RNA Bacteriophages: Diversity, Abundance, and Applications

A thesis presented to the National University of Ireland

for the Degree of

Doctor of Philosophy

by

Julie Callanan B.Sc.

Student ID No. 113337961

School of Microbiology

University College Cork

Supervisors: Prof. Colin Hill and Prof. Paul Ross

Head of School: Prof. Paul O'Toole

2021

Contents

Declaration 5
Abbreviations 6
Thesis Abstract 8
Chapter I
RNA Bacteriophage Biology in a Metagenomic Era 12
1.1 Abstract 13
1.2 Introduction 13
1.3 *Cystoviridae* 17
1.4 *Leviviridae* 24
1.4.1 *Levivirus* 27
1.4.2 *Allolevivirus* 32
1.5 Discussion 34
1.6 References 37
Chapter II
Biases in Viral Metagenomics-Based Detection, Cataloguing and Quantification of Bacteriophage Genomes in Human Faeces, a Review 57
2.1 Abstract 58
2.2 Introduction 58
2.3 Sample handling 64
2.4 VLP isolation 66
2.5 Nucleic acid extraction and library preparation 69
2.6 Bioinformatic pipelines 76
2.7 Discussion 80
2.8 References 80
Chapter III
Expansion of Known +ssRNA Bacteriophage Genomes: From Tens to over a Thousand 95
3.1 Abstract 97
3.2 Introduction 97
3.3 Materials and Methods 101
3.3.1 Assembly of metatranscriptome samples 101
3.3.2 Generation of profile hidden Markov models 103
3.3.3 Validating HMM detection of +ssRNA phages 104
3.3.4 Detecting +ssRNA within metatranscriptome samples 105
3.3.5 Analysis of +ssRNA phage proteins 105
3.4 Results and Discussion 107
3.4.1 Existing +ssRNA phage sequences 107

3.4.2 Expansion of known +ssRNA phage sequences 109
3.4.3 Examination of genome-associated proteins and architecture 113
3.4.4 Phylogenetic assessment of near-complete +ssRNA phage genomes 119
3.4.5 Examination of phage-host interactions 123
3.5 Conclusion 128
3.6 References 129
Chapter IV
Leviviricetes: Expanding and Restructuring the Taxonomy of Bacteria-Infecting Single-Stranded RNA Viruses 140
4.1 Abstract 141
4.2 Open access data 141
4.3 Introduction 142
4.4 Profile HMMs to detect and classify bacteria-infecting +ssRNA viruses 146
4.5 Taxonomy of class *Leviviricetes* 149
4.6 Discussion 156
4.7 References 157
Chapter V
Examination of Enrichment and Nucleic Acid Extraction Methods for RNA Bacteriophage Detection, Isolation, and Characterisation. 163
5.1 Abstract 164
5.2 Introduction 164
5.3 Materials and Methods 167
5.3.1 Sample collection and storage 167
5.3.2 Propagation of +ssRNA phages 168
5.3.3 Extraction methods 168
5.3.4 Nucleic acid extraction 171
5.3.5 Sequencing of VLP nucleic acids 172
5.3.6 Bioinformatic analysis 172
5.4 Results 175
5.4.1 Overview of the three extraction methods 175
5.4.2 Examination of the recovery of +ssRNA phages per extraction method 178
5.4.3 Application of Biopsy Method in Biogeography Study 182
5.4.4 Future applications of this method to large-scale human gut phageome studies 186
5.5 Conclusion 186
5.6 References 187
Chapter VI
An Assessment of *Cystovirus* phi6 as a Surrogate for SARS-CoV-2 in Lipopeptide Exposure and Thermotolerance Assays. 195

6.1 Abstract 196
6.2 Introduction 196
6.3 Materials and Methods 202
6.3.1 Bacterial strains, bacteriophage stocks and media 202
6.3.2 ART24 analysis 202
6.3.3 SDS-PAGE analysis 203
6.3.4 Thermotolerance evaluation 204
6.4 Results 206
6.4.1 ART24 activity 206
6.4.2 Thermotolerance 210
6.5 Conclusion 212
6.6 References 213
Thesis summary and future work 217
Appendix I
An Insight into the Elusive RNA Bacteriophages 221
Appendix II
Rename one class (*Leviviricetes* - formerly *Allassoviricetes*), rename one order (*Norzivirales* - formerly *Levivirales*), create one new order (*Timlovirales*), and expand the class to a total of six families, 420 genera and 883 species. 225
Appendix III
Guidance on Creating Individual and Bulk Latinized Binomial Virus Species and Other Taxon Names. 238
Acknowledgements 247

Declaration

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

Signed:  Date: __11/10/2021__

Julie Callanan

This thesis was derived from research conducted with the financial support of Science Foundation Ireland (SFI) for the APC Microbiome Ireland under grant numbers SFI/12/RC/2273 and 12/RC/2273_P2.

Abbreviations

amino acid – aa

base pair – bp

Basic Local Alignment Search Tool – BLAST

caesium chloride – CsCl

coat protein – CP

complementary DNA – cDNA

deoxyribonuclease – DNase

deoxyribonucleic acid – DNA

double stranded – ds

gastrointestinal tract – GIT

gigabyte – Gb

hidden Markov model – HMM

International Code of Virus Classification and Nomenclature – ICVCN

International Committee on Taxonomy of Viruses – ICTV

kilobases – kb

Luria Bertani broth – LBB

maturation protein – MP

minimum inhibitory concentration – MIC

multiple displacement amplification – MDA

National Center for Biotechnology Information – NCBI

NCBI Reference Sequence database – RefSeq database

open reading frame – ORF

pairwise amino acid identity – PAAI

phosphate-buffered saline buffer – PBS buffer

plaque forming units – pfu

polyethylene glycol – PEG

polymerase chain reaction – PCR

reverse transcription – RT

ribonuclease – RNase

ribonucleic acid – RNA

RNA-directed RNA polymerase – RdRP

single stranded – ss

sodium chloride and magnesium sulphate buffer – SM buffer

tryptic soy broth – TSB

virus-like particle – VLP

Thesis Abstract

We live in a world created by and dominated by microbes, yet we are only beginning to understand this complex and diverse realm. This vast collection of microorganisms (the microbiome) is composed of bacteria, viruses, fungi, archaea, protozoa, and algae, and is essential to every aspect of life and are involved in countless natural processes. While interest in the microbiome has exploded in recent years, studies regarding the viral fraction has lagged behind. This viral component (virome) is dominated by bacteriophages (phages) - viruses that target and infect prokaryotes. They are intrinsically linked to the bacterial community of every ecosystem and potentially dictate the bacterial composition, function, and dynamics through a series of complex interactions. Their roles across global environments, from human gut to marine and terrestrial settings, are only just beginning to be described. While gradual improvements in virome and phageome research have provided us with some insights into the function of these viruses, much more work is needed to gauge their full importance.

Of the known phages, those that encode either a double-stranded DNA (dsDNA) or single-stranded DNA (ssDNA) genome have been more intensively studied than their RNA counterparts. The RNA phages are either positive-sense, single-stranded RNA (+ssRNA; *Leviviricetes*) or double-stranded RNA (dsRNA; *Cystoviridae*) and have been understudied and underrepresented in publicly available databases. This thesis tackled the limited knowledge of these phages, their lifecycles, and our current understandings of their taxonomy. It also explored the potential biases associated with isolating and extracting RNA phages from human faecal samples which may have contributed to their under-representation in many virome studies. Properly isolating and identifying RNA phage is crucial to better understand the diversity of the global microbiome.

Given that only limited numbers of +ssRNA phages are present in databases, it was timely to explore their true abundance in different environments by exploiting advances in the

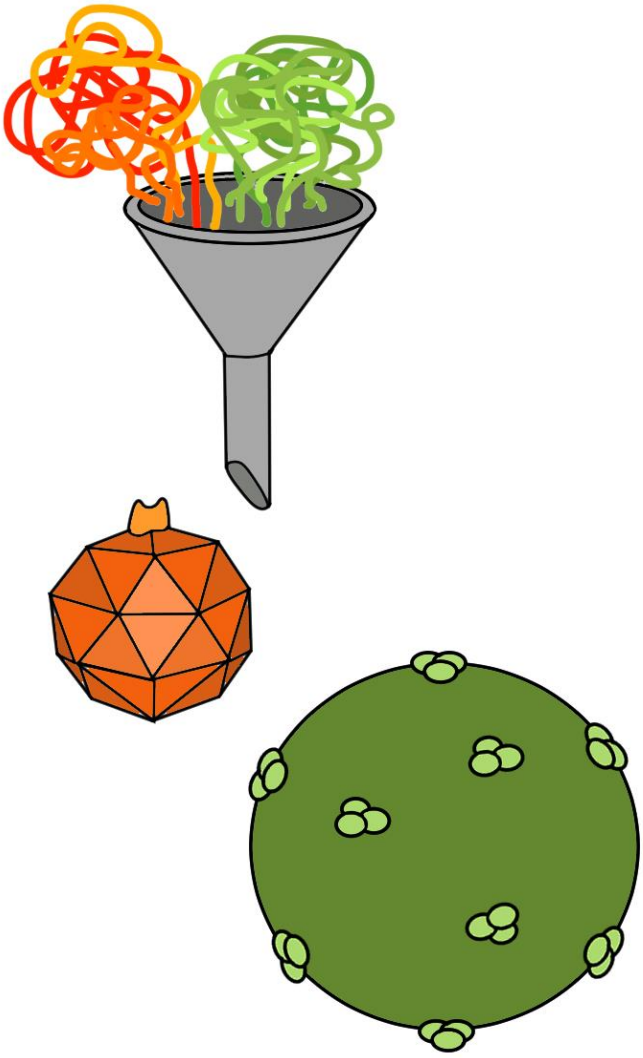
science of bioinformatics. Our method, utilizing specific profile hidden Markov model (HMM) search tools, is described in detail. This work greatly expanded the numbers of +ssRNA phage genomes and resulted in the submission to and acceptance of an updated taxonomy by the International Committee on Taxonomy of Viruses (ICTV). The framework depicted in this thesis allows for the expected expansion of these phages in future work. It also offers an example for potential studies looking to combine both cultured and metagenomic-derived genomes in taxonomic updates.

It is important that future studies not only optimize the bioinformatic approaches used but also target and improve the isolation and extraction methods applied to enhance the recovery of RNA phages. Since the biases associated with different extraction methodologies have been pinpointed as a crucial factor, three methods were examined and assessed for their efficacy using controls spiked with MS2 and Qbeta. This work was coupled with an in-house study that, using one of these alternative phage-extraction methods, isolated +ssRNA phages from a mammalian gut for the first time in our laboratory.

Over the past year, the importance of studying RNA viruses has never been so apparent as a result of the global pandemic due to Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) and its associated disease. However, given that this virus is classified as being extremely pathogenic and subject to high rates of mutation, the idea of using another virus as a safe surrogate has been previously suggested. One such candidate virus is phi6, a dsRNA phage of the *Cystoviridae* family. With its enveloped structure it offers a reliable model in the examination of different treatments and therapies in terms of their potential in combating SARS-CoV-2. A chapter of this thesis is dedicated to exploring phi6 as a surrogate in lipopeptide exposure and thermotolerance assays.

Overall, this thesis investigated the realm of RNA phages. From examining the current literature on their basic biology to biases associated with their recovery from virome studies,

the initial two chapters offer a foundation for the four subsequent chapters. In addition, RNA phage numbers were expanded, taxonomically restructured, tracked through different extraction methods, and assessed as a surrogate for SARS-CoV-2. In 1980, Norton Zinder wrote "as long as there are bacteria, there will be RNA phage" and it is suspected that we are just beginning to realize how accurate he was.



Chapter I

RNA Bacteriophage Biology in a Metagenomic Era

This chapter was published as a review in *Viruses.*

Callanan, Julie, Stephen R. Stockdale, Andrey Shkoporov, Lorraine A. Draper, R. Paul Ross, and Colin Hill. "RNA phage biology in a metagenomic era." Viruses 10, no. 7 (2018): 386.

https://doi.org/10.3390/v10070386

1.1 Abstract

The number of novel bacteriophage sequences has expanded significantly as a result of many metagenomic studies of phage populations in diverse environments. Most of these novel sequences bear little or no homology to existing databases (referred to as the "viral dark matter"). Also, these sequences are primarily derived from DNA-encoded bacteriophages (phages) with few RNA phages included. Despite the rapid advancements in high-throughput sequencing, few studies enrich for RNA viruses, i.e., target viral rather than cellular fraction and/or RNA rather than DNA via a reverse transcriptase step, in an attempt to capture the RNA viruses present in a microbial communities. It is timely to compile existing and relevant information about RNA phages to provide an insight into many of their important biological features, which should aid in sequence-based discovery, and in their subsequent annotation. Without comprehensive studies, the biological significance of RNA phages has been largely ignored. Future phage studies should be adapted to ensure they are properly represented in viromic and phageomic studies.

1.2 Introduction

Bacteriophages, commonly known as phages, are the most abundant biological entities on the planet, with approximately 10^{31} in the biosphere (Hatfull 2015). Phages were independently identified in 1915 by Twort and in 1917 by d'Hérelle (Twort 1915; d'Herelle 1917). They are viruses which can alter microbial populations, with a major role in diversity patterns of microbial populations (Rodriguez-Valera *et al*. 2009). They were first recorded as antibacterial agents by d'Hérelle and quickly developed into clinical aids against bacterial infections, particularly across Eastern Europe (d'Herelle 1917; Carlton 1999; Summers 2012). The first known RNA phage, f2, which infects *Escherichia coli* (*E. coli*), was described more than 40

years after the discovery of DNA phages (Loeb and Zinder 1961). Weissmann (1974), suggested that RNA phages offered a means to examine basic biological processes at an in-depth molecular level (Weissmann 1974). Since their identification, RNA phages have served as valuable models for understanding not just essential viral processes but also fundamental molecular mechanisms such as RNA genome replication, translational control, and gene regulation (D. Brown and Gold 1996; Gytz *et al*. 2015; Lodish 1968; Stock-Ley, Stonehouse, and Valegård 1994).

The RNA phage MS2, isolated by Alvin John Clark in 1961 and highly similar phage f2 (Davis, Strauss, and Sinsheimer 1961), have become key models in molecular biology and genetics. The MS2 phage coat protein gene was the first gene to be completely sequenced in 1972 by Fiers and his colleagues (Jou *et al*. 1972). In addition, the genome of the MS2 phage was the first to be fully sequenced in 1976, also by Walter Fiers and colleagues (Fiers *et al*. 1976). This preceded the sequencing of the first DNA based genome of phage phiX174 in 1977 (Sanger *et al*. 1977). RNA phages have also provided scientists with a model system for understanding the biology of many human pathogenic viruses such as hepatitis, influenza, and human immunodeficiency virus (HIV) (Adcock *et al*. 2009; Kenyon, Prestwood, and Lever 2015; Wang *et al*. 2016). For example, the stem loop structure of MS2 phage RNA has become a common tool for studying the key group antigen (Gag) polyprotein of HIV by replicating the protein–protein interactions (Becker and Sherer 2017).

While there is a substantial amount of literature and studies involving the bacterial component of the gut microbial system, there is still relatively little known about the human virome, and in particular the phage fraction of the microbiome. In addition, most newly identified phage sequences do not have known counterparts in viral databases, and these unknown sequences are sometimes referred to as the “viral dark matter” (Hatfull 2015). Phages

influence microbial populations by infecting and destroying specific species of bacteria. Alternatively, temperate or lysogenic phage are capable of integrating their genomes into the host bacterium's chromosome, often providing bacteria with a fitness advantage while the phage remains dormant and replicates in tandem with the host chromosome (Harrison and Brockhurst 2017).

The genomic composition of phages is extremely diverse and are composed of either DNA or RNA, which can in turn be either single-stranded (ss) or double stranded (ds). Single-stranded RNA genomes can exist in two variants: negative sense (−) and positive sense (+). This depends on their orientation and whether there is a prerequisite for transcription prior to translation. Some eukaryotic RNA viruses use reverse transcriptase to replicate their genetic material through a DNA intermediate, while no DNA stage has been observed amongst bacterial RNA phages to date. In addition, the genomes of phages are described as being "mosaic", composed of individual modules that may appear in other phages but in an alternative arrangement (Vasiljeva *et al*. 1998; Hatfull 2008). While phages can evolve through the accumulation of mutations, within environments they are responsible for vast amounts of genetic recombination and horizontal gene transfer events (de la Cruz and Davies 2000). Altogether, these attributes make both DNA and RNA phage genomes very diverse and difficult to classify.

Typically, phage infection proceeds via adsorption, penetration, replication, assembly, and release. Briefly, phages use specialized surface receptor-binding proteins to interact with and adhere to their specific cognate host receptor. Phages then use various mechanisms to breach the cell wall of bacteria and inject their genomes into the cytoplasm of the host. Infection can then proceed via a lysogenic (temperate) or lytic (virulent) lifecycle. Virulent phages hijack the host's cellular components to direct the replication of the phage's genome and produce the

necessary viral encoded proteins. Once the phage genome is replicated, it is packed into self-assembled viral particles. Phages induce host cell lysis, and the assembled phage progeny are released into the surrounding environment for successive infections. Temperate phages that can replicate through lytic or lysogenic lifecycles are typically able to integrate into the bacterial chromosome and are subsequently replicated in tandem with the host genome. Temperate phages can also be maintained through formation of an episome within the host, which is disseminated through a population via cell division (Cenens *et al*. 2015). Temperate phages can respond to host cues from environmental stresses to initiate the lytic cycle and release phage progeny. They have been shown to dramatically affect susceptible bacterial populations through transfer of novel genes to their host, they can provide resistance to subsequent phage predation and can also alter host gene expression (Howard-Varona *et al*. 2017).

Although studies into the phage component of the microbiome have increased rapidly in recent years, databases are dominated by phages with DNA genomes. According to the latest (2017) report by the International Committee for the Taxonomy of Viruses (ICTV), viruses are separated into 134 families. The same report separated RNA phages into only two families; *Cystoviridae* (dsRNA phage) with 1 genus, *Cystovirus*, with 7 recognised species, and *Leviviridae* (+ssRNA phage), with 2 genera, *Levivirus* and *Allolevivirus*, each of which contain two species (Olsthoorn and van Duin 2017; Poranen and Mäntynen 2017).

A recent examination of RNA phage populations in 2016 by Krishnamurthy and Wang, through metagenomic dataset analysis, led to the identification of 122 partial genomes of novel RNA phages (Krishnamurthy *et al*. 2016). The host range of DNA phages typically varies greatly in contrast to that of RNA phages, which were all thought to target members of the Proteobacteria phylum. However, in that study, an RNA phage was identified from a

transcriptome of pure culture of *Streptomyces avermitilis*, a Gram-positive bacterium, known as *Streptomyces* phage phi0. This phage is thought to belong to the *Cystoviridae* family based on RNA-directed RNA-polymerase (RdRP) analysis. This was the first report of an RNA phage with a natural affinity for a Gram-positive host. In addition, a recent pre-print has described an RNA virus, a planarian-infecting *Nidovirales*, with a genome of 41.1 kb in length, significantly longer than the previous largest RNA virus genome of 30 kb (Saberi *et al*. 2018). These findings highlight that there are certainly many RNA viruses yet to be discovered and described, including RNA phages.

This review focuses on examining known RNA phages, both dsRNA and +ssRNA, which target bacterial cells. Outlined are their mechanisms of adsorption through to the release of progeny. Future endeavours may use conserved features of RNA phages as genetic signatures to aid in prospective metagenomic exploration of RNA phages in the "viral dark matter" via sequence-based targeting.

1.3 *Cystoviridae*

Currently there are seven recognized species of *Cystoviridae* listed in the 2017 ICTV report. The type species of Cystovirus family is phi6, which for a long time was thought to be unique as a dsRNA phage. The *Cystoviridae* have a tri-segmented, linear dsRNA genome, with the concatenated genome varying size from 12.7 kb (phi2954) to 15.0 kb (phi8). Individual genome segments range in size from 2.9 kb to 6.4 kb (Figure 1). The three genome segments, large (L), medium (M), and small (S) are transcribed into separate polycistronic mRNAs that are predicted to be translated by the host machinery into 12 proteins. A lipid membrane envelops a double-layered proteinaceous nucleocapsid (NC) (Etten *et al*. 1976).

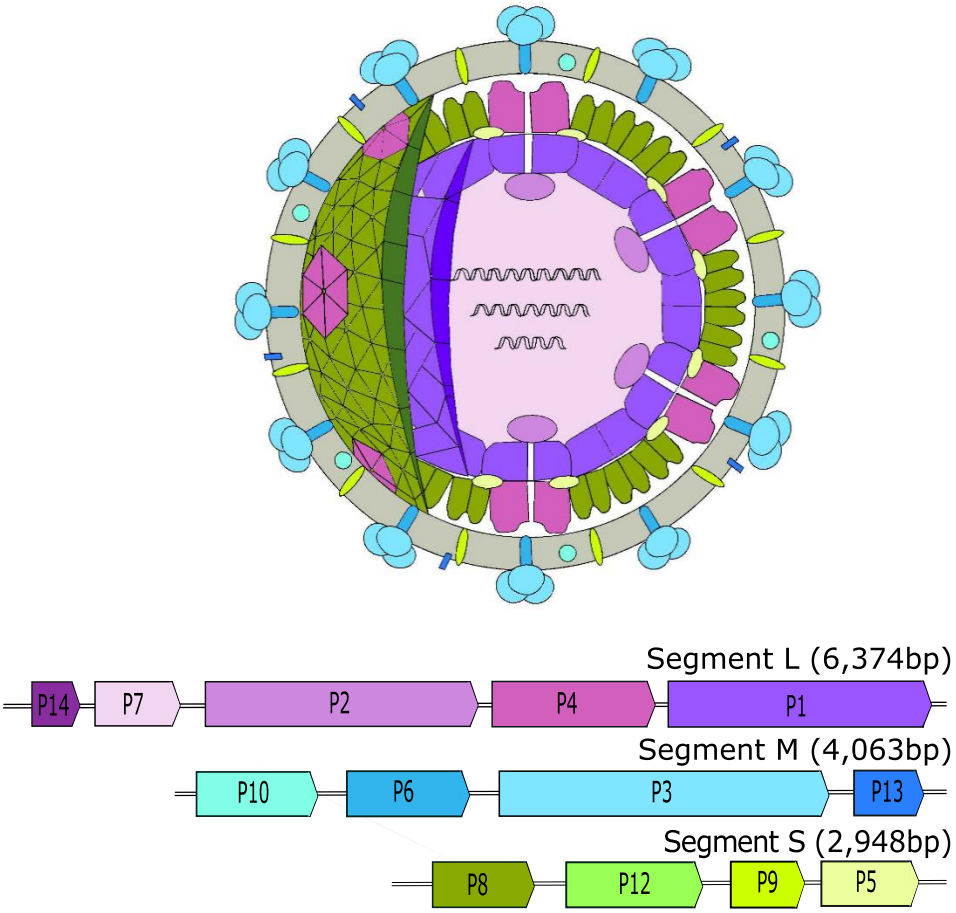


Figure 1. Virion of *Pseudomonas* phage phi6, the type-virus of the *Cystoviridae* family. The virion and genes encoded by the tri-segmented genome of this phage are color coordinated. The grey circle represents the membrane encapsulating the virion. See text regarding gene information. (This figure was reproduced based on other images (Alphonse and Ghose 2017; Poranen and Mäntynen 2017; "Viral Zone: *Cystoviridae*" 2018)).

Cystoviridae genes are ordered into functional units within the segments: L-segment contains genes for the virion core (P1, P2, P4, and P7), the M-segment encodes the complex essential for host recognition (P3 and P6), and the S-segment is responsible for the shell protein of the nucleocapsid (P8 (except in phi8), P9, P12, and P5) (Gottlieb *et al*. 2002; 1988;

Hoogstraten *et al*. 2000; Mäntynen *et al*. 2015; McGraw, Mindich, and Frangione 1986; L. Mindich *et al*. 1988; X. Qiao *et al*. 2010; 2000; Y. Yang *et al*. 2016). P5 and P11 are transcript variants of the same gene (Carpino 2014). The noncoding regions that flank the coding sequences within the segments are required for efficient genome replication and packaging. The 5′ untranslated region (UTR) of the plus strand region encodes a *cis*-acting RNA sequence known as the *pac* sequence (Leonard Mindich 1999). The segment-specific *pac* sequence is composed of 200 nucleotides located within several stem-loop structures. The *pac* sequences act in unison with other fundamental structural elements to ensure the correct packaging of the genome when required.

The integral-membrane, fusogenic P6 protein is responsible for securing the receptor-binding protein of *Cystoviridae*, P3, to the viral envelope. It is this multimeric spike protein, P3, which enables the recognition of the host bacteria receptor pilin, the protein monomer making up bacterial pili, by the phi6 phage. The P3 protein of phages phi8, phi12, phi13, and phiYY have been suggested to be a single polypeptide or a multimer (Mäntynen, Sundberg, and Poranen 2017). The P3 protein of phi6 adsorbs to host type IV pili of its target, *Pseudomonas syringae*, which then retracts to bring the phage into close proximity of the host membrane (Bamford, Palva, and Lounatmaa 1976; Roine *et al*. 1998). This form of attachment is also exploited by phiNN and phi2954 (Mäntynen *et al*. 2015; X. Qiao *et al*. 2010). Other members of *Cystoviridae*, such as phi8, phi12, phi13, and phiYY, utilize their heteromeric P3 protein to attach to the lipopolysaccharide (LPS) on the cell surface (Leonard Mindich *et al*. 1999). The P3 protein of these species differs in its composition as it contains two or three different polypeptides (P3a, P3b and, in some cases, P3c). The P6 protein is activated following the removal of P3 and then mediates the fusion of the viral membrane with the host membrane to release the NC into the periplasmic space.

The loss of viral membrane around the NC enables the muralytic (peptidoglycan-degrading) enzyme P5, located on the NC surface, to degrade the peptidoglycan layer of the bacterial cell wall (Leonard Mindich and Lehman 1979; Caldentey and Bamford 1992). The permeabilization of the host plasma membrane facilitates the translocation of the NC across the cytoplasmic membrane of the host cell through an endocytosis-like process, driven by P8 (Poranen *et al*. 1999; Romantschuk, Olkkonen, and Bamford 1988). Upon entry into the cytoplasm, the P8 shell of the NC dissociates to reveal the naked dodecahedral polymerase complex (PC). The release of P8 stimulates the PC, which is transcriptionally active. This is the characteristic mechanism dsRNA viruses exploit in order to replicate their genome—delivery of the nucleic acids in a specialized icosahedral capsule containing the necessary RNA metabolism enzymes such as mRNA synthesizing enzymes. This nano-compartment enables the dsRNA genome to remain "hidden" from any antiviral mechanisms of the host and avoids dsRNA induced host responses (Poranen and Bamford 2012). It also provides a safe environment for phage replication and translation. The dimeric P7 protein acts as an assembly and packaging cofactor by accelerating the rate of immature PC assembly through stabilization of the entire complex (Juuti and Bamford 1997; Poranen *et al*. 2001).

The core particle is composed of P1, P2, P4, and P7. These proteins are involved in the transcription of the phi6 genome. The monomeric RdRP of the P2 gene is activated by PC entry into the cytoplasm. This enzyme catalyzes the semi-conservative transcription of polycistronic mRNAs within the core particle (Usala, Brownstein, and Haselkorn 1980). Bacterial hosts lack the capability to synthesize complementary strands from the RNA template, so all characterized RNA viruses, including phages, encode their own enzymes. The RdRP attaches to the 3′ end of the single-stranded mRNA transcripts and through primer-independent de novo initiation it efficiently replicates and transcribes the phage genome (T. Blumenthal 1980; Silverman 1973). The suggested transcription mechanism involves the dsRNA genome unwinding as it is pulled

through one channel of P2 and nucleotide triphosphates (NTPs), oligonucleotides, manganese (Mn^{2+}) and magnesium (Mg^{2+}) ions entering through another (Butcher *et al*. 2001). Initially the template strand overextends and it locks into a "specificity pocket" (Butcher *et al*. 2001). The strand then reverses, in the presence of two cognate NTPs, to form the functional initiation complex. The reaction is primed through the activity of one of the NTPs as it serves as the carboxyl-terminal domain of the protein. It has been suggested that *Cystoviridae* control transcription through an interchange of two independent mechanisms (a) plus-sense initiation sites are preferred by the polymerase and (b) initiation competent ssRNA templates have more available transcription initiation sites (H. Yang *et al*. 2003). Initiation is the rate-limiting step of transcription, located at the 3′-terminal cytidine nucleotide of the −ssRNA template.

By directly releasing the mRNA transcripts into the cytoplasm, the dsRNA genome is never exposed to the host cytoplasm which helps the phage to avoid host defense mechanism activation. The mRNA transcripts are used as templates for translation of the necessary proteins. The early stage of infection is characterized by equal amounts of mRNA from L, M, and S segments (Coplin *et al*. 1975; Emori, Iba, and Okada 1983). However, only the L-segment transcripts are efficiently produced in this early stage, to give rise to an increased level of PC proteins and the formation of empty PCs.

The large free-strand +ssRNA is then translated to form P1, P2, P4, and P7, which are subsequently assembled to form empty PCs (Carpino 2014). The hexameric nucleoside triphosphatase (NTPase) motor of P4 directs the bundling of the three genome segments in the form of +ssRNA into the empty PCs by recognition of 5′ *pac* sequences (Frilander and Bamford 1995; Leonard Mindich 1999). This packaging is controlled through the expression of segment-specific binding sites on the PC. Binding sites specific for the S-segment are exposed initially to allow P4 to package the S-segment into the empty PC. A conformational change of the PC

alters the binding site to become M-segment specific to package this segment of the genome into the viral progeny. Another change allows the packaging of the L-segment. Once the PC expands to a threshold size, these +ssRNA transcripts are then converted into dsRNA by a single round of negative strand synthesis of RdRP P2 (Poranen, Tuma, and Bamford 2005). Studies by Pirttimaa and colleagues (2002) found that of the 12 P4 hexamers, one is both functionally and structurally unique (Pirttimaa *et al*. 2002). Although studies focused on the basic molecular mechanisms of phages have exploded in recent years, the exact transcriptional and translational processes of *Cystoviridae* are yet to be fully described in exact detail.

The size and organization of this PC is regulated through the activity of inner capsid protein P1 and P4 (Poranen *et al*. 2001; J. Qiao *et al*. 2003; X. Qiao, Qiao, and Mindich 2003). P1 is conserved throughout dsRNA viruses, although it appears to vary in multimeric status (Kainov *et al*. 2003; Poranen *et al*. 2001). Transcription is initiated following effective replication of the dsRNA genome. As the infection progresses, the M and S segment mRNA predominate to produce the proteins essential to virion assembly. The naked PC is encapsulated in a newly synthesized NC shell. The membrane protein P9, along with morphogenic P12, have crucial roles in construction of a new phospholipid membrane around the NC particle from the host plasma membrane (Stitt and Mindich 1983). The spike protein complex of P3 and P6 is the last component attached to the surface, to ensure the progeny are capable of receptor recognition.

Cystoviridae are categorized as virulent phages as they induce lysis of their host bacterium at the end of the infection cycle in order to release viral progeny, through P5 and P10 activity (Leonard Mindich and Lehman 1979; Caldentey and Bamford 1992). However, recent findings have shown that phi6 is capable of forming a pseudolysogenic carrier state within its host (Onodera *et al*. 1992). *Cystoviridae* species phage phi6 targets the Gram-

negative bacterium, *Pseudomonas syringae*, an important plant pathogen. This phage was first isolated in the 1970's in the USA from *Pseudomonas*-infected bean straw (Vidaver, Koski, and Etten 1973).

There have recently been six additional *Cystoviridae* isolated and characterized in the 2017 ICTV report with another five requiring further analysis. Their genetic and structural similarities with phage phi6 suggest that there will be an expansion of this phage taxonomic family with further classification required. Sampling of various legumes in the USA have resulted in the isolation of additional dsRNA phages but these have not been characterized beyond their sequences (Leonard Mindich *et al*. 1999; O'Keefe *et al*. 2010; Silander *et al*. 2005). Assorted environmental sources in Europe and Asia have yielded more novel dsRNA phages: *Pseudomonas* phage phiNN was isolated from a freshwater sample in Finland, while *Pseudomonas* phage phiYY came from hospital sewage waste in China (Mäntynen *et al*. 2015; Y. Yang *et al*. 2016). Phage isolate phiYY has been found to target *P. aeruginosa* strains, an opportunistic pathogen of immuno-compromised individuals. This suggests there may be potential to develop a phage therapy to combat *Pseudomonas* infections in these individuals.

It is clear from the recent isolations of *Cystoviridae* from multiple environments, with only a single member infecting a Gram-positive host, that there are many more RNA phages yet to be discovered. Recently, Alphonse and Ghose (2017) examined known *Cystoviridae* using their encoded RdRP (Alphonse and Ghose 2017). While ssRNA phage genomes have high mutation rates (Drake 1993), RdRP appears to be conserved amongst RNA phage genomes and thus might be a good candidate as a genetic signature to identify further RNA phage sequences. However, identification of *Cystoviridae* in metagenomic datasets using a marker such as the RdRP is complicated by the tri-segmented nature of the *Cystoviridae* genomes. Therefore, sequence-based detection of all three genomic segments of *Cystoviridae*,

particularly if they are divergent from sequences present in public repositories, will be challenging. Incorporation of genetic tags from each of the three segments will greatly enhance de novo efforts of finding *Cystoviridae* members.

1.4 *Leviviridae*

The *Leviviridae* family encompasses phages with a positive-sense single stranded, monopartite RNA genome of 3.3–4.3 kb in length. The non-enveloped, somewhat spherical virion capsid is composed of 178 copies of the dimeric coat protein (CP) and a single copy of the maturation protein (Figure 2). The 5′ end of the genome carries a triphosphate cap.

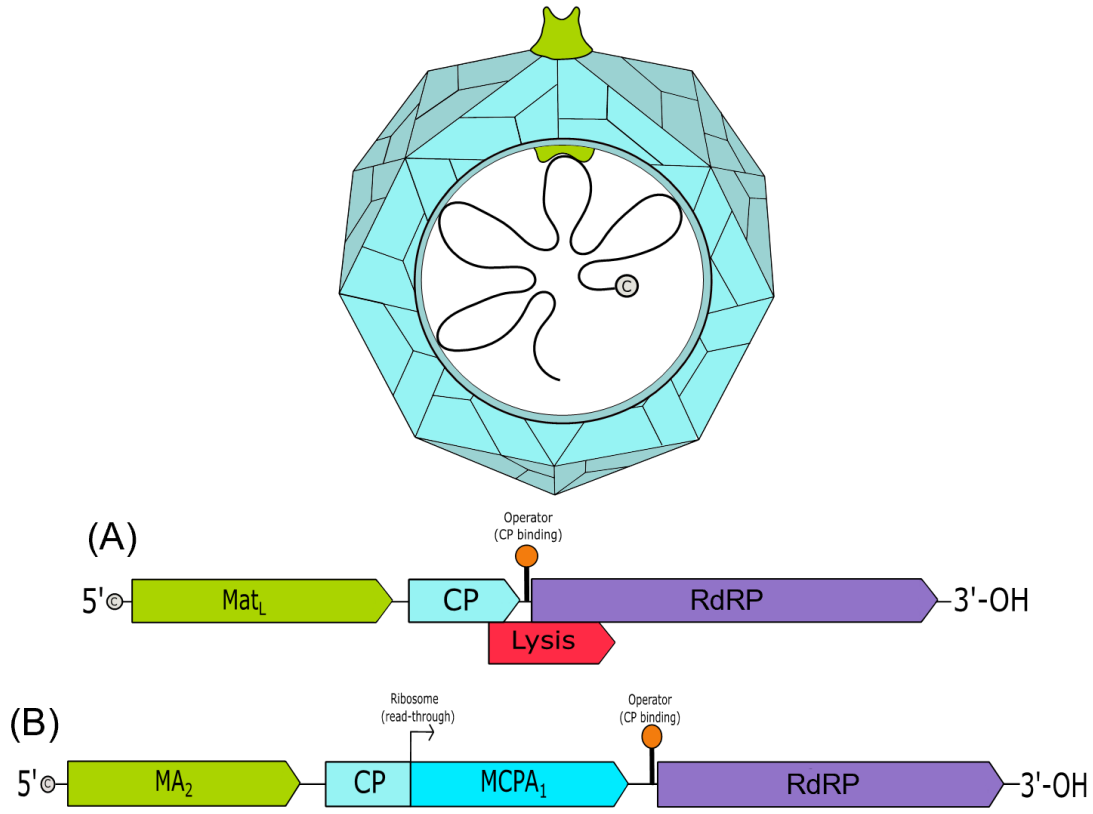


Figure 2. Virion of typical *Leviviridae* family. **(A)** Genome of Enterobacteria phage MS2, an example of a *Levivirus* (3,569 bp). **(B)** Genome of Enterobacteria phage Qbeta, an example of an *Allolevivirus* (4,215 bp). The genomes and the virion structures are color-coded. (Mat_L = maturation protein of *Levivirus;* MA_2 = maturation protein A_2 of *Allolevivirus;* CP = Coat Protein; $MCPA_1$ = Minor-CP A_1 of *Allolevivirus*; RdRP = RNA-directed RNA polymerase). (These figures were created based on a previous depiction (Olsthoorn and van Duin 2017; "Viral Zone: *Leviviridae*" 2018)).

There are two genera of *Leviviridae*: *Levivirus* and *Allolevivirus*. These genera were historically differentiated through serological cross-reactivity, sedimentation, molecular weight and density (Olsthoorn and van Duin 2011). More recently, the number of known genes in their genomes have been used to distinguish between *Levivirus* and *Allolevivirus* members, with three and four, respectively (Figure 2). These genera are subdivided into genogroups; *Levivirus* has MS2-like (genogroup I) and BZ13-like (genogroup II) and *Allolevivirus* has Qbeta-like (genogroup III) and F1-like (genogroup IV) (Olsthoorn and van Duin 2017).

Leviviridae phages that target *E. coli*, known as coliphages, are male-specific, adsorb along the fertility (F) pilus, coded by the F-plasmid of *Escherichia coli*, or the chromosomal marker Hfr, whereas in non-coliphage species alternative pili are exploited (Zinder 1965). Alternatively, coliphages that can infect cells via the cell wall are classified as somatic (Dryden *et al*. 2006). The presence of enteroviruses in water from pollution is often detected through the identification of RNA coliphages as biomarkers (Cole, Long, and Sobsey 2003). Phages that utilize F-pili are classified as male-specific phages. The way in which the *Leviviridae* phages induce lysis of their host is a notable difference between the genera; *Levivirus* phages encode a separate lysis polypeptide, whereas *Allolevivirus* phages utilise their maturation protein in lysis mediation (Karnik and Billeter 1983; Young 1992). These proteins are two canonical "single gene lysis" (SGL) systems that are utilised by small phages, the third is the *E* lysin from phage φX174, a ssDNA *Microviridae* representative (Chamakura, Edwards, and Young 2017). The lysis mechanism, and specific protein where applicable, is fundamental to the lifecycle of the phage.

1.4.1 *Levivirus*

The type species of *Levivirus* is the *Enterobacteria* phage MS2, a member of the MS2-like phages (genogroup I). Phages of the *Levivirus* genus infect their host targets through the initial adsorption of the virion along the sides of pili using the maturation A-protein (Mat_L) as the receptor binding protein (Roberts and Steitz 1967). This results in the self-proteolytic cleavage of the A-protein into at least two fragments and a structural change of the F-pilus (Krahn, O'Callaghan, and Paranchych 1972). This induces the release of the phage RNA into the host bacterium. Studies have reported that the two largest polypeptide components are transferred into the host along with the genomic RNA (Krahn, O'Callaghan, and Paranchych 1972). The fragmented Mat_L binds the RNA at two distinct regions: the Mat_L coding region and the 3′-UTR (Shiba and Suzuki 1981). It appears that Mat_L-RNA complex may be injected into the cell as opposed to free RNA, suggesting that the Mat_L protein may have a greater biological role than originally envisaged (Krahn, O'Callaghan, and Paranchych 1972). The exact mechanism of how the Mat_L-RNA complex gains entry to the host remains undescribed, but could involve a type IV secretion system (T4SS) homolog (Zechner, Lang, and Schildbach 2012). It has been postulated that Mat_L may also contribute to the replication process of the RNA genome.

As the nucleic acid is a single copy of +ssRNA, it functions both as the genome template and mRNA upon infection. Thus, there is constant competition between replication and translation processes as the ribosome and replicase run in opposite directions along the template strand (Eigen *et al*. 1991). The two events are independent of each other with the secondary structures of the +ssRNA strand and formation of a complementary negative strand of RNA maintaining this equilibrium. It has been noted that in the 3′-terminal sequence of the *Levivirus* genomes, there is a signature sequence of 5′-ACCACCCA-3′ (Friedman *et al*. 2009).

For effective genome replication, leviviruses encode a copy of RdRP that codes for the catalytic ß-subunit of the replicase. This protein associates with three host proteins: ribosomal protein S1 (Wahba *et al*. 1974) and the translational elongation factors EF-Tu and EF-Ts (Blumenthal, Landers, and Weber 1972), to form a functional polymerase unit, the holoenzyme. The role of EF-Tu has been established as delivering an aminoacyl-tRNA to the ribosome when in its GTP-bound form (Agirrezabala and Frank 2009). This GTP is hydrolyzed to form GDP-bound EF-Tu following a codon anti-codon match within the ribosomal complex. This displaces the EF-Tu and EF-Ts binds to the GDP-bound EF-Tu and removes the GDP molecule. This allows the EF-Tu to be recycled for further elongation rounds (Schmeing *et al*. 2009; Schuette *et al*. 2009). Sequestration of these elongation factors inhibits initiation of translation. The S1 protein functions as a translational initiation factor. The sole purpose of this protein is to recognize the template plus strand, the core-complex of the three remaining proteins is sufficient to synthesize new +ssRNA strands (Kamen *et al*. 1972).

Studies have shown that there are two internal sequences which are key to the recognition of the plus strand by the replicase, the S site and M site. (Meyer, Weber, and Weissmann 1981). The S site is described as being a uracil rich sequence of approximately 100 nucleotides, located just before the initiation codon of the coat protein. The secondary structure of the S site is poorly defined. The M site is of similar length, forms a branched stem-loop structure and resides within the replicase coding region (Schuppli *et al*. 1998). These two sites are simultaneously bound by the S1 protein to allow for effective replication by the replicase through enhanced recognition of the template to the active site (Miranda *et al*. 1997).

The RNA template is protected from cellular nuclease degradation through an unknown mechanism. There is an additional host factor required for successful translation that has been isolated but not genetically identified in the case of *Levivirus* species. This protein does not

interact with the polymerase machinery but instead binds directly to the 3′ terminal of the mRNA template (Schuppli, Georgijevic, and Weber 2000). The replicase will associate to the start site and initiate negative-strand synthesis by replicating through the genome. This strand is used to synthesize new +ssRNA genomes for the viral progeny.

As the infection cycle reaches the end-stage, the CP-dimers bind the replicase gene start site, located within a hairpin-structured operator, and act as translational repressors (H. Robertson, Webster, and Zinder 1968; Valegrad *et al*. 1994). This results in a packaging signal that stimulates the assembly of functional viral progeny. At the same time, there is an increase in quantities of the lysis protein, with a single lysis protein required for each phage progeny. Since the lysis protein lyses the cell without affecting the integrity of the peptidoglycan network, and in the absence of muralytic enzyme activity, it is referred to as an amurin (Bernhardt *et al*. 2002). Research focused on this protein has revealed that it is primarily localized in Bayer's patches, the periplasmic zones of adhesion between the inner and outer membrane (Walderich and Höltje 1989). The exact mechanism by which this 75-amino acid lysin induces host lysis is not exactly known (Beremand and Blumenthal 1979). However, the current proposal is that the lysis protein forms lesions and hydrophobic pores in the inner membrane that dissipates the proton motive force (PMF) (Goessens *et al*. 1988). This alteration in PMF activates autolysis of the bacterial host through certain enzymes such as DD-endopeptidases and lytic transglycosylases. Supporting research has shown alteration in the average length in glycan strands and degree of cross-linkage, suggesting the activation of the aforementioned enzymes (Walderich *et al*. 1988).

Nonetheless, the molecular information and functioning schema of such an autolytic pathway have yet to be identified (Chamakura, Edwards, and Young 2017). Recent findings have indicated that the lysis of host cells by MS2 lysin is dependent on a range of host factors,

including host chaperone DnaJ (Chamakura, Tran, and Young 2017). This post-translational regulator allows for another level of control of both quantity and activity of the lysis protein of MS2.

Translation of the phage genes requires the ribosome to associate with the RNA through a Shine–Dalgarno sequence, the start codon and the host ribosomal S1 protein. The S1 protein can bind the S site, as mentioned above. This creates a situation whereby the S1 protein of the ribosome and replicase are competing for the same RNA binding site.

There are a variety of systems that regulate protein synthesis, including: RNA secondary structure, ribosome access to the initiation codon, and folding kinetics (Lodish 1970; Kozak 1983; Poot *et al*. 1997). The secondary structures of the +ssRNA are the predominant factors in determining different protein yields; e.g., the CP gene is free from any secondary structures as it is required in high copy numbers (178 per virion), whereas the replicase gene is trapped in tight secondary structures as only one copy per progeny is required (Hans Weber 1976). The open reading frame (ORF) of the coat protein is readily available for the ribosomal translation. As the ribosome moves along the RNA transcript, it disrupts the secondary structure to allow hidden genes to be translated. Following CP gene translation, the initiation codon of the replicase gene becomes available, resulting in the synthesis of the replicase ß-catalytic subunit. The translation of the lysis and replicase gene is dependent on successful translation of the CP gene.

Newly synthesized viral particles require only one copy of the lysis protein and the Mat_L protein (Groeneveld, Thimon, and van Duin 1995; Reed *et al*. 2013; Rumnieks and Tars 2017). The ORF of the lysis gene overlaps the replicase gene in a +1 frameshift, with the termination sequence of the lysis protein located in the coding region of the replicase gene (Atkins *et al*. 1979; Beremand and Blumenthal 1979). Studies by van Duin and his colleagues

(1990) on the translational control of the lysis protein provided key information as to the role of secondary structures in transcriptional regulation. Their work demonstrated that the formation of a stable hairpin in the RNA between the Shine–Dalgarno sequence and the start codon of the lysis gene, represses the expression of the lysis gene. Following successful transcription, during translation there is incomplete dissociation of the ribosome from the mRNA as it creates the CP protein (Berkhout *et al*. 1987; Schmidt *et al*. 1987). The ribosome backtracks to reinitiate at the start codon for the lysis protein in approximately 5% of translational cycles (Khazaie, Buchanan, and Rosenberger 1984; Weiner and Weber 1971). The lysis protein is produced at low levels towards the ends of the infection cycle. This allows for gradual accumulation of the lysis protein to ensure that the viral progeny have sufficient time to mature.

The Mat_L protein is only transcribed from newly synthesized genome templates (H. D. Robertson and Lodish 1970). The strong secondary structure formed by the Shine–Dalgarno sequence and the S1-binding sequence prevent translation of the 5′-end in normal mRNA structure, where the Mat_L gene is positioned. In nascent RNA strands, there is an alternate, shorter hairpin structure created that enables translation of the maturation gene in the 5′ terminal by allowing access and binding of the ribosome. This RNA-folding intermediate of newly synthesized strands enable the ribosome access to the start codon of the A-protein.

A recently isolated RNA phage for *Acinetobacter* species, AP205, was found to have an unusual genome structure with the lysis gene located in the 5′ terminal (Klovins *et al*. 2002). Although the genome of AP205 mirrors the typical *Levivirus* genome map, the secondary structure and 3′-UTR follows that of *Allolevivirus.* This phage has yet to be approved as a *Levivirus.* Potential *Leviviruses* of *Pseudomonas,* phages PPR7 and PRR1, have also been

isolated and characterized, both exhibit particular hallmarks of *Levivirus* phages (Bradley 1966; Olsen and Thomas 1973; R. C. L. Olsthoorn *et al*. 1995; Ruokoranta *et al*. 2006).

1.4.2 *Allolevivirus*

The type species of *Allolevivirus* is Qbeta, the representative of the Qbeta-like phages (genogroup III). Species of *Allolevivirus* contain a longer version of the genome with an extension of the C-terminal of the CP gene (Olsthoorn and van Duin 2011). The presence of this minor-CP A_1 ($MCPA_1$) protein, also known as the read-through protein, is a feature unique to *Allolevivirus* phages (Weissmann *et al*. 1973). Both the $MCPA_1$ and the maturation A_2 (MA_2) proteins are essential for host attachment (Olsthoorn and van Duin 2011). The majority of *Allolevivirus* members were found to encode an Arg-Gly-Asp (RGD) motif, essential for host cell recognition and attachment, within their $MCPA_1$ and/or MA_2 (Friedman *et al*. 2009). This motif is absent in the *Levivirus* phages. Similar to the signature 3′-terminal sequence of *Levivirus*, *Allolevivirus* species contain a 5′-TCCTCCCA-3′ within the 3′-terminal of their genome (Friedman *et al*. 2009).

