative CrAss-like Phages with Unique Genomic Features. *Nature Communications* **2021**, *12*, 1044.

[8] Camarillo-Guerrero, L. F.; Almeida, A.; Rangel-Pineros, G.; Finn, R. D.; Lawley, T. D. Massive Expansion of Human Gut Bacteriophage Diversity. *Cell* **2021**, *184*, 1098–1109.e9.

[9] Guerin, E.; Shkoporov, A. N.; Stockdale, S. R.; Clooney, A. G.; Ryan, F. J.; Sutton, T. D. S.; Draper, L. A.; Gonzalez-Tortuero, E.; Ross, R. P.; Hill, C. Biology and Taxonomy of CrAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. *Cell Host & Microbe* **2018**, *24*, 653–664.e6.

[10] Edwards, R. A.; Vega, A. A.; Norman, H. M.; Ohaeri, M.; Levi, K.; Dinsdale, E. A.; Cinek, O.; Aziz, R. K.; McNair, K.; Barr, J. J., *et al.* Global Phylogeography and Ancient Evolution of the Widespread Human Gut Virus CrAssphage. *Nature Microbiology* **2019**, *4*, 1727–1736.

[11] Honap, T. P.; Sankaranarayanan, K.; Schnorr, S. L.; Ozga, A. T.; Warinner, C.; Lewis, C. M. Biogeographic Study of Human Gut-Associated CrAssphage Suggests Impacts from Industrialization and Recent Expansion. *PLoS ONE* **2020**, *15*, e0226930.

[12] Shkoporov, A. N.; Clooney, A. G.; Sutton, T. D. S.; Ryan, F. J.; Daly, K. M.; Nolan, J. A.; McDonnell, S. A.; Khokhlova, E. V.; Draper, L. A.; Forde, A., *et al.* The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. *Cell Host & Microbe* **2019**, *26*, 527–541.e5.

[13] Gulyaeva, A.; Garmaeva, S.; Ruigrok, R. A. A. A.; Wang, D.; Riksen, N. P.; Netea, M. G.; Wijmenga, C.; Weersma, R. K.; Fu, J.; Vila, A. V., *et al.* Discovery, Diversity, and Functional Associations of CrAss-like Phages in Human Gut Metagenomes from Four Dutch Cohorts. *Cell Reports* **2022**, *38*, 110204.

[14] Cervantes-Echeverría, M.; Gallardo-Becerra, L.; Cornejo-Granados, F.; Leyva, A. O. A Loss of CrAssphage Stability in the Human Gut Virome Is Associated with Obesity and Metabolic Syndrome. *bioRxiv* **2022**, Preprint.

[15] Clooney, A. G.; Sutton, T. D. S.; Shkoporov, A. N.; Holohan, R. K.; Daly, K. M.; O'Regan, O.; Ryan, F. J.; Draper, L. A.; Plevy, S. E.; Ross, R. P., *et al.* Whole-Virome Analysis Sheds Light on Viral Dark Matter in Inflammatory Bowel Disease. *Cell Host & Microbe* **2019**, *26*, 764–778.e5.

[16] Zafar, H.; Saier, M. H. Gut *Bacteroides* Species in Health and Disease. *Gut Microbes* **2021**, *13*, 1848158.

[17] Shkoporov, A. N.; Khokhlova, E. V.; Fitzgerald, C. B.; Stockdale, S. R.; Draper, L. A.; Ross, R. P.; Hill, C. ΦCrAss001 Represents the Most Abundant Bacteriophage Family in the Human Gut and Infects *Bacteroides intestinalis*. *Nature Communications* **2018**, *9*, 4781.

[18] Hryckowian, A. J.; Merrill, B. D.; Porter, N. T.; Van Treuren, W.; Nelson, E. J.; Garlena, R. A.; Russell, D. A.; Martens, E. C.; Sonnenburg, J. L. *Bacteroides thetaiotaomicron*-Infecting Bacteriophage Isolates Inform Sequence-Based Host Range Predictions. *Cell Host & Microbe* **2020**, *28*, 371–379.e5.

[19] Guerin, E.; Shkoporov, A. N.; Stockdale, S. R.; Comas, J. C.; Khokhlova, E. V.; Clooney, A. G.; Daly, K. M.; Draper, L. A.; Stephens, N.; Scholz, D., *et al.* Isolation and Characterisation of ΦcrAss002, a CrAss-like Phage from the Human Gut That Infects *Bacteroides xylanisolvens*. *Microbiome* **2021**, *9*, 89.

[20] Hedžet, S.; Rupnik, M.; Accetto, T. Broad Host Range May Be a Key to Long-Term Persistence of Bacteriophages Infecting Intestinal *Bacteroidaceae* Species. *Scientific Reports* **2022**, *12*, 21098.

[21] Shkoporov, A. N.; Khokhlova, E. V.; Stephens, N.; Hueston, C.; Seymour, S.; Hryckowian, A. J.; Scholz, D.; Ross, R. P.; Hill, C. Long-Term Persistence of CrAss-like Phage CrAss001 Is Associated with Phase Variation in *Bacteroides intestinalis*. *BMC Biology* **2021**, *19*, 163.

[22] Fitzgerald, C. B.; Shkoporov, A. N.; Upadrasta, A.; Khokhlova, E. V.; Ross, R. P.; Hill, C. Probing the "Dark Matter" of the Human Gut Phageome: Culture Assisted Metagenomics Enables Rapid Discovery and Host-Linking for

Novel Bacteriophages. *Frontiers in Cellular and Infection Microbiology* **2021**, *11*, 616918.

[23] Hsu, B. B.; Gibson, T. E.; Yeliseyev, V.; Liu, Q.; Bry, L.; Silver, P. A.; Gerber, G. K. Bacteriophages Dynamically Modulate the Gut Microbiota and Metabolome. *bioRxiv* **2018**, Preprint.

[24] Campbell, D. E.; Ly, L. K.; Ridlon, J. M.; Hsiao, A.; Whitaker, R. J.; Degnan, P. H. Infection with *Bacteroides* Phage BV01 Alters the Host Transcriptome and Bile Acid Metabolism in a Common Human Gut Microbe. *Cell Reports* **2020**, *32*, 108142.

[25] Balcazar, J. L. Bacteriophages as Vehicles for Antibiotic Resistance Genes in the Environment. *PLoS Pathogens* **2014**, *10*, e1004219.

[26] Penadés, J. R.; Chen, J.; Quiles-Puchalt, N.; Carpena, N.; Novick, R. P. Bacteriophage-Mediated Spread of Bacterial Virulence Genes. *Current Opinion in Microbiology* **2015**, *23*, 171–178.

[27] Koonin, E. V.; Yutin, N. The CrAss-like Phage Group: How Metagenomics Reshaped the Human Virome. *Trends in Microbiology* **2020**, *28*, 349–359.

[28] Reyes, A.; Wu, M.; McNulty, N. P.; Rohwer, F. L.; Gordon, J. I. Gnotobiotic Mouse Model of Phage–Bacterial Host Dynamics in the Human Gut. *Proceedings of the National Academy of Sciences of the United States of America* **2013**, *110*, 20236–20241.

[29] Shkoporov, A. N.; Turkington, C. J.; Hill, C. Mutualistic Interplay between Bacteriophages and Bacteria in the Human Gut. *Nature Reviews Microbiology* **2022**, *20*, 737–749.