The underlying translation and replication mechanisms are similar to *Levivirus* with minor variations. The host factor that associates with the functional replicase has been isolated, purified and genetically characterized for Qbeta as the protein encoded by the *host factor of Qbeta* (*hfq*) gene of *E. coli* (Gottesman and Storz 2015; Schuppli *et al*. 1997). This nonspecific ssRNA binding protein, Hfq, aids polymerase association to the 3′ end of the +ssRNA template. The start of the 5′-terminal begins with a GG sequence. There is a nontranslated A residue attached to the extreme 3′ terminus in a CCA sequence, following activity of the terminal nucleotidyl transferase (TNTase) domain of RdRP (Weber and Weissmann 1970; Blumenthal

and Carmichael 1979; Bausch *et al*. 1983). This does not serve as a template nucleotide, instead RNA synthesis begins at the penultimate C residue.

The transition between replication and translation is similar to the above mentioned *Levivirus*, with a slight change; the translation of the MA_2 is controlled in a temporal manner as opposed to a structural intermediate. This is dictated by the length of time it takes for the polymerase to move from the start site of the maturation gene to the complement of the Shine–Dalgarno sequence (Beekwilder, Nieuwenhuizen, and Poot 1996; Staples *et al*. 1971). Once it has been translated, these two sequences bind to form a strong secondary structure to prevent continuous translation of the same gene. The additional $MCPA_1$ protein is formed following ribosomal read-through of the leaky-stop codon (UGA) of the CP gene (Weiner and Weber 1971). It is read as a tryptophan codon (UGG), which promotes gene expression of the $MCPA_1$ protein. The ribosome occasionally, in approximately 5% of cases, translates past this leaky termination sequence for an additional 600 nucleotides to form a C-terminal extension of the CP (Rumnieks and Tars 2011). This protein is incorporated in low quantities into viral progeny and is essential for successful infection. Studies of the amino acid sequence and the three-dimensional structure of the $MCPA_1$ protein, have shown it to be unique to the small group of *Allolevivirus* phages (Rumnieks and Tars 2011). The MA_2 and the $MCPA_1$ protein, whose exact role is unknown yet, are essential for successful infection of pili-positive hosts

Another notable difference in the infection pattern of *Allolevivirus* is the absence of a lysis gene in the genome. Instead, the MA_2 protein has a secondary function to induce the lysis of the host cell for release of viral progeny (Kastelein *et al*. 1982). The MA_2 protein is referred to as an amurin as it does not destroy the peptidoglycan layer directly through muralytic activity. It is also known as a "protein antibiotic" due to the similarity in function to antibacterial agents which target cell walls (Bernhardt *et al*. 2001). It has been reported that

MA_2 induces host cell lysis by inhibiting the enzymatic activity of MurA, a UDP-*N*-acetylglucosamine-enolpyruvyl transferase. This is an essential enzyme in the production of peptidoglycan as it catalyses the first committed step, the biosynthesis of murein precursor (Brown *et al*. 1995). At the next stage of cell division, the inhibition of cell wall biosynthesis leads to host lysis and release of the phage progeny.

A study by Friedman *et al*. (2009), noted that the sequences of both *Levivirus* and *Allolevivirus* genera had strong homogeny across position of ORF, length of proteins and the catalytic ß-domains of the RdRP (Friedman *et al*. 2009). The conservation of the YGDD motif of the replicase protein across all +ssRNA viruses was recorded throughout the *Leviviridae.*

Although both *Levivirus* and *Allolevivirus* phages target the pilus of their hosts as receptors to initiate, the fact there is no conserved infection mechanism suggests that there may be varying mechanisms for the RNA to enter the cell. Originally thought to only affect plasmid-encoded appendages, there have been *Leviviridae* specific for genome-encoded pili of Gram-negative bacteria, such as *Pseudomonas* phage PP7 and *Acinetobacter* phage AP205.

1.5 Discussion

Although there have only been a limited number of RNA phages identified to date, their true diversity and abundance in nature remains unknown. Current approaches used for the isolation, selection, and purification of viral particles, including precipitation by polyethylene glycol (PEG) and caesium chloride (CsCl) gradient purification, are almost certainly biased against RNA phages (Grasis 2018). The selection of DNA phages in these methods goes a long way to explaining why RNA phages are under-represented in genome databases.

The fragile nature of RNA and the widespread presence of RNases in human and animal derived samples also hinders studies involving RNA phages. The development of RNA phage-selective isolation protocols will also greatly enhance our endeavors. For example, separation of DNA and RNA fractions of samples and complete eradication of unwanted RNase is recommended. It should also be noted that the low abundance of RNA phages in databases will result in reduced hits for novel sequences. As research into the RNA section of the phage community is expanded, the databases are expected to become more representative of the wider RNA phage community.

An interesting paper recently proposed that members of the *Picobirnaviridae* family may not be eukaryotic viruses, as originally thought, but may in fact represent a novel family of RNA phages (Krishnamurthy and Wang 2018). This research involved analysis of bacterial ribosome binding sites (RBS) upstream of the coding sequences in their bi-segmented, dsRNA genomes. It was noted that an RBS motif, thought to be unique to prokaryotic-infecting viruses, was enriched in the picobirnaviruses. This finding suggests that these dsRNA viruses could be classified as putative bacteriophages. Furthermore, an additional study has supported this hypothesis by proposing that picobirnaviruses are in fact a novel RNA phage family of high genomic diversity (Adriaenssens *et al*. 2018). This type of analysis demonstrates the possibility that more members of RNA virus populations may in fact be mischaracterized. A more robust method for classification of RNA phages would help to resolve this issue.

Identifiable RNA phage-specific domains, such as the RdRP gene, capsid gene, maturation protein gene, or the NTPase gene, can serve as features which one could use to mine metagenomic databases for RNA phages. However, since the RdRP gene is conserved amongst RNA viruses, unique genetic elements of *Leviviridae* and *Cystoviridae* families should also be used in specific studies. Contigs with homologs to both the leviviral and cystoviral RdRP gene

are potential RNA phages and should be subjected to further analysis. Based on the recent studies mentioned above, homologs to the RdRP gene of picobirnaviruses should also be included (Krishnamurthy and Wang 2018). The study by Krishnamurthy and colleagues which identified 20 unique RNA phage phylotypes utilized nucleotide identity to the RdRP and the maturation gene to categorize these phages (Krishnamurthy *et al*. 2016). The specific 3′-terminal sequences of *Levivirus* and *Allolevivirus* members could be used to further classify these phages. Signature features of *Cystoviridae* members, such as the muralytic enzyme gene or the nucleocapsid shell protein gene, could also serve as genetic signatures when screening the databases for RNA phages (Yang *et al*. 2016).

A common theme of this review is the need for greater efforts to be directed towards the discovery of more RNA phages for all potential applications, such as tools for advancing molecular biology and as potential phage therapeutics. The rise in antimicrobial resistance across bacteria is not a novel problem but it is alarming. The host range of RNA phages could offer therapeutic potential against some of the World Health Organizations' (WHO) list of deadly pathogens, including some of the Gram-negative members of the ESKAPE pathogens, such as *Klebsiella pneumoniae*, *Acinetobacter baumannii*, and *Pseudomonas aeruginosa*. Clinical isolates of *P. aeruginosa* have been found to be resistant to most of the antibiotics normally used to treat this infection (Lister, Wolter, and Hanson 2009). An unclassified *Levivirus P. aeruginosa* phage PP7 has been identified which targets this bacterium via a pilin-specific mechanism (Kim, Bae, and Cho 2018). Further studies regarding the therapeutic parameters of RNA phages, such as PP7, should be done to examine their efficiency to control these pathogens and to explore their potential use as components of cocktails used in phage therapy.

1.6 References

Adcock, Noreen J., Eugene W. Rice, Mano Sivaganesan, Justin D. Brown, David E. Stallknecht, and David E. Swayne. 2009. "The Use of Bacteriophages of the Family *Cystoviridae* as Surrogates for H5N1 Highly Pathogenic Avian Influenza Viruses in Persistence and Inactivation Studies." *Journal of Environmental Science and Health. Part A, Toxic/Hazardous Substances & Environmental Engineering* 44 (13): 1362–66. https://doi.org/10.1080/10934520903217054.

Adriaenssens, Evelien, Kata Farkas, Christian Harrison, David Jones, Heather E Allison, and Alan J McCarthy. 2018. "Viromic Analysis of Wastewater Input to a River Catchment Reveals a Diverse Assemblage of RNA Viruses," February. https://doi.org/10.1101/248203.

Agirrezabala, Xabier, and Joachim Frank. 2009. "Elongation in Translation as a Dynamic Interaction among the Ribosome, TRNA, and Elongation Factors EF-G and EF-Tu." *Quarterly Reviews of Biophysics* 42 (3): 159–200. https://doi.org/10.1017/S0033583509990060.

Alphonse, Sébastien, and Ranajeet Ghose. 2017. "Cystoviral RNA-Directed RNA Polymerases: Regulation of RNA Synthesis on Multiple Time and Length Scales." *Virus Research*, Viral polymerases, 234 (April): 135–52. https://doi.org/10.1016/j.virusres.2017.01.006.

Atkins, John F., Joan A. Steitz, Carl W. Anderson, and Peter Model. 1979. "Binding of Mammalian Ribosomes to MS2 Phage RNA Reveals an Overlapping Gene Encoding a Lysis Function." *Cell* 18 (2): 247–56. https://doi.org/10.1016/0092-8674(79)90044-8.

Bamford, D. H., E. T. Palva, and K. Lounatmaa. 1976. "Ultrastructure and Life Cycle of the Lipid-Containing Bacteriophage Φ6." *Journal of General Virology* 32 (2): 249–59. https://doi.org/10.1099/0022-1317-32-2-249.

Bausch, James N, Fred Russell Kramer, Eleanor A Miele, Carl Dobkin, and Donald R Mills. 1983. “Terminal Adenylation in the Synthesis of RNA by Qβ Replicase” 258 (3): 1978–84.

Becker, Jordan T., and Nathan M. Sherer. 2017. “Subcellular Localization of HIV-1 Gag-Pol MRNAs Regulates Sites of Virion Assembly.” Edited by Frank Kirchhoff. *Journal of Virology* 91 (6): e02315-16. https://doi.org/10.1128/JVI.02315-16.

Beekwilder, Jules, Rob Nieuwenhuizen, and Raymond Poot. 1996. “Secondary Structure Model for the First Three Domains of Qβ RNA. Control of A-Protein Synthesis.” *Journal of Molecular Biology* 256: 8–19.

Beremand, Marian N., and Thomas Blumenthal. 1979. “Overlapping Genes in RNA Phage: A New Protein Implicated in Lysis.” *Cell* 18 (2): 257–66. https://doi.org/10.1016/0092-8674(79)90045-X.

Berkhout, Ben, Brian F. Schmidt, Anja van Strien, Jacques van Boom, Jeroen van Westrenen, and Jan van Duin. 1987. “Lysis Gene of Bacteriophage MS2 Is Activated by Translation Termination at the Overlapping Coat Gene.” *Journal of Molecular Biology* 195 (3): 517–24. https://doi.org/10.1016/0022-2836(87)90180-X.

Bernhardt, Thomas G., Ing-Nang Wang, Douglas K. Struck, and Ry Young. 2002. “Breaking Free: ‘Protein Antibiotics’ and Phage Lysis.” *Research in Microbiology* 153 (8): 493–501. https://doi.org/10.1016/S0923-2508(02)01330-X.

Bernhardt, Thomas, Ing-Nang Wang, Douglas K. Struck, and Ryland Young. 2001. “A Protein Antibiotic in the Phage Qβ Virion: Diversity in Lysis Targets.” *Science* 292 (5525): 2326–29. https://doi.org/10.1126/science.1058289.

Blumenthal, Thomas 1980. “Qbeta Replicase Template Specificity: Different Templates Require Different GTP Concentrations for Initiation.” *Proceedings of the National Academy of Sciences of the United States of America* 77 (5): 2601.

Blumenthal, Thomas, and G. G. Carmichael. 1979. “RNA Replication: Function and Structure of Qbeta-Replicase.” *Annual Review of Biochemistry* 48: 525–48. https://doi.org/10.1146/annurev.bi.48.070179.002521.

Blumenthal, Thomas, Terry A. Landers, and Klaus Weber. 1972. “Bacteriophage Qβ Replicase Contains the Protein Biosynthesis Elongation Factors EF Tu and EF Ts.” *Proceedings of the National Academy of Sciences* 69 (5): 1313–17. https://doi.org/10.1073/pnas.69.5.1313.

Bradley, D. 1966. “The Structure and Infective Process of a Pseudomonas Aeruginosa Bacteriophage Containing Ribonucleic Acid.” *Journal of General Microbiology* 45 (1): 83–96. https://doi.org/10.1099/00221287-45-1-83.

Brown, D., and L. Gold. 1996. “RNA Replication by Q Beta Replicase: A Working Model.” *Proceedings of the National Academy of Sciences* 93 (21): 11558–62. https://doi.org/10.1073/pnas.93.21.11558.

Brown, D., E. I. Vivas, C. T. Walsh, and R Kolter. 1995. “MurA (MurZ), the Enzyme That Catalyzes the First Committed Step in Peptidoglycan Biosynthesis, Is Essential in Escherichia Coli.” *Journal of Bacteriology* 177 (14): 4194–97. https://doi.org/10.1128/jb.177.14.4194-4197.1995.

Butcher, Sarah J, Jonathan M Grimes, Eugeny V Makeyev, Dennis H Bamford, and David I Stuart. 2001. “A Mechanism for Initiating RNA-Dependent RNA Polymerization” 410: 6.

Caldentey, Javier, and Dennis H. Bamford. 1992. “The Lytic Enzyme of the Pseudomonas Phage Φ6. Purification and Biochemical Characterization.” *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology* 1159 (1): 44–50. https://doi.org/10.1016/0167-4838(92)90073-M.

Carlton, RM. 1999. “Phage Therapy: Past History and Future Prospects.” *ARCHIVUM IMMUNOLOGIAE ET THERAPIAE EXPERIMENTALIS-ENGLISH EDITION* 47 (June): 267–74.

Carpino, James. 2014. “Structure and Function in Bacteriophage Phi6.” *CUNY Academic Works.*, June. https://academicworks.cuny.edu/gc_etds/183.

Cenens, William, Angela Makumi, Sander K. Govers, Rob Lavigne, and Abram Aertsen. 2015. “Viral Transmission Dynamics at Single-Cell Resolution Reveal Transiently Immune Subpopulations Caused by a Carrier State Association.” Edited by Josep Casadesús. *PLOS Genetics* 11 (12): e1005770. https://doi.org/10.1371/journal.pgen.1005770.

Chamakura, Karthik R., Garrett B. Edwards, and Ry Young. 2017. “Mutational Analysis of the MS2 Lysis Protein L.” *Microbiology* 163 (7): 961–69. https://doi.org/10.1099/mic.0.000485.

Chamakura, Karthik R., Jennifer S. Tran, and Ry Young. 2017. “MS2 Lysis of *Escherichia Coli* Depends on Host Chaperone DnaJ.” *Journal of Bacteriology* 199 (12): e00058-17. https://doi.org/10.1128/JB.00058-17.

Cole, Dana, Sharon C. Long, and Mark D. Sobsey. 2003. “Evaluation of F+ RNA and DNA Coliphages as Source-Specific Indicators of Fecal Contamination in Surface Waters.” *Applied and Environmental Microbiology* 69 (11): 6507–14. https://doi.org/10.1128/AEM.69.11.6507-6514.2003.

Coplin, D. L., J. L. Van Etten, R. K. Koski, and A. K. Vidaver. 1975. “Intermediates in the Biosynthesis of Double-Stranded Ribonucleic Acids of Bacteriophage Phi 6.” *Proceedings of the National Academy of Sciences* 72 (3): 849–53. https://doi.org/10.1073/pnas.72.3.849.

Cruz, Fernando de la, and Julian Davies. 2000. “Horizontal Gene Transfer and the Origin of Species: Lessons from Bacteria.” *Trends in Microbiology* 8 (3): 128–33. https://doi.org/10.1016/S0966-842X(00)01703-0.

Davis, James E, James H. Strauss, and Robert L Sinsheimer. 1961. “Bacteriophage MS2: Another RNA Phage.” *Science* 14 (March): 1427.

Drake, J. W. 1993. “Rates of Spontaneous Mutation among RNA Viruses.” *Proceedings of the National Academy of Sciences* 90 (9): 4171–75. https://doi.org/10.1073/pnas.90.9.4171.

Dryden, S.K., B. Ramaswami, Z. Yuan, D.E. Giammar, and L.T. Angenent. 2006. “A Rapid Reverse Transcription-PCR Assay for F+ RNA Coliphages to Trace Fecal Pollution in Table Rock Lake on the Arkansas–Missouri Border.” *Water Research* 40 (20): 3719–24. https://doi.org/10.1016/j.watres.2006.09.003.

Eigen, Manfred, Christof K. Biebricher, Michael Gebinoga, and William C. Gardiner. 1991. “The Hypercycle. Coupling of RNA and Protein Biosynthesis in the Infection Cycle of an RNA Bacteriophage.” *Biochemistry* 30 (46): 11005–18. https://doi.org/10.1021/bi00110a001.

Emori, Y, H Iba, and Y Okada. 1983. “Transcriptional Regulation of Three Double-Stranded RNA Segments of Bacteriophage Phi 6 in Vitro.” *Journal of Virology* 46 (1): 196–203.

Etten, James van, Les Lane, Carlos Gonzalez, James Partridge, and Anne Vidaver. 1976. “Comparative Properties of Bacteriophage 46 and 46 Nucleocapsid” 18: 7.

Fiers, W., R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaert, W. Min Jou, *et al*. 1976. “Complete Nucleotide Sequence of Bacteriophage MS2 RNA: Primary and Secondary Structure of the Replicase Gene.” *Nature* 260 (5551): 500–507. https://doi.org/10.1038/260500a0.

Friedman, Stephanie D., Fred J. Genthner, Jennifer Gentry, Mark D. Sobsey, and Jan Vinjé. 2009. “Gene Mapping and Phylogenetic Analysis of the Complete Genome from 30 Single-Stranded RNA Male-Specific Coliphages (Family Leviviridae).” *Journal of Virology* 83 (21): 11233–43. https://doi.org/10.1128/JVI.01308-09.

Frilander, Mikko, and Dennis H. Bamford. 1995. “In VitroPackaging of the Single-Stranded RNA Genomic Precursors of the Segmented Double-Stranded RNA Bacteriophage ψ: The Three Segments Modulate Each Other’s Packaging Efficiency.” *Journal of Molecular Biology* 246 (3): 418–28. https://doi.org/10.1006/jmbi.1994.0096.

Goessens, W H, A J Driessen, J Wilschut, and J van Duin. 1988. “A Synthetic Peptide Corresponding to the C-Terminal 25 Residues of Phage MS2 Coded Lysis Protein Dissipates the Protonmotive Force in Escherichia Coli Membrane Vesicles by Generating Hydrophilic Pores.” *The EMBO Journal* 7 (3): 867–73.

Gottesman, Susan, and Gisela Storz. 2015. “RNA Reflections: Converging on Hfq.” *RNA* 21 (4): 511–12. https://doi.org/10.1261/rna.050047.115.

Gottlieb, Paul, Shulamit Metzger, Martin Romantschuk, Jacob Carton, Jeffrey Strassman, Dennis H. Bamford, Nisse Kalkkinen, and Leonard Mindich. 1988. “Nucleotide Sequence of the Middle DsRNA Segment of Bacteriophage Φ6: Placement of the Genes of Membrane-Associated Proteins.” *Virology* 163 (1): 183–90. https://doi.org/10.1016/0042-6822(88)90245-0.

Gottlieb, Paul, Hui Wei, Christiaan Potgieter, and Igor Toporovsky. 2002. “Characterization of Φ12, a Bacteriophage Related to Φ6: Nucleotide Sequence of the Small and Middle Double-Stranded RNA.” *Virology* 293 (1): 118–24. https://doi.org/10.1006/viro.2001.1288.

Grasis, Juris A. 2018. “Host-Associated Bacteriophage Isolation and Preparation for Viral Metagenomics.” In *Viral Metagenomics*, edited by Vitantonio Pantaleo and Michela

Chiumenti, 1746:1–25. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4939-7683-6_1.

Groeneveld, H, K Thimon, and J van Duin. 1995. “Translational Control of Maturation-Protein Synthesis in Phage MS2: A Role for the Kinetics of RNA Folding?” *RNA* 1 (1): 79–88.

Gytz, Heidi, Durita Mohr, Paulina Seweryn, Yuichi Yoshimura, Zarina Kutlubaeva, Fleur Dolman, Bosene Chelchessa, *et al*. 2015. “Structural Basis for RNA-Genome Recognition during Bacteriophage Qβ Replication.” *Nucleic Acids Research* 43 (22): 10893–906. https://doi.org/10.1093/nar/gkv1212.

Harrison, Ellie, and Michael A. Brockhurst. 2017. “Ecological and Evolutionary Benefits of Temperate Phage: What Does or Doesn’t Kill You Makes You Stronger.” *BioEssays* 39 (12): 1700112. https://doi.org/10.1002/bies.201700112.

Hatfull, Graham F. 2008. “Bacteriophage Genomics.” *Current Opinion in Microbiology* 11 (5): 447–53. https://doi.org/10.1016/j.mib.2008.09.004.

Hatfull, Graham F. 2015. “Dark Matter of the Biosphere: The Amazing World of Bacteriophage Diversity.” *Journal of Virology* 89 (16): 8107–10. https://doi.org/10.1128/JVI.01340-15.

Herelle, Felix d’. 1917. “Sur Un Microbe Invisible Antagoniste Des Bacilles Dysentériques.” *CR Acad. Sci. Paris* 165: 373–75.

Hoogstraten, Deborah, Xueying Qiao, Yang Sun, Aizong Hu, Shiroh Onodera, and Leonard Mindich. 2000. “Characterization of Φ8, a Bacteriophage Containing Three Double-Stranded RNA Genomic Segments and Distantly Related to Φ6.” *Virology* 272 (1): 218–24. https://doi.org/10.1006/viro.2000.0374.

Howard-Varona, Cristina, Katherine R. Hargreaves, Stephen T. Abedon, and Matthew B. Sullivan. 2017. “Lysogeny in Nature: Mechanisms, Impact and Ecology of Temperate Phages.” *The ISME Journal* 11 (7): 1511–20. https://doi.org/10.1038/ismej.2017.16.

Jou, W. Min, G. Haegeman, M. Ysebaert, and W. Fiers. 1972. “Nucleotide Sequence of the Gene Coding for the Bacteriophage MS2 Coat Protein.” *Nature* 237 (5350): 82–88. https://doi.org/10.1038/237082a0.

Juuti, J. T., and D. H. Bamford. 1997. “Protein P7 of Phage Phi6 RNA Polymerase Complex, Acquiring of RNA Packaging Activity by in Vitro Assembly of the Purified Protein onto Deficient Particles.” *Journal of Molecular Biology* 266 (5): 891–900. https://doi.org/10.1006/jmbi.1996.0817.

Kainov, Denis E, Sarah J Butcher, Dennis H Bamford, and Roman Tuma. 2003. “Conserved Intermediates on the Assembly Pathway of Double-Stranded RNA Bacteriophages.” *Journal of Molecular Biology* 328 (4): 791–804. https://doi.org/10.1016/S0022-2836(03)00322-X.

Kamen, Robert, Masatoshi Kondo, Werner Römer, and Charles Weissmann. 1972. “Reconstitution of Qβ Replicase Lacking Subunit α with Protein-Synthesis-Interference Factor i.” *European Journal of Biochemistry* 31 (1): 44–51. https://doi.org/10.1111/j.1432-1033.1972.tb02498.x.

Karnik, Sadashiva, and Martin Billeter. 1983. “The Lysis Function of RNA Bacteriophage Qβ Is Mediated by the Maturation (A2) Protein.” *The EMBO Journal* 2 (9): 1521–26.

Kastelein, R. A., E. Remaut, W. Fiers, and J. van Duin. 1982. “Lysis Gene Expression of RNA Phage MS2 Depends on a Frameshift during Translation of the Overlapping Coat Protein Gene.” *Nature* 295 (5844): 35–41. https://doi.org/10.1038/295035a0.

Kenyon, Julia C., Liam J. Prestwood, and Andrew M. L. Lever. 2015. “A Novel Combined RNA-Protein Interaction Analysis Distinguishes HIV-1 Gag Protein Binding Sites from Structural Change in the Viral RNA Leader.” *Scientific Reports* 5 (1). https://doi.org/10.1038/srep14369.

Khazaie, Khashayarsha, John H. Buchanan, and Robert F. Rosenberger. 1984. “The Accuracy of Qbeta RNA Translation. 1. Errors during the Synthesis of Qbeta Proteins by Intact Escherichia Coli Cells.” *European Journal of Biochemistry* 144 (3): 485–89. https://doi.org/10.1111/j.1432-1033.1984.tb08491.x.

Kim, Eun Sook, Hee-Won Bae, and You-Hee Cho. "A pilin region affecting host range of the Pseudomonas aeruginosa RNA phage, PP7." *Frontiers in microbiology* 9 (2018): 247.

Klovins, J., G. P. Overbeek, S. H. E. van den Worm, H.-W. Ackermann, and J. van Duin. 2002. “Nucleotide Sequence of a ssRNA Phage from Acinetobacter: Kinship to Coliphages.” *Journal of General Virology* 83 (6): 1523–33. https://doi.org/10.1099/0022-1317-83-6-1523.

Kozak, M. 1983. “Comparison of Initiation of Protein Synthesis in Procaryotes, Eucaryotes, and Organelles.” *Microbiological Reviews* 47 (1): 1–45.

Krahn, P. M., R. J. O’Callaghan, and W. Paranchych. 1972. “Stages in Phage R17 Infection. VI. Injection of A Protein and RNA into the Host Cell.” *Virology* 47 (3): 628–37.

Krishnamurthy, Siddharth R., Andrew B. Janowski, Guoyan Zhao, Dan Barouch, and David Wang. 2016. “Hyperexpansion of RNA Bacteriophage Diversity.” *PLOS Biology* 14 (3): e1002409. https://doi.org/10.1371/journal.pbio.1002409.

Krishnamurthy, Siddharth R., and David Wang. 2018. “Extensive Conservation of Prokaryotic Ribosomal Binding Sites in Known and Novel Picobirnaviruses.” *Virology* 516 (March): 108–14. https://doi.org/10.1016/j.virol.2018.01.006.

Lister, Philip D., Daniel J. Wolter, and Nancy D. Hanson. "Antibacterial-resistant Pseudomonas aeruginosa: clinical impact and complex regulation of chromosomally encoded resistance mechanisms." *Clinical microbiology reviews* 22, no. 4 (2009): 582-610.

Lodish, Harvey F. 1968. “Bacteriophage F2 RNA: Control of Translation and Gene Order.” *Nature* 220 (5165): 345–50. https://doi.org/10.1038/220345a0.

Lodish, Harvey F. 1970. “Secondary Structure of Bacteriophage F2 Ribonucleic Acid and the Initiation of in Vitro Protein Biosynthesis.” *Journal of Molecular Biology* 50 (3): 689–702. https://doi.org/10.1016/0022-2836(70)90093-8.

Loeb, Tim, and Norton D. Zinder. 1961. “A Bacteriophage Containing RNA.” *Proceedings of the National Academy of Sciences of the United States of America* 47 (March): 282–89.

Mäntynen, Sari, Elina Laanto, Annika Kohvakka, Minna M. Poranen, Jaana K. H. Bamford, and Janne J. Ravantti. 2015. “New Enveloped DsRNA Phage from Freshwater Habitat.” *Journal of General Virology* 96 (5): 1180–89. https://doi.org/10.1099/vir.0.000063.

Mäntynen, Sari, Lotta-Riina Sundberg, and Minna M. Poranen. 2017. “Recognition of Six Additional Cystoviruses: Pseudomonas Virus Phi6 Is No Longer the Sole Species of the Family Cystoviridae.” *Archives of Virology*, December, 1–8. https://doi.org/10.1007/s00705-017-3679-4.

McGraw, T., L. Mindich, and B. Frangione. 1986. “Nucleotide Sequence of the Small Double-Stranded RNA Segment of Bacteriophage Phi 6: Novel Mechanism of Natural Translational Control.” *Journal of Virology* 58 (1): 142–51.

Meyer, François, Hans Weber, and Charles Weissmann. 1981. “Interactions of Qβ Replicase with Qβ RNA.” *Journal of Molecular Biology* 153 (3): 631–60. https://doi.org/10.1016/0022-2836(81)90411-3.

Mindich, Leonard., I. Nemhauser, P. Gottlieb, M. Romantschuk, J. Carton, S. Frucht, J. Strassman, D. H. Bamford, and N. Kalkkinen. 1988. “Nucleotide Sequence of the Large Double-Stranded RNA Segment of Bacteriophage Phi 6: Genes Specifying the Viral Replicase and Transcriptase.” *Journal of Virology* 62 (4): 1180–85.

Mindich, Leonard. 1999. “Precise Packaging of the Three Genomic Segments of the Double-Stranded-RNA Bacteriophage Φ6.” *Microbiology and Molecular Biology Reviews* 63 (1): 149–60.

Mindich, Leonard, and John Lehman. 1979. “Cell Wall Lysin as a Component of the Bacteriophage Ø6 Virion.” *Journal of Virology* 30 (2): 489–96.

Mindich, Leonard, Xueying Qiao, Jian Qiao, Shiroh Onodera, Martin Romantschuk, and Deborah Hoogstraten. 1999. “Isolation of Additional Bacteriophages with Genomes of Segmented Double-Stranded RNA.” *Journal of Bacteriology* 181 (15): 4505–8.

Miranda, Giovanni, Daniel Schuppli, Imma Barrera, Christoph Hausherr, José M Sogo, and Hans Weber. 1997. “Recognition of Bacteriophage Qβ plus Strand RNA as a Template by Qβ Replicase: Role of RNA Interactions Mediated by Ribosomal Proteins S1 and Host Factor11Edited by N. Yaniv.” *Journal of Molecular Biology* 267 (5): 1089–1103. https://doi.org/10.1006/jmbi.1997.0939.

O’Keefe, Kara J., Olin K. Silander, Helen McCreery, Daniel M. Weinreich, Kevin M. Wright, Lin Chao, Scott V. Edwards, Susanna K. Remold, and Paul E. Turner. 2010. “Geographic Differences in Sexual Reassortment in RNA Phage.” *Evolution* 64 (10): 3010–23. https://doi.org/10.1111/j.1558-5646.2010.01040.x.

Olsen, Ronald H., and Deanna D. Thomas. 1973. “Characteristics and Purification of PRR1, an RNA Phage Specific for the Broad Host Range Pseudomonas R1822 Drug Resistance Plasmid.” *Journal of Virology* 12 (6): 1560–67.

Olsthoorn, René C. L., G. Garde, T. Dayhuff, J. F. Atkins, and J. Van Duin. 1995. “Nucleotide Sequence of a Single-Stranded RNA Phage from Pseudomonas Aeruginosa: Kinship to Coliphages and Conservation of Regulatory RNA Structures.” *Virology* 206 (1): 611–25. https://doi.org/10.1016/S0042-6822(95)80078-6.

Olsthoorn, René C. L., and J. Van Duin. 2017. "*Leviviridae* - Positive Sense RNA Viruses - Positive Sense RNA Viruses (2011) - International Committee on Taxonomy of Viruses (ICTV)." International Committee on Taxonomy of Viruses (ICTV). 2017. https://talk.ictvonline.org/ictv-reports/ictv_9th_report/positive-sense-rna-viruses-2011/w/posrna_viruses/263/leviviridae.

Olsthoorn, René C. L., and Jan van Duin. 2011. "Bacteriophages with ssRNA." In *ELS*. American Cancer Society. https://doi.org/10.1002/9780470015902.a0000778.pub3.

Onodera, S, V M Olkkonen, P Gottlieb, J Strassman, X Y Qiao, D H Bamford, and L Mindich. 1992. "Construction of a Transducing Virus from Double-Stranded RNA Bacteriophage Phi6: Establishment of Carrier States in Host Cells." *Journal of Virology* 66 (1): 190–96.

Pirttimaa, M. J., A. O. Paatero, M. J. Frilander, and D. H. Bamford. 2002. "Nonspecific Nucleoside Triphosphatase P4 of Double-Stranded RNA Bacteriophage 6 Is Required for Single-Stranded RNA Packaging and Transcription." *Journal of Virology* 76 (20): 10122–27. https://doi.org/10.1128/JVI.76.20.10122-10127.2002.

Poot, Raymond A., Nina V. Tsareva, Irina V. Boni, and Jan van Duin. 1997. "RNA Folding Kinetics Regulates Translation of Phage MS2 Maturation Gene." *Proceedings of the National Academy of Sciences of the United States of America* 94 (19): 10110–15.

Poranen, Minna M., and Dennis H. Bamford. 2012. "Assembly of Large Icosahedral Double-Stranded RNA Viruses." In *Viral Molecular Machines*, 379–402. Advances in Experimental Medicine and Biology. Springer, Boston, MA. https://doi.org/10.1007/978-1-4614-0980-9_17.

Poranen, Minna M., Rimantas Daugelavičius, Päivi M. Ojala, Michael W. Hess, and Dennis H. Bamford. 1999. "A Novel Virus–Host Cell Membrane Interaction: Membrane

Voltage–Dependent Endocytic-like Entry of Bacteriophage Φ6 Nucleocapsid." *The Journal of Cell Biology* 147 (3): 671–82. https://doi.org/10.1083/jcb.147.3.671.

Poranen, Minna M., and Sari Mäntynen. 2017. "ICTV Virus Taxonomy Profile: Cystoviridae." *The Journal of General Virology* 98 (10): 2423–24. https://doi.org/10.1099/jgv.0.000928.

Poranen, Minna M, Anja O Paatero, Roman Tuma, and Dennis H Bamford. 2001. "Self-Assembly of a Viral Molecular Machine from Purified Protein and RNA Constituents." *Molecular Cell* 7 (4): 845–54. https://doi.org/10.1016/S1097-2765(01)00228-3.

Poranen, Minna M., Roman Tuma, and Dennis H. Bamford. 2005. "Assembly of Double-Stranded RNA Bacteriophages." In *Advances in Virus Research*, 64:15–43. Virus Structure and Assembly. Academic Press. https://doi.org/10.1016/S0065-3527(05)64002-X.

Qiao, Jian, Xueying Qiao, Yang Sun, and Leonard Mindich. 2003. "Isolation and Analysis of Mutants of Double-Stranded-RNA Bacteriophage Φ6 with Altered Packaging Specificity." *Journal of Bacteriology* 185 (September): 4572–77. https://doi.org/10.1128/JB.185.15.4572-4577.2003.

Qiao, Xueying, Jian Qiao, and Leonard Mindich. 2003. "Analysis of Specific Binding Involved in Genomic Packaging of the Double-Stranded-RNA Bacteriophage Φ6." *Journal of Bacteriology* 185 (21): 6409–14. https://doi.org/10.1128/JB.185.21.6409-6414.2003.

Qiao, Xueying, Jian Qiao, Shiroh Onodera, and Leonard Mindich. 2000. "Characterization of Φ13, a Bacteriophage Related to Φ6 and Containing Three dsRNA Genomic Segments." *Virology* 275 (1): 218–24. https://doi.org/10.1006/viro.2000.0501.

Qiao, Xueying, Yang Sun, Jian Qiao, Fabiana Di Sanzo, and Leonard Mindich. 2010. "Characterization of F2954, a Newly Isolated Bacteriophage Containing Three dsRNA Genomic Segments," 7.

Reed, Catrina A., Carrie Langlais, Ing-Nang Wang, and Ry Young. 2013. “A2 Expression and Assembly Regulates Lysis in Qβ Infections.” *Microbiology* 159 (Pt 3): 507–14. https://doi.org/10.1099/mic.0.064790-0.

Roberts, J. W., and J. E. Steitz. 1967. “The Reconstitution of Infective Bacteriophage R17.” *Proceedings of the National Academy of Sciences* 58 (4): 1416–21. https://doi.org/10.1073/pnas.58.4.1416.

Robertson, Hugh D., and Harvey F. Lodish. 1970. “Messenger Characteristics of Nascent Bacteriophage RNA.” *Proceedings of the National Academy of Sciences* 67 (2): 710–16. https://doi.org/10.1073/pnas.67.2.710.

Robertson, Hugh, Robert E. Webster, and Norton D. Zinder. 1968. “Bacteriophage Coat Protein as Repressor.” *Nature* 218 (5141): 533–36. https://doi.org/10.1038/218533a0.

Rodriguez-Valera, Francisco, Ana-Belen Martin-Cuadrado, Beltran Rodriguez-Brito, Lejla Pašić, T. Frede Thingstad, Forest Rohwer, and Alex Mira. 2009. “Explaining Microbial Population Genomics through Phage Predation.” *Nature Reviews Microbiology* 7 (11): 828–36. https://doi.org/10.1038/nrmicro2235.

Roine, Elina, Deanna M. Raineri, Martin Romantschuk, Mark Wilson, and David N. Nunn. 1998. “Characterization of Type IV Pilus Genes in Pseudomonas Syringae Pv. Tomato DC3000.” *Molecular Plant-Microbe Interactions* 11 (11): 1048–56. https://doi.org/10.1094/MPMI.1998.11.11.1048.

Romantschuk, M., V. M. Olkkonen, and D. H. Bamford. 1988. “The Nucleocapsid of Bacteriophage Phi 6 Penetrates the Host Cytoplasmic Membrane.” *The EMBO Journal* 7 (6): 1821–29. https://doi.org/10.1002/j.1460-2075.1988.tb03014.x.

Rumnieks, Janis, and Kaspars Tars. 2011. “Crystal Structure of the Read-through Domain from Bacteriophage Qβ A1 Protein.” *Protein Science : A Publication of the Protein Society* 20 (10): 1707–12. https://doi.org/10.1002/pro.704.

Rumnieks, Janis, Kaspars Tars. 2017. “Crystal Structure of the Maturation Protein from Bacteriophage Qβ.” *Journal of Molecular Biology* 429 (5): 688–96. https://doi.org/10.1016/j.jmb.2017.01.012.

Ruokoranta, Tanja M., A. Marika Grahn, Janne J. Ravantti, Minna M. Poranen, and Dennis H. Bamford. 2006. “Complete Genome Sequence of the Broad Host Range Single-Stranded RNA Phage PRR1 Places It in the *Levivirus* Genus with Characteristics Shared with Allolevivirus.” *Journal of Virology* 80 (18): 9326–30. https://doi.org/10.1128/JVI.01005-06.

Saberi, Amir, Anastasia A. Gulyaeva, John Brubacher, Phillip A. Newmark, and Alexander Gorbalenya. 2018. “A Planarian *Nidovirus* Expands the Limits of RNA Genome Size,” April. https://doi.org/10.1101/299776.

Sanger, F., G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, J. C. Fiddes, C. A. Hutchison Iii, P. M. Slocombe, and M. Smith. 1977. “Nucleotide Sequence of Bacteriophage ΦX174 DNA.” *Nature* 265 (5596): 687–95. https://doi.org/10.1038/265687a0.

Schmeing, T. Martin, Rebecca M. Voorhees, Ann C. Kelley, Yong-Gui Gao, Frank V. Murphy, John R. Weir, and V. Ramakrishnan. 2009. “The Crystal Structure of the Ribosome Bound to EF-Tu and Aminoacyl-TRNA.” *Science* 326 (5953): 688–94. https://doi.org/10.1126/science.1179700.

Schmidt, Brian F., Ben Berkhout, Gerrit P. Overbeek, Anja van Strien, and Jan van Duin. 1987. “Determination of the RNA Secondary Structure That Regulates Lysis Gene Expression in Bacteriophage MS2.” *Journal of Molecular Biology* 195 (3): 505–16. https://doi.org/10.1016/0022-2836(87)90179-3.

Schuette, Jan-Christian, Frank V. Murphy, Ann C. Kelley, John R. Weir, Jan Giesebrecht, Sean R. Connell, Justus Loerke, *et al*. 2009. “GTPase Activation of Elongation Factor EF-

Tu by the Ribosome during Decoding." *The EMBO Journal* 28 (6): 755–65. https://doi.org/10.1038/emboj.2009.26.

Schuppli, Daniel, Jelena Georgijevic, and Hans Weber. 2000. "Synergism of Mutations in Bacteriophage Qβ RNA Affecting Host Factor Dependence of Qβ Replicase". "Edited by M. Yaniv." *Journal of Molecular Biology* 295 (2): 149–54. https://doi.org/10.1006/jmbi.1999.3373.

Schuppli, Daniel, Giovanni Miranda, Su Qiu, and Hans Weber. 1998. "A Branched Stem-Loop Structure in the M-Site of Bacteriophage Qβ RNA Is Important for Template Recognition by Qβ Replicase Holoenzyme." *Journal of Molecular Biology* 283 (3): 585–93. https://doi.org/10.1006/jmbi.1998.2123.

Schuppli, Daniel, Giovanni Miranda, Ho-Ching Tiffany Tsui, Malcolm E. Winkler, José M. Sogo, and Hans Weber. 1997. "Altered 3′-Terminal RNA Structure in Phage Qβ Adapted to Host Factor-Less *Escherichia Coli*." *Proceedings of the National Academy of Sciences* 94 (19): 10239–42. https://doi.org/10.1073/pnas.94.19.10239.

Shiba, Tadayoshi, and Yurie Suzuki. 1981. "Localization of A Protein in the RNA-A Protein Complex of RNA Phage MS2." *Biochimica et Biophysica Acta (BBA) - Nucleic Acids and Protein Synthesis* 654 (2): 249–55. https://doi.org/10.1016/0005-2787(81)90179-9.

Silander, Olin K., Daniel M. Weinreich, Kevin M. Wright, Kara J. O'Keefe, Camilla U. Rang, Paul E. Turner, and Lin Chao. 2005. "Widespread Genetic Exchange among Terrestrial Bacteriophages." *Proceedings of the National Academy of Sciences* 102 (52): 19009–14. https://doi.org/10.1073/pnas.0503074102.

Silverman, Philip M. 1973. "Replication of RNA Viruses: Specific Binding of the Qβ RNA Polymerase to Qβ RNA." *Archives of Biochemistry and Biophysics* 157 (1): 222–33. https://doi.org/10.1016/0003-9861(73)90408-6.

Staples, D. H., J. Hindley, M. A. Billeter, and C. Weissmann. 1971. “Localization of Qβ Maturation Cistron Ribosome Binding Site.” *Nature New Biology* 234: 202–4.

Stitt, B. L., and L. Mindich. 1983. “Morphogenesis of Bacteriophage Phi 6: A Presumptive Viral Membrane Precursor.” *Virology* 127 (2): 446–58.

Stock-Ley, P. G., N. J. Stonehouse, and K. Valegård. 1994. “Molecular Mechanism of RNA Phage Morphogenesis.” *International Journal of Biochemistry* 26 (10): 1249–60. https://doi.org/10.1016/0020-711X(94)90094-9.

Summers, William C. 2012. “The Strange History of Phage Therapy.” *Bacteriophage* 2 (2): 130–33. https://doi.org/10.4161/bact.20757.

Twort, F. W. 1915. “An Investigation on the Nature of Ultra-Microscopic Viruses.” *The Lancet* 186 (4814): 1241–43. https://doi.org/10.1016/S0140-6736(01)20383-3.

Usala, Stephen J., Bernard H. Brownstein, and Robert Haselkorn. 1980. “Displacement of Parental RNA Strands during in Vitro Transcription by Bacteriophage Φ6 Nucleocapsids.” *Cell* 19 (4): 855–62. https://doi.org/10.1016/0092-8674(80)90076-8.

Valegrad, Karin, James B. Murray, Peter G. Stockley, Nicola J. Stonehouse, and Lars Liljas. 1994. “Crystal Structure of an RNA Bacteriophage Coat Protein-Operator Complex.” *Nature* 371 (October): 623–26.

Vasiljeva, Inta, Tatjana Kozlovska, Indulis Cielens, Anna Strelnikova, Andris Kazaks, Velta Ose, and Paul Pumpens. 1998. “Mosaic Qβ Coats as a New Presentation Model.” *FEBS Letters* 431 (1): 7–11. https://doi.org/10.1016/S0014-5793(98)00716-9.

Vidaver, Anne K., R. K. Koski, and J. L. Van Etten. 1973. “Bacteriophage Φ6: A Lipid-Containing Virus of Pseudomonas Phaseolicola.” *Journal of Virology* 11 (5): 799–805.

“Viral Zone: *Cystoviridae*.” 2018. May 30, 2018. https://viralzone.expasy.org/165?outline=all_by_species.

"Viral Zone: *Leviviridae*." 2018. May 30, 2018. https://viralzone.expasy.org/163?outline=all_by_species.

Wahba, Albert J., Martha J. Miller, Alain Niveleau, Terry A. Landers, Gordon G. Carmichael, Klaus Weber, David A. Hawley, and Lawrence I. Slobin. 1974. "Subunit I of Qβ Replicase and 30 S Ribosomal Protein Sl of *Escherichia Coli* Evidence for the Identity of the Two Proteins." *Journal of Biological Chemistry* 249 (10): 3314–16.

Walderich, B, and J V Höltje. 1989. "Specific Localization of the Lysis Protein of Bacteriophage MS2 in Membrane Adhesion Sites of Escherichia Coli." *Journal of Bacteriology* 171 (6): 3331–36.

Walderich, B., A. Ursinus-Wössner, J. van Duin, and J. V. Höltje. 1988. "Induction of the Autolytic System of *Escherichia Coli* by Specific Insertion of Bacteriophage MS2 Lysis Protein into the Bacterial Cell Envelope." *Journal of Bacteriology* 170 (11): 5027–33. https://doi.org/10.1128/jb.170.11.5027-5033.1988.

Wang, Shen, Ying Liu, Dandan Li, Tiezhong Zhou, Shenyang Gao, Enhui Zha, and Xiqing Yue. 2016. "Preparation and Evaluation of MS2 Bacteriophage-like Particles Packaging Hepatitis E Virus RNA." *FEMS Microbiology Letters* 363 (20). https://doi.org/10.1093/femsle/fnw221.

Weber, Hans, and Charles Weissmann. 1970. "The 3'-Termini of Bacteriophage Qβ Plus and Minus Strands." *Journal of Molecular Biology*, no. 51: 215–24.

Weber, Hans. 1976. "The Binding Site for Coat Protein on Bacteriophage Qβ RNA." *Biochimica et Biophysica Acta (BBA) - Nucleic Acids and Protein Synthesis* 418 (2): 175–83. https://doi.org/10.1016/0005-2787(76)90067-8.

Weiner, A. M., and K. Weber. 1971. "Natural Read-through at the UGA Termination Signal of Q-Beta Coat Protein Cistron." *Nature: New Biology* 234 (50): 206–9.

Weissmann, Charles 1974. “The Making of a Phage.” *FEBS Letters* 40 (S1): S3–9. https://doi.org/10.1016/0014-5793(74)80684-8.

Weissmann, Charles, M A Billeter, H M Goodman, J Hindley, and H Weber. 1973. “Structure and Function of Phage RNA.” *Annual Review of Biochemistry* 42 (1): 303–28. https://doi.org/10.1146/annurev.bi.42.070173.001511.

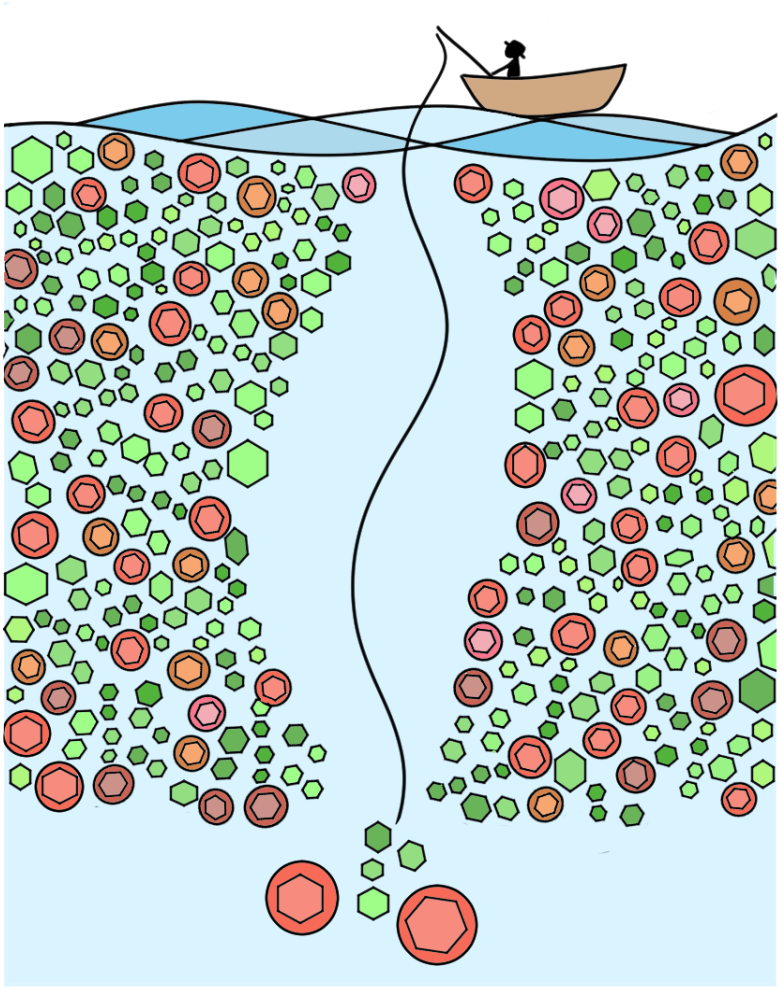
Yang, H., E. V. Makeyev, S. J. Butcher, A. Gaidelyte, and D. H. Bamford. 2003. “Two Distinct Mechanisms Ensure Transcriptional Polarity in Double-Stranded RNA Bacteriophages.” *Journal of Virology* 77 (2): 1195–1203. https://doi.org/10.1128/JVI.77.2.1195-1203.2003.

Yang, Yuhui, Shuguang Lu, Wei Shen, Xia Zhao, Mengyu Shen, Yinling Tan, Gang Li, *et al*. 2016. “Characterization of the First Double-Stranded RNA Bacteriophage Infecting *Pseudomonas Aeruginosa*.” *Scientific Reports* 6 (December): 38795. https://doi.org/10.1038/srep38795.

Young, Ry. 1992. “Bacteriophage Lysis: Mechanism and Regulation.” *Microbiological Reviews* 56 (3): 430–81.

Zechner, Ellen L., Silvia Lang, and Joel F. Schildbach. 2012. “Assembly and Mechanisms of Bacterial Type IV Secretion Machines.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 367 (1592): 1073–87. https://doi.org/10.1098/rstb.2011.0207.

Zinder, Norton D. 1965. “RNA Phages.” *Annual Review of Microbiology*, no. 19: 455–73.



Chapter II

Biases in Viral Metagenomics-Based Detection, Cataloguing and Quantification of Bacteriophage Genomes in Human Faeces, a Review

This chapter was published as a paper as part of the Special Issue 'Comparative Genomics of the Human Gut Microbiome' for *Microorganisms.*

Callanan, Julie, Stephen R. Stockdale, Andrey Shkoporov, Lorraine A. Draper, R. Paul Ross, and Colin Hill. "Biases in Viral Metagenomics-Based Detection, Cataloguing and Quantification of Bacteriophage Genomes in Human Faeces, a Review." Microorganisms 9, no. 3 (2021): 524.

https://doi.org/10.1126/sciadv.aay5981

2.1 Abstract

The human gut is colonised by a vast array of microbes that include bacteria, viruses, fungi, and archaea. While interest in these microbial entities has largely focused on the bacterial constituents, recently the viral component has attracted more attention. Metagenomic advances, compared to classical isolation procedures, have greatly enhanced our understanding of the composition, diversity, and function of viruses in the human microbiome (virome). It is highlighted that viral extraction methodologies are crucial in terms of identifying and characterising communities of viruses infecting eukaryotes and bacteria. Different viral extraction protocols, including those used in some of the most significant human virome publications to date, have introduced biases affecting their overall conclusions. It is important that protocol variations should be clearly highlighted across studies, with the ultimate goal of identifying and acknowledging biases associated with different protocols and, perhaps, the generation of an unbiased and standardised method for examining this portion of the human microbiome.

2.2 Introduction

The estimated number of bacteriophages (phages) within the human gut has been recently calculated as approximately 10^{10} per gram of faeces (Shkoporov and Hill 2019). The genetic material encapsulated within these phages is either DNA or RNA, which in turn can be double-stranded (ds) or single-stranded (ss). The single-stranded variants can exist in two different forms depending on their orientation and polarity: positive-sense or negative-sense. No negative-sense ssRNA phages have been identified to date.

It has been over a decade since the first attempts to conduct metagenomic analyses of gut viral communities (Breitbart *et al*. 2003). Many studies in this area have deposited their data in public databases. Of particular interest are the phages that may influence the

composition, turnover, and functionality of bacterial communities (Hsu *et al*. 2019; Khan Mirzaei *et al*. 2020). The number of studies focusing on this phage population, termed the phageome, has increased in recent decades (Shkoporov and Hill 2019). This surge in phageome research has been made possible by advances in contemporary sequencing technologies and specialised virome sequencing data analysis tools including VirSorter and Demovir (Roux, Enault, *et al*. 2015; Ryan 2018).

There have been efforts to create standardised protocols to study the faecal phageome through metagenomic analyses that allow for reliable comparisons between studies from different groups (Shkoporov, Ryan, *et al*. 2018; d'Humières *et al*. 2019). One such effort was that of Conceição-Neto and colleagues (2015) in which they proposed the 'Novel enrichment technique of VIRomes' (NetoVIR) protocol (Conceição-Neto *et al*. 2015). This method was designed using mock viral and bacterial communities which included both +ssRNA and dsRNA viruses, which were not phages, but does suggest an approach to optimise their recovery. Nevertheless, the search for common protocols enabling cross-study comparison should not discourage researchers from developing novel techniques to capture new phages. Most newly identified phage sequences do not have known counterparts in viral databases, and these unknown sequences are often collectively referred to as the "viral dark matter" (Shkoporov and Hill 2019; Krishnamurthy and Wang 2017; Roux, Hallam, *et al*. 2015). It has been revealed that the "viral dark matter" can account for 60-95% of the genomes identified (Roux, Hallam, *et al*. 2015; Ogilvie and Jones 2015).

The majority of newly discovered phages may be novel because (i) their bacterial hosts are recalcitrant to isolation and cultivation, (ii) they exhibit unusual or previously undescribed lifecycles which may prevent them from being detected using the typical plaque-dependent methods, or (iii) there is a strong likelihood that the methods used may not have been suitable and more effort is required to capture all types of phage (Guerin and Hill 2020; Sutton and Hill

2019; Forster *et al*. 2019). Considering that many bacteria are yet to be grown in a laboratory, culture-based methods are limited in their efficacy for isolating new phage-host pairs (Duhaime *et al*. 2012). Even when the host is culturable, the phage may not plaque as it may not infect until the host has reached a specific growth phase (Chibani-Chennoufi *et al*. 2004), the plaques may be very difficult to see if the phage diffuse poorly in agar, or it may be lysogenic or practice pseudolysogeny (the delayed development of a phage in the host cell) (Łoś and Węgrzyn 2012). It could also be a result of differences in the physiology of a bacterium in a laboratory environment compared to growth in its natural environment (Tank and Bryant 2015). In an effort to bring order to these novel sequences, collaborative efforts are required to link both metagenomic analyses and culture-based investigations. One example is the successful isolation of the first crAss-like phage. CrAss-like phages are viruses with relatively large genomes (~100kb) that were originally found in metagenomic studies and predicted to infect bacteria in the order *Bacteroidales*. Collectively, this phage family group is the most abundant human gut-associated viral clade, identified in >50% of people, and representing up to 90% of all sequencing reads in some human gut viromes (Edwards *et al*. 2019; Guerin *et al*. 2018; Dutilh *et al*. 2014). Through a combination of bioinformatic-based discovery and subsequent laboratory-based experiments, the first representative of this family of phages was isolated and propagated on its *Bacteroides intestinalis* host (Shkoporov, Khokhlova, *et al*. 2018).

The lack of a single phylogenetic marker in virology (equivalent to the 16S rRNA or the *chaperonin-60* (*cpn60*) gene in bacteria) further complicates our ability to properly assign taxonomic ranks to this “viral dark matter”. A recent publication has described Minimum Information about an Uncultivated Virus Genome (MIUViG) standards in an attempt to overcome this difficulty (Roux *et al*. 2019). These include virus origin, genome annotation and quality, taxonomic classification, and a collection of other mandatory and optional metadata.

Community-wide compliance with these standards will allow for more effective evaluation of the global virosphere and more robust comparisons between studies.

The characterisation and quantification of nucleic acids of uncultured viruses isolated from different biomes is dependent on many factors, including concentration, purification, extraction, and sequencing techniques. There is no ideal "capture-all" protocol but care and consideration is crucial in relation to the choices made at each stage of the protocol (Thurber *et al*. 2009). There are four main processes involved in the development of a phage sequencing protocol, including: i) acquisition and storage of the sample, ii) separation of viral particles, iii) the extraction of pure nucleic acids with the elimination of free nucleic acids and contaminating cells, and iv) successful sequencing and bioinformatic analysis of these nucleic acids (as depicted in Figure 1).

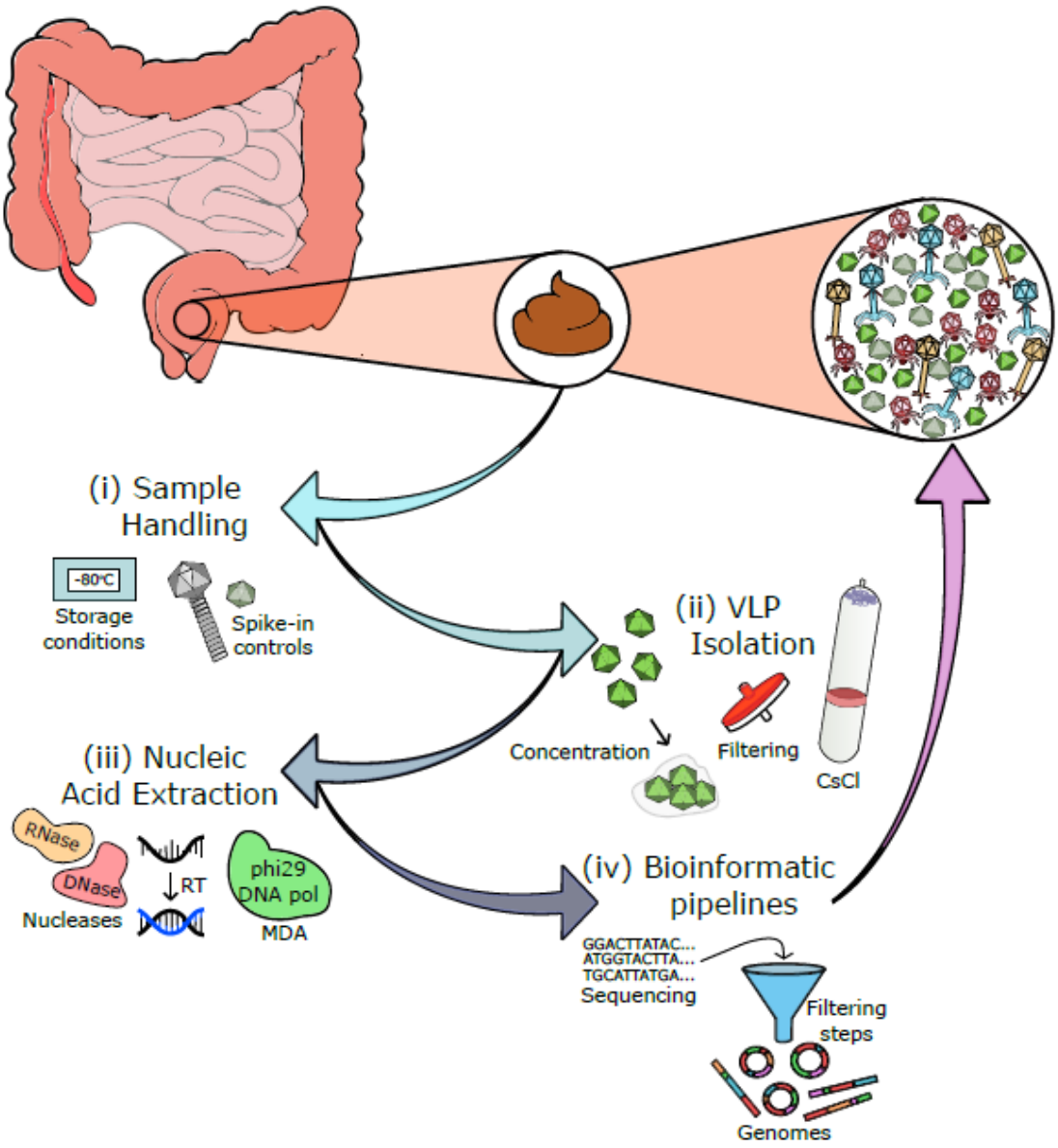


Figure 1. Basis of a viral/phage isolation protocol. Faecal samples are often used as a proxy for the human gut virome. Through four main processes, the viral and phage communities of the human gut are analysed: i) acquisition and storage of samples, ii) concentration of viral particles, iii) extraction of pure nucleic acids with the elimination of free nucleic acids, and iv) successful sequencing and bioinformatic analysis of these nucleic acids.

Despite rapid advances in high-throughput sequencing technologies, few studies detect RNA viruses in human and animal faecal samples, even when a reverse transcription (RT) step is included. This may be due to low RNA viral loads, destabilisation of the viral particle, reliance on physical virion characteristics, or as a direct result of the nucleic acid extraction method used in the study (Shkoporov, Ryan, *et al*. 2018). It has been suggested that RNA viruses form an important part of the total gut virome but most studies to date have concluded that members of RNA phage families are only a minor component (Breitbart *et al*. 2003; Minot *et al*. 2011; Reyes *et al*. 2010; Zhang *et al*. 2005). However, while RNA phages were rarely detected in environmental metagenomics, recent studies have reported logarithmic increases in the total number of known single-stranded RNA phages from these sources (Callanan *et al*. 2020; Shi *et al*. 2016; Krishnamurthy *et al*. 2016). Therefore, gut RNA phages may be underestimated and without comprehensive studies targeting these elusive phages, the biological significance of RNA phages may remain largely overlooked.

There is an unavoidable loss of some virions at almost every step of the protocol. Viral particles can become adsorbed to larger molecules such as food particles, immobilised on filters or damaged by nucleases. Certain viruses such as giant viruses that can reach 750nm in size may also be excluded in some filtering protocols, while filamentous viruses such as *Lipothrixviridae* can often reach over 2μm in length (Vestergaard *et al*. 2008; Xiao *et al*. 2005). Both these virus types will not pass through the majority of filtering processes. Conceição-Neto and colleagues (2015) also highlighted that the use of small filter pores, coupled with strict centrifugation conditions, may lead to the exclusion of these larger viruses from virome analyses (Conceição-Neto *et al*. 2015). In the development of an optimised viral isolation protocol, it can be difficult to balance increasing contamination risk with larger particles and smaller bacterial cells for the possible reward of incorporating these viral types.

Here, the biases associated with the key isolation steps used in published virome studies were examined. This was performed by surveying studies over the past decade that have started with a faecal sample and used different methods to examine the human gut virome, with particular attention on phageome composition and recordings of RNA viruses. It is hoped that by addressing any shortcomings of current methods and identifying crucial procedures in retaining the true viral diversity of the human gut, a standard or reference protocol could be developed that would be reproducible and comparable across research studies focusing on reducing method biases and including RNA phages.

2.3 Sample handling

The quantity of sample required is dependent on the efficiency at which viruses can be isolated. This review is focused on studies using faecal samples. Faecal samples have been extensively used as a starting material in order to study the complex virome and phageome associated with the human gut. Faeces is widely used as it offers a more practical and non-invasive means to access novel phages from the gut. This is compared to other sampling sites of the gastrointestinal tract (GIT) which are ethically and practically more difficult to acquire, such as biopsies from the human GIT mucosa. Nearly two decades have passed since the initial analysis of the composition and population structure of the uncultured viral community from human faeces (Breitbart *et al*. 2003). Breitbart and colleagues noted that the majority of sequences were unrelated to previously known sequences, and the most recognisable were those belonging to *Siphoviridae*, a family of dsDNA phage belonging to the order *Caudovirales*. Since then, treatments, techniques, and protocols associated with extracting the viral and phage fractions from human faeces have improved.

While storage temperatures of samples could potentially contribute to virome composition, Shkoporov *et al*. demonstrated that repeated freeze-thaw and alternative storage

temperature (4˚C vs room temperature) only had a mild effect on the dsDNA composition of the virome (Shkoporov, Ryan, *et al*. 2018). In the same study researchers assessed the affect of freeze-thaw cycles on elusive gut phages and found that this treatment affected the bacterial components more so than the phage population. It is suggested that where possible it is best to avoid repeated freeze-thaw cycles of the sample and the sample should be frozen as soon as possible (Gorzelak *et al*. 2015).

The choice of buffer also requires careful consideration. The majority of virome studies use either a buffer composed of a mixture of sodium chloride and magnesium sulphate (SM buffer) or phosphate-buffered saline (PBS) buffer as a means to resuspend the faecal sample and release viral particles. There have been suggestions that SM is the preferred option due to the potential of the different ions to inactivate the phage, for example Adams observed that phosphate ions inactivate Enterobacteria phage T5 (Adams 1949).

The inclusion of a spiked-in exogenous phage standard, such as the lactococcal phage Q33 or +ssRNA phage Qbeta, enables a semi-quantitative analysis of individual members of the virome following metagenomic sequencing by analysing the percentage of reads aligning to these spiked-in genomes (Shkoporov, Ryan, *et al*. 2018). It should be noted that no reads aligning to the spiked-in +ssRNA phage were detected in downstream analyses of that study. In 2019, d'Humières and colleagues examined the effect that four different methods, varying only in the phage concentration step, had on the overall phageome composition from the same faecal material (d'Humières *et al*. 2019). They noted that the initial mechanical agitation is essential to dissolve the phage particles in the faeces when homogenised in PBS. Therefore, protocols should note at which point in the preliminary stage of the protocol the spiked-in controls are added.

2.4 VLP isolation

The purification of virus-like particles (VLPs) is one of the most critical steps in the quantitative and qualitative metagenomic analyses of the total viral population. Ideally one would want to reduce the number of contaminating bacterial sequences present in the VLP fraction. Bacterial genomic DNA/RNA detected in downstream analysis could be due to contamination or may have been packaged within phage particles as a result of generalised transduction, specialised transduction, or incorporation in Gene Transfer Agents (GTAs) (McDaniel *et al*. 2010; Bushman 2002). In the majority of virome and phageome work, centrifugation and filtering of faecal supernatants are used to eliminate debris and remove bacteria.

Many studies report the level of contamination associated with their samples, both of bacterial and human origin. One such example is a study by Norman and colleagues (2015) where analysis of their VLP sequences revealed a low level of contamination with human sequences (0–4%) and they acknowledged that there was also possible contamination with bacterial sequences that was confounded by the presence of integrated prophages in full genome sequences of bacteria (Norman *et al*. 2015). This highlights how essential it is to identify contaminating particles from the sample following both physical and bioinformatic filtering in downstream processes. Some groups have adopted a novel approach of identifying contaminating bacterial sequences which align to *cpn60*, a highly conserved house-keeping gene. As the *cpn60* gene occurs once per genome it offers an alternative bacterial taxonomic marker to the traditional 16S rRNA and also gives finer taxonomic discrimination between bacteria (Shkoporov and Hill 2019; Shkoporov, Ryan, *et al*. 2018; Links *et al*. 2012; Hill *et al*. 2004). The analysis of the *cpn60* gene overcomes concerns that the 16S rRNA gives disproportionally high levels of bacterial contamination due to the rRNA being purified in workflows in the form of ribosomes being co-isolated along with viral particles.

It had been previously demonstrated under microscopic examination that 0.22µm filters reduced the number of viral particles from faecal samples by almost half (Hoyles *et al*. 2014). Indeed, in the previously mentioned d'Humières study, it was also found that the filter size is crucial in the early stages of phageome studies (d'Humières *et al*. 2019). Their results showed that filtration should be done using 0.45µm and 0.2µm filters and not just 0.22µm as the combination of filter sizes allows the faecal lysate to be purified of larger contaminants prior to selecting for the VLP portion.

Polyethylene glycol (PEG) precipitation is often used in protocols to concentrate VLPs in the sample prior to purification or nucleic acid extraction. In the study by d'Humières and colleagues, they deemed the method including PEG to be the best of those examined across a range of faecal samples in order to assure reproducibility and sequencing depth (d'Humières *et al*. 2019). They also discussed the efficacy of the method including PEG to concentrate phages from faecal filtrate, suggesting that despite the fact it requires an overnight incubation step, it would be a beneficial reagent to include in phage isolation protocols. The requirement for chloroform in the PEG-removal step, and at different points in nucleic acid extraction protocols, has repercussions as chloroform degrades and destroys the phospholipid membrane of some enveloped viruses, such as dsRNA phages of the family *Cystoviridae*, potentially leading to dramatic under-representation of such viral groups.

Tangential-flow filtration (TFF) can also be used to concentrate viral particles from samples. Thurber and colleagues (2009) discussed the advantages and disadvantages associated with this method and ultimately decided to exclude it from their final protocol (Thurber *et al*. 2009). The main flaw associated with TFF is that in order to maximise viral recovery, approximately two volumes of the filtrate is recirculated which results in a dilute final retentate. This approach is better suited to non-faecal samples and is routinely used to study aquatic environments. Similarly, zinc chloride and ammonium acetate precipitation protocols to

concentrate phages are more suited to non-faecal samples (Czajkowski, Ozymko, and Lojkowska 2016; Casey *et al*. 2015).

Another method commonly used for viral particle purification is caesium chloride (CsCl) density centrifugation which is based on physical properties of the virion (Fauquet *et al*. 2005). The factors associated with CsCl purification include the speed of the centrifuge, the solvent used for resuspension of faeces, and the number of gradient layers examined and are dependent on the virus buoyant density. Researchers should always make the gradients from the same buffer present in the samples and filter-purify the gradients to reduce the levels of contaminating viruses in the final fractions. It is crucial not to unsettle the borders of the layers and to fill the column tube completely prior to centrifugation. Phages are concentrated in a multi-layer gradient where they are localised at different densities and subsequently removed using a sterile needle. The CsCl gradient selection and type are crucial factors to consider prior to excision of bands as it may inadvertently exclude RNA phages and skew the outcome in favour of DNA phages. In an effort to recover RNA phages from this method, as well as the DNA phages, multiple different bands should be excised and examined. A summary table of this information is available by Fauquet *et al.* in which they note the densities and sensitivities associated with RNA phages (Fauquet *et al*. 2005). It should be noted that this may introduce biases to the resulting population as phages and other viruses outside this range may be excluded based on the selected densities.

In the d'Humières study, it was found that the method including a CsCl step gave the lowest bacterial contamination and largest contigs but also had the lowest phage diversity, was very time consuming, and showed poor reproducibility (d'Humières *et al*. 2019). These key points agreed with the findings from another study by Kleiner and colleagues where they examined the effects of different extraction methods on an artificial intestinal microbiota

sample (Kleiner, Hooper, and Duerkop 2015). Some studies specifically noted that this optional step is excluded for being too labour intensive and inappropriate for high-throughput studies.

2.5 Nucleic acid extraction and library preparation

Once the VLP fraction has been separated from the faecal material, there are a series of necessary steps to allow for the isolation of the nucleic acids. This is essential to yield nucleic acids of sufficient purity and concentration for downstream library preparation and sequencing (Thomas, Gilbert, and Meyer 2012). Although the vast majority of virome studies have solely focused on the DNA portion, metagenomes of RNA viruses have also been generated. It is crucial that RNase-free and viral-free reagents be used in the isolation of the RNA. These studies rely on creating sufficient quantities of cDNA via reverse transcription of the viral RNA.

There are some studies that have combined the use of the phenol/chloroform protocol for bacterial nucleic acid extraction and the formamide/cetyltrimethylammonium bromide (CTAB) method which was traditionally used for the extraction of viral DNA. The main advantages of these methods compared to commercially available kits are the decreased associated costs and the absence of a carrier RNA. This additional carrier RNA functions as a means to enhance the recovery of DNA/RNA by preventing the target nucleic acids in low yield samples from being irretrievably bound and increasing the success of downstream PCR processes (Shaw *et al*. 2009). This carrier RNA can often contaminate samples and requires the addition of an additional RNase step to remove it.

A recent review by Garmaeva *et al* (2019) discussed the impact of nucleic acid extraction protocols on the observed composition of the human gut virome and the apparent dominance of DNA viruses, particularly dsDNA phages, from faecal samples used in different studies (Garmaeva *et al*. 2019; Shkoporov, Ryan, *et al*. 2018; Kleiner, Hooper, and Duerkop

2015; Conceição-Neto *et al*. 2015; Minot *et al*. 2011; Reyes *et al*. 2010). They highlighted the fact that current understanding of the human gut virome composition may underestimate the abundance and importance of the RNA viral portion. The inclusion of an RNase step in the treatment of the faecal sample is commonly used to remove free non-viral contaminating RNA (Shkoporov, Ryan, *et al*. 2018; Kleiner, Hooper, and Duerkop 2015). However, recent work has demonstrated that the addition of RNase negatively affects the RNA-fraction of the virome (Adriaenssens *et al*. 2018). Some RNA viruses also contain portions of the RNA as a component of their nucleocapsid structure while others have loose capsid structures which make the virus susceptible to RNase degradation. This structure may be destroyed by the addition of RNase, as demonstrated by Acheson and Tamm with their findings that Semliki Forest virus nucleocapsid disintegrated following RNase treatment (Acheson and Tamm 1970). Despite the risk of increased contamination levels with rRNA or cellular mRNA, it may be wise to restrict the use of RNases in order to capture the true RNA viral diversity and instead implement stricter filtering steps in the bioinformatic quality control workflows. The inclusion of a DNase step is still widely accepted as an essential step in the removal of contaminating free DNA fragments from the sample.

With the low yield of phage nucleic acids from extraction methods and the high amount of DNA required for library preparation kits, many researchers have had to rely on multiple displacement amplification (MDA) prior to sequencing. It has been widely shown that this amplification protocol preferentially amplifies small and circular ssDNA (d'Humières *et al*. 2019; Džunková *et al*. 2014; Yilmaz, Allgaier, and Hugenholtz 2010). MDA is dependent on the high processivity of the phi29 DNA polymerase, an enzyme with strand-displacement activity which allows for amplification of genomic DNA using random primers with a single denaturation step (Dean *et al*. 2002). This type of practice is referred to as a whole genome amplification (WGA), a robust method to amplify the entire genome of limited extracted

nucleic acid samples. There is an ongoing controversy regarding the use of this method in phageome studies as it appears to introduce bias affecting the relative frequency of dsDNA phages of the *Caudovirales* order (including *Siphoviridae*, *Myoviridae* and *Podoviridae* families) and +ssDNA phages of the *Microviridae* family in the healthy human gut (Shkoporov, Ryan, *et al*. 2018; Minot *et al*. 2013; 2011; Waller *et al*. 2014; Reyes *et al*. 2010). Several of these studies have included an MDA step before sequencing (Shkoporov, Ryan, *et al*. 2018; McCann *et al*. 2018; Norman *et al*. 2015; Minot *et al*. 2013; Reyes *et al*. 2010), whereas some studies use kits, such as the Nextera XT DNA Library Prep kit, without the MDA step to exclude ssDNA phages (d'Humières *et al*. 2019; Roux *et al*. 2016). Another issue associated with the inclusion of an MDA step is its inability to capture RNA viruses (Lim *et al*. 2015). A study in 2010 by Reyes and colleagues examined the effects that MDA had on their samples by comparing an unamplified sample to an MDA/WGA processed sample (Reyes *et al*. 2010). They determined that 98.4% of unamplified sequences were present in the WGA, while 91.96% of the WGA were reciprocally found in the unamplified sample. These discrepancies may be due to the preferential bias in the amplification of small ssDNA viruses, as also corroborated by other studies (Roux *et al*. 2016; Norman *et al*. 2015). Recent work by Gregory *et al.*, examining age-dependent patterns of the human gut virome using pre-existing datasets, found that 96% of studies were MDA treated (Gregory *et al*. 2020) It has been suggested that MDA should be avoided where possible, as it can result in less diversity and less reproducible outputs. Some studies have completely avoided this step to evade potential amplification biases (Manrique *et al*. 2016). To overcome this bias, improvements in the library preparation protocols are required which will also lead to metagenomic studies of the human gut phageome becoming more representative of the true composition (Roux *et al*. 2016).

A study by Lim and associates in 2015 utilised both MDA and sequence-independent amplification (SIA) of the DNA and RNA, although they noted that the SIA method is less

sensitive for DNA viruses (Lim *et al*. 2015). It was used to balance the MDA as it can capture RNA viruses, although it is generally less sensitive in terms of DNA virus representation. The SIA method involved in this study incorporated base-balanced specific 16nt sequence upstream of a random 15-mer for random priming. It is based on the flanking of unknown sequences with known sequences to enable PCR amplification (Bohlander *et al*. 1992). Nonetheless, there were no RNA phages detected in the SIA-generated data and the authors focused on the MDA-generated data to make their results comparable with other virome and phageome studies (Lim *et al*. 2015). They identified picobirnaviruses that were previously classified as eukaryotic viruses but have more recently have been suggested to be dsRNA phages as they contain conserved prokaryotic ribosomal binding sites (Krishnamurthy and Wang 2018). Lim *et al*. also identified eukaryotic RNA viruses including *Caliciviridae*, *Picornaviridae* and *Astroviridae* which are +ssRNA viruses (with non-segmented genomes) from the infant faecal samples. This is consistent with results from other PCR-based studies (Kapusinszky, Minor, and Delwart 2012). In a 2015 study by Norman and colleagues, there was also quite a high relative abundance of dsRNA viruses, retro-transcribing viruses (these contain ssRNA), and unclassified phages, some of which may represent undefined RNA phages (Norman *et al*. 2015). These viral and phage contigs may be detected as a result of alterations made to a previous protocol including the removal of an RNase step.

Some studies have attempted to incorporate the typical protocols for RNA phage isolation from the VLP fraction, such as reverse transcription (RT), but not all record recovering RNA phage (Figure 2). The resulting DNA, known as complementary DNA (cDNA) can then be used as a template for PCR reactions. There may also be a MDA step coupled with the RT step in order to convert a sample from an ssDNA/ssRNA heteroduplex into dsDNA. This is crucial for the enzyme associated with the library preparation kit to work

efficiently as they are selective to the nucleic acid sample, in the case of the transposase in the Nextera XT Library Preparation Kit which requires dsDNA input.

Study ID	Method	Buffer	Filter	Nuclease	PEG	MDA	RT	TFF	CsCl	Chloroform	Sequencing	RNA viruses
d'Humières *et al.* (2019)	Method I	PBS	2.0µM, 0.45µM, 0.22µM, Ultrafiltration	DNase	×	×	×	×	×	×	Illumina	×
	Method II	PBS	2.0µM, 0.45µM, 0.22µM, Ultrafiltration	DNase	×	×	×	×	×	×	Illumina	×
	Method III	PBS	2.0µM, 0.45µM, 0.22µM	DNase	×	✓	×	×	✓	×	Illumina	×
	Method IV	PBS	2.0µM, 0.45µM	DNase	✓	×	×	×	×	×	Illumina	×
	Method V	PBS	2.0µM, 0.45µM	DNase	×	×	×	×	×	×	Illumina	×
Shkoporov *et al.* (2018)		SM	0.45µM	DNase, RNase	✓	✓	✓	×	×	✓	Illumina	✓
Manrique *et al.* (2016)		SM	0.45µM	DNase	×	×	×	×	×	×	Illumina	×
Ly *et al.* (2016)		SM	0.45µM, 0.2µM	DNase	×	✓	×	×	✓	×	Ion Semiconductor sequencing	×
Monaco *et al.* (2016)		SM	0.45µM, 0.22µM (x2)	DNase	×	✓	×	×	×	✓	Illumina	×
Lim *et al.* (2015)		PBS	0.45µM	No	×	✓	✓	×	×	×	Illumina	✓
Norman *et al.* (2015)		SM	None	DNase	×	✓	×	×	×	✓	Illumina	×
Reyes *et al.* (2015)		SM	0.45µM, 0.22µM	DNase	×	✓	×	×	✓	✓	454 sequencing	×
Minot *et al.* (2013)		SM	0.22µM, Ultrafiltration	DNase	×	✓	×	×	×	✓	Illumina	×
Minot *et al.* (2011)		SM	0.22µM	DNase	×	✓	×	×	✓	✓	454 sequencing	×
Reyes *et al.* (2010)		SM	0.45µM, 0.22µM	DNase	×	✓	×	×	✓	✓	454 sequencing	×
Zhang *et al.* (2005)		SM	Nitrex filter	DNase, RNase	×	×	×	✓	×	×	Invitrogen	✓
Breitbart *et al.* (2003)		PBS	Nitrex filter	No	×	×	×	✓	✓	×	Psmart vector, MC12 cells and AmpL2 forward primer	×

Figure 2. Summarised extraction method comparisons from a selection of recent human gut virome papers. All studies included in this review analysed the composition of the human gut virome based on faeces as the starting material (d'Humières *et al*. 2019; Shkoporov, Ryan, *et al*. 2018; Manrique *et al*. 2016; Ly *et al*. 2016; Monaco *et al*. 2016; Lim *et al*. 2015; Norman *et al*. 2015; Reyes *et al*. 2015; Minot *et al*. 2013; 2011; Reyes *et al*. 2010; Zhang *et al*. 2005; Breitbart *et al*. 2003). The different sample handling methods, procedure for extracting the VLP fraction, and sequencing technology are examined. In the 2019 study by d'Humières and colleagues, method I includes ultrafiltration whereas method II involves both ultrafiltration and ultracentrifugation.

Shkoporov and colleagues acknowledged the fact most studies on the human phageome have neglected to study the RNA fraction, with a few notable exceptions, and in their study, attempted to incorporate the RNA viral consortium by including an RT step (Shkoporov, Ryan, *et al*. 2018). Interestingly, the authors highlight the fact the protocol may have failed to quantitatively recover the small +ssRNA phage that was deliberately spiked in the sample.

In a preliminary small-scale RNA-focused study, Zhang and colleagues performed a metagenomic study of the uncultured RNA viruses residing in the human gut and indicated that the majority (>95%) of these were plant viruses, the most abundant found to be pepper mild mottle virus (PMMV) (Zhang *et al*. 2005). They did detect a large amount of hits for the animal virus *Picobirnavirus*, which, as previously mentioned, may in fact be a dsRNA phage (Krishnamurthy and Wang 2018). However, in this study, the method of extraction was based on TFF and both DNase and RNase were added which may affect the sensitivity of this method to find other RNA viruses and phages. This finding led to a debate of to what extent RNA viruses and phages inhabit the human gut and if there is any significance in attempting to capture these entities. Several studies have been published that demonstrate the abundance of RNA phages in non-faecal metatranscriptomic samples such as activated sludge, seawater, insect, and avian samples (Callanan *et al*. 2020; Starr *et al*. 2019; Shi *et al*. 2016; Krishnamurthy *et al*. 2016).

Next-generation sequencing (NGS) includes approaches that are non-Sanger method-based high-throughput DNA or cDNA sequencing methods which, in brief, operate by the initial fragmentation of the DNA/RNA into shorter fragments, the ligation of terminal adapter sequences, amplification and sequencing of these libraries (based on one out of several available chemical or physical principles), and, finally, an attempt to assemble these short sequences into larger contigs, or even complete genomes. Following the advances made in NGS in recent years, cost-effective and rapid sequencing platforms such as Illumina HiSeq,

and third generation long-read sequencing platforms such as Oxford Nanopore, have become more accessible. Prior to sequencing, NGS libraries are prepared from the isolated viral nucleic acids which have been fragmented to particular lengths to comply with the specific sequencing platform chosen. Subsequently, there may be a series of preparation steps in which special adapters are added to allow single entities to be identified when the samples are pooled for sequencing runs. These adapters also provide priming sites for amplification after ligation, priming sites during isothermal bridge amplification inside flow cells and at the sequencing step. Enhanced library preparation, such as Nextera XT and Accel-NGS® 1S DNA library kits, have allowed for quicker and more efficient sequencing from limited amounts of the DNA/cDNA starting material. Prior to library preparation, the proper type and amount input material is essential, for example Nextera XT requires dsDNA input so any sample with a potential DNA/RNA heteroduplex will require additional treatment. Once the library has been prepared, sequencing using a NGS-specific platform, such as Illumina MiSeq or HiSeq, is performed and the results are then analysed.

The sequencing depth of the samples is also something that needs to be considered prior to selecting the sequencing platform as different platforms offer varying length (bp), throughput, and number of reads and can often dramatically range in cost per gigabyte (Gb) (Hölzer and Marz 2017). For example, considering short-read NGS using Illumina with long-read NGS by Oxford Nanopore (MinION), there are notable difference in length (25-300bp vs up to 200kb, respectively), throughput (2-900Gb vs up to 1.5Gb, respectively) and number of reads (10M-4B vs >100k, respectively) and these factors need to be addressed in order to select the best platform for the specific study. The majority of studies examined in this review used Illumina (usually either HiSeq or MiSeq) as the preferred sequencing platform.

A study by Castro-Mejía and colleagues attempted to optimise the extraction and purification of phages from human faecal samples prior to metagenomic analysis (Castro-Mejía

et al. 2015). They separated the process into two parts; the pre-processing, which included the spiking of three phages, and the purification which included PEG, TFF, and an adapted method from the literature. Despite the fact these protocols were found to be highly efficient in the purification of DNA phages prior to high-throughput sequencing for phage-metavirome studies, their efficacy at recovering RNA viruses from such samples is yet to be tested.

2.6 Bioinformatic pipelines

In order to validate the various isolation protocols and combinations, a specific and robust bioinformatic pipeline is essential. Following on from the VLP isolation, and extraction and sequencing of the nucleic acids, the resulting viral sequences are analysed in order to identify and characterise the viral contigs. A recent literature review by Nooij and colleagues examined 49 bioinformatic workflows for viral metagenomics which led to the creation of two decision trees which can be applied to a variety of viral analyses (Nooij *et al*. 2018). The vast amount of data derived from NGS has resulted in challenges with the quality analysis and the processing of the sequences. To help circumvent the demanding nature of some of these processes and to make metagenomic analyses more accessible, online tools and resources have been developed. There are virome-specific programs such as Viral MetaGenome Annotation Pipeline (VMGAP) (Lorenzi *et al*. 2011), Viral Informatics Resource for Metagenomic Exploration (VIROME) (Wommack *et al*. 2012), and Metavir 2 (Roux *et al*. 2014). The aforementioned tools are dependent on reference databases as they operate on a similarity-based system, but there are several similarity-independent resources that have also been developed such as PHAge Communities from Contig Spectrum (PHACCS) (Angly *et al*. 2005). This enables the user to bypass issues that may arise due to a lack of sequence similarity in databases (Lorenz *et al*. 2005).

The complete collection of bioinformatic resources has been reviewed and new tools are constantly emerging in an attempt to better analyse the sequences (Sharma, Priyadarshini, and Vrati 2015). These include VirSorter (Roux, Enault, *et al*. 2015), DemoVir (Ryan 2018), DeepVirFinder (Ren *et al*. 2020), Detection & Analysis of viral and Microbial Infectious Agents by NGS (DAMIAN) (Alawi *et al*. 2019) and numerous others. There are also studies that use tailor-made pipelines, such as that by Monaco and colleagues who used a bioinformatics pipeline, VirusSeeker, to analyse their viral sequences (Monaco *et al*. 2016).

The choice of assembly software used in different studies may offer a source of differentiation in studies as it has been recently shown that this has a critical impact on the recovery of the viral contigs (see Figure 3 for basic bioinformatic pipeline) (Sutton and Hill 2019). Certain criteria, such as genome circularity, contig length, presence of particular phage proteins, and percentage identity to known viruses, are also applied to further filter the viral contigs. However, it should be noted that many of these filters could remove the RNA viruses, for instance, the particular step detecting circular genomes would exclude all known RNA phages. In a 2015 paper by Reyes and colleagues, where the DNA gut virome of Malawian twins was analysed, circular contigs were used as a criterion and revealed three distinct size ranges for circular contigs: (i) >30kb (the reported size range for circular dsDNA phages belonging to the *Caudovirales* order); (ii) 6-7 kb (size reported for ssDNA phages in the *Microviridae* family, particularly the *Alpavirinae*); and (iii) 3-4 kb (expected size for ssDNA eukaryotic viruses in the *Anelloviridae* family) (Reyes *et al*. 2015). It is also important that assembly statistics are reported in studies to evaluate the quality of the assembly. One such example is the N50 which, in simple terms, denotes the shortest contig used to represent 50% of the assembled genome (Salzberg *et al*. 2012). Therefore, it is a measure of the quality of assembled genomes and the degree of fragmentation.

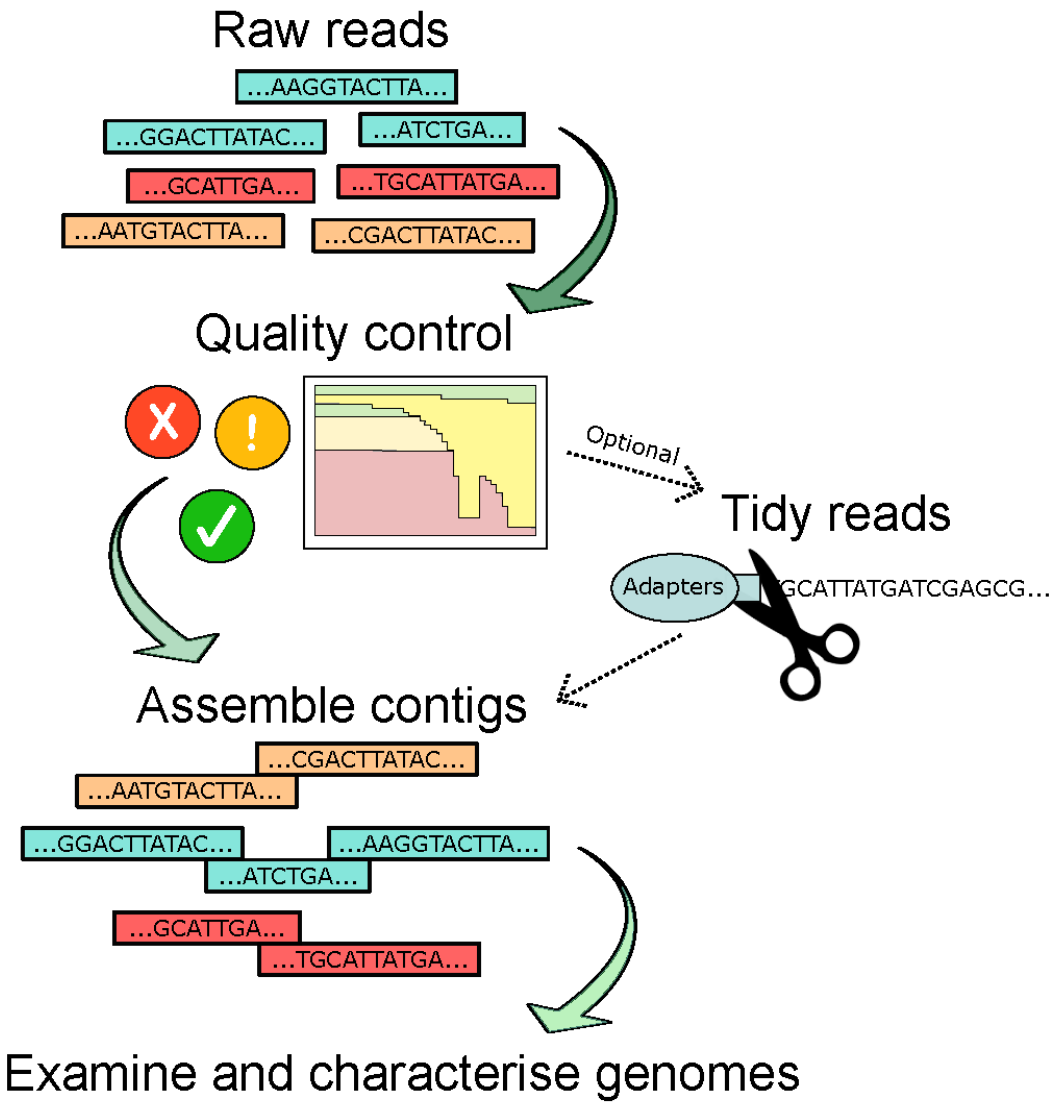


Figure 3. Overview of genome assembly in as part of the bioinformatic pipeline used for virome/phageome analyses.

Decontamination of samples to remove bacterial and other non-viral sequences can be done by positive or negative selection, i.e., filtration and selection of viral contigs from the total sample or the identification and removal of non-viral sequences, respectively. The compositional profile of the viral sample is often assessed by aligning the assembled reads to a reference database of known viruses using Basic Local Alignment Search Tool (BLAST) or other BLAST-based programs. Alternative sequence analysis methods such as *k*-mer

algorithms, such as VirFinder and Libra, can reduce the time required for these analyses (Choi *et al*. 2019; Ren *et al*. 2017). It is important to note that these programs can be extremely computationally heavy to work at such speeds. This *k-mer* based method is rare but has been used in some virome studies, including the gut virome study by Norman and colleagues (Norman *et al*. 2015). These searches are restricted to reference databases like the NCBI Reference Sequence (RefSeq) database that are limited in the level of annotation they offer as they represent only a modest proportion of the total global virome ("Viral Genomes" 2021). Other references databases also exist such as Reference Viral Database (RVDB), which includes all viral sequences except for bacterial viruses (Goodacre *et al*. 2018), ViPR database (Pickett *et al*. 2012), and GenBank, which is a collection of all annotated sequences ("GenBank Overview" 2021).

There are also custom-built profile hidden Markov model (HMM) databases generated through the collection of conserved viral proteins e.g. the Prokaryotic Virus Orthologous Groups (pVOGs) database (Grazziotin, Koonin, and Kristensen 2017). Another example is the recent publication which utilised a profile-HMM database of conserved +ssRNA phage proteins to expand the number of these entities from tens to thousands (Callanan *et al*. 2020). This tool will enable the identification of +ssRNA phages that are somewhat closely related to those already known and more distantly related strains from future studies.

In the MIUViG paper, Roux and colleagues discuss how the numbers of viral reference databases are being created at extraordinary rates but these are rarely deeply examined for inflated dataset novelty (Roux *et al*. 2019). As the number of virome studies increases year on year, the number and breath of the databases available should increase which will allow for more robust assigning of identity to viral contigs.

2.7 Discussion

Efforts to examine the phageome and overall virome of the human gut through the analysis of the viruses and phages from a faecal sample are reliant on the accuracy and reproducibility of various protocols and subsequent analyses. Shkoporov *et al*. noted that the available data on the conclusive concentration of the VLP fraction of human faecal samples are repeatedly contradictory and heavily dependent on the extraction and quantification procedures applied (Shkoporov and Hill 2019). Throughout the various studies there have been notable sources of bias, ranging from the inclusion of MDA treatment to the addition of RNase. By highlighting these biases, efforts to avoid such procedures can be made and alternative techniques can be tested. It is also essential that for all viruses and phages to be equally represented, both the RNA and DNA portions need to be examined in future phage-metavirome studies.

2.8 References

Acheson, Nicholas H., and Igor Tamm. 1970. "Ribonuclease Sensitivity of Semliki Forest Virus Nucleocapsids." Journal of Virology 5 (6): 714–17. https://doi.org/10.1128/JVI.5.6.714-717.1970.

Adams, M. H. 1949. "The Stability of Bacterial Viruses in Solutions of Salts." The Journal of General Physiology 32 (5): 579–94. https://doi.org/10.1085/jgp.32.5.579.

Adriaenssens, Evelien M., Kata Farkas, Christian Harrison, David L. Jones, Heather E. Allison, and Alan J. McCarthy. 2018. "Viromic Analysis of Wastewater Input to a River Catchment Reveals a Diverse Assemblage of RNA Viruses." MSystems 3 (3). https://doi.org/10.1128/mSystems.00025-18.

Alawi, Malik, Lia Burkhardt, Daniela Indenbirken, Kerstin Reumann, Maximilian Christopeit, Nicolaus Kröger, Marc Lütgehetmann, Martin Aepfelbacher, Nicole Fischer, and Adam Grundhoff. 2019. "DAMIAN: An Open Source Bioinformatics Tool for Fast, Systematic

and Cohort Based Analysis of Microorganisms in Diagnostic Samples." Scientific Reports 9 (1): 16841. https://doi.org/10.1038/s41598-019-52881-4.

Angly, Florent, Beltran Rodriguez-Brito, David Bangor, Pat McNairnie, Mya Breitbart, Peter Salamon, Ben Felts, James Nulton, Joseph Mahaffy, and Forest Rohwer. 2005. "PHACCS, an Online Tool for Estimating the Structure and Diversity of Uncultured Viral Communities Using Metagenomic Information." BMC Bioinformatics 6 (1): 41. https://doi.org/10.1186/1471-2105-6-41.

Bohlander, Stefan K., Rafael Espinosa, Michelle M. Le Beau, Janet D. Rowley, and Manuel O. Díaz. 1992. "A Method for the Rapid Sequence-Independent Amplification of Microdissected Chromosomal Material." Genomics 13 (4): 1322–24. https://doi.org/10.1016/0888-7543(92)90057-Y.

Breitbart, Mya, Ian Hewson, Ben Felts, Joseph M. Mahaffy, James Nulton, Peter Salamon, and Forest Rohwer. 2003. "Metagenomic Analyses of an Uncultured Viral Community from Human Feces." Journal of Bacteriology 185 (20): 6220–23. https://doi.org/10.1128/jb.185.20.6220-6223.2003.

Bushman, Frederic. 2002. Lateral DNA Transfer. Cold Spring Harbor Laboratory Press. https://agris.fao.org/agris-search/search.do?recordID=US201300089728.

Callanan, Julie, Stephen R. Stockdale, Andrey N. Shkoporov, Lorraine A. Draper, R. Paul Ross, and Colin Hill. 2020. "Expansion of Known ssRNA Phage Genomes: From Tens to over a Thousand." Science Advances 6 (6): eaay5981. https://doi.org/10.1126/sciadv.aay5981.