[30] Holmfeldt, K.; Solonenko, N.; Shah, M.; Corrier, K.; Riemann, L.; Verberkmoes, N. C.; Sullivan, M. B. Twelve Previously Unknown Phage Genera Are Ubiquitous in Global Oceans. *Proceedings of the National Academy of Sciences of the United States of America* **2013**, *110*, 12798–12803.

[31] Rosario, K.; Breitbart, M. Exploring the Viral World through Metagenomics. *Current Opinion in Virology* **2011**, *1*, 289–297.

[32] Koonin, E. V.; Krupovic, M.; Dolja, V. V. The Global Virome: How Much Diversity and How Many Independent Origins? *Environmental Microbiology* **2022**, *25*, 40–44.

[33] Shkoporov, A. N.; Stockdale, S. R.; Lavelle, A.; Kondova, I.; Heuston, C.; Upadrasta, A.; Khokhlova, E. V.; van der Kamp, I.; Ouwerling, B.; Draper, L. A., *et al.* Viral Biogeography of the Mammalian Gut and Parenchymal Organs. *Nature Microbiology* **2022**, *7*, 1301–1311.

[34] López-Bueno, A.; Tamames, J.; Velázquez, D.; Moya, A.; Quesada, A.; Alcamí, A. High Diversity of the Viral Community from an Antarctic Lake. *Science* **2009**, *326*, 858–861.

[35] Edwards, R. A.; McNair, K.; Faust, K.; Raes, J.; Dutilh, B. E. Computational Approaches to Predict Bacteriophage–Host Relationships. *FEMS Microbiology Reviews* **2016**, *40*, 258–272.

[36] Gupta, V. K.; Paul, S.; Dutta, C. Geography, Ethnicity or Subsistence-Specific Variations in Human Microbiome Composition and Diversity. *Frontiers in Microbiology* **2017**, *8*, 1162.

[37] Draper, L. A.; Ryan, F. J.; Smith, M. K.; Jalanka, J.; Mattila, E.; Arkkila, P. A.; Ross, R. P.; Satokari, R.; Hill, C. Long-Term Colonisation with Donor Bacteriophages Following Successful Faecal Microbial Transplantation. *Microbiome* **2018**, *6*, 220.

[38] Yutin, N.; Makarova, K. S.; Gussow, A. B.; Krupovic, M.; Segall, A.; Edwards, R. A.; Koonin, E. V. Discovery of an Expansive Bacteriophage Family That Includes the Most Abundant Viruses from the Human Gut. *Nature Microbiology* **2018**, *3*, 38–46.

[39] Li, Y.; Gordon, E.; Shean, R. C.; Idle, A.; Deng, X.; Greninger, A. L.; Delwart, E. CrAssphage and Its Bacterial Host in Cat Feces. *Scientific Reports* **2021**, *11*, 815.

[40] Scanlan, P. D. Bacteria–Bacteriophage Coevolution in the Human Gut: Implications for Microbial Diversity and Functionality. *Trends in Microbiology* **2017**, *25*, 614–623.

[41] Rossi, A.; Treu, L.; Toppo, S.; Zschach, H.; Campanaro, S.; Dutilh, B. E. Evolutionary Study of the Crassphage Virus at Gene Level. *Viruses* **2020**, *12*, 1035.

[42] Morozova, V.; Fofanov, M.; Tikunova, N.; Babkin, I.; Morozov, V. V.; Tikunov, A. First CrAss-Like Phage Genome Encoding the Diversity-Generating Retroelement (DGR). *Viruses* **2020**, *12*, 573.

[43] Guo, H.; Arambula, D.; Ghosh, P.; Miller, J. F. Diversity-Generating Retroelements in Phage and Bacterial Genomes. *Microbiology Spectrum* **2014**, *2*, 1237–1252.

[44] Brown, B. P.; Chopera, D.; Havyarimana, E.; Wendoh, J.; Jaumdally, S.; Nyangahu, D. D.; Gray, C. M.; Martin, D. P.; Varsani, A.; Jaspan, H. B. CrAssphage Genomes Identified in Fecal Samples of an Adult and Infants with Evidence of Positive Genomic Selective Pressure within Tail Protein Genes. *Virus Research* **2021**, *292*, 198219.

[45] de Jonge, P. A.; von Meijenfeldt, F. A. B.; van Rooijen, L. E.; Brouns, S. J. J.; Dutilh, B. E. Evolution of BACON Domain Tandem Repeats in CrAssphage and Novel Gut Bacteriophage Lineages. *Viruses* **2019**, *11*, 1085.

[46] Shkoporov, A.; Stockdale, S. R.; Adriaenssens, E. M.; Yutin, N.; Koonin, E. V.; Dutilh, B. E.; Krupovic, M.; Edwards, R. A.; Tolstoy, I.; Hill, C. Taxonomic Proposal: Create One New Order (Crassvirales) Including Four New Families, Ten New Subfamilies, 42 New Genera and 73 New

Species (Caudoviricetes). 2020; ICTV Taxonomic Proposal. Available at https://ictv.global/filebrowser/download/5740.

[47] Roux, S.; Adriaenssens, E. M.; Dutilh, B. E.; Koonin, E. V.; Kropinski, A. M.; Krupovic, M.; Kuhn, J. H.; Lavigne, R.; Brister, J. R.; Varsani, A., *et al.* Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nature Biotechnology* **2019**, *37*, 29–37.

[48] Borges, A. L.; Lou, Y. C.; Sachdeva, R.; Al-Shayeb, B.; Penev, P. I.; Jaffe, A. L.; Lei, S.; Santini, J. M.; Banfield, J. F. Widespread Stop-Codon Recoding in Bacteriophages May Regulate Translation of Lytic Genes. *Nature Microbiology* **2022**, *7*, 918–927.

[49] Edgell, D. R.; Chalamcharla, V. R.; Belfort, M. Learning to Live Together: Mutualism between Self-Splicing Introns and Their Hosts. *BMC Biology* **2011**, *9*, 22.

[50] Gogarten, J. P.; Senejani, A. G.; Zhaxybayeva, O.; Olendzenski, L.; Hilario, E. Inteins: Structure, Function, and Evolution. *Annual Review of Microbiology* **2002**, *56*, 263–287.

[51] Frost, L. S.; Leplae, R.; Summers, A. O.; Toussaint, A. Mobile Genetic Elements: The Agents of Open Source Evolution. *Nature Reviews Microbiology* **2005**, *3*, 722–732.

[52] Peters, S. L.; Borges, A. L.; Giannone, R. J.; Morowitz, M. J.; Banfield, J. F.; Hettich, R. L. Experimental Validation That Human Microbiome Phages Use Alternative Genetic Coding. *Nature Communications* **2022**, *13*, 5710.

[53] Crisci, M. A.; Chen, L.-X.; Devoto, A. E.; Borges, A. L.; Bordin, N.; Sachdeva, R.; Tett, A.; Sharrar, A. M.; Segata, N.; Debenedetti, F., *et al.* Closely Related Lak Megaphages Replicate in the Microbiomes of Diverse Animals. *iScience* **2021**, *24*, 102875.

[54] Stachler, E.; Kelty, C.; Sivaganesan, M.; Li, X.; Bibby, K.; Shanks, O. C. Quantitative CrAssphage PCR Assays for Human Fecal Pollution Measurement. *Environmental Science & Technology* **2017**, *51*, 9146–9154.