Casey, Aidan, Kieran Jordan, Horst Neve, Aidan Coffey, and Olivia McAuliffe. 2015. "A Tail of Two Phages: Genomic and Functional Analysis of Listeria Monocytogenes Phages VB_LmoS_188 and VB_LmoS_293 Reveal the Receptor-Binding Proteins Involved in

Host Specificity." Frontiers in Microbiology 6. https://doi.org/10.3389/fmicb.2015.01107.

Castro-Mejía, Josué L., Musemma K. Muhammed, Witold Kot, Horst Neve, Charles M. A. P. Franz, Lars H. Hansen, Finn K. Vogensen, and Dennis S. Nielsen. 2015. "Optimizing Protocols for Extraction of Bacteriophages Prior to Metagenomic Analyses of Phage Communities in the Human Gut." Microbiome 3 (1): 64. https://doi.org/10.1186/s40168-015-0131-4.

Chibani-Chennoufi, Sandra, Anne Bruttin, Marie-Lise Dillmann, and Harald Brüssow. 2004. "Phage-Host Interaction: An Ecological Perspective." Journal of Bacteriology 186 (12): 3677–86. https://doi.org/10.1128/JB.186.12.3677-3686.2004.

Choi, Illyoung, Alise J Ponsero, Matthew Bomhoff, Ken Youens-Clark, John H Hartman, and Bonnie L Hurwitz. 2019. "Libra: Scalable k- Mer–Based Tool for Massive All-vs-All Metagenome Comparisons." GigaScience 8 (2). https://doi.org/10.1093/gigascience/giy165.

Conceição-Neto, Nádia, Mark Zeller, Hanne Lefrère, Pieter De Bruyn, Leen Beller, Ward Deboutte, Claude Kwe Yinda, *et al*. 2015. "Modular Approach to Customise Sample Preparation Procedures for Viral Metagenomics: A Reproducible Protocol for Virome Analysis." Scientific Reports 5 (November): 16532. https://doi.org/10.1038/srep16532.

Czajkowski, Robert, Zofia Ozymko, and Ewa Lojkowska. 2016. "Application of Zinc Chloride Precipitation Method for Rapid Isolation and Concentration of Infectious Pectobacterium Spp. and Dickeya Spp. Lytic Bacteriophages from Surface Water and Plant and Soil Extracts." Folia Microbiologica 61: 29–33. https://doi.org/10.1007/s12223-015-0411-1.

Dean, Frank B., Seiyu Hosono, Linhua Fang, Xiaohong Wu, A. Fawad Faruqi, Patricia Bray-Ward, Zhenyu Sun, *et al*. 2002. "Comprehensive Human Genome Amplification Using

Multiple Displacement Amplification." Proceedings of the National Academy of Sciences 99 (8): 5261–66. https://doi.org/10.1073/pnas.082089499.

Duhaime, Melissa B., Li Deng, Bonnie T. Poulos, and Matthew B. Sullivan. 2012. "Towards Quantitative Metagenomics of Wild Viruses and Other Ultra-Low Concentration DNA Samples: A Rigorous Assessment and Optimization of the Linker Amplification Method." Environmental Microbiology 14 (9): 2526–37. https://doi.org/10.1111/j.1462-2920.2012.02791.x.

Dutilh, Bas E., Noriko Cassman, Katelyn McNair, Savannah E. Sanchez, Genivaldo G. Z. Silva, Lance Boling, Jeremy J. Barr, *et al*. 2014. "A Highly Abundant Bacteriophage Discovered in the Unknown Sequences of Human Faecal Metagenomes." Nature Communications 5 (1): 4498. https://doi.org/10.1038/ncomms5498.

Džunková, Mária, Marc Garcia-Garcerà, Llúcia Martínez-Priego, Giussepe D'Auria, Francesc Calafell, and Andrés Moya. 2014. "Direct Squencing from the Minimal Number of DNA Molecules Needed to Fill a 454 Picotiterplate." PloS One 9 (6): e97379. https://doi.org/10.1371/journal.pone.0097379.

Edwards, Robert A., Alejandro A. Vega, Holly M. Norman, Maria Ohaeri, Kyle Levi, Elizabeth A. Dinsdale, Ondrej Cinek, *et al*. 2019. "Global Phylogeography and Ancient Evolution of the Widespread Human Gut Virus CrAssphage." Nature Microbiology 4 (10): 1727–36. https://doi.org/10.1038/s41564-019-0494-6.

Fauquet, Claude, M.A. Mayo, J. Maniloff, U. Desselberger, and L.A. Ball. 2005. "Virus Taxonomy - Eighth Report of the International Committee on the Taxonomy of Viruses." The Viruses 83 (July): 988–92.

Forster, Samuel C., Nitin Kumar, Blessing O. Anonye, Alexandre Almeida, Elisa Viciani, Mark D. Stares, Matthew Dunn, *et al*. 2019. "A Human Gut Bacterial Genome and

Culture Collection for Improved Metagenomic Analyses." Nature Biotechnology 37 (2): 186–92. https://doi.org/10.1038/s41587-018-0009-7.

Garmaeva, Sanzhima, Trishla Sinha, Alexander Kurilshikov, Jingyuan Fu, Cisca Wijmenga, and Alexandra Zhernakova. 2019. "Studying the Gut Virome in the Metagenomic Era: Challenges and Perspectives." BMC Biology 17 (1): 84. https://doi.org/10.1186/s12915-019-0704-y.

"GenBank Overview." 2021 Accessed January 28, 2021. https://www.ncbi.nlm.nih.gov/genbank/.

Goodacre, Norman, Aisha Aljanahi, Subhiksha Nandakumar, Mike Mikailov, and Arifa S. Khan. 2018. "A Reference Viral Database (RVDB) To Enhance Bioinformatics Analysis of High-Throughput Sequencing for Novel Virus Detection." MSphere 3 (2). https://doi.org/10.1128/mSphereDirect.00069-18.

Gorzelak, Monika A., Sandeep K. Gill, Nishat Tasnim, Zahra Ahmadi-Vand, Michael Jay, and Deanna L. Gibson. 2015. "Methods for Improving Human Gut Microbiome Data by Reducing Variability through Sample Processing and Storage of Stool." PloS One 10 (8): e0134802. https://doi.org/10.1371/journal.pone.0134802.

Grazziotin, Ana Laura, Eugene V. Koonin, and David M. Kristensen. 2017. "Prokaryotic Virus Orthologous Groups (PVOGs): A Resource for Comparative Genomics and Protein Family Annotation." Nucleic Acids Research 45 (Database issue): D491–98. https://doi.org/10.1093/nar/gkw975.

Gregory, Ann C., Olivier Zablocki, Ahmed A. Zayed, Allison Howell, Benjamin Bolduc, and Matthew B. Sullivan. 2020. "The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut." Cell Host & Microbe 28 (5): 724-740.e8. https://doi.org/10.1016/j.chom.2020.08.003.

Guerin, Emma, and Colin Hill. 2020. “Shining Light on Human Gut Bacteriophages.” Frontiers in Cellular and Infection Microbiology 10: 481. https://doi.org/10.3389/fcimb.2020.00481.

Guerin, Emma, Andrey Shkoporov, Stephen R. Stockdale, Adam G. Clooney, Feargal J. Ryan, Thomas D. S. Sutton, Lorraine A. Draper, Enrique Gonzalez-Tortuero, R. Paul Ross, and Colin Hill. 2018. “Biology and Taxonomy of CrAss-like Bacteriophages, the Most Abundant Virus in the Human Gut.” Cell Host & Microbe 24 (5): 653-664.e6. https://doi.org/10.1016/j.chom.2018.10.002.

Hill, Janet E., Susanne L. Penny, Kenneth G. Crowell, Swee Han Goh, and Sean M. Hemmingsen. 2004. “CpnDB: A Chaperonin Sequence Database.” Genome Research 14 (8): 1669–75. https://doi.org/10.1101/gr.2649204.

Hölzer, Martin, and Manja Marz. 2017. “Software Dedicated to Virus Sequence Analysis ‘Bioinformatics Goes Viral.’” Advances in Virus Research 99: 233–57. https://doi.org/10.1016/bs.aivir.2017.08.004.

Hoyles, Lesley, Anne L. McCartney, Horst Neve, Glenn R. Gibson, Jeremy D. Sanderson, Knut J. Heller, and Douwe van Sinderen. 2014. “Characterization of Virus-like Particles Associated with the Human Faecal and Caecal Microbiota.” Research in Microbiology 165 (10): 803–12. https://doi.org/10.1016/j.resmic.2014.10.006.

Hsu, Bryan B., Travis E. Gibson, Vladimir Yeliseyev, Qing Liu, Lorena Lyon, Lynn Bry, Pamela A. Silver, and Georg K. Gerber. 2019. “Dynamic Modulation of the Gut Microbiota and Metabolome by Bacteriophages in a Mouse Model.” Cell Host & Microbe 25 (6): 803-814.e5. https://doi.org/10.1016/j.chom.2019.05.001.

Humières, Camille d’, Marie Touchon, Sara Dion, Jean Cury, Amine Ghozlane, Marc Garcia-Garcera, Christiane Bouchier, Laurence Ma, Erick Denamur, and Eduardo P.C.Rocha. 2019. “A Simple, Reproducible and Cost-Effective Procedure to Analyse Gut Phageome:

From Phage Isolation to Bioinformatic Approach." Scientific Reports 9 (1): 11331. https://doi.org/10.1038/s41598-019-47656-w.

Kapusinszky, Beatrix, Philip Minor, and Eric Delwart. 2012. "Nearly Constant Shedding of Diverse Enteric Viruses by Two Healthy Infants." Journal of Clinical Microbiology 50 (11): 3427–34. https://doi.org/10.1128/JCM.01589-12.

Khan Mirzaei, Mohammadali, Md. Anik Ashfaq Khan, Prakash Ghosh, Zofia E. Taranu, Mariia Taguer, Jinlong Ru, Rajashree Chowdhury, *et al*. 2020. "Bacteriophages Isolated from Stunted Children Can Regulate Gut Bacterial Communities in an Age-Specific Manner." Cell Host & Microbe 27 (2): 199-212.e5. https://doi.org/10.1016/j.chom.2020.01.004.

Kleiner, Manuel, Lora V. Hooper, and Breck A. Duerkop. 2015. "Evaluation of Methods to Purify Virus-like Particles for Metagenomic Sequencing of Intestinal Viromes." BMC Genomics 16 (January): 7. https://doi.org/10.1186/s12864-014-1207-4.

Krishnamurthy, Siddharth R., Andrew B. Janowski, Guoyan Zhao, Dan Barouch, and David Wang. 2016. "Hyperexpansion of RNA Bacteriophage Diversity." PLOS Biology 14 (3): e1002409. https://doi.org/10.1371/journal.pbio.1002409.

Krishnamurthy, Siddharth R., and David Wang. 2017. "Origins and Challenges of Viral Dark Matter." Virus Research 239 (July): 136–42. https://doi.org/10.1016/j.virusres.2017.02.002.

Krishnamurthy, Siddharth R., and David Wang. 2018. "Extensive Conservation of Prokaryotic Ribosomal Binding Sites in Known and Novel Picobirnaviruses." Virology 516 (March): 108–14. https://doi.org/10.1016/j.virol.2018.01.006.

Lim, Efrem S., Yanjiao Zhou, Guoyan Zhao, Irma K. Bauer, Lindsay Droit, I. Malick Ndao, Barbara B. Warner, Phillip I. Tarr, David Wang, and Lori R. Holtz. 2015. "Early Life Dynamics of the Human Gut Virome and Bacterial Microbiome in Infants." Nature Medicine 21 (10): 1228–34. https://doi.org/10.1038/nm.3950.

Links, Matthew G., Tim J. Dumonceaux, Sean M. Hemmingsen, and Janet E. Hill. 2012. “The Chaperonin-60 Universal Target Is a Barcode for Bacteria That Enables De Novo Assembly of Metagenomic Sequence Data.” PLOS ONE 7 (11): e49755. https://doi.org/10.1371/journal.pone.0049755.

Lorenz, Joseph G., Whitney E. Jackson, Jeanne C. Beck, and Robert Hanner. 2005. “The Problems and Promise of DNA Barcodes for Species Diagnosis of Primate Biomaterials.” Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 360 (1462): 1869–77. https://doi.org/10.1098/rstb.2005.1718.

Lorenzi, Hernan A., Jeff Hoover, Jason Inman, Todd Safford, Sean Murphy, Leonid Kagan, and Shannon J. Williamson. 2011. “TheViral MetaGenome Annotation Pipeline(VMGAP):An Automated Tool for the Functional Annotation of Viral Metagenomic Shotgun Sequencing Data.” Standards in Genomic Sciences 4 (3): 418–29. https://doi.org/10.4056/sigs.1694706.

Łoś, Marcin, and Grzegorz Węgrzyn. 2012. “Pseudolysogeny.” Advances in Virus Research 82: 339–49. https://doi.org/10.1016/B978-0-12-394621-8.00019-4.

Ly, Melissa, Marcus B. Jones, Shira R. Abeles, Tasha M. Santiago-Rodriguez, Jonathan Gao, Ivan C. Chan, Chandrabali Ghose, and David T. Pride. 2016. “Transmission of Viruses via Our Microbiomes.” Microbiome 4 (1): 64. https://doi.org/10.1186/s40168-016-0212-z.

Manrique, Pilar, Benjamin Bolduc, Seth T. Walk, John van der Oost, Willem M. de Vos, and Mark J. Young. 2016. “Healthy Human Gut Phageome.” Proceedings of the National Academy of Sciences 113 (37): 10400–405. https://doi.org/10.1073/pnas.1601060113.

McCann, Angela, Feargal J. Ryan, Stephen R. Stockdale, Marion Dalmasso, Tony Blake, C. Anthony Ryan, Catherine Stanton, Susan Mills, Paul R. Ross, and Colin Hill. 2018.

"Viromes of One Year Old Infants Reveal the Impact of Birth Mode on Microbiome Diversity." PeerJ 6: e4694. https://doi.org/10.7717/peerj.4694.

McDaniel, Lauren D., Elizabeth Young, Jennifer Delaney, Fabian Ruhnau, Kim B. Ritchie, and John H. Paul. 2010. "High Frequency of Horizontal Gene Transfer in the Oceans." Science (New York, N.Y.) 330 (6000): 50. https://doi.org/10.1126/science.1192243.

Minot, Samuel, Alexandra Bryson, Christel Chehoud, Gary D. Wu, James D. Lewis, and Frederic D. Bushman. 2013. "Rapid Evolution of the Human Gut Virome." Proceedings of the National Academy of Sciences 110 (30): 12450–55. https://doi.org/10.1073/pnas.1300833110.

Minot, Samuel, Rohini Sinha, Jun Chen, Hongzhe Li, Sue A. Keilbaugh, Gary D. Wu, James D. Lewis, and Frederic D. Bushman. 2011. "The Human Gut Virome: Inter-Individual Variation and Dynamic Response to Diet." Genome Research 21 (10): 1616–25. https://doi.org/10.1101/gr.122705.111.

Monaco, Cynthia L., David B. Gootenberg, Guoyan Zhao, Scott A. Handley, Musie S. Ghebremichael, Efrem S. Lim, Alex Lankowski, *et al*. 2016. "Altered Virome and Bacterial Microbiome in Human Immunodeficiency Virus-Associated Acquired Immunodeficiency Syndrome." Cell Host & Microbe 19 (3): 311–22. https://doi.org/10.1016/j.chom.2016.02.011.

Nooij, Sam, Dennis Schmitz, Harry Vennema, Annelies Kroneman, and Marion P. G. Koopmans. 2018. "Overview of Virus Metagenomic Classification Methods and Their Biological Applications." Frontiers in Microbiology 9. https://doi.org/10.3389/fmicb.2018.00749.

Norman, Jason M., Scott A. Handley, Megan T. Baldridge, Lindsay Droit, Catherine Y. Liu, Brian C. Keller, Amal Kambal, *et al*. 2015. "Disease-Specific Alterations in the Enteric

Virome in Inflammatory Bowel Disease." Cell 160 (3): 447–60. https://doi.org/10.1016/j.cell.2015.01.002.

Ogilvie, Lesley A., and Brian V. Jones. 2015. "The Human Gut Virome: A Multifaceted Majority." Frontiers in Microbiology 6: 918. https://doi.org/10.3389/fmicb.2015.00918.

Pickett, Brett E., Eva L. Sadat, Yun Zhang, Jyothi M. Noronha, R. Burke Squires, Victoria Hunt, Mengya Liu, *et al*. 2012. "ViPR: An Open Bioinformatics Database and Analysis Resource for Virology Research." Nucleic Acids Research 40 (Database issue): D593-598. https://doi.org/10.1093/nar/gkr859.

Ren, Jie, Nathan A. Ahlgren, Yang Young Lu, Jed A. Fuhrman, and Fengzhu Sun. 2017. "VirFinder: A Novel k-Mer Based Tool for Identifying Viral Sequences from Assembled Metagenomic Data." Microbiome 5 (1): 69. https://doi.org/10.1186/s40168-017-0283-5.

Ren, Jie, Kai Song, Chao Deng, Nathan A. Ahlgren, Jed A. Fuhrman, Yi Li, Xiaohui Xie, Ryan Poplin, and Fengzhu Sun. 2020. "Identifying Viruses from Metagenomic Data Using Deep Learning." Quantitative Biology 8 (1): 64–77. https://doi.org/10.1007/s40484-019-0187-4.

Reyes, Alejandro, Laura V. Blanton, Song Cao, Guoyan Zhao, Mark Manary, Indi Trehan, Michelle I. Smith, *et al*. 2015. "Gut DNA Viromes of Malawian Twins Discordant for Severe Acute Malnutrition." Proceedings of the National Academy of Sciences 112 (38): 11941–46. https://doi.org/10.1073/pnas.1514285112.

Reyes, Alejandro, Matthew Haynes, Nicole Hanson, Florent E. Angly, Andrew C. Heath, Forest Rohwer, and Jeffrey I. Gordon. 2010. "Viruses in the Fecal Microbiota of Monozygotic Twins and Their Mothers." Nature 466 (7304): 334–38. https://doi.org/10.1038/nature09199.

Roux, Simon, Evelien M. Adriaenssens, Bas E. Dutilh, Eugene V. Koonin, Andrew M. Kropinski, Mart Krupovic, Jens H. Kuhn, *et al*. 2019. "Minimum Information about an

Uncultivated Virus Genome (MIUViG).” Nature Biotechnology 37 (1): 29–37. https://doi.org/10.1038/nbt.4306.

Roux, Simon, Francois Enault, Bonnie L. Hurwitz, and Matthew B. Sullivan. 2015. “VirSorter: Mining Viral Signal from Microbial Genomic Data.” PeerJ 3 (May): e985. https://doi.org/10.7717/peerj.985.

Roux, Simon, Steven J Hallam, Tanja Woyke, and Matthew B Sullivan. 2015. “Viral Dark Matter and Virus–Host Interactions Resolved from Publicly Available Microbial Genomes.” Edited by Richard A Neher. ELife 4 (July): e08490. https://doi.org/10.7554/eLife.08490.

Roux, Simon, Natalie E. Solonenko, Vinh T. Dang, Bonnie T. Poulos, Sarah M. Schwenck, Dawn B. Goldsmith, Maureen L. Coleman, Mya Breitbart, and Matthew B. Sullivan. 2016. “Towards Quantitative Viromics for Both Double-Stranded and Single-Stranded DNA Viruses.” PeerJ 4 (December): e2777. https://doi.org/10.7717/peerj.2777.

Roux, Simon, Jeremy Tournayre, Antoine Mahul, Didier Debroas, and François Enault. 2014. “Metavir 2: New Tools for Viral Metagenome Comparison and Assembled Virome Analysis.” BMC Bioinformatics 15 (1): 76. https://doi.org/10.1186/1471-2105-15-76.

Ryan, Fergal 2018. Demovir. R. https://github.com/feargalr/Demovir.

Salzberg, Steven L., Adam M. Phillippy, Aleksey Zimin, Daniela Puiu, Tanja Magoc, Sergey Koren, Todd J. Treangen, *et al*. 2012. “GAGE: A Critical Evaluation of Genome Assemblies and Assembly Algorithms.” Genome Research 22 (3): 557–67. https://doi.org/10.1101/gr.131383.111.

Sharma, Deepak, Pragya Priyadarshini, and Sudhanshu Vrati. 2015. “Unraveling the Web of Viroinformatics: Computational Tools and Databases in Virus Research.” Journal of Virology 89 (3): 1489–1501. https://doi.org/10.1128/JVI.02027-14.

Shaw, Kirsty J., Lauren Thain, Peter T. Docker, Charlotte E. Dyer, John Greenman, Gillian M. Greenway, and Stephen J. Haswell. 2009. “The Use of Carrier RNA to Enhance DNA Extraction from Microfluidic-Based Silica Monoliths.” Analytica Chimica Acta, Fundamental and Applied Analytical Science. A Special Issue In Honour of Alan Townshend., 652 (1): 231–33. https://doi.org/10.1016/j.aca.2009.03.038.

Shi, Mang, Xian-Dan Lin, Jun-Hua Tian, Liang-Jun Chen, Xiao Chen, Ci-Xiu Li, Xin-Cheng Qin, *et al*. 2016. “Redefining the Invertebrate RNA Virosphere.” Nature 540 (7634): 539–43. https://doi.org/10.1038/nature20167.

Shkoporov, Andrey N., and Colin Hill. 2019. “Bacteriophages of the Human Gut: The ‘Known Unknown’ of the Microbiome.” Cell Host & Microbe 25 (2): 195–209. https://doi.org/10.1016/j.chom.2019.01.017.

Shkoporov, Andrey N., Ekaterina V. Khokhlova, C. Brian Fitzgerald, Stephen R. Stockdale, Lorraine A. Draper, R. Paul Ross, and Colin Hill. 2018. “ΦCrAss001 Represents the Most Abundant Bacteriophage Family in the Human Gut and Infects Bacteroides Intestinalis.” Nature Communications 9 (1): 4781. https://doi.org/10.1038/s41467-018-07225-7.

Shkoporov, Andrey N., Feargal J. Ryan, Lorraine A. Draper, Amanda Forde, Stephen R. Stockdale, Karen M. Daly, Siobhan A. McDonnell, *et al*. 2018. “Reproducible Protocols for Metagenomic Analysis of Human Faecal Phageomes.” Microbiome 6 (1): 68. https://doi.org/10.1186/s40168-018-0446-z.

Starr, Evan P., Erin E. Nuccio, Jennifer Pett-Ridge, Jillian F. Banfield, and Mary K. Firestone. 2019. “Metatranscriptomic Reconstruction Reveals RNA Viruses with the Potential to Shape Carbon Cycling in Soil.” Proceedings of the National Academy of Sciences 116 (51): 25900–908. https://doi.org/10.1073/pnas.1908291116.

Sutton, Thomas D. S., and Colin Hill. 2019. “Gut Bacteriophage: Current Understanding and Challenges.” Frontiers in Endocrinology 10: 784. https://doi.org/10.3389/fendo.2019.00784.

Tank, Marcus, and Donald A. Bryant. 2015. “Nutrient Requirements and Growth Physiology of the Photoheterotrophic Acidobacterium, Chloracidobacterium Thermophilum.” Frontiers in Microbiology 6 (March). https://doi.org/10.3389/fmicb.2015.00226.

Thomas, Torsten, Jack Gilbert, and Folker Meyer. 2012. “Metagenomics - a Guide from Sampling to Data Analysis.” Microbial Informatics and Experimentation 2 (1): 3. https://doi.org/10.1186/2042-5783-2-3.

Thurber, Rebecca V, Matthew Haynes, Mya Breitbart, Linda Wegley, and Forest Rohwer. 2009. “Laboratory Procedures to Generate Viral Metagenomes.” Nature Protocols 4 (4): 470–83. https://doi.org/10.1038/nprot.2009.10.

Vestergaard, Gisle, Ricardo Aramayo, Tamara Basta, Monika Häring, Xu Peng, Kim Brügger, Lanming Chen, *et al*. 2008. “Structure of the Acidianus Filamentous Virus 3 and Comparative Genomics of Related Archaeal Lipothrixviruses.” Journal of Virology 82 (1): 371–81. https://doi.org/10.1128/JVI.01410-07.

“Viral Genomes.” 2021 Accessed January 28, 2021. https://www.ncbi.nlm.nih.gov/genome/viruses/.

Waller, Alison S., Takuji Yamada, David M. Kristensen, Jens Roat Kultima, Shinichi Sunagawa, Eugene V. Koonin, and Peer Bork. 2014. “Classification and Quantification of Bacteriophage Taxa in Human Gut Metagenomes.” The ISME Journal 8 (7): 1391–1402. https://doi.org/10.1038/ismej.2014.30.

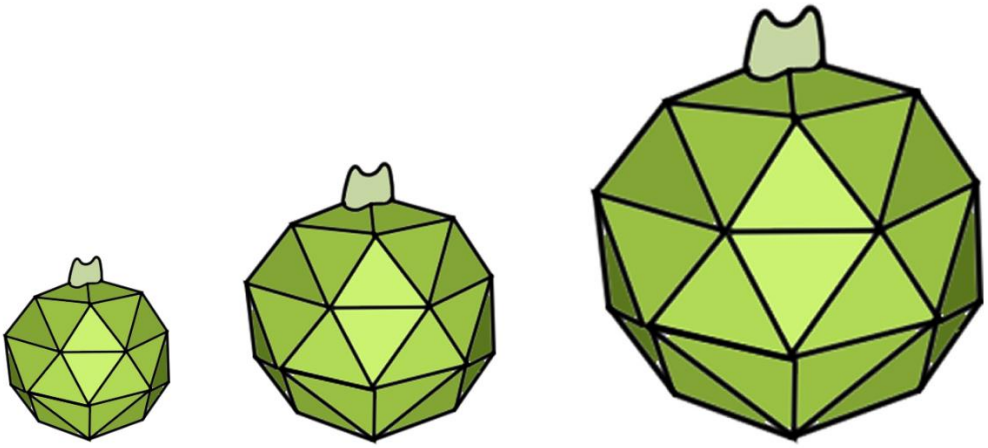
Wommack, K. Eric, Jaysheel Bhavsar, Shawn W. Polson, Jing Chen, Michael Dumas, Sharath Srinivasiah, Megan Furman, Sanchita Jamindar, and Daniel J. Nasko. 2012. “VIROME:

A Standard Operating Procedure for Analysis of Viral Metagenome Sequences." Standards in Genomic Sciences 6 (3): 427–39. https://doi.org/10.4056/sigs.2945050.

Xiao, Chuan, Paul R. Chipman, Anthony J. Battisti, Valorie D. Bowman, Patricia Renesto, Didier Raoult, and Michael G. Rossmann. 2005. "Cryo-Electron Microscopy of the Giant Mimivirus." Journal of Molecular Biology 353 (3): 493–96. https://doi.org/10.1016/j.jmb.2005.08.060.

Yilmaz, Suzan, Martin Allgaier, and Philip Hugenholtz. 2010. "Multiple Displacement Amplification Compromises Quantitative Analysis of Metagenomes." Nature Methods 7 (12): 943–44. https://doi.org/10.1038/nmeth1210-943.

Zhang, Tao, Mya Breitbart, Wah Heng Lee, Jin-Quan Run, Chia Lin Wei, Shirlena Wee Ling Soh, Martin L. Hibberd, Edison T. Liu, Forest Rohwer, and Yijun Ruan. 2005. "RNA Viral Community in Human Feces: Prevalence of Plant Pathogenic Viruses." PLOS Biology 4 (1): e3. https://doi.org/10.1371/journal.pbio.0040003.



Chapter III

Expansion of Known +ssRNA Bacteriophage Genomes: From Tens to over a Thousand

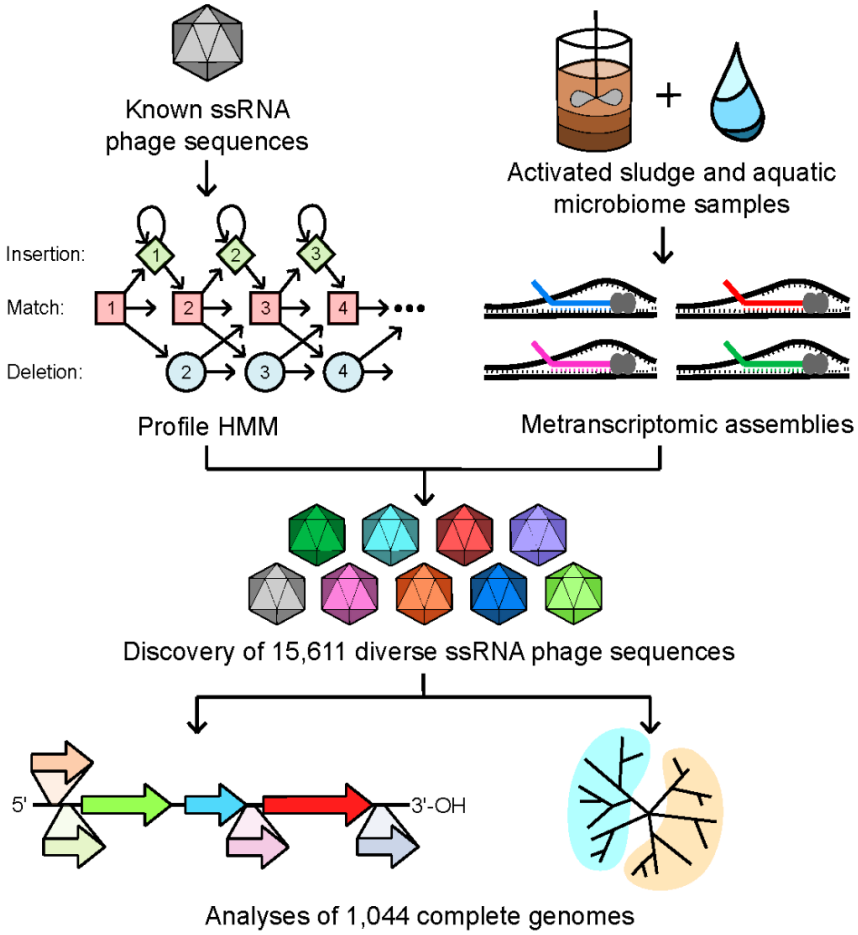
Running title: Metatranscriptome mining unearths RNA phages

A version of this chapter was published as a paper in *Science Advances,* in which I conceived the study, performed the analysis, produced the images, and wrote the manuscript along with my co-author Dr Stephen Stockdale. Edits were made to include the supplementary work with the main manuscript.

Callanan, Julie, Stephen R. Stockdale, Andrey N. Shkoporov, Lorraine A. Draper, R. Paul Ross, and Colin Hill. "Expansion of known ssRNA phage genomes: from tens to over a thousand." Science Advances 6, no. 6 (2020): eaay5981.

https://doi.org/10.1126/sciadv.aay5981

Graphical abstract



3.1 Abstract

The first sequenced genome was that of the 3,569nt positive-sense, single-stranded (+ss) RNA bacteriophage (phage) MS2. Despite the recent accumulation of vast amounts of DNA and RNA sequence data, only 12 representative +ssRNA phage genome sequences are available from the NCBI Genome database (June 2019). The difficulty in detecting RNA phages in metagenomic datasets raises questions as to their abundance, taxonomic structure, and ecological importance. In this study, profile hidden Markov models (HMMs) were iteratively applied to detect conserved +ssRNA phage proteins in 82 publicly available metatranscriptomic datasets generated from activated sludge and aquatic environments. Following this, 15,611 non-redundant +ssRNA phage sequences were identified, including 1,015 near-complete genomes. This expansion in the number of known sequences allowed for a phylogenetic assessment of both novel and known +ssRNA phage genomes. This expansion of these viruses from two environments suggest they have been significantly overlooked within microbiome studies.

3.2 Introduction

Viruses, in particular bacteriophages targeting prokaryotes, are the most diverse biological entities in the biosphere (Cobián Güemes *et al*. 2016; Clokie *et al*. 2011). Currently there are 11,489 genome sequences available in the NCBI Viral RefSeq database (version 94). The vast majority of known phage possess a double-stranded (ds) DNA genome (Manrique *et al*. 2016; Norman *et al*. 2015). Recent metagenomic analysis of 145 marine virome sampling sites identified 195,728 DNA viral populations, highlighting that only a fraction of the Earth's viral diversity has been characterised (Gregory *et al*. 2019). An additional expansion of known phage populations by Roux *et al*. revealed not only dsDNA phages but ssDNA *Inoviridae* are far more diverse than previously considered (Roux *et al*. 2019). The rapid expansion in viral

discovery through metagenomics is enabling a greater understanding of their roles within environments and their evolutionary relationships, which is subsequently causing a revolution in phage taxonomy (Barylski *et al*. 2019).

Despite the identification of +ssRNA phages over 50 years ago (Loeb and Zinder 1961), there are few representative sequences available. The International Committee on Taxonomy of Viruses (ICTV) has currently categorised approximately 5,500 viruses (Walker *et al*. 2019). Yet, their classification only applies to 25 +ssRNA phage sequences (complete or partial) across two genera, *Levivirus* and *Allolevivirus*, as well as an additional 32 sequences unclassified below a family taxonomic rank (Olsthoorn and van Duin 2017). Historically, methods for classifying *Leviviridae* depended on molecular weight, density, sedimentation and serological cross-reactivity (Olsthoorn and van Duin 2011). A subsequent classification method separated the two genera, with the Alloleviviruses containing a fourth unique gene predicted to encode a lysin (Atkins *et al*. 1979). Recently, an analysis of the evolution origin of all currently known RNA viruses by Wolf *et al.* suggested +ssRNA phages may actually be two distinct lineages, which they termed *Leviviridae* and 'Levi-like' viruses (Wolf *et al*. 2018).

The +ssRNA phage MS2 is a non-enveloped virus with a positive-sense monopartite genome of 3,569nt and was the first biological entity to have its entire genome sequenced (Fiers *et al*. 1976). MS2 and its relatives were assigned to the family *Leviviridae* and were generally isolated against Proteobacteria. With additional studies, it can be anticipated that +ssRNA phages will be found which target additional bacterial phyla. Genomes of +ssRNA phages encode a maturation protein (MP) responsible for host recognition, a coat protein (CP) for genome encapsulation, and an RNA-directed RNA polymerase (RdRP) required for viral replication. During the phage replication process, there is a negative-sense template produced for genome replication, although it does not persist and no negative-sense +ssRNA phages have been isolated or characterised to date (Koonin, Senkevich, and Dolja 2006).

An analysis of the evolution of all RNA viruses recently proposed their primordial origin from reverse transcriptases. ICTV have recently established a new viral realm, *Ribovira*, to incorporate all known RNA viruses, as they all encode an RdRP for replication (Gorbalenya *et al*. 2019). The origin of +ssRNA phages followed the acquisition of a CP, potentially allowing them to survive *ex vivo* and prey on the first cellular microbes (Wolf *et al*. 2018). Despite their small genome size (encoding only three or four genes), +ssRNA phages have served as models for understanding some of nature's most widespread fundamental processes, including genome secondary structure to mechanisms of controlling gene expression and genome replication (Gytz *et al*. 2015; Lodish 1968).

Identification of phages was traditionally dependent on culture-based methods (Kannoly, Shao, and Wang 2012). In recent years, there has been a shift to culture-independent metagenomic approaches which aim to capture all microbial genomes within a given environment (Dantas *et al*. 2013). An analysis by Krishnamurthy *et al*. identified 158 +ssRNA phage sequences (complete and partial), significantly expanding the previously recognised diversity of this group (Krishnamurthy *et al*. 2016). A more recent study by Starr *et al*. demonstrated that metatranscriptomics will advance +ssRNA phage discovery, with 1,338 +ssRNA phage RdRP sequences detected in soil (Starr *et al*. 2019). Metatranscriptomics is indeed well suited to capturing +ssRNA phage sequences in complex biological samples, given that their genomes resemble the mRNA transcripts which are targeted by this method.

The actual abundance and diversity of +ssRNA phages has remained unknown despite recent advancements to better study the phage populations of different environments. Databases are dominated by DNA phage genomes and novel +ssRNA phages may not be recognised. Isolation and purification techniques for phages, such as caesium chloride (CsCl) gradient purification and polyethylene glycol (PEG), are biased towards isolating specific phage types (Kleiner, Hooper, and Duerkop 2015). Even accepting that specific

metatranscriptomic approaches will introduce their own biases in the process of removing ribosomal RNA (Alberti *et al*. 2014), it is likely to be more representative of the RNA composition of a specific microbiome, including the RNA viral contingents.

RNA phages have served as key models in understanding some of biology's most intricate pathways such as gene regulation. These phages also offer a potential option in terms of phage therapy, as they have been isolated against many pathogenic bacteria including *Acinetobacter* and *Pseudomonas*. Fundamentally, the expansion in +ssRNA phage genomes reported here demonstrates that their contributions to the diversity of ecological niches and their impacts on their associated hosts may have been underestimated. Given that scientists are just starting to explore Earth's "viral dark matter" through metagenomics, it seems fitting that a portion of this unexplored viral diversity is represented by phages that are not encoded by DNA.

In this study, the identification of 15,611 near-complete and partial +ssRNA phage sequences was reported. Of these, 1,015 were defined as near-complete in that they encode all three MP, CP, and RdRP genes that form the recognised +ssRNA phage core genome. The identification of +ssRNA phage sequences was performed by iteratively developing and applying HMMs based on conserved +ssRNA phage proteins. These HMMs were applied to ever increasing samples from 70 activated sludge and 12 aquatic environments. This expansion in the number of +ssRNA phage genomes enabled the phylogenetic relationships between novel and known sequences to be examined and a preliminary investigation of phage-host interactions to be performed.

3.3 Materials and Methods

3.3.1 Assembly of metatranscriptome samples

The assembly of metatranscriptome samples is portrayed in Figure 1A. Fastq raw reads were downloaded from the NCBI SRA database using accession numbers provided in Supplementary Data, with files separated into forward and reverse reads using the '--split-files' option. Illumina adapter sequences were removed using Cutadapt (version 1.9.1; (Martin 2011)). The overall read quality was improved using Trimmomatic (version 0.32;(Bolger, Lohse, and Usadel 2014)), pruning sequences where the read quality dropped below a Phred score of 30 for a 4bp sliding window. Reads less than 70bp were discarded, with surviving reads assembled using rnaSPAdes (version 3.12.0; (Bushmanova *et al*. 2018)). Only metatranscriptome sample SRR5466337, which generated an error during rnaSPAdes assembly, was assembled differently using Megahit (version 1.1.1-2; (D. Li *et al*. 2015)). This one sample of the total 82 samples, failed to assemble using rnaSPAdes. The reasons were not investigated further. All contig assemblies less than 500bp were discarded. Only the rnaSPAdes 'hard filtered transcript' outputs were examined for the presence of +ssRNA phages.

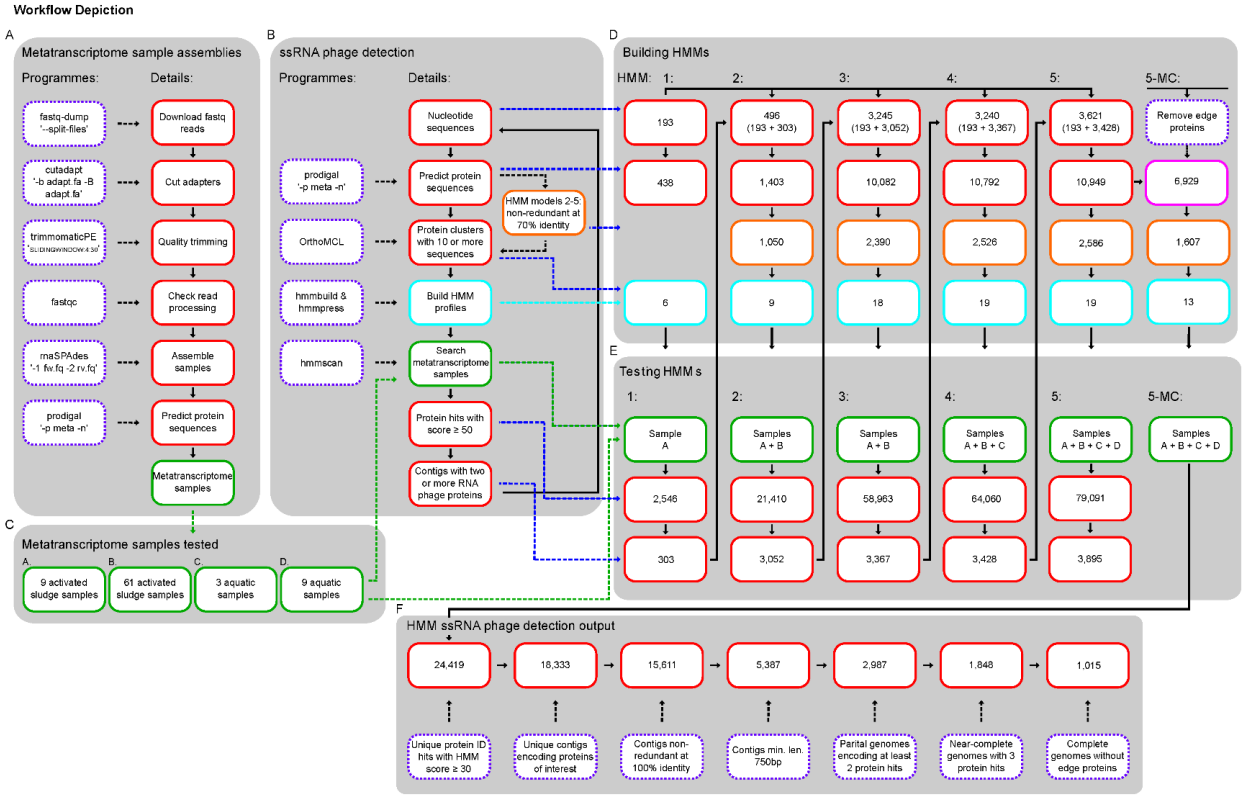


Figure 1. Workflow depiction of the study pipeline, outlining: (A) metatranscriptome sample assemblies, **(B)** the detection of +ssRNA phages, **(C)** samples tested, and the breakdown of the **(D)** building, **(E)** testing, and **(F)** output of the HMM iterations.

3.3.2 Generation of profile hidden Markov models

The pipeline for generating profile hidden Markov models (HMMs) is depicted in Figure 1, with the numerical breakdown of the HMM building and testing stages depicted in Figures 1D & 1E, respectively. In order to generate the first HMM, 'HMM 1', all +ssRNA phage near-complete and partial genome sequences were downloaded from NCBI Taxonomy database (October 2018) and previous published studies (Krishnamurthy *et al*. 2016). The encoded proteins of all identifiable +ssRNA phage sequences (n=193) were predicted using Prodigal with the '-p meta' option enabled for small contigs, and '-n' option specified in order to do a full motif scan per nucleotide sequence (version 2.6.3; (Hyatt *et al*. 2010)). Predicted proteins were clustered using OrthoMCL using a BLASTp all-v-all E-value 1E-05 and default settings (version 2.0; (Fischer *et al*. 2011)). Clusters of +ssRNA phage proteins with 10 or more sequences were aligned using MUSCLE (version 3.8.31; (Edgar 2004)), and used to generate HMMs via hmmbuild (version 3.1b1; (Finn, Clements, and Eddy 2011)). Multiple HMMs were combined into a single HMM search tool through hmmpress (version 3.1b1).

The number of samples tested by each HMM iteration is outlined in Figure 1C. HMMs 2-5 were built in a similar fashion to HMM 1 with the following alterations. Subsequent to the detection of contigs in metatranscriptome samples encoding two or more functionally distinct +ssRNA phage proteins (hmmscan score of 50 or greater), the predicted proteins were combined with those from the initial 193 +ssRNA phage sequences, obtained from NCBI and a previous publication (Krishnamurthy *et al*. 2016). Using a BLAST all-v-all approach, the proteins used to generate HMMs 2-5 were made non-redundant at 70% amino acid identity, removing the shorter of two protein sequences when the overlap exceeded 70%. Prior to the generation of HMM 5-MC, proteins were manually curated to remove sequences encoded at the edge of contigs (termed "edge proteins").

3.3.3 Validating HMM detection of +ssRNA phages

The metatranscriptome sample SRR1027978, which was an activated sludge sample previously shown by Krishnamurthy *et al.* (2016) as containing +ssRNA phage sequences using a tBLASTn approach, was downloaded as a positive control and examined for the presence of +ssRNA phage proteins. Briefly, a random subset of 10 million reads were extracted from the SRA file with the seqtk 'sample' command (version 1.0-r31; (Li 2019)) using a user defined seed ('-s13'). Adaptor and read trimming were performed as described above, with surviving reads assembled using Megahit. Proteins were predicted in all contigs greater than 500bp, using options '-p meta -n', before scanning with HMM 1.

After manual curation of +ssRNA phage hits, it was decided to adopt a conservative approach for the remainder of the study (results not shown). Only hmmscan hits with a score of 50 or greater were considered during the generating of HMM iterations, with hmmscan scores of 30 further investigated during metatranscriptome sample analyses. Future studies may benefit from less stringent +ssRNA phage discovery cut-offs, by lowering the hmmscan score requirements and/or using rnaSPAdes 'soft filtered transcripts'. However, results would need to be treated cautiously to avoid false positives.

A comparison between a BLAST and HMM-based approach to identify +ssRNA phages was performed using the complete +ssRNA phage proteins which built the final HMM model 5-MC. The BLAST and HMM approaches were applied to the 2,308 unique viral sequences described by Shi and colleagues (Shi *et al*. 2016). This database contains 67 +ssRNA levi-like viruses. Using a relaxed BLASTp E-value of 1E-05, 78 viral sequences were considered +ssRNA phages (11 false positives). However, with a more stringent BLASTp E-value of 1E-15, only the expected 67 sequences were returned. Utilising a HMM scan with a score of 30 identified the 67 levi-like viruses without any false positives identified.

When the strict BLASTp search approach (E-value 1E-15) was applied to the assembled contigs from the metatranscriptome sample SRR1027978, 12 +ssRNA phages were identified. The HMM-based approach identified 13 +ssRNA phages. Reducing the BLASTp stringency to 1E-05 did identify 13 putative +ssRNA phages. However, due the false positives noted while using a less-strict BLASTp approach against a curated database, only HMM searches were used through-out this study.

3.3.4 Detecting +ssRNA within metatranscriptome samples

After confirming that HMM 1 could detect +ssRNA phage proteins in a positive control sample, HMM 1 was implemented against 9 previously untested metatranscriptome samples of activated sludge. This environment was chosen as Krishnamurthy *et al.* (2016) demonstrated sewage as a rich-source for +ssRNA phages (Krishnamurthy *et al*. 2016). These 9 SRA files analysed represent 3 activated sludge samples from each of the Austria, Illinois, and Japan study locations (available in Supplementary Data online). The total collection of activated sludge and aquatic samples cumulatively analysed during this study are outlined in Figure 1C. The remaining samples tested represent; 13 activated sludge samples from Austria, 39 activated sludge samples from Illinois, 9 activated sludge samples from Japan, 4 freshwater aquatic samples from Lake Mendota (Wisconsin), 4 aquatic samples from the Mississippi river (Louisiana), and 4 freshwater aquatic samples from Singapore.

3.3.5 Analysis of +ssRNA phage proteins

Analyses were conducted using the R programming language (version 3.5.3) implemented through RStudio (RStudio Team 2020). Images were generated using the 'ggplot2' package (Wickham *et al*. 2021), with additional colours obtained from the 'RColorBrewer' (Neuwirth 2014), the 'wesanderson' (Ram *et al*. 2018), and the 'YaRrr' package (Phillips 2017). The bipartite network of +ssRNA phage proteins, for sequences containing two or three core proteins, was generated using the 'igraph' package (Gabor Csardi and Tamas Nepusz 2006).

The distance between core proteins (squares) is automatically calculated based on the number of +ssRNA sequences (circles) that share similar protein profiles. The +ssRNA phage partial genomes are coloured based on the associated CP. The Sankey plot demonstrating the connection patterns of +ssRNA phage-encoded proteins was illustrated using the R package 'networkD3'(Allaire *et al*. 2017).

Phylogeny of +ssRNA phage proteins was performed as follows. Proteins fulfilling the same functions amongst +ssRNA phages were assigned the name of their originating contig, and subsequently aligned using MUSCLE. The alignments of the three core proteins were concatenated using MEGA (version 10.0.5; (Kumar *et al*. 2018)). After the three proteins were concatenated, the MUSCLE alignment was performed with default settings – no alignment trimming, all positions were retained, and the substitution model was applied to all proteins together. These alignments were imported into R using the 'seqinr' package (Charif *et al*. 2017; Charif and Lobry 2007) with 'ape' package dependencies (Paradis and Schliep 2019), before conversion to a phyDat format using the 'phangorn' package (Schliep 2011). The best evolutionary model was estimated using the phangorn 'modelTest' function, with the model yielding the lowest Akaike information criterion (AIC) score selected for maximum likelihood tree construction. Blosum62 was determined as the best amino acid substitution model. Phylogenetic trees were bootstrapped 100 times and saved using the 'treeio' package (Yu 2019), before visualization using 'ggtree' (Yu *et al*. 2017).

The R scripts and input data used to generate this study's images and infer results were provided in the Supplementary Data, which is available online (http://advances.sciencemag.org/content/suppl/2020/02/03/6.6.eaay5981.DC1).