[55] Park, G. W.; Ng, T. F. F.; Freeland, A. L.; Marconi, V. C.; Boom, J. A.; Staat, M. A.; Montmayeur, A. M.; Browne, H.; Narayanan, J.; Payne, D. C., *et al.* CrAssphage as a Novel Tool to Detect Human Fecal Contamination on Environmental Surfaces and Hands. *Emerging Infectious Diseases* **2020**, *26*, 1731–1740.

[56] García-Aljaro, C.; Ballesté, E.; Muniesa, M.; Jofre, J. Determination of CrAssphage in Water Samples and Applicability for Tracking Human Faecal Pollution. *Microbial Biotechnology* **2017**, *10*, 1775–1780.

[57] Reyes, A.; Haynes, M.; Hanson, N.; Angly, F. E.; Heath, A. C.; Rohwer, F. Viruses in the Faecal Microbiota of Monozygotic Twins and Their Mothers. *Nature* **2010**, *466*, 334–338.

[58] Fitzgerald, C. B.; Shkoporov, A. N.; Sutton, T. D. S.; Chaplin, A. V.; Velayudhan, V.; Ross, R. P.; Hill, C. Comparative Analysis of *Faecalibacterium prausnitzii* Genomes Shows a High Level of Genome Plasticity and Warrants Separation into New Species-Level Taxa. *BMC Genomics* **2018**, *19*, 931.

[59] Cornuault, J. K.; Moncaut, E.; Loux, V.; Mathieu, A.; Sokol, H.; Petit, M.-A.; De Paepe, M. The Enemy from Within: A Prophage of *Roseburia intestinalis* Systematically Turns Lytic in the Mouse Gut, Driving Bacterial Adaptation by CRISPR Spacer Acquisition. *The ISME Journal* **2020**, *14*, 771–787.

[60] Zheng, W.; Zhao, S.; Yin, Y.; Zhang, H.; Needham, D. M.; Evans, E. D.; Dai, C. L.; Lu, P. J.; Alm, E. J.; Weitz, D. A. High-Throughput, Single-Microbe Genomics with Strain Resolution, Applied to a Human Gut Microbiome. *Science* **2022**, *376*, eabm1483.

[61] Asnicar, F.; Manara, S.; Zolfo, M.; Truong, D. T.; Scholz, M.; Armanini, F.; Ferretti, P.; Gorfer, V.; Pedrotti, A.; Tett, A., *et al.* Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling. *mSystems* **2017**, *2*, e00164–16.

[62] Sugimoto, R.; Nishimura, L.; Nguyen, P. T.; Ito, J.; Parrish, N. F.; Mori, H.; Kurokawa, K.; Nakaoka, H.; Inoue, I. Comprehensive Discovery of CRISPR-Targeted Terminally Redundant Sequences in the Human Gut Metagenome: Viruses, Plasmids, and More. *PLoS Computational Biology* **2021**, *17*, e1009428.

[63] Tomofuji, Y.; Kishikawa, T.; Maeda, Y.; Ogawa, K.; Nii, T.; Okuno, T.; Oguro-Igashira, E.; Kinoshita, M.; Yamamoto, K.; Sonehara, K., *et al.* Whole Gut Virome Analysis of 476 Japanese Revealed a Link between Phage and Autoimmune Disease. *Annals of the Rheumatic Diseases* **2022**, *81*, 278–288.

[64] Cinek, O.; Kramna, L.; Lin, J.; Oikarinen, S.; Kolarova, K.; Ilonen, J.; Simell, O.; Veijola, R.; Autio, R.; Hyöty, H. Imbalance of Bacteriome Profiles within the Finnish Diabetes Prediction and Prevention Study: Parallel Use of 16S Profiling and Virome Sequencing in Stool Samples from Children with Islet Autoimmunity and Matched Controls. *Pediatric Diabetes* **2017**, *18*, 588–598.

[65] Coclet, C.; Roux, S. Global Overview and Major Challenges of Host Prediction Methods for Uncultivated Phages. *Current Opinion in Virology* **2021**, *49*, 117–126.

[66] Bickhart, D. M.; Watson, M.; Koren, S.; Panke-Buisse, K.; Cersosimo, L. M.; Press, M. O.; Van Tassell, C. P.; Van Kessel, J. A. S.; Haley, B. J.; Kim, S. W., *et al.* Assignment of Virus and Antimicrobial Resistance Genes to Microbial Hosts in a Complex Microbial Community by Combined Long-Read Assembly and Proximity Ligation. *Genome Biology* **2019**, *20*, 153.

[67] Marbouty, M.; Cournac, A.; Flot, J.-F.; Marie-Nelly, H.; Mozziconacci, J.; Koszul, R. Metagenomic Chromosome Conformation Capture (Meta3C) Unveils the Diversity of Chromosome Organization in Microorganisms. *eLife* **2014**, *3*, e03318.

[68] Marbouty, M.; Baudry, L.; Cournac, A.; Koszul, R. Scaffolding Bacterial Genomes and Probing Host–Virus Interactions in Gut Microbiome by Proximity Ligation (Chromosome Capture) Assay. *Science Advances* **2017**, *3*, e1602105.

[69] Yaffe, E.; Relman, D. A. Tracking Microbial Evolution in the Human Gut Using Hi-C Reveals Extensive Horizontal Gene Transfer, Persistence and Adaptation. *Nature Microbiology* **2020**, *5*, 343–353.

[70] Labonté, J. M.; Swan, B. K.; Poulos, B.; Luo, H.; Koren, S.; Hallam, S. J.; Sullivan, M. B.; Woyke, T.; Wommack, K. E.; Stepanauskas, R. Single-Cell Genomics-Based Analysis of Virus–Host Interactions in Marine Surface Bacterioplankton. *The ISME Journal* **2015**, *9*, 2386–2399.

[71] Swan, B. K.; Tupper, B.; Sczyrba, A.; Lauro, F. M.; Martinez-Garcia, M.; González, J. M.; Luo, H.; Wright, J. J.; Landry, Z. C.; Hanson, N. W., *et al.* Prevalent Genome Streamlining and Latitudinal Divergence of Planktonic Bacteria in the Surface Ocean. *Proceedings of the National Academy of Sciences of the United States of America* **2013**, *110*, 11463–11468.

[72] Munson-McGee, J. H.; Peng, S.; Dewerff, S.; Stepanauskas, R.; Whitaker, R. J.; Weitz, J. S.; Young, M. J. A Virus or More in (Nearly) Every Cell: Ubiquitous Networks of Virus–Host Interactions in Extreme Environments. *The ISME Journal* **2018**, *12*, 1706–1714.

[73] Martínez-Hernández, F.; Fornas, O.; Lluesma Gomez, M.; Bolduc, B.; de la Cruz Peña, M. J.; Martínez, J. M.; Anton, J.; Gasol, J. M.; Rosselli, R.; Rodriguez-Valera, F., *et al.* Single-Virus Genomics Reveals Hidden Cosmopolitan and Abundant Viruses. *Nature Communications* **2017**, *8*, 15892.

[74] Allen, L. Z.; Ishoey, T.; Novotny, M. A.; McLean, J. S.; Lasken, R. S.; Williamson, S. J. Single Virus Genomics: A New Tool for Virus Discovery. *PLoS ONE* **2011**, *6*, e17722.

[75] de Jonge, P. A.; von Meijenfeldt, F. A. B.; Costa, A. R.; Nobrega, F. L.; Brouns, S. J. J.; Dutilh, B. E. Adsorption Sequencing as a Rapid Method to Link Environmental Bacteriophages to Hosts. *iScience* **2020**, *23*, 101439.

[76] Deng, L.; Ignacio-Espinoza, J. C.; Gregory, A. C.; Poulos, B. T.; Weitz, J. S.; Hugenholtz, P.; Sullivan, M. B. Viral Tagging Reveals Discrete Populations in *Synechococcus* Viral Genome Sequence Space. *Nature* **2014**, *513*, 242–245.

[77] Džunková, M.; Low, S. J.; Daly, J. N.; Deng, L.; Rinke, C.; Hugenholtz, P. Defining the Human Gut Host–Phage Network through Single-Cell Viral Tagging. *Nature Microbiology* **2019**, *4*, 2192–2203.

[78] Hindson, B. J.; Ness, K. D.; Masquelier, D. A.; Belgrader, P.; Heredia, N. J.; Makarewicz, A. J.; Bright, I. J.; Lucero, M. Y.; Hiddessen, A. L.; Legler, T. C., *et al.* High-Throughput Droplet Digital PCR System for Absolute Quantitation of DNA Copy Number. *Analytical Chemistry* **2011**, *83*, 8604–8610.

[79] Morella, N. M.; Yang, S. C.; Hernandez, C. A.; Koskella, B. Rapid Quantification of Bacteriophages and Their Bacterial Hosts *in Vitro* and *in Vivo* Using Droplet Digital PCR. *Journal of Virological Methods* **2018**, *259*, 18–24.

[80] Tadmor, A. D.; Ottesen, E. A.; Leadbetter, J. R.; Phillips, R. Probing Individual Environmental Bacteria for Viruses by Using Microfluidic Digital PCR. *Science* **2011**, *333*, 58–62.

[81] Sakowski, E. G.; Arora-Williams, K.; Tian, F.; Zayed, A. A.; Zablocki, O.; Sullivan, M. B.; Preheim, S. P. Interaction Dynamics and Virus–Host Range for Estuarine Actinophages Captured by EpicPCR. *Nature Microbiology* **2021**, *6*, 630–642.

[82] Spencer, S. J.; Tamminen, M. V.; Preheim, S. P.; Guo, M. T.; Briggs, A. W.; Brito, I. L.; Weitz, D. A.; Pitkänen, L. K.; Vigneault, F.; Virta, M. P. J., *et al.* Massively Parallel Sequencing of Single Cells by EpicPCR Links Functional Genes with Phylogenetic Markers. *The ISME Journal* **2016**, *10*, 427–436.

[83] Bayfield, O.; Shkoporov, A.; Yutin, N.; Khokhlova, E.; Smith, J.; Hawkins, D.; Koonin, E.; Hill, C.; Antson, A. Structural Atlas of the Most Abundant Human Gut Virus. 2022; Research Square Preprint.

[84] Barr, J. J.; Auro, R.; Furlan, M.; Whiteson, K. L.; Erb, M. L.; Pogliano, J.; Stotland, A.; Wolkowicz, R.; Cutting, A. S.; Doran, K. S., *et al.* Bacteriophage Adhering to Mucus Provide a Non–Host-Derived Immunity. *Proceedings of the National Academy of Sciences of the United States of America* **2013**, *110*, 10771–10776.

[85] Knecht, L. E.; Veljkovic, M.; Fieseler, L. Diversity and Function of Phage Encoded Depolymerases. *Frontiers in Microbiology* **2020**, *10*, 2949.

[86] Siranosian, B. A.; Tamburini, F. B.; Sherlock, G.; Bhatt, A. S. Acquisition, Transmission and Strain Diversity of Human Gut-Colonizing CrAss-like Phages. *Nature Communications* **2020**, *11*, 280.

[87] Wittmann, J.; Turner, D.; Millard, A. D.; Mahadevan, P.; Kropinski, A. M.; Adriaenssens, E. M. From Orphan Phage to a Proposed New Family–The Diversity of N4-Like Viruses. *Antibiotics* **2020**, *9*, 663.

[88] Choi, K.; McPartland, J.; Kaganman, I.; Bowman, V. D.; Rothman-Denes, L.; Rossmann, M. G. Insight into DNA and Protein Transport in Double-Stranded DNA Viruses: The Structure of Bacteriophage N4. *Microscopy and Microanalysis* **2008**, *14*, 1574–1575.

[89] Faith, J. J.; Guruge, J. L.; Charbonneau, M.; Subramanian, S.; Seedorf, H.; Goodman, A. L.; Clemente, J. C.; Knight, R.; Heath, A. C.; Leibel, R. L., *et al.* The Long-Term Stability of the Human Gut Microbiota. *Science* **2013**, *341*, 1237439.

[90] Chen, L.; Wang, D.; Garmaeva, S.; Kurilshikov, A.; Vich Vila, A.; Gacesa, R.; Sinha, T.; Segal, E.; Weersma, R. K.; Wijmenga, C., *et al.* The Long-Term Genetic Stability and Individual Specificity of the Human Gut Microbiome. *Cell* **2021**, *184*, 2302–2315.e12.

[91] Minot, S.; Bryson, A.; Chehoud, C.; Wu, G. D.; Lewis, J. D.; Bushman, F. D. Rapid Evolution of the Human Gut Virome. *Proceedings of the National Academy of Sciences of the United States of America* **2013**, *110*, 12450–12455.

[92] Hsu, B. B.; Gibson, T. E.; Yeliseyev, V.; Liu, Q.; Lyon, L.; Bry, L.; Silver, P. A.; Gerber, G. K. Dynamic Modulation of the Gut Microbiota and Metabolome by Bacteriophages in a Mouse Model. *Cell Host & Microbe* **2019**, *25*, 803–814.e5.

[93] Porter, N. T.; Hryckowian, A. J.; Merrill, B. D.; Fuentes, J. J.; Gardner, J. O.; Glowacki, R. W. P.; Singh, S.; Crawford, R. D.; Snitkin, E. S.; Sonnenburg, J. L., *et al.* Phase-Variable Capsular Polysaccharides and Lipoproteins Modify Bacteriophage Susceptibility in *Bacteroides thetaiotaomicron*. *Nature Microbiology* **2020**, *5*, 1170–1181.

[94] Porter, N. T.; Canales, P.; Peterson, D. A.; Martens, E. C. A Subset of Polysaccharide Capsules in the Human Symbiont *Bacteroides thetaiotaomicron* Promote Increased Competitive Fitness in the Mouse Gut. *Cell Host & Microbe* **2017**, *22*, 494–506.e8.

[95] Chen, W.; Xiao, H.; Wang, L.; Wang, X.; Tan, Z.; Han, Z.; Li, X.; Yang, F.; Liu, Z.; Song, J., *et al.* Structural Changes in Bacteriophage T7 upon Receptor-Induced Genome Ejection. *Proceedings of the National Academy of Sciences of the United States of America* **2021**, *118*, e2102003118.

[96] Drobysheva, A. V.; Panafidina, S. A.; Kolesnik, M. V.; Klimuk, E. I.; Minakhin, L.; Yakunina, M. V.; Borukhov, S.; Nilsson, E.; Holmfeldt, K.; Yutin, N., *et al.* Structure and Function of Virion RNA Polymerase of a CrAss-like Phage. *Nature* **2021**, *589*, 306–309.