3.4 Results and Discussion

3.4.1 Existing +ssRNA phage sequences

Due to nomenclature/accession number disparities associated with sequences deposited to different databases/organisations, all the identifiable +ssRNA phage sequences were graphically depicted for simplicity (Figure 2). The latest ICTV report (Olsthoorn and van Duin 2017) was used to initially identify 57 *Leviviridae*, although not all had identifiable genomes within public sequence repositories (Figure 2A). The single largest source of +ssRNA phage sequences was from the recent metagenomic study of Krishnamurthy *et al*. (2016), where they identified 158 +ssRNA phage sequences across invertebrate, vertebrate, sewage, aquatic and soil samples (Figure 2B; (Krishnamurthy *et al*. 2016)). By investigating the NCBI Taxonomy database, an additional 35 unique +ssRNA phage genome sequences were identified (including those from the ICTV report).

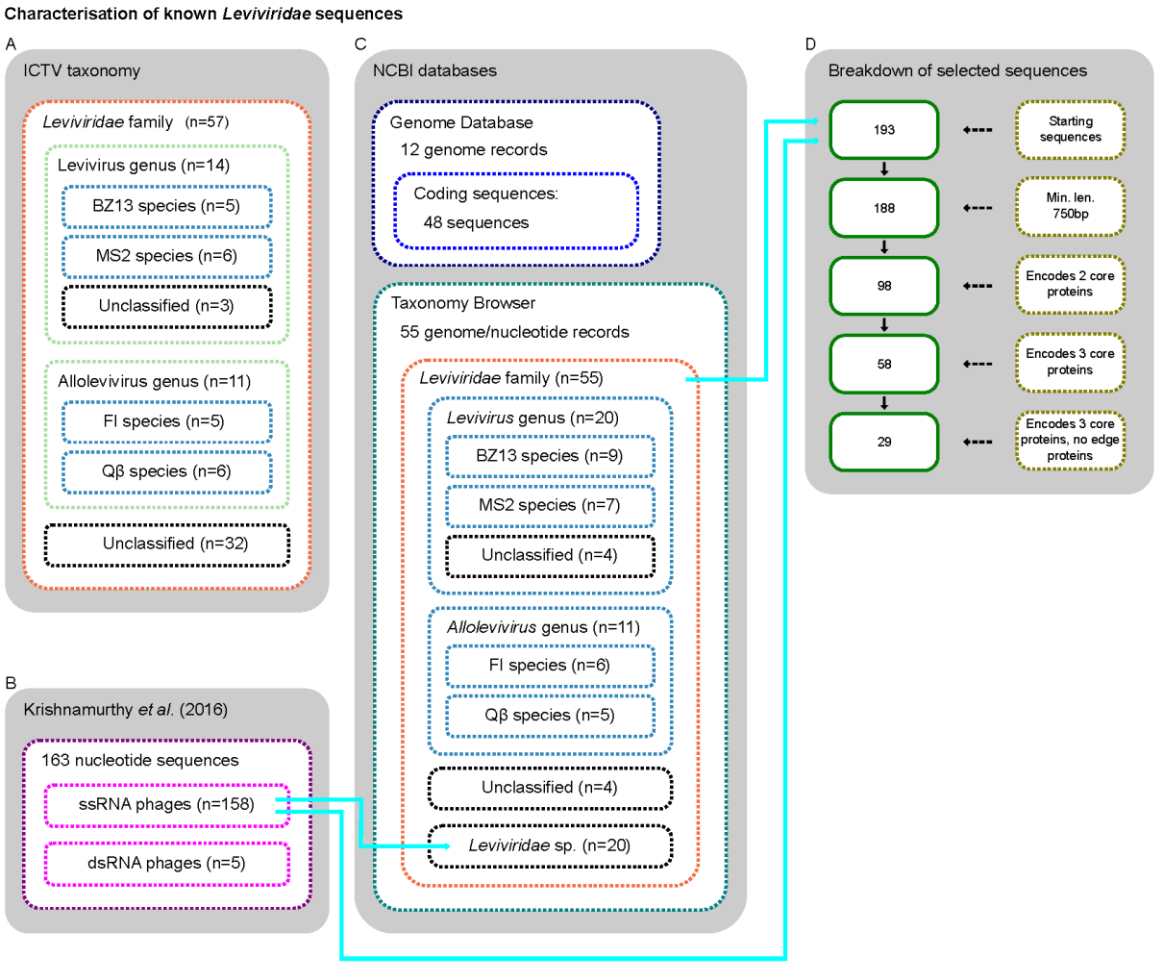


Figure 2. Workflow depiction of known +ssRNA phage sequences, outlining: **(A)** ICTV taxonomy, **(B)** Krishnamurthy *et al*. (2016), **(C)** NCBI Genome and Taxonomy database available sequences, and **(D)** the breakdown of the identifiable +ssRNA phage sequences used in this study.

While a total of 193 previously described unique +ssRNA phage sequences were identified at this study's onset, how many of these represented complete or near-complete genomes was characterised (Figure 2D). Using HMM 5-MC followed by manual curation, only 29 sequences fulfilled the requirement of encoding all three of the +ssRNA phage core proteins (MP, CP, and RdRP) without their premature termination by the edge of a phage contig. Determining phage "edge proteins" was performed by analysing the encoded start and stop

codons of the MP and RdRP genes. Only proteins beginning with canonical or cognate start codons (AUG, GUG or UUG) or stop codons (UAG, UAA, or UGA) were considered full length.

3.4.2 Expansion of known +ssRNA phage sequences

From publicly available databases and relevant studies, 193 identifiable unique, partial +ssRNA phage genome sequences were collected (Figure 2). An additional 67 Levi-like sequences, described by Shi *et al.* (2016), were used to validate the identification of +ssRNA phages from an RNA viral database (see Material and Methods). The encoded proteins of the 193 +ssRNA phage genomes were predicted and used a graph-based clustering method to build a database of HMM sequence profiles representative of their protein sequences (Figure 1). Four subsequent HMM iterations were built, each using the previous HMM output, and were applied to a final total of 82 publicly available environmental metatranscriptome samples generated from globally sourced activated sludge and aquatic samples. A final manually curated HMM, designated 5-MC, was developed by removing all partial protein sequences.

A caveat with searches tools that provide users with an expected-value output is that E-values are dependent on the length of the query sequence, and the size of the searched database. Therefore, for consistent implementation across studies of different sizes, where possible hmmscan scores and not E-values were recorded. During the iterative development of this +ssRNA phage detecting HMMs, only hmmscan scores greater than 50 were considered for continuation into the subsequent model (Figure 2). However, during the final analysis and detection of +ssRNA phages across metatranscriptome samples, a less stringent hmmscan score of 30 was adopted.

The implementation of a strict, uncompromising set of parameters was decided early in this study as RdRP proteins are conserved across all RNA viruses. However, as RNA viruses have high nucleotide mutation rates (Drake *et al*. 1998; Holland *et al*. 1982), often little or no

sequence similarity is observed between evenly closely related RNA viruses. Therefore, future studies could benefit from lowering the hmmscan score thresholds to find more diverse sequences, albeit at the risk of finding false positives.

In total, 15,611 +ssRNA phage genomes or partial sequences were identified (Figure 3B). This represents an approximately 60-fold increase in the number of partial genome sequences. Of the 15,611 identified sequences, there were 5,387 +ssRNA phage sequences which had a minimum length of 750bp and included at least one core gene (MP, CP or RdRP), 2,987 included two core genes and 1,848 had sequences from all three core genes. Of these, 1,015 are predicted to encode full length core genes. Only 29 of the currently publicly identifiable 193 +ssRNA phage sequences meet this same criterion (Figure 2D).

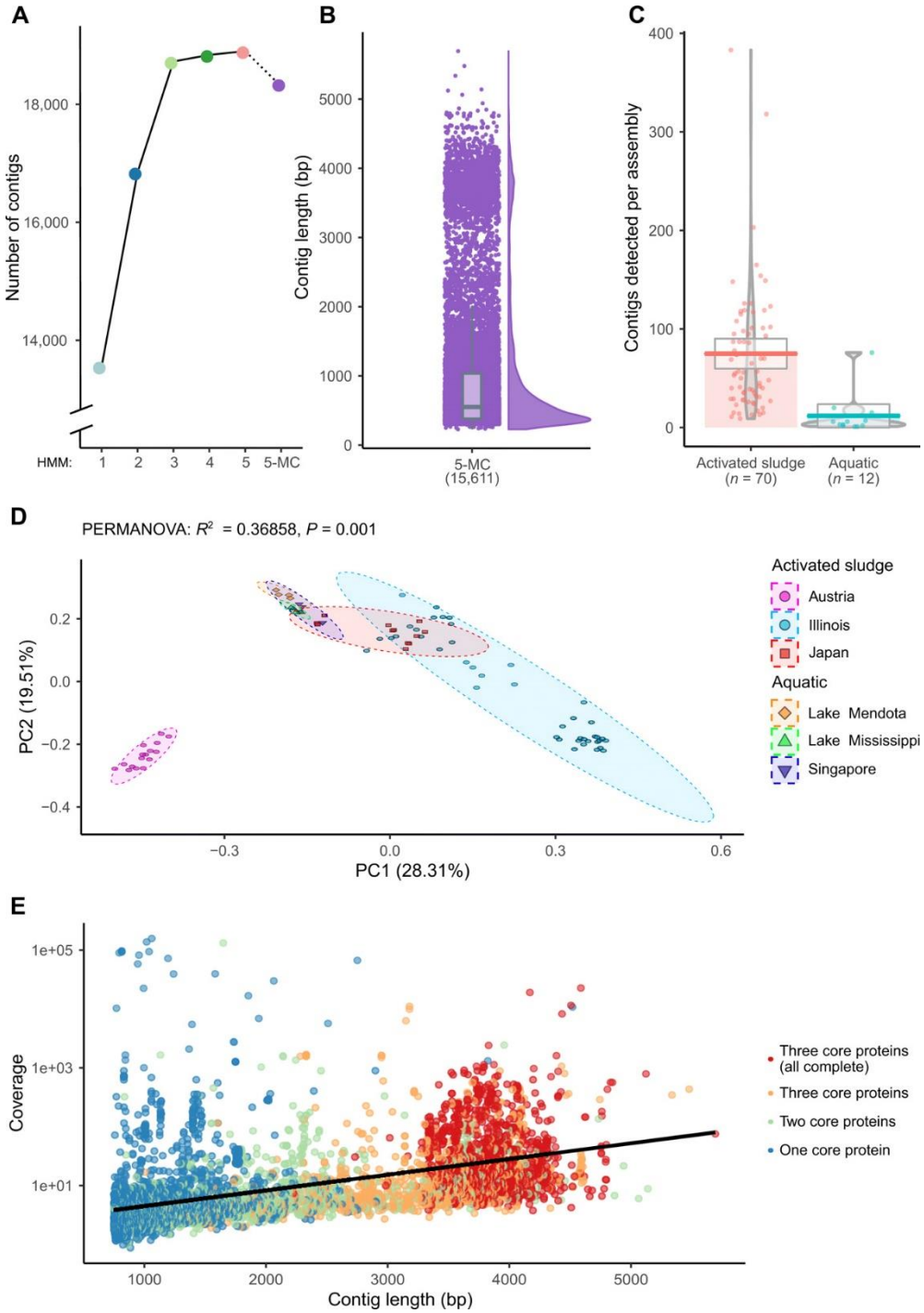


Figure 3. **Identification of +ssRNA phages in metatranscriptome samples**. **(A)** The total number of redundant contigs detected per HMM search. **(B)** The manually curated HMM 5-MC detected 15,611 non-redundant +ssRNA phage sequences. Boxplot displays the median value within the 25th and 75th quartiles, with whiskers representing +/- 1.5 the interquartile range. **(C)** The number of contigs (near-complete or partial) detected per assembly in activated sludge and aquatic samples. Boxplot horizontal lines indicate the mean, while the grey boxes

represent 95% highest density intervals. **(D)** Two-dimensional ordination of +ssRNA compositional abundance across different geographical locations using the Bray-Curtis Dissimilarity index. The colours and shapes of individual samples differentiate study location and environment, respectively. **(E)** Linear model of metatranscriptome sequencing coverage and contig length. Contigs included are of minimum length 750bp, and the number of core proteins encoded are indicated.

The previously isolated +ssRNA phage AP205 (NCBI accession NC_002700.2), which has previously been characterised (Klovins *et al*. 2002), did not make it into the final curated dataset of 29 full length sequences. A detailed examination of phage AP205 highlighted that its CP is so diverse from other identified CP sequences it did not achieve an hmmscan score of 30 using HMM 5-MC. Therefore, as it did not fulfil the criteria for containing the three recognisable core proteins of +ssRNA phages, the pipeline excluded it from the 1,044 sequences analysed in detail during this study. However, the CP of AP205 was investigated further and found it is similar to a potential CP encoded by another previously characterised +ssRNA phage, ESE002 (NCBI accession KT462711.1;(Krishnamurthy *et al*. 2016); with a BLASTp E-value 1.00e-5).

It was found that there were six additional putative CP sequences similar to AP205's encoded within the genomes of the 15,611 +ssRNA phage sequences detected in this study. Five of the six hits demonstrate only weak sequence similarity to the CP of AP205 (BLASTp E-value ranges 9.00e-06 ≤ 1.00e-04), with a single closely related protein sequence (BLASTp E-value 1.00e-57). Due to this HMM development pipeline requiring a minimum of 10 similar protein sequences in order to generate a HMM, an additional ninth CP cluster was not generated that is specific for the third core protein associated with AP205, ESE002 and the aforementioned 6 additional +ssRNA phage contigs. This demonstrates that there is clearly still

undiscovered +ssRNA phage diversity that additional studies, potentially from alternative environments, will no doubt capture.

Significantly more +ssRNA phage sequences were detected in activated sludge than in aquatic samples (Krustal-Wallis, p-value = $1.847e^{-06}$; Figure 3C). It is possible that activated sludge provides an environment in which Proteobacteria, the only known hosts for +ssRNA phages, can grow and support phage enrichment. The higher levels of detection could also be due to a variety of technical factors such as increased sequencing depth, microbiome complexity, and metatranscriptome sampling protocols. Indeed, the ability to detect longer +ssRNA phage sequences correlates with metatranscriptome sequencing depth (Figure 3E).

3.4.3 Examination of genome-associated proteins and architecture

The +ssRNA phage encoded proteins were predicted using Prodigal, which was designed for the annotation of bacterial genomes. It is a well reported feature that the compact genomes of +ssRNA phages use diverse and atypical mechanisms for the control and production of additional functional proteins. These mechanisms include the formation of RNA secondary structures (Beekwilder, Nieuwenhuizen, and van Duin 1995; Duin 1988), translational frame-shifting (Rumnieks and Tars 2012), and encoding protein sequences within the boundaries of another larger gene sequence (Kazaks *et al*. 2011). Therefore, new tools are needed to fully predict coding sequences within not just +ssRNA phages, but across all phages.

The 15,611 +ssRNA phage sequences encoded 24,419 predicted proteins that could be grouped into three MP, eight CP, and two RdRP clusters (Figure 2, Figure 4A). It is evident that the RdRP is the most conserved protein, forming only two clusters, whereas the CP is the most diverse of the +ssRNA phage-associated core protein, splitting into eight clusters. Next all 2,987 +ssRNA sequences encoding at least two core proteins were examined, which revealed two highly distinct groups (Figure 4B). Only five of the almost three thousand assembled sequences bridge the two groups, and these were investigated further. In brief, the

five outliers only encode partial rather than complete proteins and their relatedness to a specific protein cluster may be driven by local rather than global sequence similarity.

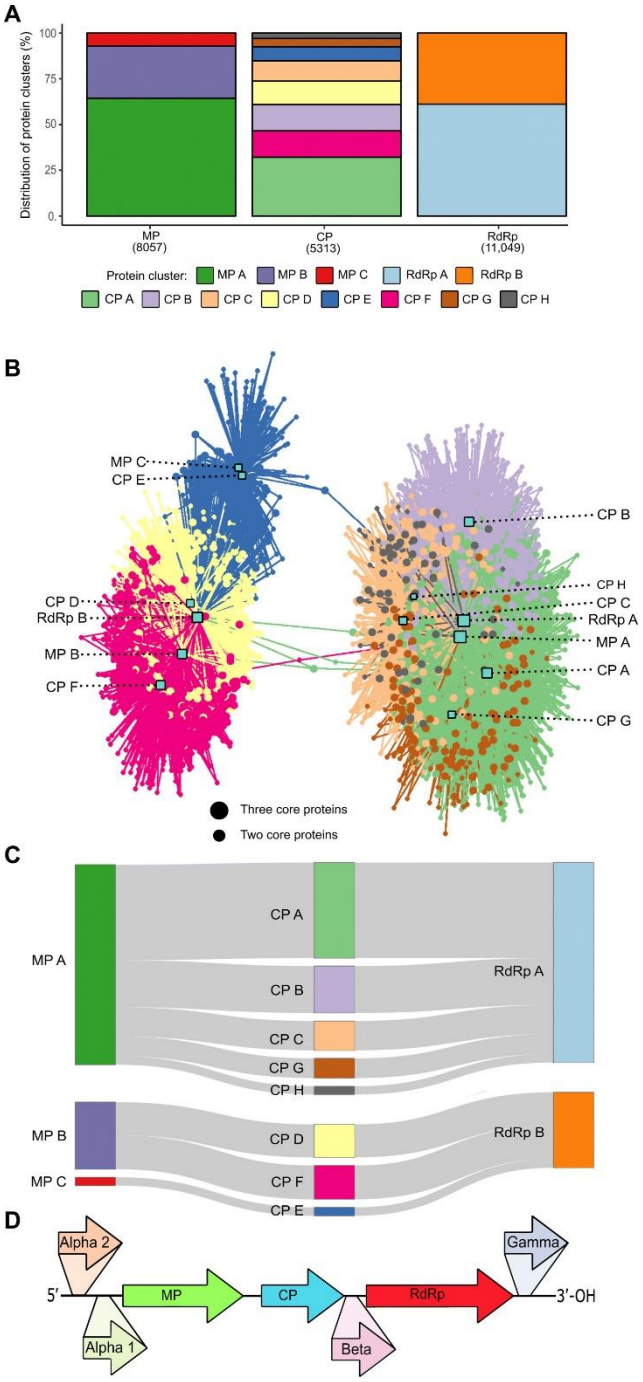


Figure 4. Examination of +ssRNA phage proteins. (A) Distribution of protein hits (in brackets) across MP, CP and RdRP clusters, identified using HMM 5-MC. **(B)** Bipartite connection network of contigs (circles) with proteins (squares). Colours are based on the associated CP, from panel A. **(C)** Protein cluster co-occurring profiles of +ssRNA phages

possessing all three full-length core proteins, and **(D)** the frequently observed positions of hypothetical proteins (not drawn to scale due to the various lengths associated with these genes).

All 1,015 near-complete +ssRNA phage genomes were analysed and observed strictly conserved protein associations (Figure 4C). In contrast to other viruses, there are no obvious instances of homologous recombination and mosaicism amongst the identified +ssRNA phages. Both mosaicism and horizontal gene transfer are well noted for dsDNA phages, with single genes and whole modules exchanged (Hatfull and Hendrix 2011; Rokyta *et al*. 2006). Recombination frequencies of RNA viruses are reported to vary dramatically during co-infection, influenced by various factors such as sequence identity, kinetics of transcription, and RNA genome secondary structure (Simon-Loriere and Holmes 2011). Only eight protein connection profiles between the three MP, eight CP, and two RdRP protein clusters of +ssRNA phages were recorded. If their genomes underwent extensive recombination events, it would be expected that the number of core-protein connection profiles would be closer to the theoretical maximum of 48 (3 MP x 8 CP x 2 RdRP). However, as this +ssRNA phage discovery pipeline is restricted to finding viruses encoding core proteins similar to those previously identified, future studies with less stringent search criteria may uncover unexplored biodiversity.

With such a tremendous expansion in the quantity of identifiable complete +ssRNA phages, an examination of their genome structure was conducted. Firstly, the specific order of MP, CP and RdRP core proteins were investigated. Notably, on no occasion was the recognisable CP identified as being situated either before the MP or after the RdRP encoding genes. In all 1,015 instances, a CP was situated between the MP and RdRP genes. Of the non-core proteins predicted within +ssRNA phage genomes, three specific locations were

frequently noted (Figure 4D). Hypothetical proteins could exist before the MP, after the CP, or following the RdRP. The locations were termed the Alpha position (preceding the MP and closest to the genomes' 5' termini), the Beta position (ensuing the CP), and the Gamma position (following the RdRP at the genomes' 3' termini). Open reading frames (ORFs) located at the Alpha and Beta positions have previously been shown to encode a lysin protein in several isolated +ssRNA phages, such as AP205 (Alpha), and MS2, PP7 and PRR1 (Beta). On 20 instances, there were two hypothetical proteins situated before the MP (termed Alpha 1 and Alpha 2, the former closest to the genomes' termini). However, mapping of the occurrence of hypothetical ORFs on the phylogram of +ssRNA phages showed no specific clustering of the Alpha 2 hypothetical protein, but several clades of related sequences encoding a hypothetical in the Alpha 1 position (Figure 5). The conservation of an ORF at this Alpha 1 position suggests these hypothetical genes indeed have a biological function.

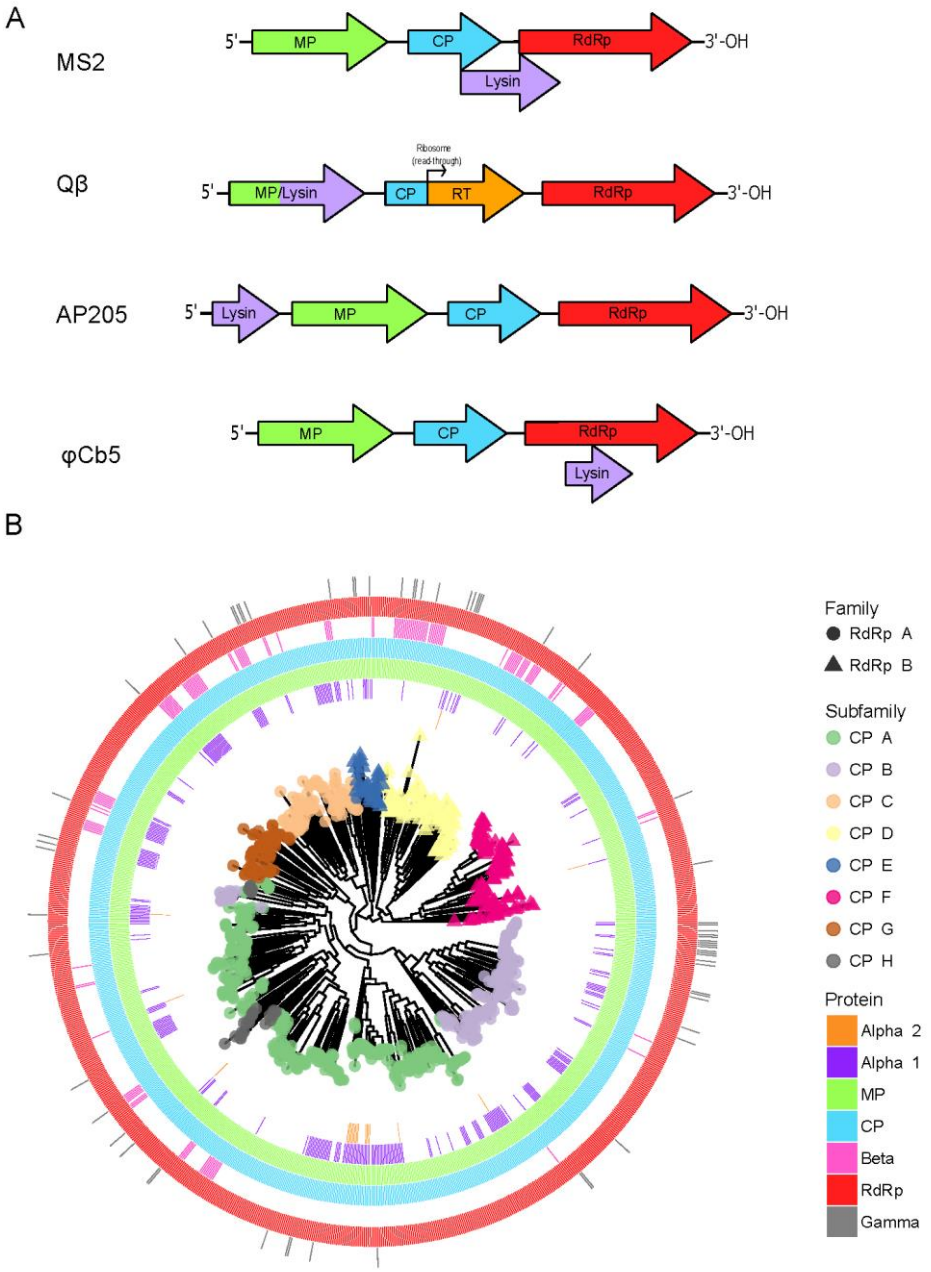


Figure 5. Genome architecture of +ssRNA phages. (A) Genome architecture described for previously known +ssRNA phages. Notably, the position and method of translation of the lysin protein is the most variable between sequences. Due to the overlapping, compact nature of MS2 and φCb5 lysin, current computational approaches did not detect these coding sequences. **(B)** The position of predicted hypothetical proteins mapped onto the suggested

phylogeny of +ssRNA phages. Specific clades are observed which share hypothetical proteins in the same genomic architectural position.

Genome architecture examples of previously described +ssRNA phages highlight the diverse methods to produce their associated lysin proteins (Figure 5A). For phage MS2, its encoded lysin overlaps with the 3'-end of the CP and 5'-edge of the RdRP. During manual curation of MS2-encoded non-core proteins, no lysin was predicted using Prodigal. Therefore, atypical approaches to identify coding sequences will be required to automate the detection of non-core proteins within +ssRNA phages. Nonetheless, clades of related +ssRNA phages encoding hypothetical proteins in the same Alpha and Beta positions were noted (Figure 5B).

Following further investigation, the ORFs predicted to occur after the RdRP were found to have weak sequence similarity to the RdRP clusters (hmmscan score < 30) and may have arisen due to insertion of a premature stop codon during sequencing assembly. However, RNA phages have previously been shown to bypass stop codons as a mechanism to regulate the translational frequency of CP (Weiner and Weber 1973). It was investigated if related +ssRNA phages all encode a hypothetical protein downstream of the RdRP but observed no specific clades of +ssRNA phage with hypotheticals in the Gamma position (Figure 5B), suggesting these ORFs are not conserved and likely a computational artefact. Nevertheless, only biochemical investigations will completely determine if these hypotheticals are indeed functional proteins, such as the lysin, or an alternative viral replication control mechanism.

A total of 506 hypothetical proteins were predicted across the 1,015 full-length +ssRNA phage genomes. These proteins were clustered using CD-HIT at a 70% sequence identity threshold using a word length of 5 (version 4.6, (Li, Jaroszewski, and Godzik 2001)). A total of 29 clusters, representing 262 protein sequences, were generated using CD-HIT that contained 5 or more sequences. All 262 sequences were queried locally against the prokaryotic

virus orthologous groups (pVOGs) profile hidden Markov model database (Grazziotin, Koonin, and Kristensen 2017). All sequences associated with the 29 CD-HIT clusters were also aligned using MUSCLE and the alignment queried against a PDB database (version 23 Feb) using the online MPI HHpred Bioinformatics Toolkit (Zimmermann *et al*. 2018). Only a single CD-HIT cluster, containing 5 protein sequences all from previously identified phages (Qbeta and closely related phages), were found as similar to previously characterised lysins of +ssRNA phages. Both the local pVOG and online HHpred queries identified the lysin (available in the Supplementary Data).

3.4.4 Phylogenetic assessment of near-complete +ssRNA phage genomes

Comparisons of RNA viruses infecting all kingdoms of life have previously been undertaken using the RdRP protein (Wolf *et al*. 2018). To validate previous taxonomic cut-offs against the aforementioned 29 known +ssRNA phages with complete genomes, pairwise amino acid identity (PAAI) comparisons of the RdRP sequences were performed (Figure 6). It was found that the current cut-offs for current +ssRNA genera and species equated to 50% and 80% pairwise PAAIs, respectively. Applying these cut-offs in a bottom-up approach to classifying the 1,044 +ssRNA phages (the 1,015 of this study and the 29 previously identified), 331 species and 247 genera were predicted (Figure 7). As these species and genera taxa were defined independently, and not in a hierarchical fashion, it was subsequently verified that no members of a single species were detected across two or more genera.

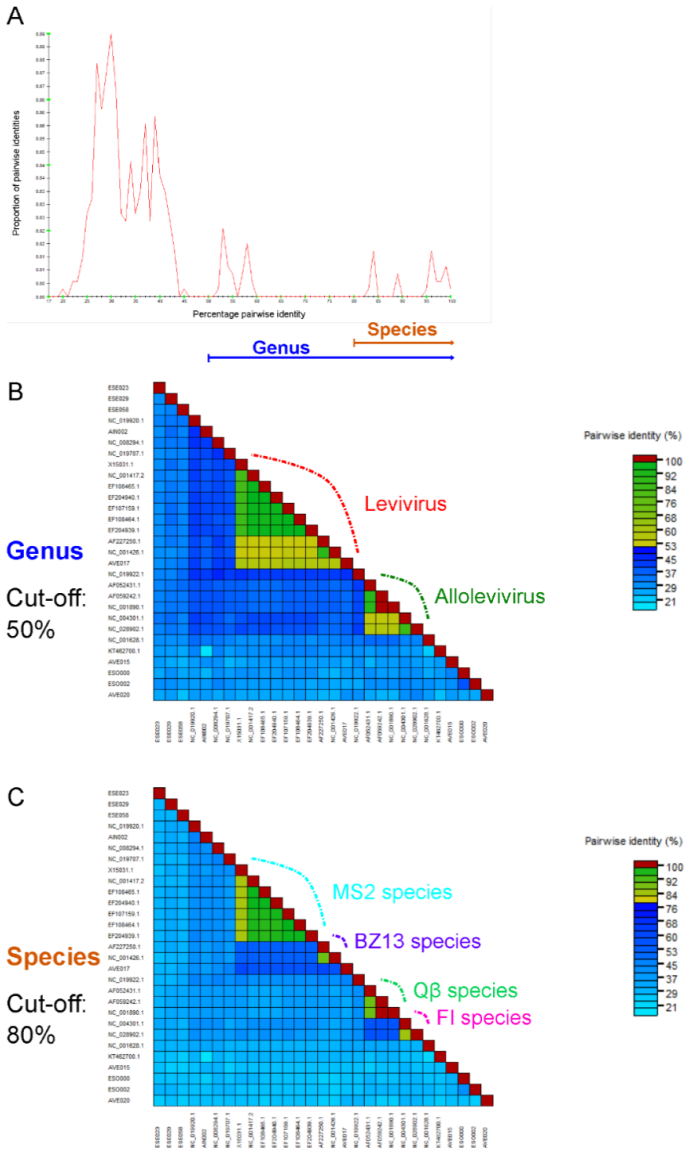


Figure 6. Taxonomic cut-off values for +ssRNA phage genera and species. **(A)** Pairwise RdRP amino acid identity (PAAI) comparisons. Four currently recognised species had PAAI ≥80%, while the two genera had PAAI ≥50%. Visualisation of PAAI at **(B)** genera level, and **(C)** species level with taxonomic groups depicted.

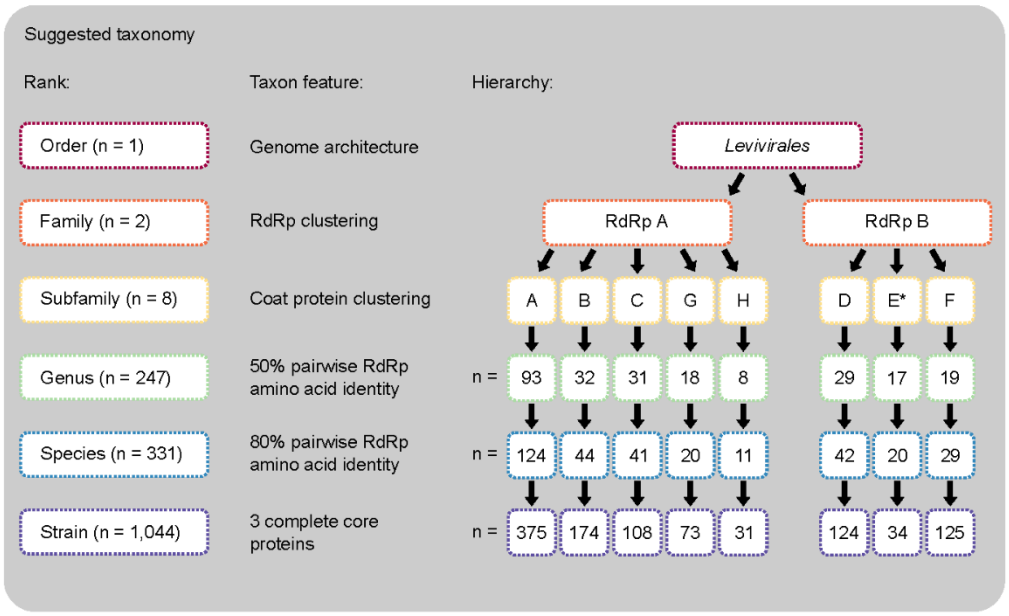


Figure 7. Potential taxonomic restructuring for +ssRNA phages. An outline of defining features for taxonomic ranks and their numerical breakdown is detailed for all 1,044 near-complete +ssRNA phage genomes. Asterisk denotes the AVE006 outlier.

For higher taxonomic divisions, where little or no nucleotide or amino acid sequence similarities are observed, the graph-based clustering of the most variable core protein, the CP, was adopted as a feature to distinguish potential subfamilies. Once more, sequences were assessed to ensure no genera or species were found across multiple subfamilies. The most conserved core protein, the RdRP, was subsequently used to distinguish the most distant relationships between +ssRNA phages at a potential family taxonomic rank.

A single phage, AVE006, which was identified in a previous study (Krishnamurthy *et al*. 2016), did not adhere to the taxonomic defining features outlined in this study (Figure 7, indicated by asterisk). The CP of AVE006 is most similar to cluster E (hmmscan score 36.4), while its replicase is most similar to RdRP cluster B (hmmscan score 229.6). All other phages with CP cluster E group by RdRP cluster A. However, AVE006's RdRP is also very similar to RdRP cluster A (hmmscan score 205.2). In agreement with this observation is the position of

AVE006 in the phylogenetic analysis (Figure 8, highlighted with green arrowhead), where AVE006 is situated on the border between suggested +ssRNA phage families.

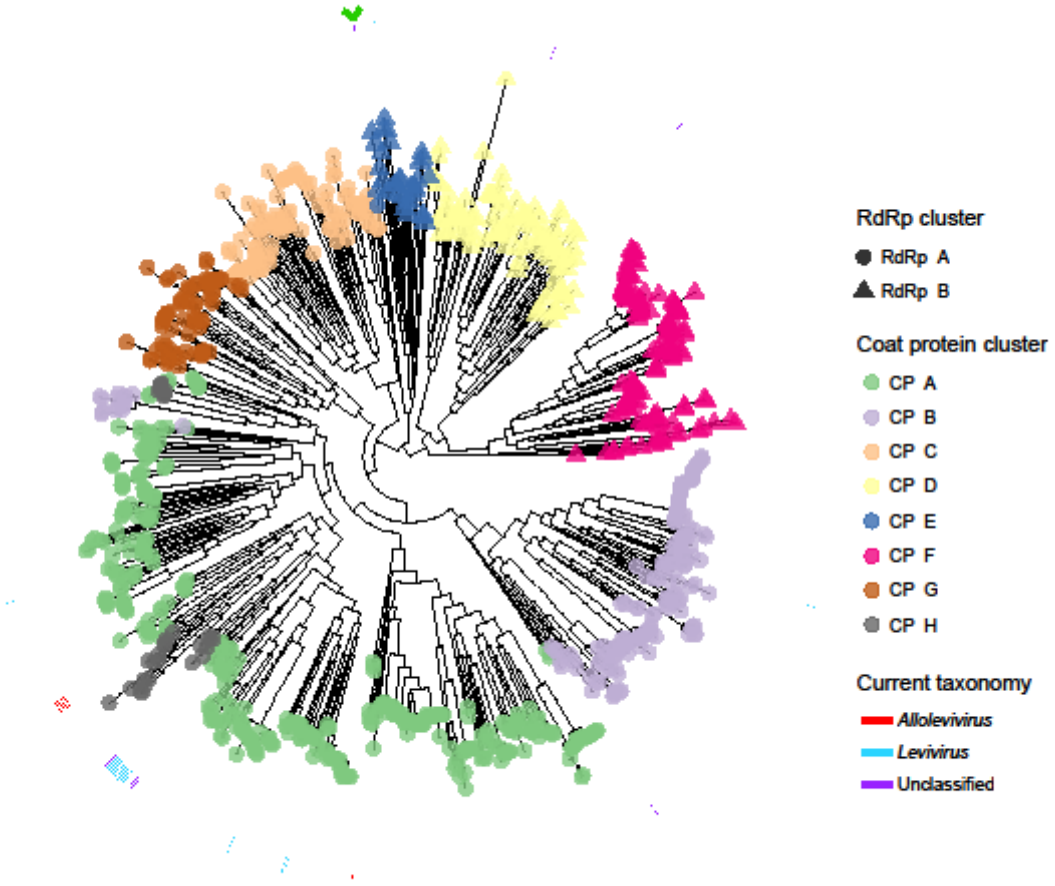


Figure 8. Phylogenetic assessment of +ssRNA phages. Phylogeny of +ssRNA phages using their core protein sequences (MP, CP and RdRP). The 29 previously characterised and 1,015 newly identified phages were included. Branch tip shapes highlight specific RdRP protein clusters, while colour indicates coat protein clustering. The encircling annotation-ring depicts current ICTV taxonomy. A green arrowhead represents AVE006, which encodes a unique RdRP and CP association. Bootstrap support values shown are for 100 iterations.

The phylogenetic divergence of +ssRNA phages by their core proteins supports the hypothesis of Wolf *et al.* (2018) that the current *Leviviridae* family is in fact two distinct lineages (Wolf *et al*. 2018). However, this analysis further classifies +ssRNA phages into eight subfamilies (currently denoted A-H) based on CP clustering. While this suggested classification system can be applied to previously identified +ssRNA phages, it does not support the current *Levivirus* and *Allolevivirus* taxonomic division (Figures 7 & 8).

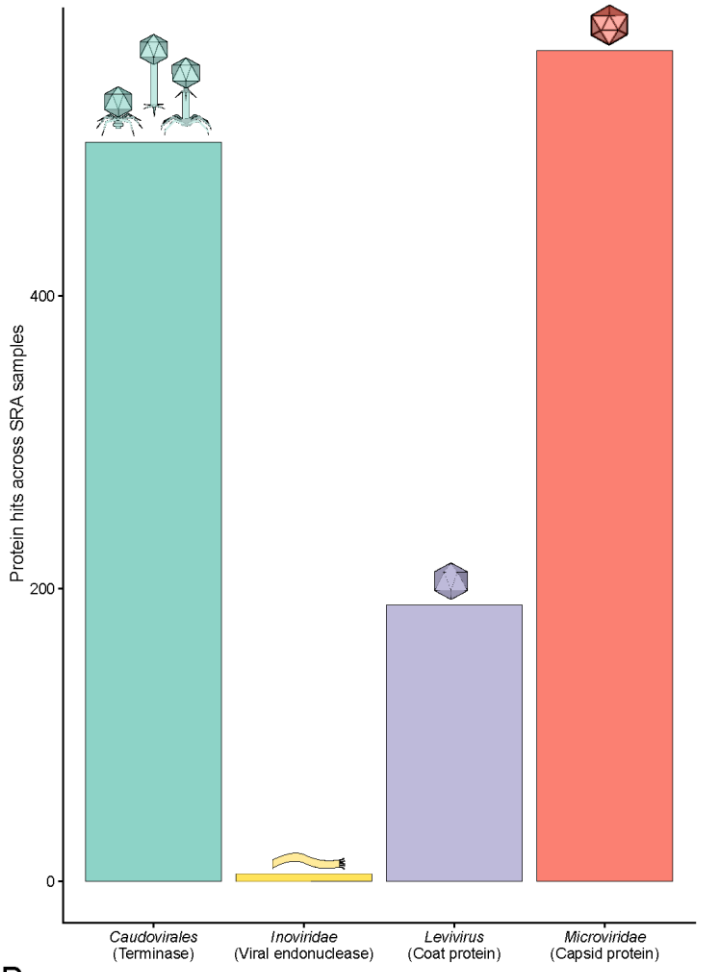
Correlation analysis between the newly proposed taxa and the source locations identified a possible link. The +ssRNA subfamilies were statistical different by geographical location (Kruskal-Wallis test; p-value < 0.001). This may signify those specific ecological niches are occupied by specific phage taxa. For example, CP A was strongly associated with +ssRNA phages identified from the Illinois study site (254 of 1,015; 25.0%) whereas it was infrequently observed amongst Singapore-associated phages (0.5%). A specific global distribution of dsDNA phages was recently detailed for crAssphage (Edwards *et al*. 2019). However, due to the inherent differences introduced through different study protocols and sequencing methodologies, a single study investigating multiple geographical locations is necessary to confirm the potential global localisation of specific +ssRNA phage taxa.

3.4.5 Examination of phage-host interactions

In recent years, there has been a greater effort to understand environmental microbes and their impact on various biogeochemical and nutrient cycles. This includes efforts to better understand the role phages play in shaping microbial community structures and metabolic pathways. For instance, while phages are capable of infecting and killing their microbial hosts, they have also been shown within aquatic environments to augment the photosynthetic capacity of cyanobacteria (Millard 2009). As +ssRNA phages have been overlooked within the majority of microbiomes until now, their ecological importance remains to be fully elucidated.

Using this newly expanded repertoire of +ssRNA phage sequences, the Bacterial RefSeq database was queried (release 89) and, not surprisingly, found no evidence for +ssRNA phage lysogens within bacterial genomes. Therefore, alternative approaches may be required to assign phage-host pairs. To characterize the association of +ssRNA phages with alternative community members, evidence of phages co-existing and potentially co-infecting bacteria was searched for. In order to perform like-for-like comparisons of +ssRNA phages with *Caudovirales*, *Inoviridae* and *Microviridae*, profile-HMMs were built using the currently available Pfam (version 32.0) protein families; PF01819 (*Levivirus* coat protein), PF04466 (caudoviral terminase), PF11726 (inoviral viral endonuclease), and PF02305 (microviral capsid protein), respectively. As hits against +ssRNA phages could potentially occur against either a native *ex vivo* virion or an actively transcribed genome, it is not accurate to perform direct comparisons against phage genomes composed of DNA. Nonetheless, there is clear evidence for transcription of caudoviral terminases and microviral capsids within activated sludge and aquatic environments (Figure 9A). Only low levels of *Inoviridae* transcription were observed within the assembled metatranscriptome samples of this study. However, this limited detection may be the result of a poor inovirus-detecting HMM, as it was recently shown they are far more prevalent within environmental samples than previously appreciated (Roux *et al*. 2019). As +ssRNA phages were detected alongside replicating +ssDNA and dsDNA phages of differing morphologies, this highlights the complex challenges faced at understanding all facets of a microbiome's viral constituents.

A



B

CRISPR spacer ID	Viral RefSeq hit	Spacer length (bp)	% Identity
SRR6049586_34554_length_594_cov_4.930902_g32347_i0_16659	Staphylococcus phage SpaA1	29	100
SRR6049586_34554_length_594_cov_4.930902_g32347_i0_16659	Bacillus phage phi4J1	29	100
SRR6049586_34554_length_594_cov_4.930902_g32347_i0_16659	Bacillus phage Waukesha92	29	100
SRR6254353_131237_length_313_cov_1.625000_g128065_i0_20879	Synechococcus phage S-RSM4	28	100
SRR6960509_109618_length_418_cov_6.208696_g104094_i0_25281	Streptococcus phage Str-PAP-1	33	100

Figure 9. Analysis of microbial community complexity. (A) Detection of diverse phages with differing morphology, infection strategies and encoded using alternative genetic material. The search was performed using HMMs built from Pfam protein families of caudoviral terminase (PF04466), microviral capsid protein (PF02305), inoviral viral endonuclease (PF11726), and *Levivirus* coat protein (PF01819), against the 82 metatranscriptomic samples. **(B)** CRISPR analysis using Viral RefSeq database (version 89).

Evidence of CRISPR spacers against +ssRNA phages that could implicate potential host bacteria was also looked for. Using an in-house pipeline, a database of all potential CRISPR spacers predicted was built using 'pilercr' (version 1.06; (Edgar 2007)). A total of 37,095 CRISPR spacers, between 20-75bp in length, were predicted from the 82 metatranscriptome sample assemblies. The BLASTn was adapted for short sequences ("-evalue 1 -word_size 7 -gapopen 10 -gapextend 2 -penalty -1 -dust no"). When the predicted CRISPR spacers were queried locally against the Viral RefSeq database (version 89), they perfectly matched sequences observed in *Staphylococcus*, *Bacillus*, *Synechococcus* and *Streptococcus*-infecting phages (Figure 9B). However, no CRISPR spacers were found to target the 15,611 new +ssRNA phages identified within this study. A similar observation was noted by Silas and colleagues (Silas *et al.* 2017). Therefore, advances in alternative techniques may be required to identify +ssRNA phage-host partners, as has been demonstrated for dsDNA phages using Hi-C sequencing and single-cell viral tagging (Marbouty *et al*. 2017; Džunková *et al*. 2019).

With the availability of multiple +ssRNA phage sequences, investigations of the regions within +ssRNA phage genomes under evolutionary selective pressure were conducted. These hotspots are often involved in phage-host interactions. As the MP of +ssRNA phages is the host receptor-binding protein, attention was focused on this specific protein. When 15 +ssRNA phage's MP protein sequences of cluster A were investigated, which varied in their BLASTp similarities (sharing 92 to 52% identity), three regions across these MPs with high variability were found which had also been highlighted in the case of a +ssRNA phage AP205 (Figure 10A; (Klovins *et al*. 2002)). Through protein homology modelling with PyMOL (version 2.2.2) using PDB model 5TC1 (Dai *et al*. 2017), it was found these three regions, when folded, formed the beta-sheet domain involved in host-binding, as found in a previous study focused on Qbeta by Gorzelnik *et al* (Figure 10B; (Gorzelnik *et al*. 2016)). In addition, the specific MP variable region is on the exposed virion surface (Figure 10C), with the more

conserved alpha-helical domain in contact with CP subunits and the viral +ssRNA genome (Figure 10D).

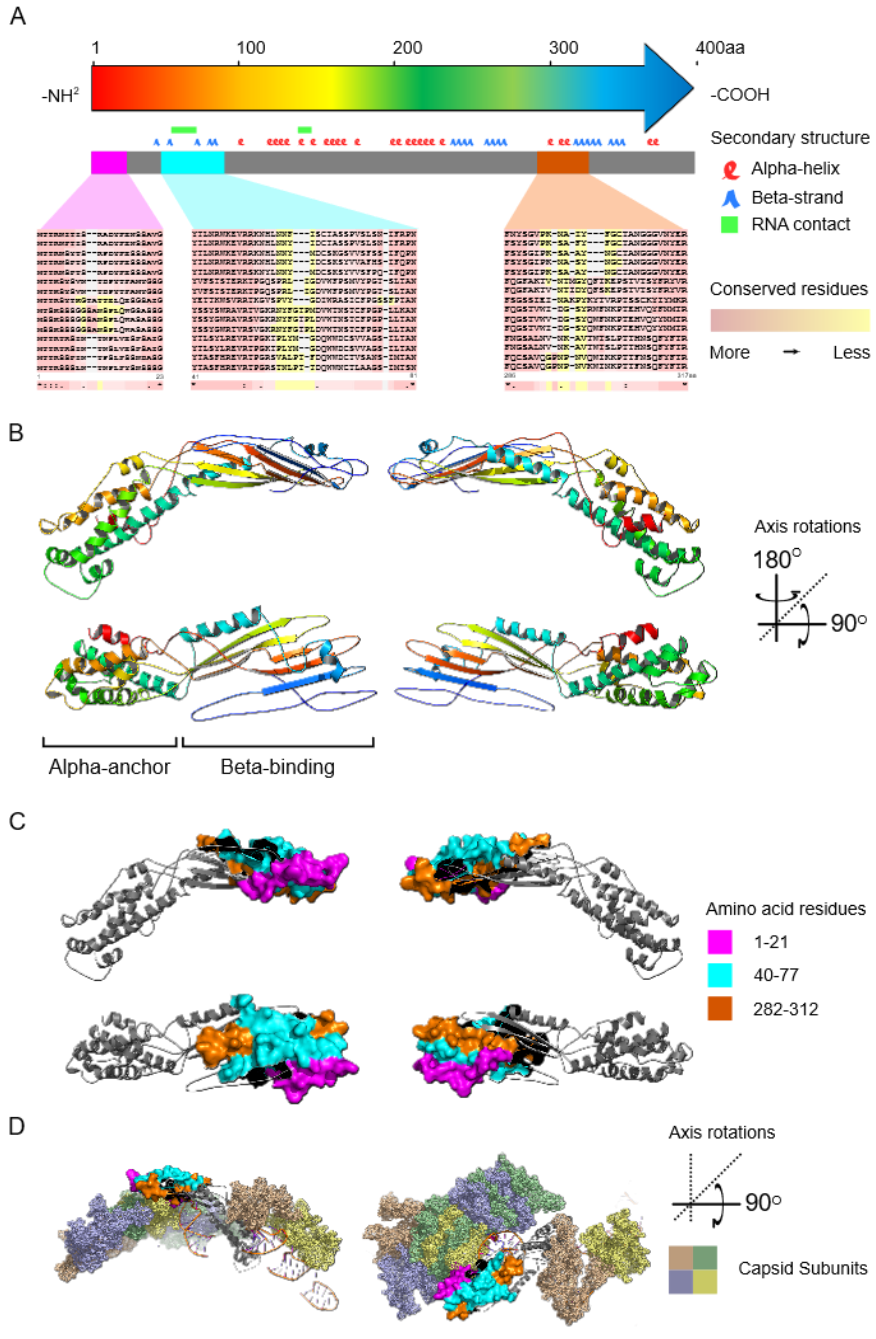


Figure 10. Structural investigation of +ssRNA phage-host interactions. **(A)** A cartoon representation of the MP of +ssRNA phages, with amino acid (aa) length, predicted secondary structure, RNA contact sites, and variable regions highlighted. **(B)** The MP has an alpha-helical

and beta-sheet domain involved in anchoring the protein within the phage virion, and binding host bacteria, respectively. **(C)** The three variable regions of the MP are in close proximity in the folded MP protein. The exposed variable surface area of the beta-sheet domain is displayed, and uses consistent colours to panel A. The corresponding panel images displayed in B and C are identical, but the individual panels are rotated 180˚ around the y-axis, and 90˚ around the x-axis. **(D)** The MP incorporated in a partial reconstruction of an +ssRNA phage virion, emphasising the variable regions with respect to CP subunits and the +ssRNA genome.