[97] Lourenço, M.; Chaffringeon, L.; Lamy-Besnier, Q.; Pédron, T.; Campagne, P.; Eberl, C.; Bérard, M.; Stecher, B.; Debarbieux, L.; De Sordi, L. The Spatial Heterogeneity of the Gut Limits Predation and Fosters Coexistence of Bacteria and Bacteriophages. *Cell Host & Microbe* **2020**, *28*, 390–401.e5.

[98] Heilmann, S.; Sneppen, K.; Krishna, S. Coexistence of Phage and Bacteria on the Boundary of Self-Organized Refuges. *Proceedings of the National Academy of Sciences of the United States of America* **2012**, *109*, 12828–12833.

[99] Maslov, S.; Sneppen, K. Population Cycles and Species Diversity in Dynamic Kill-the-Winner Model of Microbial Ecosystems. *Scientific Reports* **2017**, *7*, 39642.

[100] Voigt, E.; Rall, B. C.; Chatzinotas, A.; Brose, U.; Rosenbaum, B. Phage Strategies Facilitate Bacterial Coexistence under Environmental Variability. *PeerJ* **2021**, *9*, e12194.

[101] Koskella, B.; Brockhurst, M. A. Bacteria–Phage Coevolution as a Driver of Ecological and Evolutionary Processes in Microbial Communities. *FEMS Microbiology Reviews* **2014**, *38*, 916–931.

[102] De Sordi, L.; Lourenço, M.; Debarbieux, L. "I Will Survive": A Tale of Bacteriophage–Bacteria Coevolution in the Gut. *Gut Microbes* **2019**, *10*, 92–99.

[103] Iranzo, J.; Krupovic, M.; Koonin, E. V. The Double-Stranded DNA Virosphere as a Modular Hierarchical Network of Gene Sharing. *mBio* **2016**, *7*, e00978–16.

[104] Koonin, E. V.; Kuhn, J. H.; Dolja, V. V.; Krupovic, M. Megataxonomy and Global Ecology of the Virosphere. *The ISME Journal* **2024**, *18*, wrad042.

[105] Weinheimer, A. R.; Ha, A. D.; Aylward, F. O. Towards a Unifying Phylogenomic Framework for Tailed Phages. *PLoS Genetics* **2025**, *21*, e1011595.

[106] Koonin, E. V.; Dolja, V. V.; Krupovic, M.; Varsani, A.; Wolf, Y. I.; Yutin, N.; Zerbini, F. M.; Kuhn, J. H. Global Organization and Proposed Megataxonomy of the Virus World. *Microbiology and Molecular Biology Reviews* **2020**, *84*, e00061–19.

[107] Hendrix, R. W. Bacteriophages: Evolution of the Majority. *Theoretical Population Biology* **2002**, *61*, 471–480.

[108] Casjens, S. R. Comparative Genomics and Evolution of the Tailed-Bacteriophages. *Current Opinion in Microbiology* **2005**, *8*, 451–458.

[109] Hatfull, G. F. Bacteriophage Genomics. *Current Opinion in Microbiology* **2008**, *11*, 447–453.

[110] Casjens, S. R.; Thuman-Commike, P. A. Evolution of Mosaically Related Tailed Bacteriophage Genomes Seen through the Lens of Phage P22 Virion Assembly. *Virology* **2011**, *411*, 393–415.

[111] Hatfull, G. F.; Jacobs-Sera, D.; Lawrence, J. G.; Pope, W. H.; Russell, D. A.; Ko, C.-C.; Weber, R. J., *et al.* Comparative Genomic Analysis of 60 Mycobacteriophage Genomes. *Journal of Molecular Biology* **2010**, *397*, 119–143.

[112] Klein-Sousa, V., *et al.* Towards a Complete Phage Tail Fiber Structure Atlas. *bioRxiv* **2024**, Preprint.

[113] Krishnamurthy, S. R.; Wang, D. Origins and Challenges of Viral Dark Matter. *Virus Research* **2017**, *239*, 136–142.

[114] Santiago-Rodriguez, T. M.; Hollister, E. B. Unraveling the Viral Dark Matter through Viral Metagenomics. *Frontiers in Immunology* **2022**, *13*, 1005107.

[115] Luo, E.; Aylward, F. O.; Mende, D. R.; DeLong, E. F. Bacteriophage Distributions and Temporal Variability in the Ocean's Interior. *mBio* **2017**, *8*, e01903–17.

[116] Ashelford, K. E.; Day, M. J.; Fry, J. C. Elevated Abundance of Bacteriophage Infecting Bacteria in Soil. *Applied and Environmental Microbiology* **2003**, *69*, 285–289.

[117] Batinovic, S.; Wassef, F.; Knowler, S. A., *et al.* Bacteriophages in Natural and Artificial Environments. *Pathogens* **2019**, *8*, 100.

[118] Roux, S.; Brum, J. R.; Dutilh, B. E., *et al.* Ecogenomics and Potential Biogeochemical Impacts of Globally Abundant Ocean Viruses. *Nature* **2016**, *537*, 689–693.

[119] Shkoporov, A. N.; Hill, C. Bacteriophages of the Human Gut: The "Known Unknown" of the Microbiome. *Cell Host & Microbe* **2019**, *25*, 195–209.

[120] Guerin, E.; Hill, C. Shining Light on Human Gut Bacteriophages. *Frontiers in Cellular and Infection Microbiology* **2020**, *10*, 481.

[121] Turner, D.; Shkoporov, A. N.; Lood, C., *et al.* Abolishment of Morphology-Based Taxa and Change to Binomial Species Names. *Archives of Virology* **2023**, *168*, 74.

[122] Smith, L.; Goldobina, E.; Govi, B.; Shkoporov, A. N. Bacteriophages of the Order Crassvirales. *Biomolecules* **2023**, *13*, 584.

[123] Jumper, J.; Evans, R., *et al.* High Accuracy Protein Structure Prediction Using Deep Learning. *CASP14 Abstracts* **2020**,

[124] Senior, A. W.; Evans, R.; Jumper, J., *et al.* Protein Structure Prediction Using Multiple Deep Neural Networks. *Proteins* **2019**, *87*, 1141–1148.

[125] Lin, Z.; Akin, H.; Rao, R., *et al.* Evolutionary-Scale Prediction of Atomic-Level Protein Structure. *Science* **2023**, *379*, 1123–1130.

[126] van Kempen, M.; Kim, S., *et al.* Foldseek: Fast and Accurate Protein Structure Search. *bioRxiv* **2022**,

[127] Moi, D.; Bernard, C.; Steinegger, M., *et al.* Structural Phylogenetics Unravels the Evolutionary Diversification of Communication Systems. *bioRxiv* **2023**,

[128] Mifsud, J. C.; Lytras, S.; Oliver, M. R.; Toon, K.; Costa, V. A.; Holmes, E. C.; Grove, J. Mapping glycoprotein structure reveals Flaviviridae evolutionary history. *Nature* **2024**, *633*, 695–703.

[129] Evseev, P.; Gutnik, D.; Shneider, M.; Miroshnikov, K. Use of an integrated approach involving AlphaFold predictions for the evolutionary taxonomy of Duplodnaviria viruses. *Biomolecules* **2023**, *13*, 110.

[130] Alves, D. R.; Nzakizwanayo, J.; Dedi, C.; Olympiou, C.; Hanin, A.; Kot, W.; Hansen, L.; Lametsch, R.; Gahan, C. G.; Schellenberger, P.; Ogilvie, L. A. Genomic and ecogenomic characterization of Proteus mirabilis bacteriophages. *Frontiers in Microbiology* **2019**, *10*, 1783.

[131] Dutilh, B. E.; Cassman, N.; McNair, K.; Sanchez, S. E.; Silva, G. G. Z.; Boling, L.; Barr, J. J.; Speth, D. R.; Seguritan, V.; Aziz, R. K.; Felts, B. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nature Communications* **2014**, *5*, 4498.

[132] Liu, Y.; Demina, T. A.; Roux, S.; Aiewsakun, P.; Kazlauskas, D.; Simmonds, P.; Prangishvili, D.; Oksanen, H. M.; Krupovic, M. Diversity, taxonomy, and evolution of archaeal viruses of the class Caudoviricetes. *PLOS Biology* **2021**, *19*, e3001442.

[133] Medvedeva, S.; Sun, J.; Yutin, N.; Koonin, E. V.; Nunoura, T.; Rinke, C.; Krupovic, M. Three families of Asgard archaeal viruses identified in metagenome-assembled genomes. *Nature Microbiology* **2022**, *7*, 962–973.

[134] Lein, K.; Paquette, K.; Pederson, T.; Switalski, M.; Smiley, C.; Oonk, E.; Baird, D.; Van Wyk, B.; Hemesath, L.; Meder, A.; Orndorff, M.; Grimes, M. Genomic Analysis of the Novel Mycobacteriophage Bromden. Poster, 17th Annual Celebration of Undergraduate Research and Creative Performance, Hope College, 2018; Hope College Digital Commons.

[135] Kabwe, M.; Brown, T. L.; Dashper, S.; Speirs, L.; Ku, H.; Petrovski, S.; Chan, H. T.; Lock, P.; Tucci, J. Genomic, morphological and functional characterisation of novel bacteriophage FNU1 capable of disrupting Fusobacterium nucleatum biofilms. *Scientific Reports* **2019**, *9*, 9107.

[136] Veesler, D.; Cambillau, C. A common evolutionary origin for tailed-bacteriophage functional modules and bacterial machineries. *Microbiology and Molecular Biology Reviews* **2011**, *75*, 423–433.

[137] Nonaka, A.; Yamamoto, H.; Kamiya, N.; Kotani, H.; Yamakawa, H.; Tsujimoto, R.; Fujita, Y. Accessory proteins of the nitrogenase assembly, NifW, NifX/NafY, and NifZ, are essential for diazotrophic growth in the nonheterocystous cyanobacterium Leptolyngbya boryana. *Frontiers in Microbiology* **2019**, *10*, 495.

[138] Hale, C. A.; Shiomi, D.; Liu, B.; Bernhardt, T. G.; Margolin, W.; Niki, H.; De Boer, P. A. J. Identification of Escherichia coli ZapC (YcbW) as a component of the division apparatus that binds and bundles FtsZ polymers. *Journal of Bacteriology* **2011**, *193*, 1393–1404.

[139] Zhang, L.; Xu, D.; Huang, Y.; Zhu, X.; Rui, M.; Wan, T.; Zheng, X.; Shen, Y.; Chen, X.; Ma, K.; Gong, Y. Structural and functional characterization of deep-sea thermophilic bacteriophage GVE2 HNH endonuclease. *Scientific Reports* **2017**, *7*, 42542.

[140] Barth, Z. K.; Dunham, D. T.; Seed, K. D. Nuclease genes occupy boundaries of genetic exchange between bacteriophages. *NAR Genomics and Bioinformatics* **2023**, *5*, lqad076.

[141] Parks, D. H.; Chuvochina, M.; Waite, D. W.; Rinke, C.; Skarshewski, A.; Chaumeil, P.-A.; Hugenholtz, P. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology* **2018**, *36*, 996–1004.

[142] Yutin, N.; Tolstoy, I.; Mutz, P.; Wolf, Y. I.; Krupovic, M.; Koonin, E. V. DNA polymerase swapping in Caudoviricetes bacteriophages. *Virology Journal* **2024**, *21*, 200.

[143] Johnson, J.; Douze, M.; Jégou, H. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* **2019**, *7*, 535–547.

[144] DeLano, W. L. PyMOL: An open-source molecular graphics tool. *CCP4 Newsletter on Protein Crystallography* **2002**, *40*, 82–92.

[145] Zhang, C.; Shine, M.; Pyle, A. M.; Zhang, Y. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nature Methods* **2022**, *19*, 1109–1115.

[146] van Rossum, G. Python programming language. Proceedings of the USENIX Annual Technical Conference. 2007; pp 1–36.

[147] Kluyver, T.; Ragan-Kelley, B.; Pérez, F.; Granger, B.; Bussonnier, M.; Frederic, J.; Kelley, K.; Hamrick, J.; Grout, J.; Corlay, S.; Ivanov, P. *Positioning and Power in Academic Publishing: Players, Agents and Agendas*; IOS Press, 2016; pp 87–90.

[148] Ihaka, R.; Gentleman, R. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **1996**, *5*, 299–314.

[149] McKinney, W. *Python for High Performance and Scientific Computing*; 2011; pp 1–9.

[150] Harris, C. R.; Millman, K. J.; van der Walt, S. J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N. J.; Kern, R. Array programming with NumPy. *Nature* **2020**, *585*, 357–362.

[151] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

[152] Waskom, M. L. Seaborn: statistical data visualization. *Journal of Open Source Software* **2021**, *6*, 3021.

[153] Barrett, P.; Hunter, J.; Miller, J. T.; Hsu, J. C.; Greenfield, P. matplotlib–A portable Python plotting package. Astronomical Data Analysis Software and Systems XIV. 2005; p 91.

[154] Buchmann, J. P.; Holmes, E. C. Entrezpy: a Python library to dynamically interact with the NCBI Entrez databases. *Bioinformatics* **2019**, *35*, 4511–4514.

[155] Gu, Z. Complex heatmap visualization. *iMeta* **2022**, *1*, e43.

[156] Wickham, H. ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics* **2011**, *3*, 180–185.

[157] Wickham, H.; François, R.; Henry, L.; Müller, K.; Vaughan, D. dplyr: A Grammar of Data Manipulation. R package version 1.1.1, 2023; https://CRAN.R-project.org/package=dplyr.

[158] Wickham, H.; Wickham, M. H. tidyr: Easily Tidy Data with spread() and gather() Functions. R package, 2017; https://CRAN.R-project.org/package=tidyr.

[159] Neuwirth, E. RColorBrewer: ColorBrewer Palettes. R package version 1.1-3, 2022; https://CRAN.R-project.org/package=RColorBrewer.

[160] Janson, G.; Paiardini, A. PyMod 3: a complete suite for structural bioinformatics in PyMOL. *Bioinformatics* **2021**, *37*, 1471–1472.

[161] Grant, B. J.; Skjærven, L.; Yao, X.-Q. The Bio3D packages for structural bioinformatics. *Protein Science* **2021**, *30*, 20–30.

[162] Letunic, I.; Bork, P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Research* **2024**, *52*, W78–W82.