3.5 Conclusion

In summary, a HMM-based +ssRNA phage discovery pipeline was iteratively optimised. Through intensive data mining of multiple metatranscriptomic datasets from just two environmental ecosystems, 15,611 near-complete and partial genomes were identified. These samples originated from America, Austria, Japan, and Singapore, highlighting the global distribution of these viruses. This represents an approximate 60-fold expansion of previously known genome sequences. Phylogenetic comparison of 1,044 near-complete genomes allowed a robust, yet elastic, taxonomic scheme to be constructed that provides a hierarchal foundation which will accommodate the expected increase in +ssRNA phage discoveries. Given the amount of the +ssRNA phages identified in this study from two environments, it is suspected that their low abundance in metagenomic studies of other ecosystems may be attributed to a variety of factors, including isolation protocols and computational shortcomings.

3.6 References

Alberti, Adriana, Caroline Belser, Stéfan Engelen, Laurie Bertrand, Céline Orvain, Laura Brinas, Corinne Cruaud, *et al*. 2014. "Comparison of Library Preparation Methods Reveals Their Impact on Interpretation of Metatranscriptomic Data." *BMC Genomics* 15 (1). https://doi.org/10.1186/1471-2164-15-912.

Allaire, J.J, Christopher Gandrud, Kenton Russell, and CJ Yetman. 2017. "NetworkD3: D3 JavaScript Network Graphs from R Version 0.4 from CRAN." 2017. https://rdrr.io/cran/networkD3/.

Atkins, John F., Joan A. Steitz, Carl W. Anderson, and Peter Model. 1979. "Binding of Mammalian Ribosomes to MS2 Phage RNA Reveals an Overlapping Gene Encoding a Lysis Function." *Cell* 18 (2): 247–56. https://doi.org/10.1016/0092-8674(79)90044-8.

Barylski, Jakub, François Enault, Bas E Dutilh, Margo B P Schuller, Robert A Edwards, Annika Gillis, Jochen Klumpp, *et al*. 2019. "Analysis of Spounaviruses as a Case Study for the Overdue Reclassification of Tailed Phages." *Systematic Biology*, May, syz036. https://doi.org/10.1093/sysbio/syz036.

Beekwilder, M. J., R. Nieuwenhuizen, and J. van Duin. 1995. "Secondary Structure Model of the Last Two Domains of Single-Stranded RNA Phage Qβ." *Journal of Molecular Biology* 247 (5): 903–17. https://doi.org/10.1006/jmbi.1995.0189.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics* 30 (15): 2114–20. https://doi.org/10.1093/bioinformatics/btu170.

Bushmanova, Elena, Dmitry Antipov, Alla Lapidus, and Andrey D Przhibelskiy. 2018. "RnaSPAdes: A de Novo Transcriptome Assembler and Its Application to RNA-Seq Data." *BioRxiv*, September. https://doi.org/10.1101/420208.

Charif, Delphine, Olivier Clerc, Carolin Frank, Jean R. Lobry, Anamaria Necşulea, Leonor Palmeira, Simon Penel, and Guy Perrière. 2017. *Seqinr: Biological Sequences Retrieval and Analysis* (version 3.4-5). https://CRAN.R-project.org/package=seqinr.

Charif, Delphine, and Jean R. Lobry. 2007. "SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis." In *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations,* edited by Ugo Bastolla, Markus Porto, H. Eduardo Roman, and Michele Vendruscolo, 207–32. Biological and Medical Physics, Biomedical Engineering. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-35306-5_10.

Clokie, Martha RJ, Andrew D Millard, Andrey V Letarov, and Shaun Heaphy. 2011. "Phages in Nature." *Bacteriophage* 1 (1): 31–45. https://doi.org/10.4161/bact.1.1.14942.

Cobián Güemes, Ana Georgina, Merry Youle, Vito Adrian Cantú, Ben Felts, James Nulton, and Forest Rohwer. 2016. "Viruses as Winners in the Game of Life." *Annual Review of Virology* 3 (1): 197–214. https://doi.org/10.1146/annurev-virology-100114-054952.

Dai, Xinghong, Zhihai Li, Mason Lai, Sara Shu, Yushen Du, Z. Hong Zhou, and Ren Sun. 2017. "In Situ Structures of the Genome and Genome-Delivery Apparatus in an SsRNA Virus." *Nature* 541 (7635): 112–16. https://doi.org/10.1038/nature20589.

Dantas, Gautam, Morten O.A. Sommer, Patrick H. Degnan, and Andrew L. Goodman. 2013. "Experimental Approaches for Defining Functional Roles of Microbes in the Human Gut." *Annual Review of Microbiology* 67 (1): 459–75. https://doi.org/10.1146/annurev-micro-092412-155642.

Drake, J W, B Charlesworth, D Charlesworth, and J F Crow. 1998. "Rates of Spontaneous Mutation." *Genetics* 148 (4): 1667–86.

Duin, Jan van. 1988. "Single-Stranded RNA Bacteriophages." In *The Bacteriophages*, 117–67. The Viruses. Springer, Boston, MA. https://doi.org/10.1007/978-1-4684-5424-6_4.

Džunková, Mária, Soo Jen Low, Joshua N. Daly, Li Deng, Christian Rinke, and Philip Hugenholtz. 2019. “Defining the Human Gut Host–Phage Network through Single-Cell Viral Tagging.” *Nature Microbiology*, August. https://doi.org/10.1038/s41564-019-0526-2.

Edgar, Robert C. 2004. “MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput.” *Nucleic Acids Research* 32 (5): 1792–97. https://doi.org/10.1093/nar/gkh340.

Edgar, Robert C. 2007. “PILER-CR: Fast and Accurate Identification of CRISPR Repeats.” *BMC Bioinformatics* 8 (1): 18. https://doi.org/10.1186/1471-2105-8-18.

Edwards, Robert A., Alejandro A. Vega, Holly M. Norman, Maria Ohaeri, Kyle Levi, Elizabeth A. Dinsdale, Ondrej Cinek, *et al*. 2019. “Global Phylogeography and Ancient Evolution of the Widespread Human Gut Virus CrAssphage.” *Nature Microbiology*, July. https://doi.org/10.1038/s41564-019-0494-6.

Fiers, W., R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaert, W. Min Jou, *et al*. 1976. “Complete Nucleotide Sequence of Bacteriophage MS2 RNA: Primary and Secondary Structure of the Replicase Gene.” *Nature* 260 (5551): 500–507. https://doi.org/10.1038/260500a0.

Finn, Robert D., Jody Clements, and Sean R. Eddy. 2011. “HMMER Web Server: Interactive Sequence Similarity Searching.” *Nucleic Acids Research* 39 (suppl_2): W29–37. https://doi.org/10.1093/nar/gkr367.

Fischer, Steve, Brian P. Brunk, Feng Chen, Xin Gao, Omar S. Harb, John B. Iodice, Dhanasekaran Shanmugam, David S. Roos, and Christian J. Stoeckert. 2011. “Using OrthoMCL to Assign Proteins to OrthoMCL-DB Groups or to Cluster Proteomes Into New Ortholog Groups.” In *Current Protocols in Bioinformatics*, edited by David S.

Goodsell, bi0612s35. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/0471250953.bi0612s35.

Gabor Csardi and Tamas Nepusz. 2006. “The Igraph Software Package for Complex Network Research.” InterJournal, Complex Systems. http://igraph.org.

Gorbalenya, Alexander, Mart Krupovic, Stuart Siddell, Arvind Varsani, and Jens H. Kuhn. 2019. “*Riboviria*: Establishing a Single Taxon That Comprises RNA Viruses at the Basal Rank of Virus Taxonomy.” International Committee on Taxonomy of Viruses (ICTV).

Gorzelnik, Karl V., Zhicheng Cui, Catrina A. Reed, Joanita Jakana, Ry Young, and Junjie Zhang. 2016. “Asymmetric Cryo-EM Structure of the Canonical *Allolevivirus* Qβ Reveals a Single Maturation Protein and the Genomic ssRNA in Situ.” *Proceedings of the National Academy of Sciences* 113 (41): 11519–24. https://doi.org/10.1073/pnas.1609482113.

Grazziotin, Ana Laura, Eugene V. Koonin, and David M. Kristensen. 2017. “Prokaryotic Virus Orthologous Groups (PVOGs): A Resource for Comparative Genomics and Protein Family Annotation.” *Nucleic Acids Research* 45 (Database issue): D491–98. https://doi.org/10.1093/nar/gkw975.

Gregory, Ann C., Ahmed A. Zayed, Nádia Conceição-Neto, Ben Temperton, Ben Bolduc, Adriana Alberti, Mathieu Ardyna, *et al*. 2019. “Marine DNA Viral Macro- and Microdiversity from Pole to Pole.” *Cell* 177 (5): 1109-1123.e14. https://doi.org/10.1016/j.cell.2019.03.040.

Gytz, Heidi, Durita Mohr, Paulina Seweryn, Yuichi Yoshimura, Zarina Kutlubaeva, Fleur Dolman, Bosene Chelchessa, *et al*. 2015. “Structural Basis for RNA-Genome Recognition during Bacteriophage Qβ Replication.” *Nucleic Acids Research* 43 (22): 10893–906. https://doi.org/10.1093/nar/gkv1212.

Hatfull, Graham F, and Roger W Hendrix. 2011. “Bacteriophages and Their Genomes.” *Current Opinion in Virology* 1 (4): 298–303. https://doi.org/10.1016/j.coviro.2011.06.009.

Holland, J., K. Spindler, F. Horodyski, E. Grabau, S. Nichol, and S. VandePol. 1982. “Rapid Evolution of RNA Genomes.” *Science* 215 (4540): 1577–85. https://doi.org/10.1126/science.7041255.

Hyatt, Doug, Gwo-Liang Chen, Philip F. LoCascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. 2010. “Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification.” *BMC Bioinformatics* 11 (1): 119. https://doi.org/10.1186/1471-2105-11-119.

Kannoly, Sherin, Yongping Shao, and Ing-Nang Wang. 2012. “Rethinking the Evolution of Single-Stranded RNA (ssRNA) Bacteriophages Based on Genomic Sequences and Characterizations of Two R-Plasmid-Dependent ssRNA Phages, C-1 and Hgal1.” *Journal of Bacteriology* 194 (18): 5073–79. https://doi.org/10.1128/JB.00929-12.

Kazaks, Andris, Tatyana Voronkova, Janis Rumnieks, Andris Dishlers, and Kaspars Tars. 2011. “Genome Structure of Caulobacter Phage PhiCb5.” *Journal of Virology* 85 (9): 4628–31. https://doi.org/10.1128/JVI.02256-10.

Kleiner, Manuel, Lora V Hooper, and Breck A Duerkop. 2015. “Evaluation of Methods to Purify Virus-like Particles for Metagenomic Sequencing of Intestinal Viromes.” *BMC Genomics* 16 (1). https://doi.org/10.1186/s12864-014-1207-4.

Klovins, J., G. P. Overbeek, S. H. E. van den Worm, H.-W. Ackermann, and J. van Duin. 2002. “Nucleotide Sequence of a ssRNA Phage from *Acinetobacter*: Kinship to Coliphages.” *Journal of General Virology* 83 (6): 1523–33. https://doi.org/10.1099/0022-1317-83-6-1523.

Koonin, Eugene V., Tatiana G. Senkevich, and Valerian V. Dolja. 2006. “The Ancient Virus World and Evolution of Cells.” *Biology Direct* 1 (September): 29. https://doi.org/10.1186/1745-6150-1-29.

Krishnamurthy, Siddharth R., Andrew B. Janowski, Guoyan Zhao, Dan Barouch, and David Wang. 2016. “Hyperexpansion of RNA Bacteriophage Diversity.” *PLOS Biology* 14 (3): e1002409. https://doi.org/10.1371/journal.pbio.1002409.

Kumar, Sudhir, Glen Stecher, Michael Li, Christina Knyaz, and Koichiro Tamura. 2018. “MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms.” *Molecular Biology and Evolution* 35 (6): 1547–49. https://doi.org/10.1093/molbev/msy096.

Li, Dinghua, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. 2015. “MEGAHIT: An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct de Bruijn Graph.” *Bioinformatics* 31 (10): 1674–76. https://doi.org/10.1093/bioinformatics/btv033.

Li, Heng. 2019. *Toolkit for Processing Sequences in FASTA/Q Formats: Lh3/Seqtk*. C. https://github.com/lh3/seqtk.

Li, Weizhong, Lukasz Jaroszewski, and Adam Godzik. 2001. “Clustering of Highly Homologous Sequences to Reduce the Size of Large Protein Databases.” *Bioinformatics* 17 (3): 282–83. https://doi.org/10.1093/bioinformatics/17.3.282.

Lodish, Harvey F. 1968. “Bacteriophage F2 RNA: Control of Translation and Gene Order.” *Nature* 220 (5165): 345–50. https://doi.org/10.1038/220345a0.

Loeb, Tim, and Norton D. Zinder. 1961. “A Bacteriophage Containing RNA.” *Proceedings of the National Academy of Sciences of the United States of America* 47 (3): 282–89.

Manrique, Pilar, Benjamin Bolduc, Seth T. Walk, John van der Oost, Willem M. de Vos, and Mark J. Young. 2016. “Healthy Human Gut Phageome.” *Proceedings of the National Academy of Sciences* 113 (37): 10400–405. https://doi.org/10.1073/pnas.1601060113.

Marbouty, Martial, Lyam Baudry, Axel Cournac, and Romain Koszul. 2017. “Scaffolding Bacterial Genomes and Probing Host-Virus Interactions in Gut Microbiome by Proximity Ligation (Chromosome Capture) Assay.” *Science Advances* 3 (2): e1602105. https://doi.org/10.1126/sciadv.1602105.

Martin, Marcel. 2011. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads.” *EMBnet.Journal* 17 (1): 10–12. https://doi.org/10.14806/ej.17.1.200.

Millard, Andrew D. 2009. “Isolation of Cyanophages from Aquatic Environments.” In *Bacteriophages: Methods and Protocols, Volume 1: Isolation, Characterization, and Interactions*, edited by Martha R.J. Clokie and Andrew M. Kropinski, 33–42. Methods in Molecular Biology™. Totowa, NJ: Humana Press. https://doi.org/10.1007/978-1-60327-164-6_4.

Neuwirth, Erich. 2014. *RColorBrewer: ColorBrewer Palettes* (version 1.1-2). https://CRAN.R-project.org/package=RColorBrewer.

Norman, Jason M., Scott A. Handley, Megan T. Baldridge, Lindsay Droit, Catherine Y. Liu, Brian C. Keller, Amal Kambal, *et al*. 2015. “Disease-Specific Alterations in the Enteric Virome in Inflammatory Bowel Disease.” *Cell* 160 (3): 447–60. https://doi.org/10.1016/j.cell.2015.01.002.

Olsthoorn, René C. L., and Jan van Duin. 2017. “*Leviviridae* - Positive Sense RNA Viruses - Positive Sense RNA Viruses (2011) - International Committee on Taxonomy of Viruses (ICTV).” International Committee on Taxonomy of Viruses (ICTV). 2017.

https://talk.ictvonline.org/ictv-reports/ictv_9th_report/positive-sense-rna-viruses-2011/w/posrna_viruses/263/leviviridae.

Olsthoorn, René C. L., and Jan van Duin. 2011. “Bacteriophages with ssRNA.” *E LS*, July. https://doi.org/10.1002/9780470015902.a0000778.pub3.

Paradis, Emmanuel, and Klaus Schliep. 2019. “Ape 5.0: An Environment for Modern Phylogenetics and Evolutionary Analyses in R.” *Bioinformatics* 35 (3): 526–28. https://doi.org/10.1093/bioinformatics/bty633.

Phillips, Nathaniel. 2017. *Yarrr: A Companion to the e-Book “YaRrr!: The Pirate’s Guide to R”* (version 0.1.5). https://CRAN.R-project.org/package=yarrr.

Ram, Karthik, Hadley Wickham, Clark Richards, and Aaron Baggett. 2018. *Wesanderson: A Wes Anderson Palette Generator* (version 0.3.6). https://CRAN.R-project.org/package=wesanderson.

Rokyta, D. R., C. L. Burch, S. B. Caudle, and H. A. Wichman. 2006. “Horizontal Gene Transfer and the Evolution of Microvirid Coliphage Genomes.” *Journal of Bacteriology* 188 (3): 1134–42. https://doi.org/10.1128/JB.188.3.1134-1142.2006.

Roux, Simon, Mart Krupovic, Rebecca A. Daly, Adair L. Borges, Stephen Nayfach, Frederik Schulz, Jan-Fang Cheng, *et al*. 2019. “Cryptic Inoviruses Are Pervasive in Bacteria and Archaea across Earth’s Biomes.” *BioRxiv*, February, 548222. https://doi.org/10.1101/548222.

Roux, Simon, Mart Krupovic, Rebecca A. Daly, Adair L. Borges, Stephen Nayfach, Frederik Schulz, Allison Sharrar, *et al*. 2019. “Cryptic Inoviruses Revealed as Pervasive in Bacteria and Archaea across Earth’s Biomes.” *Nature Microbiology*, July, 1–12. https://doi.org/10.1038/s41564-019-0510-x.

RStudio Team. 2020. “RStudio: Integrated Development for R. RStudio, PBC, Boston, MA.” 2020. https://rstudio.com/.

Rumnieks, Janis, and Kaspars Tars. 2012. “Diversity of Pili-Specific Bacteriophages: Genome Sequence of IncM Plasmid-Dependent RNA Phage M.” *BMC Microbiology* 12 (1): 277. https://doi.org/10.1186/1471-2180-12-277.

Schliep, Klaus Peter. 2011. “Phangorn: Phylogenetic Analysis in R.” *Bioinformatics* 27 (4): 592–93. https://doi.org/10.1093/bioinformatics/btq706.

Shi, Mang, Xian-Dan Lin, Jun-Hua Tian, Liang-Jun Chen, Xiao Chen, Ci-Xiu Li, Xin-Cheng Qin, *et al*. 2016. “Redefining the Invertebrate RNA Virosphere.” *Nature* 540 (7634): 539–43. https://doi.org/10.1038/nature20167.

Silas, Sukrit, Kira S. Makarova, Sergey Shmakov, David Páez-Espino, Georg Mohr, Yi Liu, Michelle Davison, *et al*. 2017. “On the Origin of Reverse Transcriptase-Using CRISPR-Cas Systems and Their Hyperdiverse, Enigmatic Spacer Repertoires.” Edited by Stephen P. Goff. *MBio* 8 (4): e00897-17, /mbio/8/4/e00897-17.atom. https://doi.org/10.1128/mBio.00897-17.

Simon-Loriere, Etienne, and Edward C. Holmes. 2011. “Why Do RNA Viruses Recombine?” *Nature Reviews Microbiology* 9 (8): 617–26. https://doi.org/10.1038/nrmicro2614.

Starr, Evan P., Erin E. Nuccio, Jennifer Pett-Ridge, Jillian F. Banfield, and Mary K. Firestone. 2019. “Metatranscriptomic Reconstruction Reveals RNA Viruses with the Potential to Shape Carbon Cycling in Soil.” Preprint. Microbiology. https://doi.org/10.1101/597468.

Walker, Peter J., Stuart G. Siddell, Elliot J. Lefkowitz, Arcady R. Mushegian, Donald M. Dempsey, Bas E. Dutilh, Balázs Harrach, *et al*. 2019. “Changes to Virus Taxonomy and the International Code of Virus Classification and Nomenclature Ratified by the International Committee on Taxonomy of Viruses (2019).” *Archives of Virology*, June. https://doi.org/10.1007/s00705-019-04306-w.

Weiner, Alan M., and Klaus Weber. 1973. “A Single UGA Codon Functions as a Natural Termination Signal in the Coliphage Qβ Coat Protein Cistron.” *Journal of Molecular Biology* 80 (4): 837–55. https://doi.org/10.1016/0022-2836(73)90213-1.

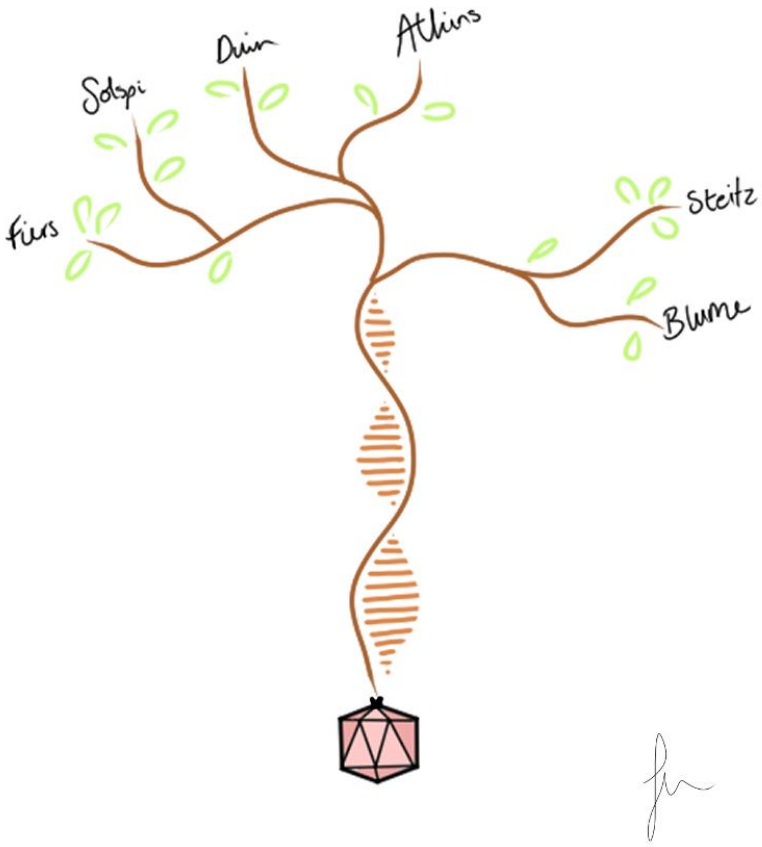
Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, Dewey Dunnington, and RStudio. 2021. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics* (version 3.3.5). https://CRAN.R-project.org/package=ggplot2.

Wolf, Yuri I., Darius Kazlauskas, Jaime Iranzo, Adriana Lucía-Sanz, Jens H. Kuhn, Mart Krupovic, Valerian V. Dolja, and Eugene V. Koonin. 2018. “Origins and Evolution of the Global RNA Virome.” *MBio* 9 (6): e02329-18. https://doi.org/10.1128/mBio.02329-18.

Yu, Guangchuang. 2019. “Treeio: Base Classes and Functions for Phylogenetic Tree Input and Output Version 1.6.2 from Bioconductor.” 2019. https://rdrr.io/bioc/treeio/.

Yu, Guangchuang, David K. Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk Lam. 2017. “Ggtree: An r Package for Visualization and Annotation of Phylogenetic Trees with Their Covariates and Other Associated Data.” *Methods in Ecology and Evolution* 8 (1): 28–36. https://doi.org/10.1111/2041-210X.12628.

Zimmermann, Lukas, Andrew Stephens, Seung-Zin Nam, David Rau, Jonas Kübler, Marko Lozajic, Felix Gabler, Johannes Söding, Andrei N. Lupas, and Vikram Alva. 2018. “A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at Its Core.” *Journal of Molecular Biology*, Computation Resources for Molecular Biology, 430 (15): 2237–43. https://doi.org/10.1016/j.jmb.2017.12.007.


Athens
Duin
Solspi
fiers
Steitz
Blume

Chapter IV

Leviviricetes: Expanding and Restructuring the Taxonomy of Bacteria-Infecting Single-Stranded RNA Viruses

A modified version of this chapter was published in *Microbial Genomics,* in which I conceived the study, performed the analysis, produced the images, and wrote the manuscript with my co-author Dr Stephen R. Stockdale. ICTV-specific changes were made in the accepted manuscript.

Callanan, Julie, Stephen R. Stockdale, Evelien M. Adriaenssens, Jens H. Kuhn, Janis Rumnieks, Mark J. Pallen, Andrey N. Shkoporov, Lorraine A. Draper, R. Paul Ross, and Colin Hill. 2021 "*Leviviricetes*: Expanding and Restructuring the Taxonomy of Bacteria-Infecting Single-Stranded RNA Viruses." Microbial Genomics 7 (11): 000686.

https://doi.org/10.1099/mgen.0.000686

4.1 Abstract

Until recent years, the limited number of positive-sense, single-stranded RNA (+ssRNA) prokaryotic viruses restricted their phylogenetic analysis. Only a single family was described that encapsulated the four species classified across two genera. However, recent metagenomic and metatranscriptomic studies have recorded a greater diversity of +ssRNA viruses than ever anticipated. Subsequently, this inspired their complete reorganization. The class *Allassoviricetes* was renamed to *Leviviricetes*, the order *Levivirales* changed to *Norzivirales* with the creation of a second order, *Timlovirales*, and the class now encompasses a total of six families, 428 genera, and 882 species. Here the new taxonomy of *Leviviricetes*, approved and ratified in 2021 by the International Committee on Taxonomy of Viruses (ICTV), is outlined. Furthermore, the taxonomy-informative open-access hidden Markov models are described to support the scientific community in the identification and classification of additional +ssRNA viruses within this new taxonomic framework.

4.2 Open access data

Open access data describing the taxonomic proposal (TaxoProp) and HMMs can found at the following web addresses, respectively:

1 https://bit.ly/38fY9Rq (2020 edition)

2 https://bit.ly/2Wr5Xgc

4.3 Introduction

In their 2020 report, the International Committee on Taxonomy of Viruses (ICTV) used eight primary taxon ranks (from realm to species) and seven secondary taxon ranks (e.g., subfamily) (Walker *et al*. 2021). This hierarchical system is designed to capture the evolutionary relationships of viruses within taxonomic ranks and assign each taxon a unique and defined name. Historically, the ICTV required viruses to be isolated to be successfully characterized and categorized. However, in recent years, the ICTV has acknowledged that many viruses would remain unclassified due to the inability to cultivate their hosts in laboratory settings or because of unique viral lifecycles (Simmonds *et al*. 2017). To overcome this restriction and to take advantage of bioinformatic developments, the ICTV has allowed sequences assembled from metagenomes to be included in taxonomy proposals once they meet the Minimum Information about an Uncultivated Virus Genome (MIUViG) standards (Roux *et al*. 2019), which can be used along with the existing guidelines (Simmonds *et al*. 2017).

In 2019, the ICTV accepted a proposal to create a realm, *Riboviria*, to encompass the majority of RNA viruses (Walker *et al*. 2020). All members of the realm *Riboviria* universally encode an RNA-directed RNA polymerase (RdRP) gene, enabling phylogenetic analyses across all groups of bacteria, fungi, plant, animal, and human-infecting RNA viruses despite limited sequence identity (Wolf *et al*. 2018).

The +ssRNA virus MS2 was the first genome of any organism to ever be sequenced (Fiers *et al*. 1976). In the First Report of the International Committee on Nomenclature of Viruses (ICNV; today the recognized as the ICTV), phage MS2 was classified as a member of the "ribophage group" (Wildy 1971), which in the Second Report (1976) was considered a genus in the then-new family Leviviridae (unitalicized at the time) (Fenner 1976). Through the discovery of additional *Escherichia coli* +ssRNA phages, the family was marginally expanded. Most recently (prior to 2021), the *Leviviridae* family included two genera for a total of four

species. Those four viruses remained the only classified viruses in *Leviviridae*-including class *Allassaviricetes*—one of four classes in phylum *Lenarviricota* (Table 1), although some 50 other viruses were considered possible members of the family. This low diversity starkly contrasted that of the thousands of prokaryotic viruses, predominantly with double-stranded DNA but also with single-stranded DNA and double-stranded RNA genomes, which are classified across five of the six currently established virus realms (*Adnaviridae, Duplodnaviria*, *Monodnaviria*, *Riboviria*, and *Varidnaviriae*) (Krupovic *et al*. 2021; Koonin *et al*. 2020).

Table 1. Pre-2021 taxonomy of +ssRNA bacterial viruses (Koonin *et al*. 2020; Olsthoorn and van Duin 2009; "Taxonomy" 2021)

Realm	Kingdom	Phylum	Class	Order	Family	Genus	Species	Virus
Riboviria	*Orthnornavirae* (one of two kingdoms in the realm)	*Lenarviricota* (one of five phyla in the kingdom)	*Allassoviricetes* (one of four classes in the phylum)	*Levivirales*	*Leviviridae*	*Allolevivirus*	*Escherichia virus F1*	Enterobacteria phage FI 4184 b
							Escherichia virus Qbeta	Escherichia phage Qbeta
						Levivirus	*Escherichia virus BZ13*	Escherichia phage BZ13
							Escherichia virus MS2	Escherichia phage MS2

Bacteria-infecting +ssRNA viruses were separated across two genera, *Levivirus* and *Allolevivirus*. The original taxonomic categorization of the family *Leviviridae* into the genera *Levivirus* and *Allolevivirus* was based on a variety of factors, including the detection of either three core genes, the maturation protein (MP), coat protein (CP), and the RdRP—or the detection of a separate lysis protein as opposed to a unique read-through protein, respectively (Olsthoorn and van Duin 2009). Each of these genera included two species, *Levivirus* with Escherichia virus MS2 and Escherichia virus BZ13, and *Allolevivirus* with Escherichia virus Qbeta and Escherichia virus F1. Only 25 members of the family *Leviviridae* were classified into the two defined genera, while 32 additional genome sequences that have been recently added, could not be categorized below family.

Recent metagenomic and metatranscriptomic analyses have suggested that bacteria-infecting +ssRNA viruses are more abundant than previously thought. This is based on the number of isolates in public databases and the rate of detection in virome studies. It has been suggested that difficulties associated with +ssRNA viral identification are due to biases in virome studies such as isolation procedures, or the number of +ssRNA viruses in specific samples (Callanan *et al*. 2021; Krishnamurthy *et al*. 2016; Zhang *et al*. 2005). Nonetheless, recent studies are expanding the collection of +ssRNA viruses and a rigorous overhaul of their taxonomic structure was essential to progress our understanding of their diversity and ecological significance. Here, a new taxonomic scheme for bacteria-infecting +ssRNA viruses that includes a single class with two orders, six families, 428 genera, and 882 species is presented. To support the continued expansion of this scheme, hidden Markov models (HMMs) generated from the expanded number of available +ssRNA virus proteins have been made available on Figshare. The +ssRNA virus HMMs can detect and provide putative taxonomic information to newly identified sequences within the current framework (Callanan *et al*. 2020).

4.4 Profile HMMs to detect and classify bacteria-infecting +ssRNA viruses

To create a novel taxonomic framework for bacteria-infecting +ssRNA viruses, 1,868 genomic sequences tentatively identified as such in previous studies (Krishnamurthy *et al*. 2016; Callanan *et al*. 2020; Shi *et al*. 2016; Starr *et al*. 2019) were obtained from the US National Center for Biotechnology Information (NCBI). Grouping of bacteria-infecting +ssRNA virus proteins was achieved using orthoMCL that implements Markov clustering (Li, Stoeckert, and Roos 2003). Three MP clusters, nine CP clusters, and two RdRP clusters were generated and given alphabetical labels that reflect their original descriptions (Callanan *et al*. 2020). Profile HMMs, based on orthoMCL clusters, were used to detect distant relationships between the three core +ssRNA virus proteins. Phylogenetic analysis of bacteria-infecting +ssRNA virus RdRPs and CPs largely agreed with protein clustering. Therefore, the phylogeny of RdRPs and CPs were used as the demarcation criterion for establishing orders and families, respectively (Figure 1). A comparison of the phylogenetic and open-access HMM approaches (see Open access data) for bacteria-infected +ssRNA virus classification yielded only nine instances (out of 882 species-representing sequences) of disagreements (Figure 2). Therefore, although not perfect in its classification predictions, the HMMs will confidently identify bacteria-infecting +ssRNA virus sequences and provide end-users with additional information to continue the expansion taxonomic framework presented here.

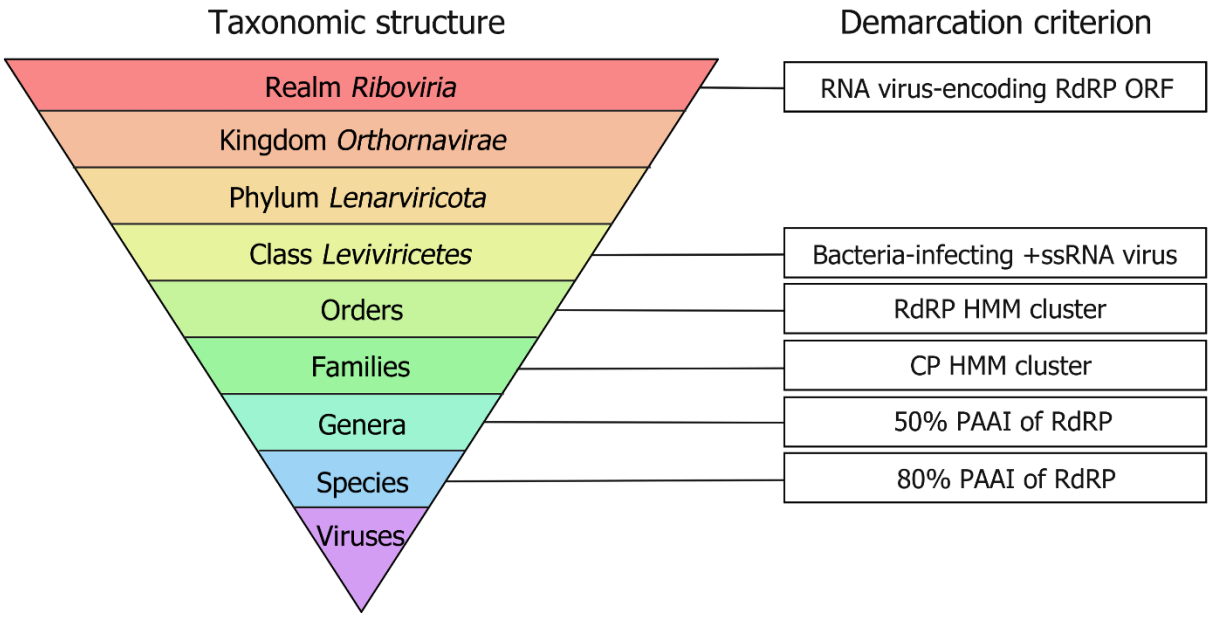


Figure 1. Taxon demarcation criteria for *Leviviricetes* classification. Taxonomic ranks for positive-sense single-stranded RNA (+ssRNA) viruses are shown alongside the demarcation criterion for each of the taxon ranks. PAAI, pairwise amino-acid sequence identity; CP, coat protein; RdRP, RNA-directed RNA polymerase; HMM, hidden Markov model; ORF, open reading frame.

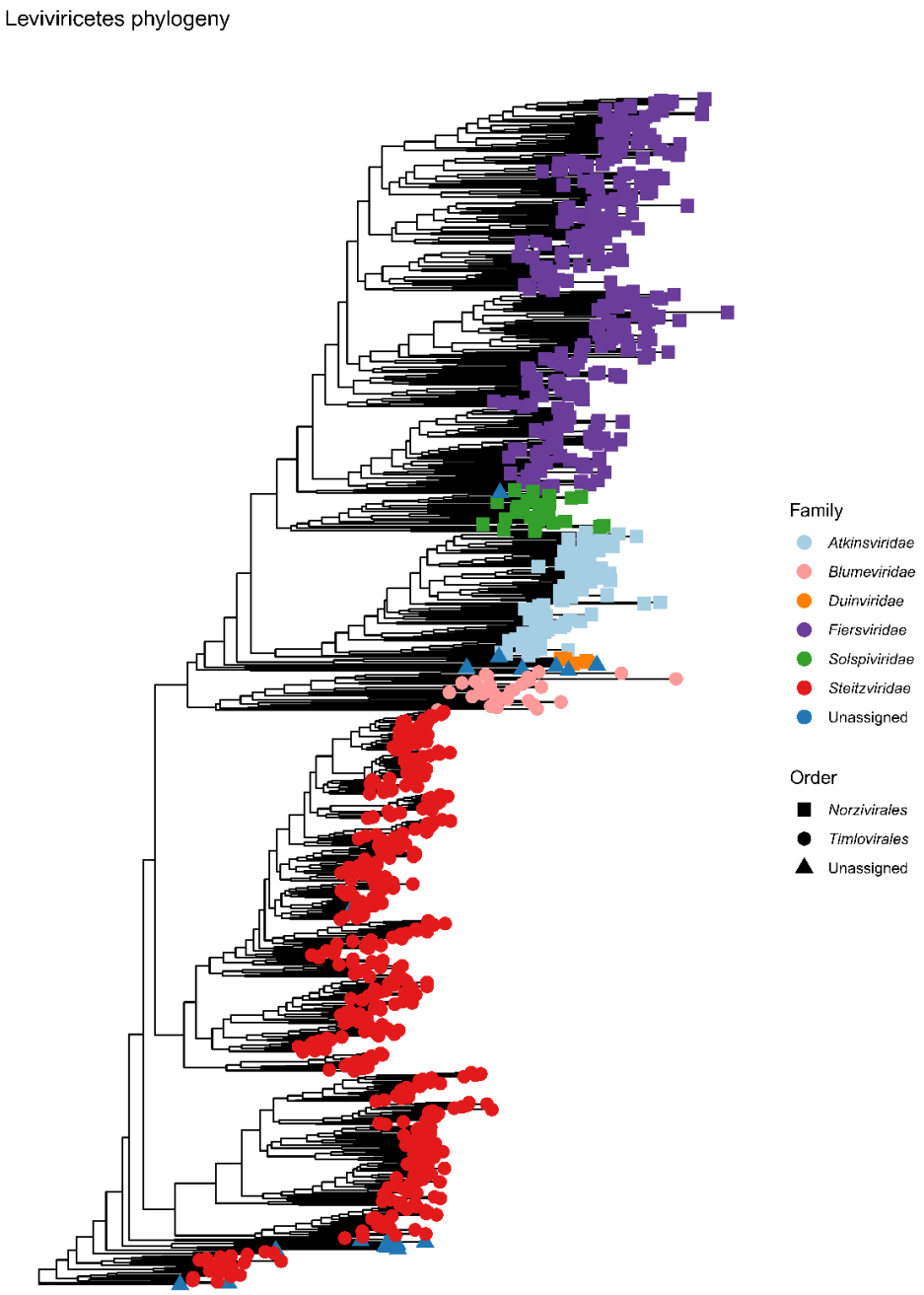


Figure 2. Hidden Markov model (HMM) predictions of leviviricetes taxonomy. While expanding and restructuring positive-sense single-stranded RNA (+ssRNA) viruses, nine genera could not be assigned with confidence to a family or an order, as the RNA-directed RNA polymerase (RdRP), and coat protein (CP) genome-encoded combinations did not adhere to established combinations. Additionally, there were ten instances (out of 882 species representative sequences) for which the hidden Markov model (HMM) predicted taxonomy of

+ssRNA viruses did not align with their phylogeny-based assignment. The color and shape aesthetics of the phylogenetic tree illustrates these taxonomic outliers.

4.5 Taxonomy of class *Leviviricetes*

A comparison of the pre-2021 and the 2021 taxonomic breakdowns of bacteria-infecting +ssRNA viruses highlights the significant expansion and restructuring of the *Leviviricetes* taxon by incorporating metagenome-assembled genomes (Tables 1, 2, and 3). The taxonomic ranks established at the order and family ranks are named after prominent +ssRNA virus biologists. The co-discoverers of +ssRNA viruses are acknowledged in the generation and assignment of the two order names, whereas family names were randomly assigned to +ssRNA virus scientists irrespective of the viruses classified at these taxonomic ranks. The description of orders and families are presented alphabetically and do not reflect the historical or future predicted contributions of specific scientists to the +ssRNA virus field. As the *Leviviricetes* taxon is adjusted over time, newly established ranks do not necessarily need to continue the presented naming system.

Table 2. 2021 taxonomy of +ssRNA bacterial viruses (Callanan, Stockdale, Adriaenssens, *et al*. 2021; "International Committee on Taxonomy of Viruses" 2021; Walker *et al*. 2020)

Realm	**Kingdom**	**Phylum**	**Class**	**Order**	**Family**	**Genus**	**Species**	**Virus**
Riboviria	*Orthnornavirae* (one of two kingdoms in the realm)	*Lenarviricota* (one of five phyla in the kingdom)	*Leviviricetes* (one of four classes in the phylum)	*Norzivirales* (one of two orders in the class)	*Fiersviridae* (one of four families in the order)	*Qubevirus*	*Qubevirus faecium*	Enterobacteria phage FI 4184 b
							Qubevirus durum	Escherichia phage Qbeta
						Emesvirus	*Emesvirus japonicum*	Escherichia phage BZ13
							Emesvirus zinderi	Escherichia phage MS2

Table 3. Numerical summary of the 2021 taxonomy of +ssRNA bacterial viruses (Callanan, Stockdale, Adriaenssens, *et al*. 2021; "International Committee on Taxonomy of Viruses" 2021; Walker *et al*. 2020)

Class	Orders	Families	Coat protein (CP) clusters	Number of genera included in family	Number of species included in family
Leviviricetes	*Norzivirales*	*Atkinsviridae*	C	56	91
		Duinviridae	AP205-like	6	6
		Fiersviridae	A, B, and H	185	298
		Solspiviridae	G	24	31
	Timlovirales	*Blumeviridae*	E	31	35
		Steitzviridae	D and F	117	412
	Unassigned	Unassigned	N/A	9	9

Class. The previously described class *Allassoviricetes* that encompassed all bacteria-infecting +ssRNA viruses has been renamed as *Leviviricetes*, reflecting the use of the term levivirus(es). This class now includes all +ssRNA viruses encoding the specific pattern of three +ssRNA virus core proteins: MP, CP, and RdRP. In a recent analysis of +ssRNA virus genomes, 1,868 sequences fit this genome architectural criterion. Additionally, the encoded MP and RdRP were required to meet a minimum length threshold of 350 and 500 amino acid residues, respectively, to ensure only near-complete (coding-complete) genomes were investigated as set out in the MIUViG criteria. The 1,868 sequences originated from sequences available through the NCBI and the studies of Callanan *et al.*, Starr *et al.*, Shi *et al.*, and Krishnamurthy *et al.* (Callanan *et al*. 2020; Starr *et al*. 2019; Shi *et al*. 2016; Krishnamurthy *et al*. 2016).

Order. Clustering and separation of the RdRP into distinct phylogenetic clades was adopted as the order demarcation criterion as the RdRP is the most conserved protein across *Leviviricetes* with the strongest phylogenetic signal (Callanan *et al*. 2020).

The order *Norzivirales* (formerly named *Levivirales*) is based on the phylogeny and clustering of bacterial +ssRNA virus RdRP protein sequences. It is named after Norton Zinder (1928–2012), who isolated the first bacterial virus with an RNA genome and continued to make crucial findings regarding these entities (Loeb and Zinder 1961; August *et al*. 1963). A total of 426 bacteria-infecting +ssRNA viral genomes are categorized as belonging to the *Norzivirales* order. Tying in with its original description, the profile HMM output additionally describes *Norzivirales* hits as cluster RdRP_A (Callanan *et al*. 2020).

Timlovirales: This order is based on the phylogeny and clustering of bacterial +ssRNA virus RdRP protein sequences (cluster RdRP_B). It is named after Timothy Loeb (1935–2016) who, with Norton Zinder, isolated the first +ssRNA bacterial virus (Loeb and Zinder 1961). There are 447 *Leviviricetes* members classified as *Timlovirales*.

Family. Familial taxonomic groups were based on the distinct phylogeny of bacterial +ssRNA virus CP sequences, as either a single cluster or collection of clusters generated from orthoMCL (Callanan *et al*. 2020). Out of 882 +ssRNA virus species representatives, there were nine instances for which the phylogeny of the CP cluster did not match its predicted corresponding RdRP cluster; no order or familial taxonomic rank was designated for these +ssRNA viruses. Once additional related viruses to these outliers are identified, it will be possible to resolve their taxonomy, which may require the formation of additional families. The families *Atkinsviridae*, *Duinviridae*, *Fiersviridae,* and *Solspiviridae* are the new families created within the *Norzivirales* order, whereas *Blumeviridae* and *Steitzviridae* are the new families in the *Timlovirales* order.

Atkinsviridae is named after John Atkins (1944–present) for his discovery of the lysin protein from *Escherichia* virus MS2 (Atkins *et al*. 1979). This family encompasses +ssRNA viruses predicted to encode a CP corresponding to CP cluster C (HMM profile CP_C). There are 91 viruses classified within *Atkinsviridae*.

Blumeviridae is named after Thomas Blumenthal (1943–present) for his findings on the replication of bacterial +ssRNA viruses, in particular the structure and function of the replicase (Blumenthal and Carmichael 1979). This family encompasses +ssRNA viruses predicted to encode a CP corresponding to CP cluster E (HMM profile CP_E). Currently, 35 +ssRNA viruses are classified within *Blumeviridae.*

Duinviridae is named after Jan van Duin (1937–2017) for his discoveries related to novel bacterial +ssRNA viruses, and the RNA folding within bacterial +ssRNA virus genomes to control gene expression (Kastelein *et al*. 1982; van Duin 1988). This family encompasses +ssRNA viruses predicted to encode a CP corresponding to CP cluster AP205-like. There are six *Leviviricetes* classified within *Duinviridae*.

Fiersviridae (formerly named *Leviviridae*) is named after Walter Fiers (1931–2019) who sequenced the first gene and genome of any organism, MS2, previously assigned to the species *Escherichia* virus MS2 (Fiers *et al*. 1976). This family encompasses +ssRNA viruses predicted to encode a CP corresponding to CP clusters A, B, and H (HMM profiles CP_A, CP_B, and CP_H, respectively). There are 298 viruses are assigned to *Fiersviridae*.

Solspiviridae is named after Sol Spiegelman (1914–1983), who discovered an RNA chain of only 218 nucleotides that could be reproduced by an RdRP (Spiegelman *et al*. 1965). This family encompasses +ssRNA viruses predicted to encode a CP corresponding to CP cluster G. There are 31 sequences classified within the *Solspiviridae* family (HMM profile CP_G).

Steitzviridae is named after Joan Argetsinger Steitz (1941–present) for her determination of an initiation sequence that is central to modern-day ribosome profiling (Steitz 1969). This family encompasses +ssRNA viruses predicted to encode a CP corresponding to CP clusters D and F (HMM profiles CP_D and CP_F, respectively). A total of 412 bacteria-infecting +ssRNA viruses are classified within *Steitzviridae.*

Genera. A 50% pairwise amino-acid identity (PAAI) of the viral encoded RdRP protein was chosen as the criterion for establishing genera based on an analysis of the previous ICTV classification of known bacteria-infecting +ssRNA viruses (Figure 3). Establishing a nomenclature for the 428 proposed genera was conducted as follows: A bacterial +ssRNA virus representing the genus was chosen if (1) it was a previously described bacterial +ssRNA virus available in the ICTV archives, (2) its sequence had been deposited in GenBank, (3) or its contig was the longest of all remaining available sequences. The full list of genera included in each family can be found ("International Committee on Taxonomy of Viruses" 2021).