[163] Fierer, N.; Jackson, R. B. The diversity and biogeography of soil bacterial communities. *Proceedings of the National Academy of Sciences* **2006**, *103*, 626–631.

[164] Naylor, D.; Fansler, S.; Brislawn, C.; Nelson, W. C.; Hofmockel, K. S.; Jansson, J. K.; McClure, R. Deconstructing the soil microbiome into reduced-complexity functional modules. *mBio* **2020**, *11*.

[165] van der Heijden, M. G. A.; Bardgett, R. D.; van Straalen, N. M. The Unseen Majority: Soil Microbes as Drivers of Plant Diversity and Productivity in Terrestrial Ecosystems. *Ecology Letters* **2008**, *11*, 296–310.

[166] Bardgett, R. D.; van der Putten, W. H. Belowground Biodiversity and Ecosystem Functioning. *Nature* **2014**, *515*, 505–511.

[167] Fierer, N. Embracing the unknown: disentangling the complexities of the soil microbiome. *Nature Reviews Microbiology* **2017**, *15*, 579–590.

[168] Sokol, N. W.; Slessarev, E.; Marschmann, G. L.; Nicolas, A.; Blazewicz, S. J.; Brodie, E. L.; Firestone, M. K.; Foley, M. M.; Hestrin, R.; Hungate, B. A.; Koch, B. J. Life and death in the soil microbiome: how ecological processes influence biogeochemistry. *Nature Reviews Microbiology* **2022**, *20*, 415–430.

[169] Hartmann, M.; Six, J. Soil structure and microbiome functions in agroecosystems. *Nature Reviews Earth & Environment* **2023**, *4*, 4–18.

[170] Trivedi, P.; Leach, J. E.; Tringe, S. G.; Sa, T.; Singh, B. K. Plant–microbiome interactions: from community assembly to plant health. *Nature Reviews Microbiology* **2020**, *18*, 607–621.

[171] Pratama, A. A.; Van Elsas, J. D. The neglected soil virome–potential role and impact. *Trends in Microbiology* **2018**, *26*, 649–662.

[172] Roux, S.; Emerson, J. B. Diversity in the soil virosphere: to infinity and beyond? *Trends in Microbiology* **2022**, *30*, 1025–1035.

[173] Ma, B.; Wang, Y.; Zhao, K.; Stirling, E.; Lv, X.; Yu, Y.; Hu, L.; Tang, C.; Wu, C.; Dong, B.; Xue, R. Biogeographic patterns and drivers of soil viromes. *Nature Ecology & Evolution* **2024**, *8*, 717–728.

[174] Kuzyakov, Y.; Mason-Jones, K. Viruses in soil: Nano-scale undead drivers of microbial life, biogeochemical turnover and ecosystem functions. *Soil Biology and Biochemistry* **2018**, *127*, 305–317.

[175] Suttle, C. A. Marine viruses–major players in the global ecosystem. *Nature Reviews Microbiology* **2007**, *5*, 801–812.

[176] Townsend, E. M.; Kelly, L.; Muscatt, G.; Box, J. D.; Hargraves, N.; Lilley, D.; Jameson, E. The human gut phageome: origins and roles in the human gut microbiome. *Frontiers in Cellular and Infection Microbiology* **2021**, *11*, 643214.

[177] Sun, M.; Yuan, S.; Xia, R.; Ye, M.; Balcázar, J. L. Underexplored viral auxiliary metabolic genes in soil: diversity and eco-evolutionary significance. *Environmental Microbiology* **2023**, *25*, 800–810.

[178] Emerson, J. B. Soil viruses: a new hope. *mSystems* **2019**, *4*.

[179] Jansson, J. K.; Wu, R. Soil viral diversity, ecology and climate change. *Nature Reviews Microbiology* **2023**, *21*, 296–311.

[180] Graham, E. B.; Camargo, A. P.; Wu, R.; Neches, R. Y.; Nolan, M.; Paez-Espino, D.; Kyrpides, N. C.; Jansson, J. K.; McDermott, J. E.; Hofmockel, K. S. A global atlas of soil viruses reveals unexplored biodiversity and potential biogeochemical impacts. *Nature Microbiology* **2024**, *9*, 1873–1883.

[181] Nishijima, S.; Fullam, A.; Schmidt, T. S.; Kuhn, M.; Bork, P. VIRE: a metagenome-derived, planetary-scale virome resource with environmental context. *Nucleic Acids Research* **2025**, gkaf1225.

[182] Liao, H.; Li, H.; Duan, C.-S.; Zhou, X.-Y.; Luo, Q.-P.; An, X.-L.; Zhu, Y.-G.; Su, J.-Q. Response of soil viral communities to land use changes. *Nature Communications* **2022**, *13*, 6027.

[183] Liu, C.; Ni, B.-J.; Wang, X.; Deng, Y.; Tao, L.; Zhou, X.; Deng, J. Effect of forest soil viruses on bacterial community succession and the implication for soil carbon sequestration. *Science of the Total Environment* **2023**, *892*, 164800.

[184] Durham, D. M.; Sieradzki, E. T.; Ter Horst, A. M.; Santos-Medellín, C.; Bess, C. W.; Geonczy, S. E.; Emerson, J. B. Substantial differences in soil viral community composition within and among four Northern California habitats. *ISME Communications* **2022**, *2*, 100.

[185] Adriaenssens, E. M.; Kramer, R.; Van Goethem, M. W.; Makhalanyane, T. P.; Hogg, I.; Cowan, D. A. Environmental drivers of viral community composition in Antarctic soils identified by viromics. *Microbiome* **2017**, *5*, 83.

[186] Casjens, S. R.; Gilcrease, E. B. Determining DNA Packaging Strategy by Analysis of the Termini of the Chromosomes in Tailed-Bacteriophage Virions. **2009**, 91–111.

[187] Deng, Z.; Norseen, J.; Wiedmer, A.; Riethman, H.; Lieberman, P. M. Telomeres and Viruses: Common Themes of Genome Maintenance. *Frontiers in Oncology* **2012**, *2*, 201.

[188] Antipov, D.; Raiko, M.; Lapidus, A.; Pevzner, P. A. Metaviral SPAdes: assembly of viruses from metagenomic data. *Bioinformatics* **2020**, *36*, 4126–4129.

[189] Berg, M.; Reiter, T.; Emerson, J.; Brown, C. T.; Roux, S. Comparison of Short-Read and Long-Read Metagenome Assemblies in a Natural Soil Community Highlights Systematic Bias in Recovery of High-Diversity Populations. *NAR Genomics and Bioinformatics* **2025**, *7*, lqaf163.

[190] Sutton, T. D. S.; Clooney, A. G.; Ryan, F. J.; Ross, R. P.; Hill, C. Choice of Assembly Software Has a Critical Impact on Virome Characterisation. *Microbiome* **2019**, *7*, 12.

[191] Hesse, U.; Van Heusden, P.; Kirby, B. M.; Olonade, I.; Van Zyl, L. J.; Trindade, M. Virome assembly and annotation: a surprise in the Namib Desert. *Frontiers in Microbiology* **2017**, *8*, 13.

[192] Chen, L.; Banfield, J. F. COBRA improves the completeness and contiguity of viral genomes assembled from metagenomes. *Nature Microbiology* **2024**, *9*, 737–750.

[193] Roux, S.; Hallam, S. J.; Woyke, T.; Sullivan, M. B. Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *eLife* **2015**, *4*, e08490.

[194] Kosmopoulos, J. C.; Anantharaman, K. Viral dark matter: illuminating protein function, ecology, and biotechnological promises. *Biochemistry* **2025**,

[195] Sorensen, J. W.; Zinke, L. A.; Ter Horst, A. M.; Santos-Medellín, C.; Schroeder, A.; Emerson, J. B. DNase treatment improves viral enrichment in agricultural soil viromes. *mSystems* **2021**, *6*.

[196] Bi, L.; He, J.-Z.; Hu, H.-W. Total metagenomes outperform viromes in recovering viral diversity from sulfuric soils. *ISME Communications* **2024**, *4*, ycae017.

[197] Nayfach, S.; Camargo, A. P.; Schulz, F.; Eloe-Fadrosh, E.; Roux, S.; Kyrpides, N. C. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nature Biotechnology* **2021**, *39*, 578–585.

[198] Gualberto, J. M.; Mileshina, D.; Wallet, C.; Niazi, A. K.; Weber-Lotfi, F.; Dietrich, A. The plant mitochondrial genome: dynamics and maintenance. *Biochimie* **2014**, *100*, 107–120.

[199] de La Higuera, I.; Kasun, G. W.; Torrance, E. L.; Pratt, A. A.; Maluenda, A.; Colombet, J.; Bisseux, M.; Ravet, V.; Dayaram, A.; Stainton, D.; Kraberger, S. Unveiling crucivirus diversity by mining metagenomic data. *mBio* **2020**, *11*, e01280–20.

[200] Desingu, P. A.; Nagarajan, K. Genetic diversity and characterization of circular replication (Rep)-encoding single-stranded (CRESS) DNA viruses. *Microbiology Spectrum* **2022**, *10*, e01057–22.

[201] Bamford, D. H.; Caldentey, J.; Bamford, J. K. H. Bacteriophage PRD1: a broad host range dsDNA tectivirus with an internal membrane. *Advances in Virus Research* **1995**, *45*, 281–319.

[202] Krupovic, M.; Varsani, A.; Kazlauskas, D.; Breitbart, M.; Delwart, E.; Rosario, K.; Yutin, N.; Wolf, Y. I.; Harrach, B.; Zerbini, F. M.; Dolja, V. V. Cressdnaviricota: a virus phylum unifying seven families of rep-encoding viruses with single-stranded, circular DNA genomes. *Journal of Virology* **2020**, *94*, e00282–20.

[203] Laufs, J.; Jupin, I.; David, C.; Schumacher, S.; Heyraud-Nitschke, F.; Gronenborn, B. Geminivirus replication: genetic and biochemical characterization of Rep protein function, a review. *Biochimie* **1995**, *77*, 765–773.

[204] Mönttinen, H. A.; Bicep, C.; Williams, T. A.; Hirt, R. P. The genomes of nucleocytoplasmic large DNA viruses: viral evolution writ large. *Microbial Genomics* **2021**, *7*, 000649.

[205] Moniruzzaman, M.; Martinez-Gutierrez, C. A.; Weinheimer, A. R.; Aylward, F. O. Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. *Nature Communications* **2020**, *11*, 1710.

[206] Janssen, P. H. Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. *Applied and Environmental Microbiology* **2006**, *72*, 1719–1728.

[207] Philippot, L.; Raaijmakers, J. M.; Lemanceau, P.; van der Putten, W. H. Going back to the roots: the microbial ecology of the rhizosphere. *Nature Reviews Microbiology* **2013**, *11*, 789–799.

[208] Breitbart, M.; Thompson, L. R.; Suttle, C. A.; Sullivan, M. B. Exploring the Vast Diversity of Marine Viruses. *Oceanography* **2007**, *20*, 135–139.

[209] Martin, C.; Emerson, J. B.; Roux, S.; Anantharaman, K., *et al.* A call for caution in the biological interpretation of viral auxiliary metabolic genes. *Nature Microbiology* **2025**, *10*, 2122–2129.

[210] Gerdes, K. Toxin-antitoxin modules may regulate synthesis of macromolecules during nutritional stress. *Journal of Bacteriology* **2000**, *182*, 561–572.

[211] Pandey, D. P.; Gerdes, K. Toxin–antitoxin loci are highly abundant in free-living but lost from host-associated prokaryotes. *Nucleic Acids Research* **2005**, *33*, 966–976.

[212] Seed, K. D.; Lazinski, D. W.; Calderwood, S. B.; Camilli, A. A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature* **2013**, *494*, 489–491.

[213] Al-Shayeb, B.; Skopintsev, P.; Soczek, K. M.; Stahl, E. C.; Li, Z.; Groover, E.; Smock, D.; Eggers, A. R.; Pausch, P.; Cress, B. F.; Huang, C. J. Diverse virus-encoded CRISPR-Cas systems include streamlined genome editors. *Cell* **2022**, *185*, 4574–4586.

[214] Dos Santos, P. C.; Igarashi, R. Y.; Lee, H. I.; Hoffman, B. M.; Seefeldt, L. C.; Dean, D. R. Substrate interactions with the nitrogenase active site. *Accounts of Chemical Research* **2005**, *38*, 208–214.

[215] Dos Santos, P. C.; Fang, Z.; Mason, S. W.; Setubal, J. C.; Dixon, R. Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. *BMC Genomics* **2012**, *13*, 162.

[216] Battchikova, N.; Eisenhut, M.; Aro, E.-M. Cyanobacterial NDH-1 complexes: novel insights and remaining puzzles. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **2011**, *1807*, 935–944.

[217] Sharon, I.; Alperovitch, A.; Rohwer, F.; Haynes, M.; Glaser, F.; Atamna-Ismaeel, N.; Pinter, R. Y.; Partensky, F.; Koonin, E. V.; Wolf, Y. I.; Nelson, N. Photosystem I gene cassettes are present in marine virus genomes. *Nature* **2009**, *461*, 258–262.

[218] Lindell, D.; Jaffe, J. D.; Johnson, Z. I.; Church, G. M.; Chisholm, S. W. Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* **2005**, *438*, 86–89.

[219] Dodonova, S. O.; Prinz, S.; Bilanchone, V.; Sandmeyer, S.; Briggs, J. A. G. Structure of the Ty3/Gypsy retrotransposon capsid and the evolution of retroviruses. *Proceedings of the National Academy of Sciences* **2019**, *116*, 10048–10057.

[220] Smith, L.; Schmidt, T. S.; Tolstoy, I.; Skvortsov, T.; Hill, C.; Shkoporov, A. N. A novel approach to *Caudoviricetes* taxonomy utilising whole proteome structure–structure comparison. *bioRxiv* **2025**, 2025.08.06, Preprint.

[221] National Energy Research Scientific Computing Center, crisprDB. https://portal.nersc.gov/cfs/m342/crisprDB, 2023; Accessed 2023.

[222] Oksanen, J.; Simpson, G. L.; Blanchet, F. G.; Kindt, R.; Legendre, P.; Minchin, P. R.; O'Hara, R. B.; Solymos, P.; Stevens, M. H. H.; Szoecs, E.; Wagner, H. *vegan: Community Ecology Package*; R package version 2.6-4, 2022.

[223] Gilchrist, C. L. M.; Chooi, Y.-H. Clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics* **2021**, *37*, 2473–2475.