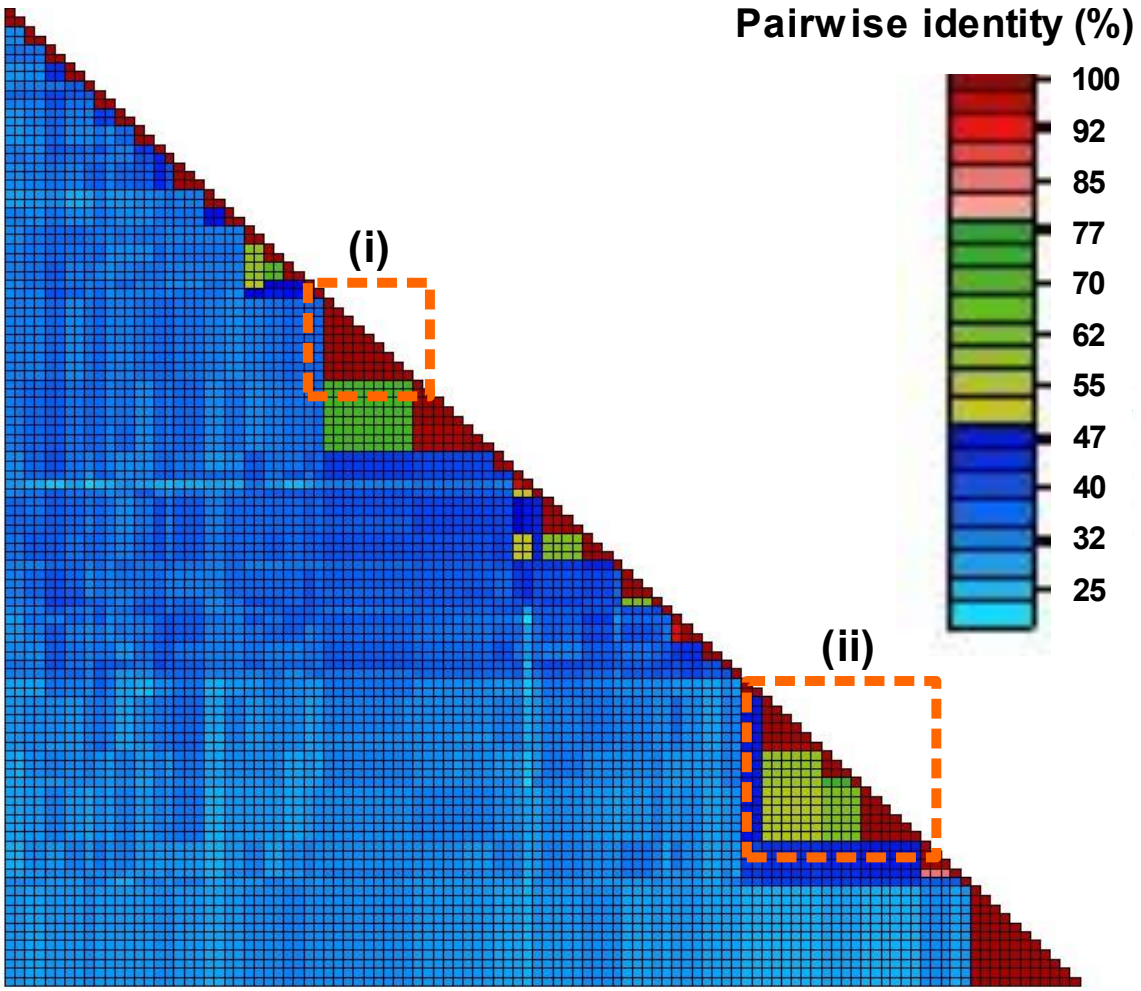


Figure 3. Examples of *Leviviricetes* genus and species demarcation cutoffs of 50% and 80%, respectively, applied to pairwise RNA-directed RNA polymerase (RdRP) amino-acid sequence comparisons for members of norziviral *Atkinsviridae*. Inset **(i)** shows a distinct species clustering (red coloring), whereas inset **(ii)** shows three species represented by multiple sequences, and a species representing a single sequence, clustered into a genus (yellow-green coloring). Pairwise comparisons in shades of blue do not meet the set genus or species clustering criteria.

Species. An 80% PAAI of the RdRP was chosen as the species demarcation criterion (Figure 3). This cut-off yielded 882 species, with all sequences assigned to specific species included in genera. Species were named following a Latinized binomial species name format in compliance with the latest International Code of Virus Classification and Nomenclature (ICVCN) iteration (Walker *et al*. 2021). The full list of species included in each genus can be found at ICTV (ictvonline.org) ("International Committee on Taxonomy of Viruses" 2021). For example, phage MS2 is now assigned to the species *Emesvirus zinderi* and phage BZ13 is now assigned to *Emesvirus japonicum*, whereas phage Qbeta is assigned to *Qubevirus durum* and FI 4184 b is assigned to *Qubevirus faecium*. The new naming scheme no longer necessitates knowledge of host bacteria and is therefore well-suited to the incorporation of sequence-only or uncultured virus genomes.

4.6 Discussion

The massive expansion in the discovery of novel bacteria-infecting +ssRNA virus genomes is now complemented with a timely update to their associated taxonomy. Fitting with phage MS2 being the first organism to have its genome completely sequenced, the presented ICTV-approved *Leviviricetes* taxonomy detailed here is the first to systematically include metagenomic sequences to build a class-rank taxonomy incorporating automated approaches. This approach demonstrates how incorporation of metagenomic sequences within ICTV's framework in future taxonomic proposals and subsequent expansion of established virus taxonomic groups—thus advancing a holistic understanding of viral diversity. At present, the expansion and restructuring of *Leviviricetes* has been the largest-ever proposal submitted to and approved by the Bacterial and Archaeal Virus Subcommittee of ICTV. However, as the incorporation of metagenome-assembled genomes into ICTV taxonomic proposals become

more frequent, due to the immense unexplored diversity of the virosphere, this record may be short-lived.

4.7 References

Atkins, John F., Joan A. Steitz, Carl W. Anderson, and Peter Model. 1979. "Binding of Mammalian Ribosomes to MS2 Phage RNA Reveals an Overlapping Gene Encoding a Lysis Function." *Cell* 18 (2): 247–56. https://doi.org/10.1016/0092-8674(79)90044-8.

August, J. Thomas, Stephen Cooper, Lucille Shapiro, and Norton D. Zinder. 1963. "RNA Phage Induced RNA Polymerase." *Cold Spring Harbor Symposia on Quantitative Biology* 28 (January): 95–97. https://doi.org/10.1101/SQB.1963.028.01.019.

Blumenthal, T, and G G Carmichael. 1979. "RNA Replication: Function and Structure of QBeta-Replicase." *Annual Review of Biochemistry* 48 (1): 525–48. https://doi.org/10.1146/annurev.bi.48.070179.002521.

Callanan, Julie, S. Stockdale, Evelien Adriaenssens, Jens H. Kuhn, M. Pallen, Janis Rumnieks, Andrey N. Shkoporov, Lorraine A. Draper, R. Ross, and Colin Hill. 2021. Rename One Class (*Leviviricetes* - Formerly *Allassoviricetes*), Rename One Order (*Norzivirales* - Formerly *Levivirales*), Create One New Order (*Timlovirales*), and Expand the Class to a Total of Six Families, 420 Genera and 883 Species. https://doi.org/10.13140/RG.2.2.25363.40481.

Callanan, Julie, Stephen R. Stockdale, Andrey N. Shkoporov, Lorraine A. Draper, R. Paul Ross, and Colin Hill. 2021. "Biases in Viral Metagenomics-Based Detection, Cataloguing and Quantification of Bacteriophage Genomes in Human Faeces, a Review." *Microorganisms* 9 (3): 524. https://doi.org/10.3390/microorganisms9030524.

Callanan, Julie, Stephen R. Stockdale, Andrey N. Shkoporov, Lorraine A. Draper, R. Paul Ross, and Colin Hill. 2020. “Expansion of Known ssRNA Phage Genomes: From Tens to over a Thousand.” *Science Advances* 6 (6): eaay5981. https://doi.org/10.1126/sciadv.aay5981.

Duin, Jan van. 1988. “Single-Stranded RNA Bacteriophages.” In *The Bacteriophages*, edited by Richard Calendar, 117–67. The Viruses. Boston, MA: Springer US. https://doi.org/10.1007/978-1-4684-5424-6_4.

Fenner, F. 1976. *Classification and Nomenclature of Viruses | Karger Book*. https://www.karger.com/Book/Home/219262.

Fiers, W., R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaert, W. Min Jou, *et al*. 1976. “Complete Nucleotide Sequence of Bacteriophage MS2 RNA: Primary and Secondary Structure of the Replicase Gene.” *Nature* 260 (5551): 500–507. https://doi.org/10.1038/260500a0.

“International Committee on Taxonomy of Viruses.” 2021. *Virus Taxonomy*. https://talk.ictvonline.org/taxonomy/.

Kastelein, R. A., E. Remaut, W. Fiers, and J. van Duin. 1982. “Lysis Gene Expression of RNA Phage MS2 Depends on a Frameshift during Translation of the Overlapping Coat Protein Gene.” *Nature* 295 (5844): 35–41. https://doi.org/10.1038/295035a0.

Koonin, Eugene V., Valerian V. Dolja, Mart Krupovic, Arvind Varsani, Yuri I. Wolf, Natalya Yutin, F. Murilo Zerbini, and Jens H. Kuhn. 2020. “Global Organization and Proposed Megataxonomy of the Virus World.” *Microbiology and Molecular Biology Reviews: MMBR* 84 (2). https://doi.org/10.1128/MMBR.00061-19.

Krishnamurthy, Siddharth R., Andrew B. Janowski, Guoyan Zhao, Dan Barouch, and David Wang. 2016. “Hyperexpansion of RNA Bacteriophage Diversity.” *PLOS Biology* 14 (3): e1002409. https://doi.org/10.1371/journal.pbio.1002409.

Krupovic, Mart, Jens H. Kuhn, Fengbin Wang, Diana P. Baquero, Valerian V. Dolja, Edward H. Egelman, David Prangishvili, and Eugene V. Koonin. 2021. "*Adnaviria*: A New Realm for Archaeal Filamentous Viruses with Linear A-Form Double-Stranded DNA Genomes." *Journal of Virology* 95 (15): e0067321. https://doi.org/10.1128/JVI.00673-21.

Li, Li, Christian J. Stoeckert, and David S. Roos. 2003. "OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes." *Genome Research* 13 (9): 2178–89. https://doi.org/10.1101/gr.1224503.

Loeb, Tim, and Norton D. Zinder. 1961. "A Bacteriophage Containing RNA." *Proceedings of the National Academy of Sciences of the United States of America* 47 (3): 282–89.

Olsthoorn, René C. L., and J. van Duin. 2009. "*Leviviridae* - Positive Sense RNA Viruses - Positive Sense RNA Viruses (2011) - ICTV." 2009. https://talk.ictvonline.org/ictv-reports/ictv_9th_report/positive-sense-rna-viruses-2011/w/posrna_viruses/263/leviviridae.

Roux, Simon, Evelien M. Adriaenssens, Bas E. Dutilh, Eugene V. Koonin, Andrew M. Kropinski, Mart Krupovic, Jens H. Kuhn, *et al*. 2019. "Minimum Information about an Uncultivated Virus Genome (MIUViG)." *Nature Biotechnology* 37 (1): 29–37. https://doi.org/10.1038/nbt.4306.

Shi, Mang, Xian-Dan Lin, Jun-Hua Tian, Liang-Jun Chen, Xiao Chen, Ci-Xiu Li, Xin-Cheng Qin, *et al*. 2016. "Redefining the Invertebrate RNA Virosphere." *Nature* 540 (7634): 539–43. https://doi.org/10.1038/nature20167.

Simmonds, Peter, Mike J. Adams, Mária Benkő, Mya Breitbart, J. Rodney Brister, Eric B. Carstens, Andrew J. Davison, *et al*. 2017. "Virus Taxonomy in the Age of Metagenomics." *Nature Reviews Microbiology* 15 (3): 161–68. https://doi.org/10.1038/nrmicro.2016.177.

Spiegelman, S, I Haruna, I B Holland, G Beaudreau, and D Mills. 1965. “The Synthesis of a Self-Propagating and Infectious Nucleic Acid with a Purified Enzyme.” *Proceedings of the National Academy of Sciences of the United States of America* 54 (3): 919–27.

Starr, Evan P., Erin E. Nuccio, Jennifer Pett-Ridge, Jillian F. Banfield, and Mary K. Firestone. 2019. “Metatranscriptomic Reconstruction Reveals RNA Viruses with the Potential to Shape Carbon Cycling in Soil.” *Proceedings of the National Academy of Sciences* 116 (51): 25900–908. https://doi.org/10.1073/pnas.1908291116.

Steitz, Joan Argetsinger. 1969. “Polypeptide Chain Initiation: Nucleotide Sequences of the Three Ribosomal Binding Sites in Bacteriophage R17 RNA.” *Nature* 224 (5223): 957–64. https://doi.org/10.1038/224957a0.

“Taxonomy.” 2021. 2021. https://talk.ictvonline.org/taxonomy/.

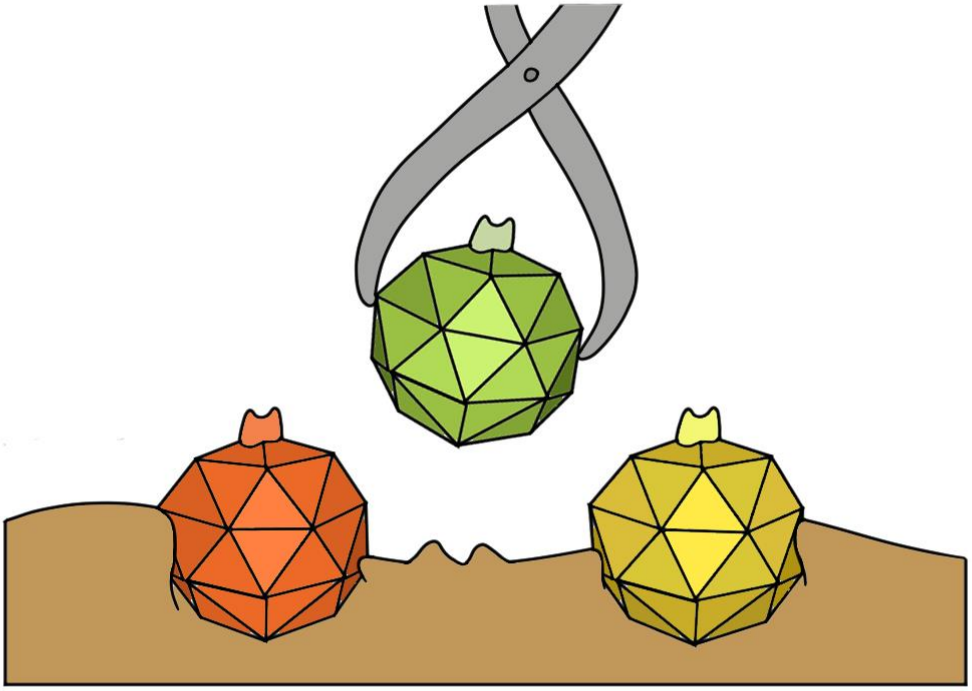
Walker, Peter J., Stuart G. Siddell, Elliot J. Lefkowitz, Arcady R. Mushegian, Evelien M. Adriaenssens, Poliane Alfenas-Zerbini, Andrew J. Davison, *et al*. 2021. “Changes to Virus Taxonomy and to the International Code of Virus Classification and Nomenclature Ratified by the International Committee on Taxonomy of Viruses (2021).” *Archives of Virology*, July. https://doi.org/10.1007/s00705-021-05156-1.

Walker, Peter J., Stuart G. Siddell, Elliot J. Lefkowitz, Arcady R. Mushegian, Evelien M. Adriaenssens, Donald M. Dempsey, Bas E. Dutilh, *et al*. 2020. “Changes to Virus Taxonomy and the Statutes Ratified by the International Committee on Taxonomy of Viruses (2020).” *Archives of Virology* 165 (11): 2737–48. https://doi.org/10.1007/s00705-020-04752-x.

Wildy, P. 1971. *Classification and Nomenclature of Viruses | Karger Book*. https://www.karger.com/Book/Home/218074.

Wolf, Yuri I., Darius Kazlauskas, Jaime Iranzo, Adriana Lucía-Sanz, Jens H. Kuhn, Mart Krupovic, Valerian V. Dolja, and Eugene V. Koonin. 2018. “Origins and Evolution of the Global RNA Virome.” *MBio* 9 (6). https://doi.org/10.1128/mBio.02329-18.

Zhang, Tao, Mya Breitbart, Wah Heng Lee, Jin-Quan Run, Chia Lin Wei, Shirlena Wee Ling Soh, Martin L. Hibberd, Edison T. Liu, Forest Rohwer, and Yijun Ruan. 2005. “RNA Viral Community in Human Feces: Prevalence of Plant Pathogenic Viruses.” *PLOS Biology* 4 (1): e3. https://doi.org/10.1371/journal.pbio.0040003.



Chapter V

Examination of Enrichment and Nucleic Acid Extraction Methods for RNA Bacteriophage Detection, Isolation, and Characterisation

For this chapter, I devised the study, performed the analyses, produced the images, and wrote the manuscript. Dr Stephen R. Stockdale assisted in the generation of some figures.

5.1 Abstract

Interest in the human gut microbiome has gained attention in recent years, leading to a greater understanding of its importance. Recent analyses have revealed the significance of the complex role bacteria-infecting viruses, known as bacteriophages, play in the human gut. However, the bulk of the detectable bacteriophages (phages) within the human microbiome possess a DNA genome. Throughout various databases and the research literature there is a lack of information regarding RNA phages isolated from the human gut. A major factor that could influence this disparity is the extraction method used in a study that can introduce major biases impacting the recovery of these entities. Here it is shown that the application of three different extraction protocols to a common faecal sample affects the recovery of RNA phages, both naturally occurring and spike-in controls. The extraction method also impacts on overall intra- and inter-sample diversity. A novel Biopsy Method developed in a recent in-house study was shown to be effective at recovering spiked-in RNA phages, as well as capturing sample complexity. Based on these findings, applying the Biopsy Method to future phageome research would assist in unveiling the true abundance and functionality of RNA phages within the human gut.

5.2 Introduction

Over the past couple of decades, there have been significant advances in describing the microbial communities of the gastrointestinal tract (GIT), unravelling the complex dynamics, and understanding its associated physiological importance. The bacterial aspect of the microbiome has overshadowed the study of other members due to their vast numbers and relative ease of analysis. Recently, interest in the viral component has gained traction due to developments in understanding the fundamental role this fraction plays with regards human health. This includes viruses that infect eukaryotes and bacterial viruses termed bacteriophages (phages).

The total community of phages within an ecosystem is referred to as the phageome. Although they are likely to play a significant role in bacterial diversity and the dynamics of an ecosystem, studies of these entities trail behind their microbial counterparts (Townsend *et al*. 2021). It is typically in marine and terrestrial environments that the roles of the phageome have been examined but it is important that studies regarding the analysis of the natural phageome of the human gut continue and expand to define their role in human health. It is crucial that phages are not overlooked in microbiome studies as they are the most abundant viral component of the human microbiome (Minot *et al*. 2011; Reyes *et al*. 2010) and have been identified as key players in bacterial composition and structure (Guerin and Hill 2020; Shkoporov and Hill 2019).

Of the described portion of the phageome, the majority of identified members have a DNA genome. The RNA viruses, especially RNA phages, have remained somewhat elusive in virome and phageome studies which has been attributed to a variety of factors, in particular biases associated with the extraction method used (Callanan *et al*. 2021). To accurately compare studies between different research teams, efforts to create a standardised protocol to study the faecal phageome have been made. Conceição-Neto *et al.* (2015) described the 'Novel enrichment technique of VIRomes' (NetoVIR) protocol which used mock bacterial and viral communities, including double-stranded RNA (dsRNA) and positive-sense, single-stranded RNA (+ssRNA) viruses (not phages) (Conceição-Neto *et al*. 2015). This type of study suggests there are potential ways to optimise the recovery of RNA viruses, including their phage equivalents, and how future studies can attempt to capture them in their viral recovery.

There are two main techniques used in isolating and characterising the phageome: culture-dependent and culture-independent approaches. This study focuses on alternative culture-independent methods for isolating the virus-like particle (VLP) fraction following the resuspension of the faecal sample in a suitable medium. This fraction is usually separated from

the complex environment of free nucleic acids and other microorganisms through a variety of methods. These methods include concentration, filtration, or fraction-based approaches which can often be used in combination (Castro-Mejía *et al*. 2015; Thurber 2009). Many of these VLP-enrichment steps, prior to nucleic acid extraction, are known to introduce biases and can skew the results (Callanan *et al*. 2021; Townsend *et al*. 2021), such as the inclusion of chloroform in VLP isolation which can degrade the phospholipid membrane of dsRNA phages, *Cystoviridae*. These biases may contribute to the absence of lowly abundant phages, such as RNA phages, or prevent us from understanding their true prevalence in the human gut. These entities are rarely recovered in human gut phageome studies with the predominant portion of RNA viruses identified as being plant-viruses contributed from the diet (Lim *et al*. 2015; Zhang *et al*. 2005).

It has been noted that RNA phages are highly underrepresented in the available viral databases due to number of factors, especially the issues associated with their isolation using typical culture-dependent methods (Callanan *et al*. 2020; Krishnamurthy *et al*. 2016). A novel sequence-based approach has been recently implemented to help recover +ssRNA phages (Callanan *et al*. 2020). This method utilises a profile-HMM search tool, built using the viral core proteins of known +ssRNA phages, to identify homologous proteins and potential genomes. This tool was utilised in the analysis of the extraction methods comparison and examination of the +ssRNA phages from the mammalian gut.

Initially three distinct extraction methods associated with human gut phageome studies were examined and their efficiency in recovering both spiked-in and naturally occurring +ssRNA phages from the same starting material was compared. These three methods include (i) a Polyethylene Glycol (PEG) Method, (ii) a Filtration Method, and (iii) a Biopsy Method. The PEG method was initially developed and published for the analysis of infant faecal viromes (McCann *et al*. 2018). It was designed to avoid the use of caesium chloride (CsCl) gradient

ultracentrifugation due to its technical difficulties and time laborious nature, and instead relies on concentrating viral particles using PEG precipitation. It has been suggested that this may introduce a bias in the recovery of RNA phages as these small, low abundant viral particles may not precipitate using PEG or the use of chloroform may destroy their virion (Callanan *et al*. 2021). The Filtration Method was designed to avoid introducing any chemicals or reagents that may damage the virion of RNA phages by exploiting ultrafiltration as a concentration method. The Biopsy Method stemmed from a recent study which examined the viral biogeography of the GIT and parenchymal organs of mammals (Shkoporov *et al*. 2021). This method was described as simple as it avoids the use of micro-filtration, VLP precipitation using PEG/sodium chloride (NaCl), chloroform, or CsCl gradient ultracentrifugation. This method was also included as there were in-house indications that it recovered RNA viruses more efficiently than the PEG Method. Unlike previous studies in our group the application of the Biopsy Method to examine the virome of mammals, including macaques and pigs, recorded novel +ssRNA phages. Subsequently, these contigs and their associated proteins were analysed.

5.3 Materials and Methods

5.3.1 Sample collection and storage

Faecal samples were collected under the study protocol APC055 which was approved by the Cork Research Ethics Committee (CREC). In short, the samples had been collected by the consenting volunteers and stored at -80°C upon arrival to the research facility until ready to be processed. The same faecal sample was used in all three method comparisons.

5.3.2 Propagation of +ssRNA phages

Two +ssRNA phages, MS2 and Qbeta, and their host, *Escherichia coli* (Migula 1895) Castellani and Chalmers 1919, were sourced from DSMZ. Both MS2 and Qbeta were propagated with agitation at 37°C in Luria Bertani broth (LB) on their host strain of *E. coli*. These phage lysates were centrifuged at for, 4,000 x g, filtered using a 0.45µm filter and stored at 4°C.

5.3.3 Extraction methods

Three differing extraction methods were tested for the recovery of +ssRNA phages. A summarized version of these methods is displayed in Figure 1.

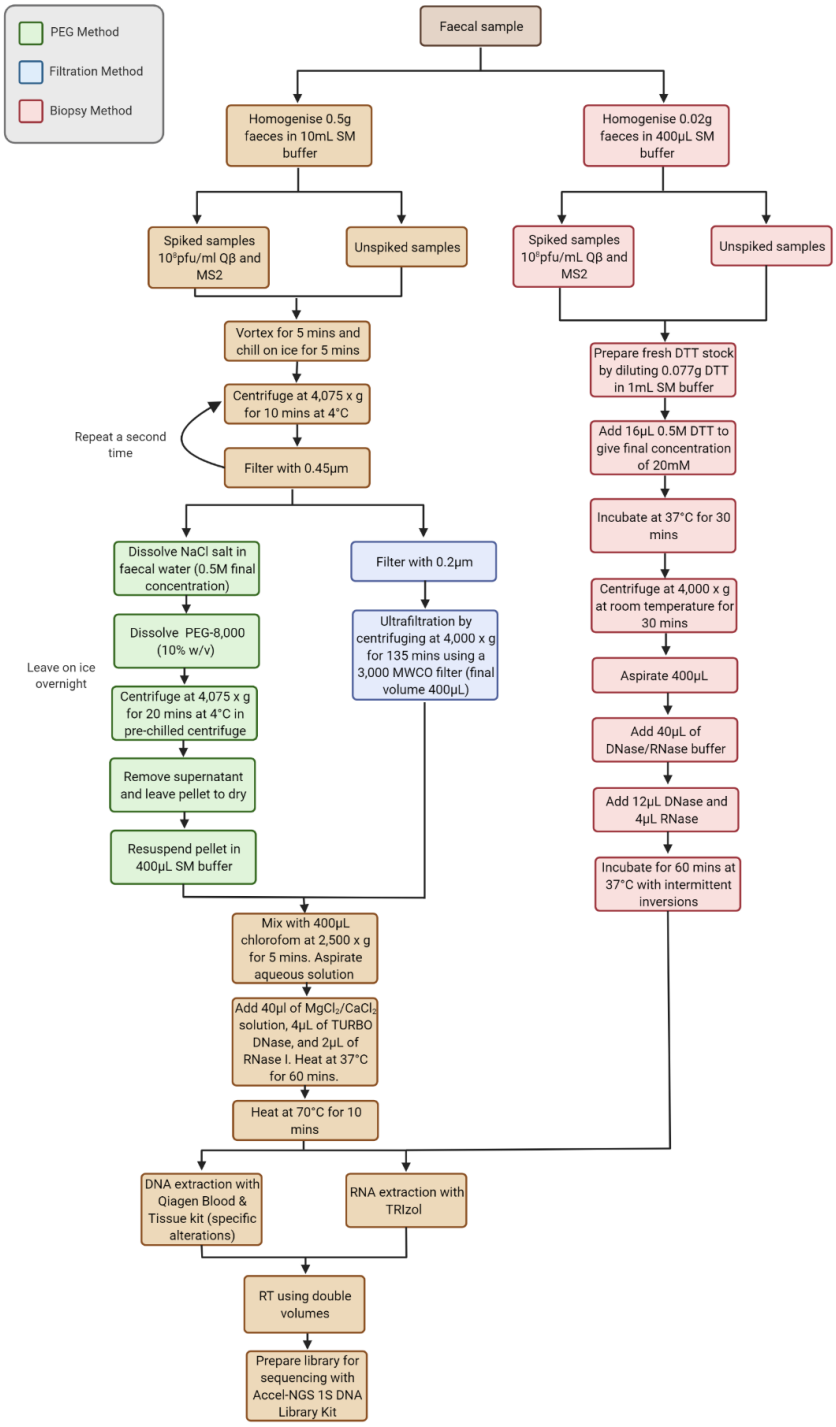


Figure 1. Overview of the different extraction methods used in this study. A common faecal sample was used in all three methods, with both a spiked (a final titre of 10^8 pfu/ml of MS2 and Qbeta) and unspiked fraction, to enable direct comparison across the results.

5.3.3a PEG Method

This method was based on in-house phageome analysis protocol (Shkoporov *et al*. 2019; 2018; McCann *et al*. 2018). Briefly, aliquots of 0.5g of faeces were resuspended in 10ml of SM buffer (50 mM Tris-HCl; 100 mM NaCl; 8.5 mM $MgSO_4$; pH 7.5). For the phage-spiked samples both MS2 and Qbeta were added (final titre of 10^8 pfu/ml) while control samples remained unspiked. The samples were homogenised by vigorous vortexing for 5 mins. The samples were then chilled on ice for 5 min before being centrifuged twice at 4,075 x g (swing bucket centrifuge) for 10 mins at 4°C to remove large particles and bacterial cells. Supernatants were subsequently filtered twice using a 0.45µm pore filter. NaCl salt (0.5 M final concentration) and 10 % (w/v) PEG-8000 was dissolved in the faecal water. This was then stored on ice overnight (approx. 16 hours). Samples were spun at 4,075 x g for 20 min at 4˚C in a pre-chilled centrifuge. The supernatant was discarded, and the tubes inverted for 5 mins to remove residual supernatant. Pellets were resuspended in 400µl SM buffer and subsequently retrieved following gentle shaking with 400µl of chloroform. The mixtures were centrifuged using a desktop centrifuge at 2,500 x g for 5 min, and the aqueous phase was aspirated into a new Eppendorf tube. This emulsion was combined with 40µl of a solution of 50 mM $MgCl_2$ and 10 mM $CaCl_2$. Free nucleic acids were removed by the addition of 40µl of DNase/RNase buffer, 12µl of DNase, and 4µl of RNase. The sample was incubated at 37˚C for 60 mins with intermittent inverting (approx. every 15 mins).

5.3.3b Filtration Method

This method was the same as the PEG-based Method, with slight modifications after the second filtration step. After the second filter with 0.45µm, the supernatant was filtered through a 0.20µm pore diameter filter. The supernatant was transferred to a 3,000 MWCO ultrafiltration tube and centrifuged at 4,000 x g for 135 mins (concentrated to 400µl). Following this, the VLPs were purified as per PEG Method using the chloroform and nucleases.

5.3.3c Biopsy Method

This method was recently described by Shkoporov and colleagues and was included as it provides a more time- and cost-effective means to analysis multiple samples (Shkoporov *et al*. 2021). In short, 0.02g of faeces was resuspended in 400µl SM buffer, again with one spiked (10^8 pfu/ml of MS2 and Qbeta) and one unspiked sample. A fresh stock of 0.5M DTT was prepared by adding 0.077g in 1ml of SM buffer and 16µl of the 0.5M DTT to the faecal samples (final concentration of 20mM). The samples were incubated at 37˚C for 30 mins. Samples were centrifuged for 30 mins at room temperature at 4,000 x g and 400µl was aspirated into a clean Eppendorf. To this portion 40µl of DNase/RNase buffer was added, followed by 12µl of DNase, and 4µl of RNase and incubate for 60 mins at 37˚C with occasional inverting (approx. every 15 mins).

5.3.4 Nucleic acid extraction

For each spiked and unspiked sample of the three different methods; both DNA and RNA fractions were extracted. For the DNA extraction, a 100µl portion of each final sample was processed with the Qiagen Blood and Tissue Kit. To this portion 180 µl buffer ATL and 20 µl Proteinase K (20 mg/ml) were added, vortexed to homogenise, and incubated at 56ºC (vortex occasionally during incubation) for 12 mins. Subsequently, 200 µl buffer AL was added, vortexed, and incubated at 56ºC for 10 mins. A 200 µl volume of absolute ethanol was combined and vortexed. The mixture was transferred to a DNA easy-spin column and centrifuged at 6,900 x g for 1 min and the flow-through discarded. To the remaining liquid 500 µl AW1 was added, centrifuged for 1 min at 6,900 x g and again the flow-through was discarded. Following this 500 µl AW2 was added, centrifuged for 3 min at 12,000 x g (max speed), and the flow-through discarded. The spin column was transferred to a 1.5 ml Eppendorf, 20 µl of buffer AE was added, and incubated for 1 min at room temperature before centrifuging at 6,100 x g for 1 min at room temperature to elute DNA. The previous step was

repeated with 10 µl of buffer AE to increase the DNA yield, which was recorded using a Qubit machine.

To extract the RNA, another 100µl of each sample was added to 1ml TRIzol and the protocol was followed accordingly. Briefly, after 5 mins incubation, the samples were lysed to separate the phases by adding 0.2µL of chloroform and incubating for 3 mins at room temperature (approximately 25°C), followed by centrifugation for 15 mins at 12,000 x g at 4˚C. The colourless, upper aqueous phase containing the RNA was carefully transferred into a new eppendorf. To precipitate the RNA, 0.5mL of isopropanol was added to the sample, incubated for 10 mins at room temperature, and centrifuged for 10 mins at 12,000 x g at 4 ˚C. At the bottom of the tube, the total RNA precipitate formed a white gel-like pellet while the supernatant was discarded. To wash the RNA, the pellet was resuspended the pellet in 1mL of 75% ethanol. The sample was vortexed briefly, centrifuged for 5 mins at 7,500 x g at 4˚C, and the supernatant was disposed of. The RNA pellet was air-dried for 10 mins and then resuspended in 40µL of RNase-free water and incubated on a heat block at 60˚C for 12 mins. The RNA yield was recorded using a Qubit machine.

5.3.5 Sequencing of VLP nucleic acids

Of the resulting faecal VLP nucleic acids samples, 16µL was subjected to reverse transcription (RT) regardless of the yield. SuperScript IV was used for first-strand cDNA synthesis, as pre the manufacturer's protocol, with double amounts of reactants and sample used to retain high concentrations. The Accel-NGS 1S Plus DNA Library Kit (Swift Biosciences) was used as it is suitable for low yields of DNA/cDNA and does not require an amplification step. Libraries were sequenced using Illumina Novaseq platform.

5.3.6 Bioinformatic analysis

The quality of the raw reads was assessed with FastQC (version 0.11.8; (Andrews 2010)) and subsequently processed using Cutadapt (version 2.4; (Martin 2011)) to remove residual

Illumina adapter sequences. To improve the overall quality of the reads, sequences with a Phred score less than 30 for a 4-bp sliding window were filtered out using Trimmomatic (version 0.36;(Bolger, Lohse, and Usadel 2014)). The resulting reads from all samples were assembled using rnaSPAdes (SPAdes version 3.11.1 (metaSPAdes mode);(Bushmanova *et al*. 2019)) and metaSPAdes (SPAdes version 3.11.1 (metaSPAdes mode);(Nurk *et al*. 2017)) and subsequently filtered to include those of a minimum length of 1kb.

The proteins associated with these reads were predicted using Prodigal with the "-p meta" choice to allow for small contigs, as well as the "-n" choice to ensure a full motif scan of each nucleotide sequence (version 2.6.3;(Hyatt *et al*. 2010)). These proteins were then scanned using the HMM 5-MC search tool, previously described in Chapter 3, with hmmscan scores of 30 or greater being investigated further (Callanan *et al*. 2020).

A viral database from the metagenomic assembled sequences was built through a series of positive and negative selection criteria. Positive selection criteria included: contigs determined as circularly permuted, contigs identified as viral through VIRSorter (Roux *et al*. 2015), or contigs returning a significant blast hit (E-value 1E-10, and a query-subject alignment length of 500bp) against the Viral RefSeq v203 database, the Gut Virome Database (GVD) v1.0, the Integrated Microbial Genomes (IMG) VR v3.0 database, or an in-house crAss-like phage database. Viral contigs were additionally identified through a database-independent method, whereby contigs were considered viral if they encoded a minimum of three prokaryotic viral orthologous group (pVOG) proteins or averaged a minimum of three pVOG proteins per 3kb. Finally, contigs from viral-enriched metagenomic assemblies with no known corresponding representative in the NCBI NT database were included in the viral database as viral dark matter.

Negative selection criteria were subsequently employed to remove potential bacterial contaminants from the compiled viral database. Contigs encoding a bacterial ribosomal protein

or a plasmid replication protein, were removed from the viral database, unless they were annotated as viral by Viral RefSeq.

Processed sequencing reads were mapped to both reference viral genomes and metagenomically-identified viral sequences using Bowtie2 in end-to-end mode (Langmead and Salzberg 2012). MS2 and Qbeta reference sequences were downloaded from NCBI using accessions NC_001417.2 and AB971354.1, respectively. The abundance of reads mapping per contig and the breadth of coverage were calculated using SAMTools and BEDTools, respectively (Quinlan and Hall 2010; Li *et al*. 2009). A minimum of 5 reads mapping to a contig, and reads covering 50% of a contig's length, were required to record the number of reads aligning to a contig. Failure to meet these criteria resulted in contigs not considered present, but potentially reads mapping to a conserved motif.

Analyses were performed as required using bash through a Linux terminal and in the R programming language (version 4.1.1) implemented through RStudio (RStudio Team 2020). Tabular and textual outputs from Linux commands were processed in R using the following functions and packages. Data was made into R readable formats using 'gsub' and 'stringr' (Wickham and RStudio 2019; Zane 2017). Count data was transformed as necessary using 'reshape2' (Wickham 2020). The relative abundance of a contig within a sample was calculated using 'funrar' (Grenié *et al*. 2020). The alpha- and beta-diversities of samples were calculated using 'vegan' and 'phyloseq' (Oksanen *et al*. 2020; McMurdie and Holmes 2013). The differential abundance of contigs between two conditions was calculated using 'DESeq2' (Love, Huber, and Anders 2014) and plotted using 'EnhancedVolcano'(Blighe *et al*. 2021). The Venn diagram was made using the 'VennDiagram' package (Chen 2018). All other images were generated using 'ggplot2' with the 'ggpubr' extension for publication quality images (Wickham *et al*. 2021; Kassambara 2020).

Statistical tests conducted were as follows. Data normality was assessed through the Shapiro-Wilks test. Non-parametric two-group mean comparisons were calculated using the Wilcoxon test. The means of three or more groups were compared using the Kruskal-Wallis test. Statistical significance between categorical variables were assessed using a Chi-squared test. False-positive discovery rates were corrected, where applicable, using the Bonferroni correction method.

5.4 Results

5.4.1 Overview of the three extraction methods

From the initial application of the three protocols, the Biopsy Method provided a more time- and cost-efficient means of examining multiple samples. This high-throughput method is less labour- and time-intensive than the PEG and Filtration Methods as it does not require multiple centrifugation, filtration, and incubation steps, instead there are just two incubation steps. One major advantage to using the Biopsy Method as opposed to the PEG Method, is that it does not entail leaving the samples overnight (16 hours) for concentration via PEG precipitation. This time was selected based on previous work and allows for increased probability of recovering all VLPs present (Shkoporov *et al.* 2018). Another crucial factor is that extractions are possible on smaller amounts of samples, which means less sample is required per extraction, and, if required, multiple replicates are possible. It also does not require as many reagents or materials to analyse samples which lends itself to being a more economical method.

The Biopsy Method successfully recovered +ssRNA phages in a recent study (Shkoporov *et al*. 2021), so given their doubling of reagents used in the RT stage, this was applied to all methods and fractions. This is a key difference between this work and previous studies. It suggests that the PEG Method in this work is an optimised version of previous studies as it gave a higher yield of cDNA from the resulting nucleic acid extracts. It also does not

include multiple displacement amplification (MDA), such as GenomiPhi (GE Healthcare), which has been found to preferentially amplify small circular +ssDNA genomes. This contributes to a loss in viral diversity, and introduces an artificial skewness in the virome composition (Callanan *et al*. 2021; Shkoporov *et al*. 2019; Roux *et al*. 2016).

The raw reads and assembled contigs (with a minimum length of 1kb) associated with each of the three methods were compared (Figure 2). A count of the number of raw reads from each sample confirmed that reads from the Biopsy Method were similar across each fraction compared to those of PEG and Filtration (Figure 2A). Given that the minimum length was set to 1kb, only the average, mean, and longest contigs from each of the samples were measured (Figure 2B - 2D). This showed that the unspiked RNA fractions of each of the methods had the lowest number of contigs that met this threshold. The spiked DNA Biopsy Method-extracted division had the largest number (16,641) of assembled contigs that met the 1kb cut-off. The Biopsy Method-isolated, unspiked DNA portion had the longest contig of 1,163kb which may be an indication of its ability to capture viruses of all sizes or a suggestion of potential contamination. The average across the DNA fractions per method represented lengths that are similar to the genome size of known +ssRNA phages (~4kb).

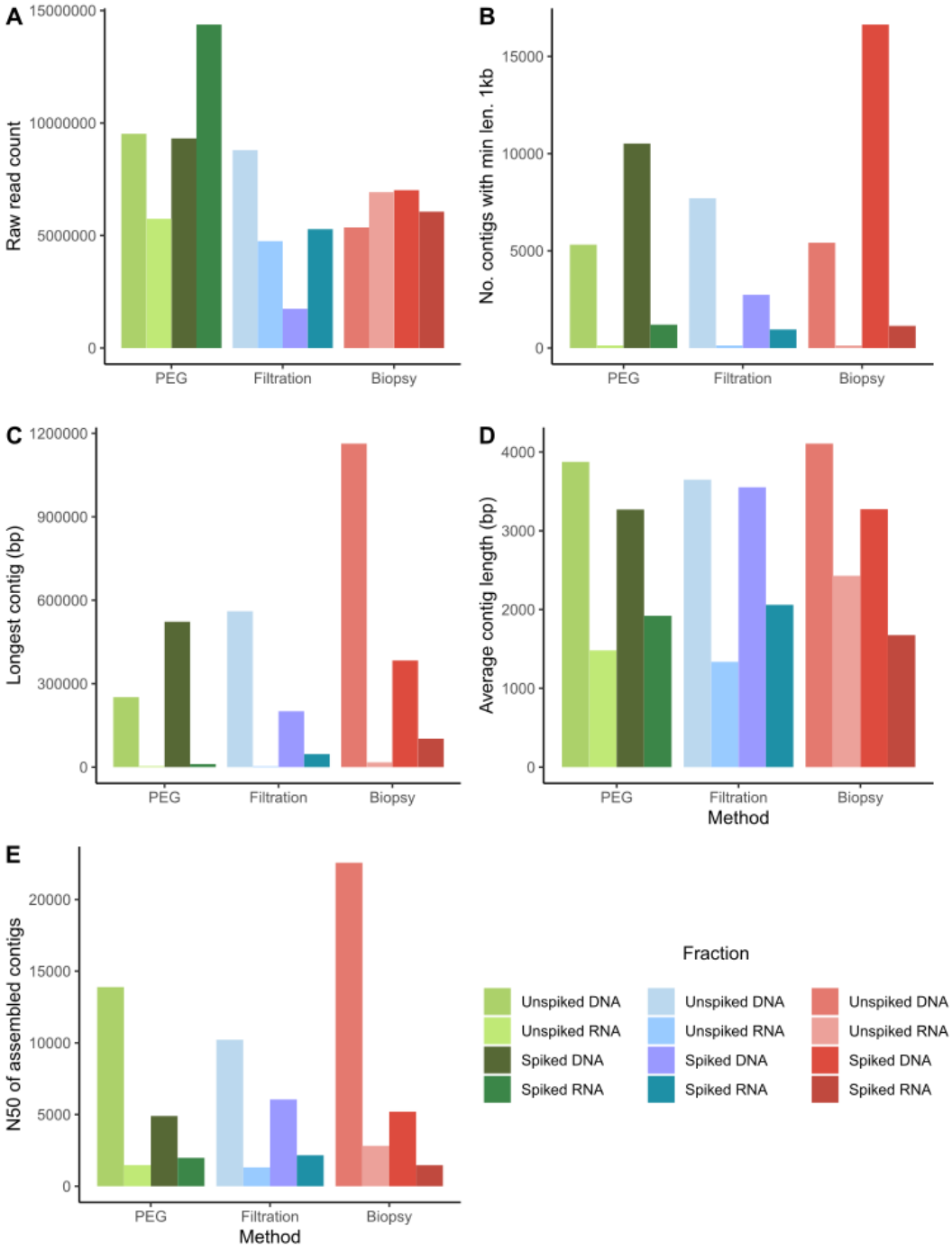


Figure 2. Analysis of raw reads and contigs per extraction method. The read-outs per extraction method of the **(A)** number of reads, **(B)** number of contigs with a minimum length of 1kb, **(C)** the longest contig per fraction of each method, **(D)** the average contig length associated with each method portion, and **(E)** N50 of reads from samples with minimum length of 1kb.

The N50 is a general assembly metric used across metagenomic analyses as it describes the shortest contig length such that 50% of the assembly is represented (Figure 2E) (Sutton *et al*. 2019; Baker 2012; Salzberg *et al*. 2012). A wide range in N50 values per method (rounded) was observed: the PEG Method ranged from 1.5kb to 13.9kb, the Filtration Method ranged from 1.3kb to 10.2kb, and the Biopsy Method ranged from 1.5kb to 22.5kb. Per extraction method, all RNA fractions have the smaller N50 values while the unspiked DNA portions had the highest values. While this assembly metric indicates that all methods resulted in slightly fragmented assemblies, it does suggest that each could capture the smaller contigs which could represent +ssRNA phages or even their associated proteins.

5.4.2 Examination of the recovery of +ssRNA phages per extraction method

The successful recovery of the spiked-in +ssRNA phages, MS2 and Qbeta, depended greatly on the combination of enrichment and nucleic acid extraction methods used (Figure 3). It was evident that linking the PEG Method with the DNA-focused extraction gave optimal recovery for the spiked-in controls, whereas the combination of the Biopsy Method with the RNA-focused extraction method was ideal for recovering these phages from unspiked controls (Figure 3A & 3B). In terms of diversity in the unspiked samples, it was revealed that the species richness from the Biopsy Method, in both the DNA and RNA fractions, was greater than that of the other two methods (Figure 3C). The unspiked Biopsy Method was also far more diverse in terms of its beta-diversity indicating that this method captures a different compositional profile of the same sample (Figure 3D). These findings suggest that the Biopsy Method allows for a more diverse representation of the sample than that observed with the PEG and Filtration Methods.

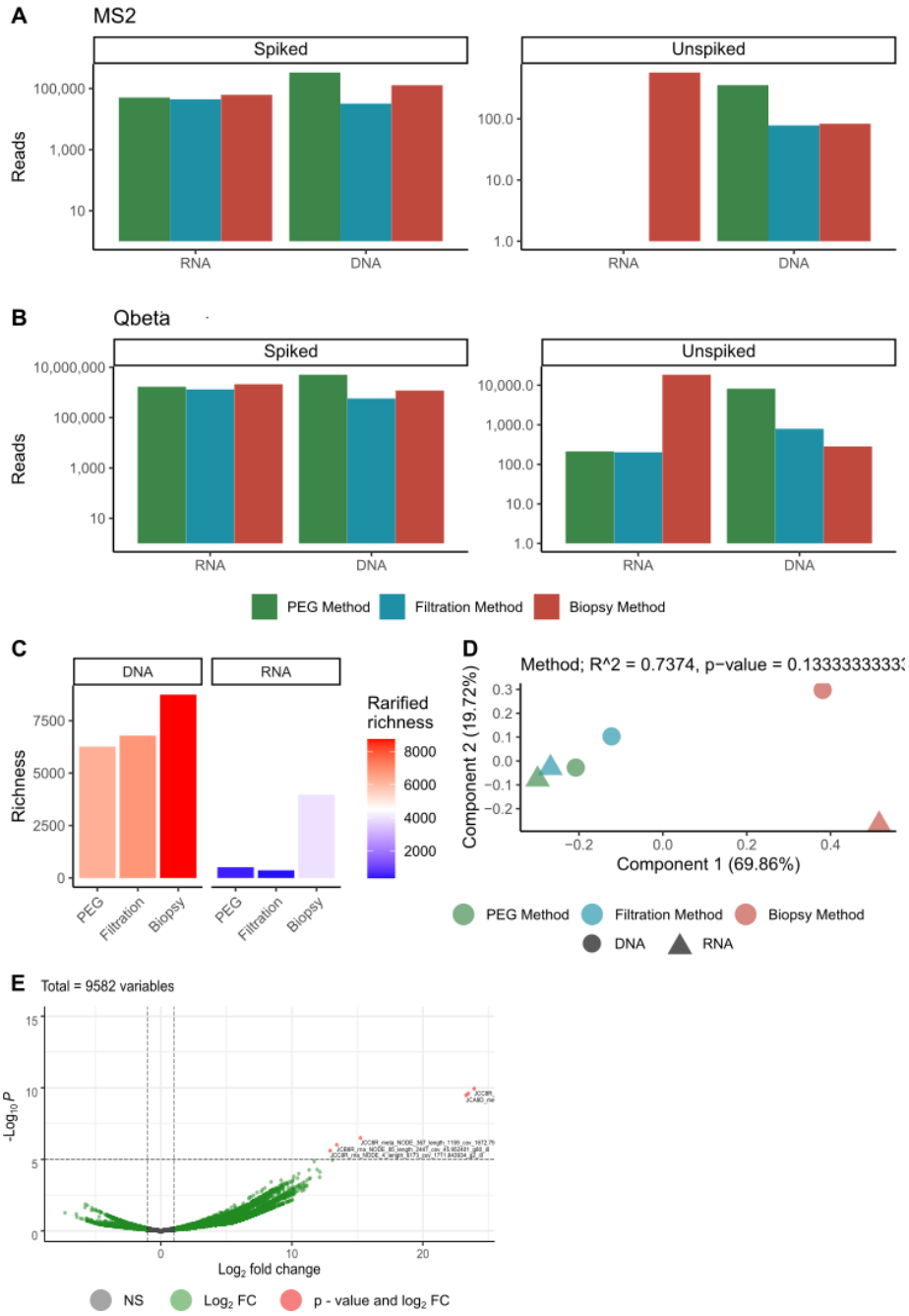


Figure 3. Comparison of method efficiency in recovering RNA phages and diversity associated parameters. (A) The analysis of MS2 and **(B)** Qbeta recovery from the different extraction samples, **(C)** richness of unspiked samples, **(D)** β-diversity of unspiked samples, and **(E)** differential abundance of the unspiked samples from the Biopsy Method compared to the PEG Method.

The PEG Method was primarily used in previous studies (Shkoporov *et al*. 2019; 2018; McCann *et al*. 2018), and despite the presence of spiked-in Qbeta, no +ssRNA phages were ever recovered. Recently the Biopsy Method was found to retrieve these entities from faecal and biopsy samples of mammals (Shkoporov *et al*. 2021) so it is fitting to compare these protocols by efficiency, richness, and diversity. Analysis of the contigs recovered in the unspiked fractions of the PEG and Biopsy Methods revealed that there are a lot of contigs differentially more abundant in the Biopsy Method-extracted portion which fits with the increased richness associated with this sample (Figure 3E).

To identify any naturally occurring +ssRNA phages, other than MS2 or Qbeta, a hmmscan was performed using the profile search tool described in Chapter III to detect any +ssRNA phage-associated proteins and related contigs (Figure 4) (Callanan *et al*. 2020). This search tool is based on the detection of the three core proteins encoded by these phages: the maturation protein (MP), the coat protein (CP), and the RNA-directed RNA polymerase (RdRP). Some of these phages are also known to encode a separate lysin protein, like MS2, however this sequence is very short and liable to a high degree of diversity so remains difficult to detect.

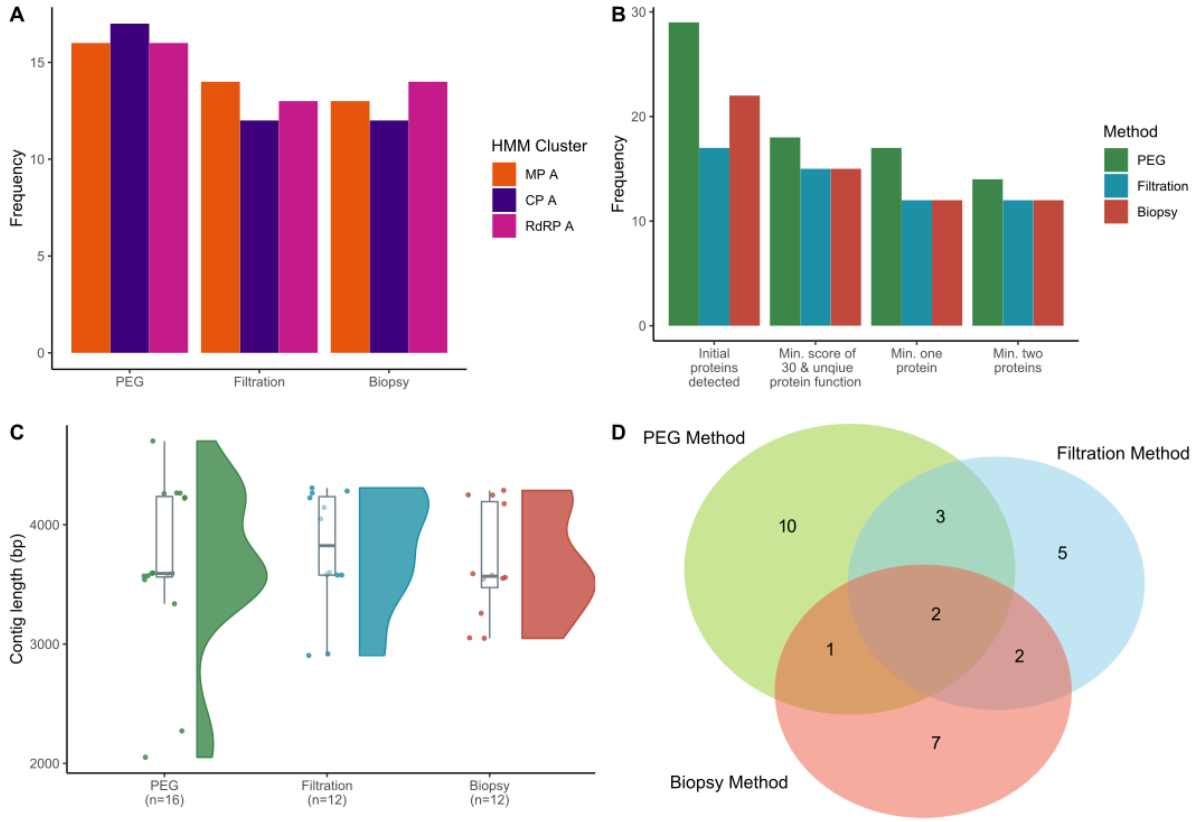


Figure 4. Identification of +ssRNA phages per method. (A) The proteins resulting from the hmmscan, **(B)** contigs per filtering step of the hmmscan, **(C)** lengths of non-redundant contigs (100% identity over 100% length) with multiple (≥ 2) proteins with a unique function and minimum score of 30, and **(D)** Venn diagram displaying shared (percentage identity of ≥ 99.5% over ≥ 100bp) non-redundant contigs across the three methods.

The recovery of proteins from the hmmscan also enabled their associated contigs to be identified. The number of contigs with various protein profiles revealed that it was the PEG Method that captured the most contigs (Figure 4B). Upon examination of the lengths of the identifiable, non-identical contigs with multiple hits (≥ 2) per method, it was apparent that the majority fell into ranges typical of +ssRNA phage genomes (Figure 4C). Subsequent analysis of the shared contigs per method revealed that there were two contigs shared across all three, as expected, which were MS2 and Qbeta (Figure 4D). Examination of the contigs' taxonomy revealed that they all belonged to *Fiersviridae* family of the *Norzivirales* order as each of their CPs and RdRPs were of the A clusters. Although the PEG Method was deemed to be the most successful in terms of recovering contigs detectable through the HMM-search tool, upon inspection, two of the PEG-derived contigs have very short genomes (~2kb) and therefore are unlikely to be true +ssRNA phages. The Biopsy Method provided comparable numbers of +ssRNA phage contigs, all of which have lengths which closely resemble that of known relatives. Future work could expand the HMM 5-MC used here to develop HMM 6 by incorporation of the proteins associated with these +ssRNA phages to recover more novel contigs, using the methodology described in Chapter III.

5.4.3 Application of Biopsy Method in Biogeography Study

As the non-redundant sequences associated with the Shkoporov *et al* (2021) study, which used the Biopsy Method, were available and indicated the presence of +ssRNA phages, a hmmscan was performed on these sequences (Figure 5). The initial scan revealed all hits to the MP, CP, and RdRP (Figure 5A). A threshold of having a minimum score of 30 was then set to assess good-quality hits and duplicated protein hits which encoded the same function were then removed. For example, if one protein hit both RdRP A and RdRP B, then the RdRP it had the higher score to would be retained (Figure 5B). This was the first time that +ssRNA phage proteins were recovered from an in-house experiment.

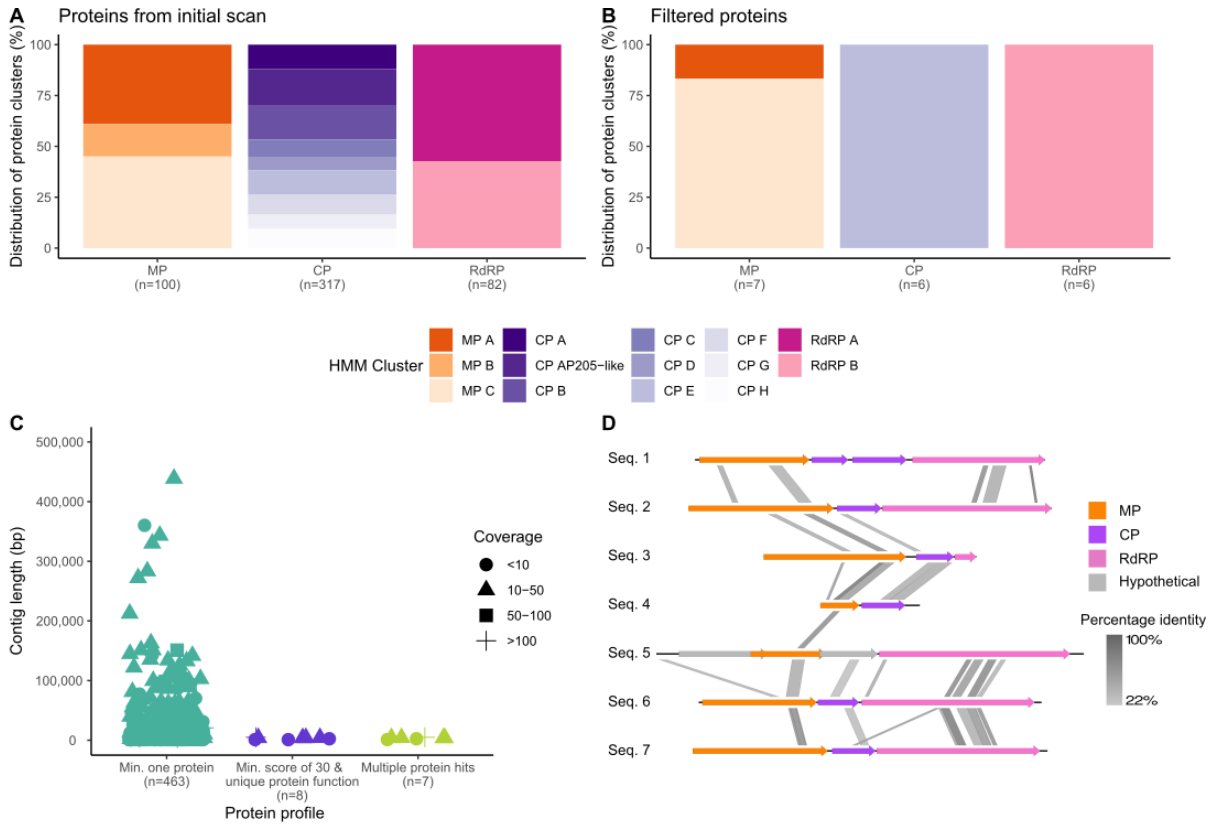


Figure 5. Examination of +ssRNA phages and associated proteins from the application of the Biopsy Method in a recent in-house biogeography study. Distribution of protein hits (in parentheses) across MP, CP, and RdRP clusters was observed using the HMM 5-MC search tool (incorporating taxonomy). **(A)** The initial scan includes all protein hits, and **(B)** proteins were filtered by setting a threshold of a minimum score of 30 and removing multiple hits to the same function. **(C)** The length (y-axis) and coverage (shape) of these contigs with different identifiable protein profiles were analysed. The number of contigs associated with the different cut-offs are displayed in the parentheses. **(D)** The genome architecture of the final seven contigs and relatedness estimated through tBLASTx. (MP = maturation protein; CP = coat protein; RdRP = RNA-directed RNA polymerase).

Following on from assessing the protein profile of the hmmscan results, the focus was redirected to investigating the related contigs. The initial analysis examined all contigs of the HMM-identified proteins, subsequently focused on those that encoded a protein which had a minimum score of 30 and were functionally unique (as previously described), and finally on those that had two or more of these proteins (Figure 5C). To gauge the abundance of these contigs in the samples, their coverage was used as a visual metric to represent the contigs. There were seven contigs of interest whose genome architecture and relatedness were examined (Figure 5D). Since there were only three +ssRNA phages classified by Demovir in the original study, the application of this phage-specific HMM search-tool enables for the detection of otherwise missed genomes.

Combining the HMM 5-MC search tool with the new taxonomy of *Leviviricetes* (described in Chapter IV) enabled the swift classification of the majority of these contigs to family level (Table 1). Of the 7 contigs, five encoded RdRP B (assigned to the *Timlovirales* order), while an RdRP protein was not detected in the other two contigs, their MP and CP clusters would also suggest they too belong to this order. There were six contigs that encoded a CP most like CP E (*Blumeviridae* family) whereas the CP of one contig could not be detected through the hmmscan. To assign a taxonomic rank of genus and species to the seven contigs, a tBLASTx of the 868 representative contigs available with a full ICTV-accepted taxonomic background was completed. Despite this, there was no good match (defined as ≥ 70% identity, ≥ 100bp) across any of these sequences which could suggest that these are a novel collection of +ssRNA phages. The mammal origin of these contigs were also available from the metadata tables of the biogeography study.

Contig ID	**Length (bp)**	**Coverage**	**MP**	**CP**	**Family**	**RdRP**	**Order**	**Mammal**
Seq. 1	4082	15	MP A	ND	ND	RdRP B	*Timlovirales*	Macaque
Seq. 2	4234	33	MP C	CP E	*Blumeviridae*	RdRP B	*Timlovirales*	Macaque
Seq. 3	2482	4.7505	MP C	CP E	*Blumeviridae*	ND	ND	Pig
Seq. 4	1160	2	MP C	CP E	*Blumeviridae*	ND	ND	Pig
Seq. 5	4983	318.3261	MP C	CP E	*Blumeviridae*	RdRP B	*Timlovirales*	Macaque
Seq. 6	3996	11.28394	MP C	CP E	*Blumeviridae*	RdRP B	*Timlovirales*	Macaque
Seq. 7	4135	33.29779	MP C	CP E	*Blumeviridae*	RdRP B	*Timlovirales*	Pig

Table 1. Description of the seven contigs detected from the biogeography study. The length, coverage, HMM-detected proteins and associated taxonomy, and mammalian origin of the final seven contigs of interest. (MP = maturation protein; CP = coat protein; RdRP = RNA-directed RNA polymerase; ND = not detected).

5.4.4 Future applications of this method to large-scale human gut phageome studies

There are many advantages to applying the Biopsy Method to future phageome studies, one being the high rate of +ssRNA phage recovery. As a high-throughput method it could be coupled with the ever-growing rate of bioinformatic discovery and advancements to enable a quicker turnaround in phageome research. Nonetheless, it should be noted that animal models should be used as a stepping-stone rather than the final evidence when it comes to inferring anything about the human gut phageome and the prevalence of the associated members. Prior to describing the +ssRNA phage consortium of the human microbiome or inferring any roles they may play in human health, a more appropriate measure of these entities in the human gut is essential. Based on this research it could be suggested that applying the Biopsy Method to a large-scale human gut phageome study would offer a more concise and accurate representation of their abundance.

5.5 Conclusion

With the goal of comprehensively describing the human phageome in as much detail and accuracy as possible, +ssRNA phages need to be accounted for. Given that interest and research of the human phageome is only set to escalate in coming years, this study provides an examination of three extraction methods that form the foundation of this research area while assessing their effectiveness at retrieving both spiked-in and naturally occurring +ssRNA phages. Previous applications of the PEG Method failed to identify any +ssRNA phages which may be attributed to the smaller volumes used for the RT or MDA-associated biases. It is apparent that the high-throughput Biopsy Method potentially offers the best means to recover +ssRNA phages when combined with RNA extraction as it is quick, efficient, and requires a smaller starting sample.

The recent application of this method in the analysis of the virome biogeography in mammals uncovered +ssRNA phages which were further investigated. These findings suggest that the extraction methods used for isolating the VLPs and nucleic acids play a crucial role in the recovery and detection of +ssRNA phages, a fundamental feature that needs to be regarded in future research. However, this work serves as only a foundation as more work is required to validate these findings in larger studies and accurately classify the recovered phages.

5.6 References

Andrews, Simon. 2010. "FastQC A Quality Control Tool for High Throughput Sequence Data." 2010. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Baker, Monya. 2012. "De Novo Genome Assembly: What Every Biologist Should Know." *Nature Methods* 9 (4): 333–37. https://doi.org/10.1038/nmeth.1935.

Blighe, Kevin, Sharmila Rana, Emir Turkes, Benjamin Ostendorf, Andrea Grioni, and Myles Lewis. 2021. *EnhancedVolcano: Publication-Ready Volcano Plots with Enhanced Colouring and Labeling* (version 1.10.0). Bioconductor version: Release (3.13). https://doi.org/10.18129/B9.bioc.EnhancedVolcano.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics (Oxford, England)* 30 (15): 2114–20. https://doi.org/10.1093/bioinformatics/btu170.

Bushmanova, Elena, Dmitry Antipov, Alla Lapidus, and Andrey D Prjibelski. 2019. "RnaSPAdes: A de Novo Transcriptome Assembler and Its Application to RNA-Seq Data." *GigaScience* 8 (9). https://doi.org/10.1093/gigascience/giz100.

Callanan, Julie, Stephen R. Stockdale, Andrey N. Shkoporov, Lorraine A. Draper, R. Paul Ross, and Colin Hill. 2021. "Biases in Viral Metagenomics-Based Detection, Cataloguing and Quantification of Bacteriophage Genomes in Human Faeces, a

Review.” *Microorganisms* 9 (3): 524. https://doi.org/10.3390/microorganisms9030524.

Callanan, Julie, Stephen R. Stockdale, Andrey N. Shkoporov, Lorraine A. Draper, R. Paul Ross, and Colin Hill. 2020. “Expansion of Known ssRNA Phage Genomes: From Tens to over a Thousand.” *Science Advances* 6 (6): eaay5981. https://doi.org/10.1126/sciadv.aay5981.

Castro-Mejía, Josué L., Musemma K. Muhammed, Witold Kot, Horst Neve, Charles M. A. P. Franz, Lars H. Hansen, Finn K. Vogensen, and Dennis S. Nielsen. 2015. “Optimizing Protocols for Extraction of Bacteriophages Prior to Metagenomic Analyses of Phage Communities in the Human Gut.” *Microbiome* 3 (1): 64. https://doi.org/10.1186/s40168-015-0131-4.

Chen, Hanbo. 2018. *VennDiagram: Generate High-Resolution Venn and Euler Plots* (version 1.6.20). https://CRAN.R-project.org/package=VennDiagram.

Conceição-Neto, Nádia, Mark Zeller, Hanne Lefrère, Pieter De Bruyn, Leen Beller, Ward Deboutte, Claude Kwe Yinda, *et al*. 2015. “Modular Approach to Customise Sample Preparation Procedures for Viral Metagenomics: A Reproducible Protocol for Virome Analysis.” *Scientific Reports* 5 (November): 16532. https://doi.org/10.1038/srep16532.

Grenié, Matthias, Pierre Denelle, Caroline Tucker, François Munoz, and Cyrille Violle. 2020. *Funrar: Functional Rarity Indices Computation* (version 1.4.1). https://CRAN.R-project.org/package=funrar.

Hyatt, Doug, Gwo-Liang Chen, Philip F. LoCascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. 2010. “Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification.” *BMC Bioinformatics* 11 (1): 119. https://doi.org/10.1186/1471-2105-11-119.

Kassambara, Alboukadel. 2020. *Ggpubr: "ggplot2" Based Publication Ready Plots* (version 0.4.0). https://CRAN.R-project.org/package=ggpubr.

Krishnamurthy, Siddharth R., Andrew B. Janowski, Guoyan Zhao, Dan Barouch, and David Wang. 2016. "Hyperexpansion of RNA Bacteriophage Diversity." *PLOS Biology* 14 (3): e1002409. https://doi.org/10.1371/journal.pbio.1002409.

Langmead, Ben, and Steven L. Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–59. https://doi.org/10.1038/nmeth.1923.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79. https://doi.org/10.1093/bioinformatics/btp352.

Lim, Efrem S., Yanjiao Zhou, Guoyan Zhao, Irma K. Bauer, Lindsay Droit, I. Malick Ndao, Barbara B. Warner, Phillip I. Tarr, David Wang, and Lori R. Holtz. 2015. "Early Life Dynamics of the Human Gut Virome and Bacterial Microbiome in Infants." *Nature Medicine* 21 (10): 1228–34. https://doi.org/10.1038/nm.3950.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (12): 550. https://doi.org/10.1186/s13059-014-0550-8.

Manrique, Pilar, Benjamin Bolduc, Seth T. Walk, John van der Oost, Willem M. de Vos, and Mark J. Young. 2016. "Healthy Human Gut Phageome." *Proceedings of the National Academy of Sciences* 113 (37): 10400–405. https://doi.org/10.1073/pnas.1601060113.

Martin, Marcel. 2011. "Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads." *EMBnet.Journal* 17 (1): 10–12. https://doi.org/10.14806/ej.17.1.200.

McCann, Angela, Feargal J. Ryan, Stephen R. Stockdale, Marion Dalmasso, Tony Blake, C. Anthony Ryan, Catherine Stanton, Susan Mills, Paul R. Ross, and Colin Hill. 2018. "Viromes of One Year Old Infants Reveal the Impact of Birth Mode on Microbiome Diversity." *PeerJ* 6: e4694. https://doi.org/10.7717/peerj.4694.

McMurdie, Paul J., and Susan Holmes. 2013. "Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data." *PLOS ONE* 8 (4): e61217. https://doi.org/10.1371/journal.pone.0061217.

Minot, Samuel, Rohini Sinha, Jun Chen, Hongzhe Li, Sue A. Keilbaugh, Gary D. Wu, James D. Lewis, and Frederic D. Bushman. 2011. "The Human Gut Virome: Inter-Individual Variation and Dynamic Response to Diet." *Genome Research* 21 (10): 1616–25. https://doi.org/10.1101/gr.122705.111.

Norman, Jason M., Scott A. Handley, Megan T. Baldridge, Lindsay Droit, Catherine Y. Liu, Brian C. Keller, Amal Kambal, *et al*. 2015. "Disease-Specific Alterations in the Enteric Virome in Inflammatory Bowel Disease." *Cell* 160 (3): 447–60. https://doi.org/10.1016/j.cell.2015.01.002.

Nurk, Sergey, Dmitry Meleshko, Anton Korobeynikov, and Pavel A. Pevzner. 2017. "MetaSPAdes: A New Versatile Metagenomic Assembler." *Genome Research* 27 (5): 824–34. https://doi.org/10.1101/gr.213959.116.

Oksanen, Jari, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, *et al*. 2020. *Vegan: Community Ecology Package* (version 2.5-7). https://CRAN.R-project.org/package=vegan.

Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26 (6): 841–42. https://doi.org/10.1093/bioinformatics/btq033.

Reyes, Alejandro, Matthew Haynes, Nicole Hanson, Florent E. Angly, Andrew C. Heath, Forest Rohwer, and Jeffrey I. Gordon. 2010. “Viruses in the Fecal Microbiota of Monozygotic Twins and Their Mothers.” *Nature* 466 (7304): 334–38. https://doi.org/10.1038/nature09199.

Roux, Simon, Francois Enault, Bonnie L. Hurwitz, and Matthew B. Sullivan. 2015. “VirSorter: Mining Viral Signal from Microbial Genomic Data.” *PeerJ* 3 (May): e985. https://doi.org/10.7717/peerj.985.

Roux, Simon, Natalie E. Solonenko, Vinh T. Dang, Bonnie T. Poulos, Sarah M. Schwenck, Dawn B. Goldsmith, Maureen L. Coleman, Mya Breitbart, and Matthew B. Sullivan. 2016. “Towards Quantitative Viromics for Both Double-Stranded and Single-Stranded DNA Viruses.” *PeerJ* 4 (December): e2777. https://doi.org/10.7717/peerj.2777.

RStudio Team. 2020. “RStudio: Integrated Development for R. RStudio, PBC, Boston, MA.” 2020. https://rstudio.com/.

Salzberg, Steven L., Adam M. Phillippy, Aleksey Zimin, Daniela Puiu, Tanja Magoc, Sergey Koren, Todd J. Treangen, *et al*. 2012. “GAGE: A Critical Evaluation of Genome Assemblies and Assembly Algorithms.” *Genome Research* 22 (3): 557–67. https://doi.org/10.1101/gr.131383.111.

Shkoporov, Andrey N., Adam G. Clooney, Thomas D. S. Sutton, Feargal J. Ryan, Karen M. Daly, James A. Nolan, Siobhan A. McDonnell, *et al*. 2019. “The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific.” *Cell Host & Microbe* 26 (4): 527-541.e5. https://doi.org/10.1016/j.chom.2019.09.009.

Shkoporov, Andrey N., Feargal J. Ryan, Lorraine A. Draper, Amanda Forde, Stephen R. Stockdale, Karen M. Daly, Siobhan A. McDonnell, *et al*. 2018. “Reproducible Protocols for Metagenomic Analysis of Human Faecal Phageomes.” *Microbiome* 6 (1): 68. https://doi.org/10.1186/s40168-018-0446-z.

Shkoporov, Andrey N., Stephen R. Stockdale, Aonghus Lavelle, Ivanela Kondova, Cara Hueston, Aditya Upadrasta, Ekaterina Khokhlova, *et al*. 2021. “Viral Biogeography of Gastrointestinal Tract and Parenchymal Organs in Two Representative Species of Mammals.” https://doi.org/10.21203/rs.3.rs-803286/v1.

Sutton, Thomas D. S., Adam G. Clooney, Feargal J. Ryan, R. Paul Ross, and Colin Hill. 2019. “Choice of Assembly Software Has a Critical Impact on Virome Characterisation.” *Microbiome* 7 (January): 12. https://doi.org/10.1186/s40168-019-0626-5.

Thurber, Rebecca Vega. 2009. “Current Insights into Phage Biodiversity and Biogeography.” *Current Opinion in Microbiology*, Antimicrobials ● Genomics, 12 (5): 582–87. https://doi.org/10.1016/j.mib.2009.08.008.

Townsend, Eleanor M., Lucy Kelly, George Muscatt, Joshua D. Box, Nicole Hargraves, Daniel Lilley, and Eleanor Jameson. 2021. “The Human Gut Phageome: Origins and Roles in the Human Gut Microbiome.” *Frontiers in Cellular and Infection Microbiology* 11: 498. https://doi.org/10.3389/fcimb.2021.643214.

Wickham, Hadley. 2020. *Reshape2: Flexibly Reshape Data: A Reboot of the Reshape Package* (version 1.4.4). https://CRAN.R-project.org/package=reshape2.

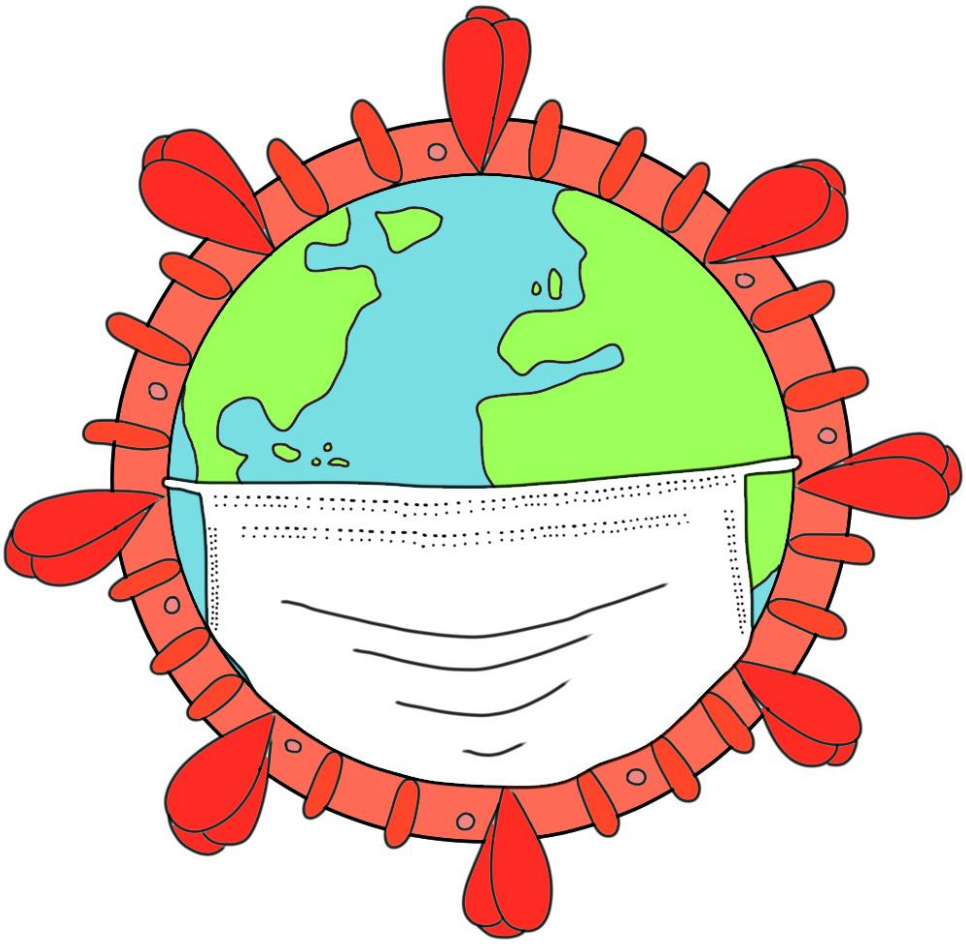
Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, Dewey Dunnington, and RStudio. 2021. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics* (version 3.3.5). https://CRAN.R-project.org/package=ggplot2.

Wickham, Hadley, and RStudio. 2019. *Stringr: Simple, Consistent Wrappers for Common String Operations* (version 1.4.0). https://CRAN.R-project.org/package=stringr.

Zane, Ray. (2017) 2017. *Gsub*. Crystal. https://github.com/rzane/gsub.

Zhang, Tao, Mya Breitbart, Wah Heng Lee, Jin-Quan Run, Chia Lin Wei, Shirlena Wee Ling Soh, Martin L. Hibberd, Edison T. Liu, Forest Rohwer, and Yijun Ruan. 2005. “RNA

Viral Community in Human Feces: Prevalence of Plant Pathogenic Viruses." *PLOS Biology* 4 (1): e3. https://doi.org/10.1371/journal.pbio.0040003.



Chapter VI

An Assessment of *Cystovirus* phi6 as a Surrogate for SARS-CoV-2 in Lipopeptide Exposure and Thermotolerance Assays

For this chapter, I devised the study, performed the analyses, produced the images, and wrote the manuscript.

6.1 Abstract

Environmental persistence of respiratory pathogens in environments is an important aspect of their transmission and infection. The emergence of one such pathogen, SARS-CoV-2 (causative agent of COVID-19), has led to the ongoing global pandemic with unprecedented impacts on global health and economic sectors. Here phi6, an enveloped dsRNA bacteriophage, was investigated as a potential surrogate for coronavirus survival to lipopeptide exposure and in thermotolerance assays. The test conditions included exposure to a lipopeptide treatment (ART24) and exposure to different temperatures that are key to dairy processing, to assess any dangers with potential outbreaks in this industry. Two other well-studied phages, MS2 (a +ssRNA bacteriophage) and phiX174 (a +ssDNA bacteriophage), were included as comparative controls. The susceptibility of the phi6 model to different treatments was examined and the effect these measures would have on *Coronaviridae* family members was evaluated. From this work, it was concluded that ART24 may serve as a potential treatment for coronaviruses.

6.2 Introduction

The ongoing pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has compelled scientists and clinicians to identify treatments and containment strategies while the vaccination programme spreads across the globe. While these experiments would provide the most reliable information if they were performed directly on coronavirus samples, this is difficult considering the requirement for biosafety level (BSL) 3 or 4 laboratories due to its pathogenicity and rate of genetic mutations for such experiments involving viral replication. SARS-CoV-2 is an enveloped virus with a monopartite, linear, positive-sense single-stranded RNA (+ssRNA) genome (Brian and Baric 2005). The bacteriophage (phage) phi6 is also an enveloped RNA virus and offers a completely safe

surrogate model for evaluating the efficacy of compounds on coronavirus as it does not infect humans and would enable more laboratories to test anti-SARS-CoV-2 compounds as a BSL-1 standard is sufficient. A recent review by Barros *et al*. details the usefulness of this phage in assessing the effectiveness of anti-SARS-CoV-2 treatments (Barros, Ferraz, and Monteiro 2021).

The capsid of enveloped viruses is generally surrounded by a lipid bilayer embedded with proteins. To infect a cell, the virus must fuse this envelope with the membrane of the host so that the envelope can connect to the cytoplasm. A selection of antiviral drugs targets this initial attachment to prevent infection (Altmeyer 2004). However, these drugs are classified as of limited efficacy as viruses can easily establish drug-resistance through their diverse membrane fusion proteins and rapid mutation rates (Wei *et al*. 2002). There are many enveloped viruses that infect and cause disease in humans such as influenza virus, hepatitis viruses (B and C), severe acute respiratory syndrome (SARS) coronavirus, and human immunodeficiency virus (HIV) which highlights the need to investigate potential treatments (Bukasov, Dossym, and Filchakova 2021).

Three well-studied, tailless phages were used in this study, each with a different virion and genome structure to offer a comprehensive comparison across the experiments. The first is the enveloped double-stranded RNA (dsRNA) bacteriophage phi6, a member of the *Cystoviridae* family that infects *Pseudomonas syringae*. It has a similar structure to coronavirus but has two additional proteinaceous capsids (nucleocapsid and polymerase complex) between the lipid envelope and the genome, as shown in Figure 1. It contains a tri-segmented dsRNA genome as opposed to the +ssRNA genome of coronaviruses but its similarity to the eukaryotic SARS-CoV-2 allows for cross-study comparisons as detailed in several studies (Vatter, Hoenes, and Hessling 2021; Fedorenko *et al*. 2020; Rockey *et al*. 2020; Gendron *et al*. 2010). MS2 is a non-enveloped +ssRNA bacteriophage that infects *Escherichia coli* and is a member

of the newly described class *Leviviricetes* (Callanan *et al*. 2021). It encodes a similar genome to that of coronavirus (+ssRNA) so serves as a control for the nucleic acid content. The third phage included was phiX174, a non-enveloped single-stranded DNA (+ssDNA) bacteriophage, which is a member of the *Microviridae* and infects *Escherichia coli*. This phage serves as a non-RNA-control.

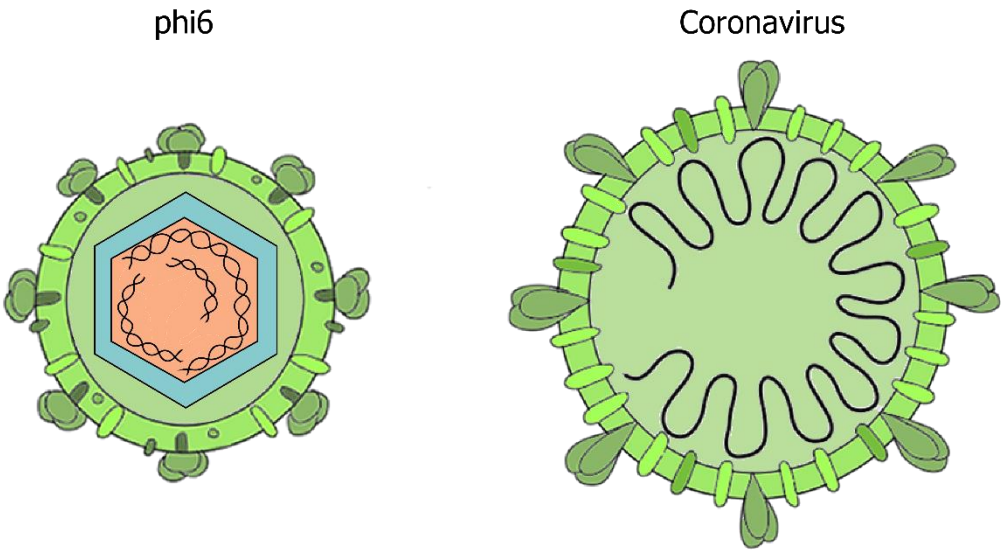


Figure 1. Virion and genome comparison of phi6 and coronavirus. There are key differences in the capsid and genome structure of these two RNA viruses, but a crucial similarity is the envelope and the envelope-associated proteins that are essential to successful attachment and infection in hosts cells. This figure was drawn based on other images (Silverman and Boehm 2020).

Recent reports have found that lipopeptides offer a potential preventative treatment for coronavirus infections as they target the initial stage of the complex viral infection by preventing the membrane fusion of the coronavirus spike (S) protein and susceptible host cell receptors (Vries *et al*. 2021; Chowdhury, Baindara, and Mandal 2021). Other work has also screened crystal structures of potential lipopeptides to identify potential treatments of COVID-

19 and detailed how they could be harnessed to directly inhibit viral activity (Chowdhury, Baindara, and Mandal 2021). The phi6 surrogate model would serve to build on the initial computer screening performed to date to test the efficacy and effectiveness of the identified therapeutic antimicrobials.

ART24 is a live biotherapeutic product (LBP) isolated from a pure strain of *Bacillus amyloliquefaciens*/*Bacillus velezensis* and previously shown to have anti-*Clostridioides difficile* activity through the production of several antimicrobials, including lipopeptides (O'Donnell *et al*. Manuscript submitted for publication; 2020). Mass spectrometry, (performed by Dr Paula O'Connor in Teagasc Food Research Centre in Moorepark, Cork) indicated that the three major components of the cell-free supernatant are surfactin, fengycin and amylocyclicin (Figure 2). Surfactin is a lipopeptide with exceptional surfactant activity, suggesting that ART24 has the potential to act as a treatment for enveloped-virus infections such as coronavirus.

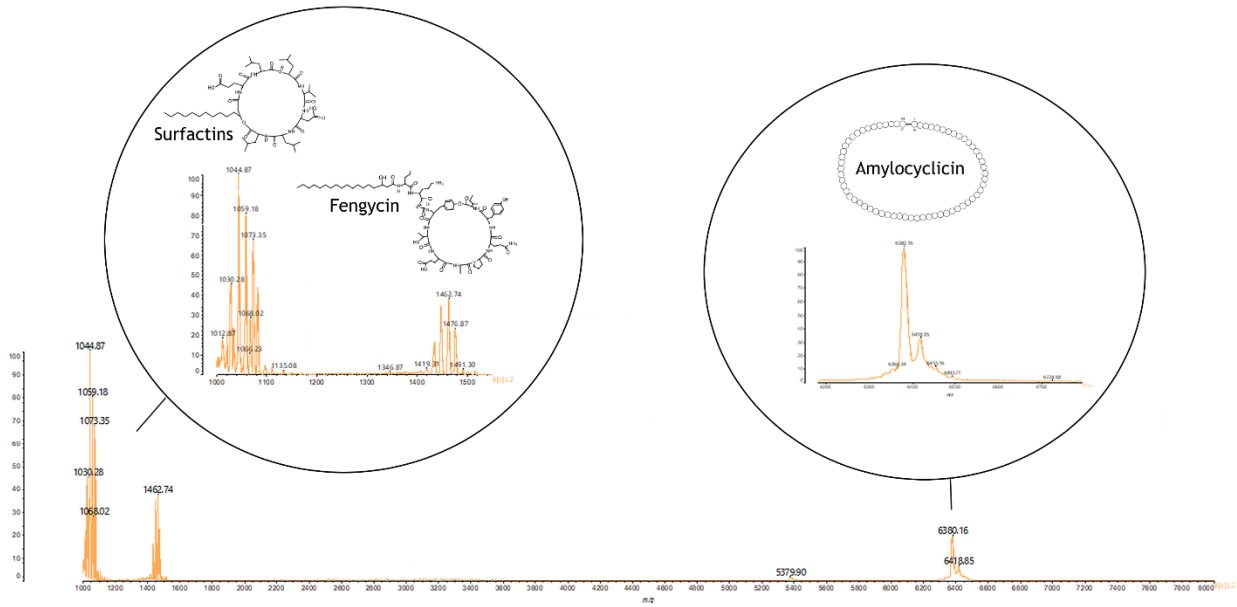


Figure 2. Mass spectrometry of ART24, including the structures of surfactin, fengycin, and amylocyclicin. There are three definitive peaks that correspond to the surfactin, fengycin and amylocyclicin profiles of ART24. It is the activity of these compounds that are of interest.

Here, the activity of ART24 on the infectivity of phi6 was compared to surfactin and fengycin individually. Amylocyclicin was not included as an individual comparator as it could not be isolated in large enough volumes and as a bacteriocin, it is not expected to have any effect on host-virus interactions. Surfactin is a cyclic lipopeptide (also synthesized by *Bacillus subtilis*) with a wide range of bioactivities including antibiotic and antiviral properties. Its antiviral activity is due to its ability to inhibit the membrane fusion between the virus and host cells, with its potent surfactant activity (Yuan *et al*. 2018). Fengycin is another cyclic lipopeptide that is produced by *B. subtilis* in response to fungal infections. It has two mechanisms, the first involves inducing systemic resistance in the root cells of plants, and the second is the lysis of the fungal cell membrane by the fengycin directly binding to the target (Sur, Romo, and Grossfield 2018).

The effect of different temperature and time combinations on the titre of these three phages in milk was also investigated. At the early stages of the pandemic, meat plants were areas of high detection levels of SARS-CoV-2 in employees which raised concerns around the safety of the associated products as this is a huge international product for Ireland. Another area of the Irish international trade of utmost importance to the economy is the dairy industry. Previous studies have shown the efficacy of different pasteurisation temperatures at neutralizing pathogens such as Middle East respiratory syndrome (MERS-CoV) (van Doremalen *et al*. 2014) and betacoronavirus (Jiang *et al*. 2021). To evaluate any potential risks from a possible rise in cases at the treatment facilities, the efficacy of the thermo-treatment of milk at inhibiting phi6 activity was examined. The standard pasteurisation temperature of 72°C and the vat pasteurisation temperature of 63°C were examined. An additional temperature of 55°C was featured as it was previously shown as an effective temperature inhibitor of coronavirus after a period of 10 mins (it destroys the nucleocapsid (N) protein) (Wang *et al*. 2004). The samples were exposed to the temperatures for a variety of times of 15 seconds, 1 min, and 5 mins.

These studies demonstrate the potential application of ART24 as a treatment for coronavirus and the efficacy of heat exposure in preventing infection using phi6 as a model system.

6.3 Materials and Methods

6.3.1 Bacterial strains, bacteriophage stocks and media

Bacterial strains of *Escherichia coli* (DSMZ 5210 and 13127) were grown in LBB (Luria Bertani broth) and LBA (Luria Bertani broth supplemented with 1.5% agar (w/v)) at 37°C, whereas *P. syringae* was grown in TSB (tryptic soy broth) and TSA (tryptic soy broth supplemented with 1.5% agar (w/v)) at 25°C. Plaque assays were carried out by overlaying agar plates with either LBA or TSA (0.4% agar (w/v)) supplemented with magnesium sulfate ($MgSO_4$, final concentration of 10mM) and plates were incubated overnight. Liquid propagation was carried out to generate a phage stock of sufficient titre (10^{10}pfu/ml of phi6 and MS2 and 10^{9}pfu/ml of phiX174) for the experimental work described. In summary, fresh sterile broth was inoculated with 2% of overnight culture and left to grow to a minimum OD_{600} of 0.1. Following this, 2% of phage lysate and $MgSO_4$ (final concentration 10mM) were added to the bacterial culture and incubated until a clearing of the bacterial culture was observed. Following centrifugation at 4,075 x g for 20 mins to pellet the bacterial debris and the supernatant was filtered using a 0.20µm filter. To determine the phage titre, a plaque assay was performed.

6.3.2 ART24 analysis

ART24 supernatant stocks were kindly provided by Dr Michelle O'Donnell (Artugen Therapeutics). Surfactin (Sigma-Aldrich) and fengycin (Sigma-Aldrich) were each prepared in stocks of a final concentration of 1mg/mL in 20% ethanol. To examine the effect these compounds had on the infectivity of the phages, 100µL of phage lysate was mixed with 100µL of either ART24, surfactin, or fengycin. These were incubated at room temperature (approx. 25°C) for one hour. Serial dilutions (10^{-1} to 10^{-8} dilutions) were then spotted in 5µL volumes onto respective hosts and incubated at necessary temperatures overnight. A control of the neat phage lysate (incubated with SM buffer in the same conditions) was also included and plates were examined the following day.

To determine the minimum inhibitory concentration (MIC) of ART24, a 1 in 2, 1 in 4 and 1 in 8 dilution sample were prepared. The same method as described above was implemented, including an untreated control, and plates were recorded the following day. These experiments were repeated in triplicate to validate results.

6.3.3 SDS-PAGE analysis

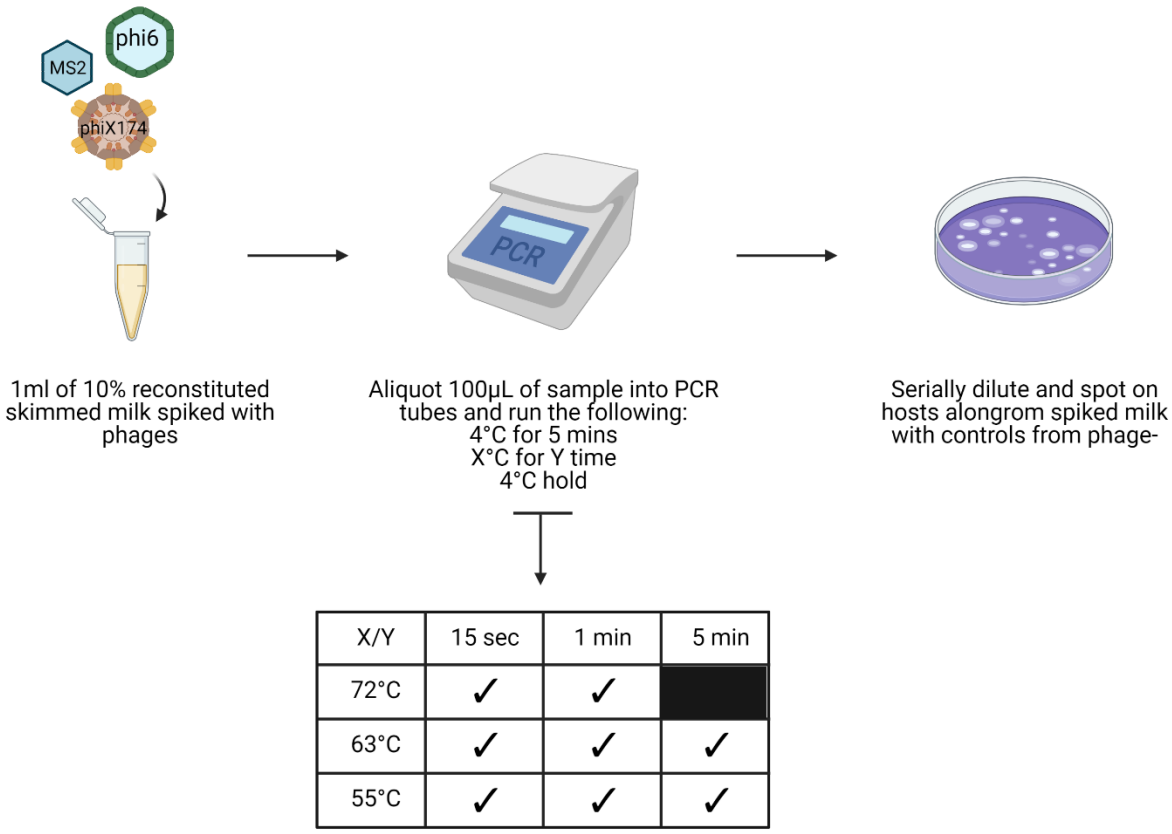
To concentrate phi6 virions, three rounds of ultracentrifugation at 49,080 x g for 2 hours at 4°C, with 10ml of fresh phi6 lysate added in between each round. The final pellet was resuspended in 1ml of Buffer A (aids the stabilization the virion once the envelope has been removed) and left overnight at room temperature. The next day the titre of this concentrated lysate was estimated by plaque assay. The ART24 and surfactin treatments were repeated; diluted in a 1:1 ratio with 200µL of both phi6 lysate and treatment and left at room temperature for 1 hour. These samples, along with a Buffer A-treated control, were purified using Microcon® Centrifugal Filters 100kDa filters (Merck) and centrifuged at 14,000 x g at 25°C for 12 mins. As per the associated protocol, the column was then immediately inverted into a new collection tube and centrifuged at 1,000 x g for 3 mins at 25°C. The resulting fraction was resuspended in 30µL of Buffer A.

To examine the effect of the treatments on the envelope protein profile of phi6, the purified samples were treated with Tricine SDS Sample Buffer (ThermoFisher/Bioscience). The obtained sample (40µL) was loaded onto a pre-made Novex 10-20% Tricine reducing SDS-PAGE gel, along with 5µL of the PageRuler™ Unstained Low Range Protein Ladder and separated at 125V for 70 mins using Novex Tricine SDS Running Buffer (1x) (all from ThermoFisher/Bioscience). The gel was stained using 25ml of Imperial Protein Stain with shaking at 160 rpm for 2 hours, followed by overnight destaining in deionized water.

To assess the infectivity of the filtered samples, both the residue and filtrate fractions were serially diluted (10^{-1} to 10^{-8} dilutions) and spotted in 5µL aliquots onto *P. syringae* overlays. These were incubated at 25°C overnight and examined the next day.

6.3.4 Thermotolerance evaluation

To ensure the phages could be mixed and separated by host, no cross-spotting across hosts was confirmed. All three phages were included in examining the effect of temperature and exposure time on titre in 10% reconstituted skimmed milk (Dairygold). This was made by dissolving reconstituted skimmed milk powder (10% (w/v)) in deionized water and autoclaving at 110°C for 10 mins. Both phi6 and MS2 were added to a final concentration of 10^8 pfu/ml while phiX174 was added at 10^7 pfu/ml. Aliquots of 100µL were treated to various temperature-time conditions in a PCR machine (2720 Thermal Cycler Applied Biosciences, Figure 3). These samples were then 10-fold serially diluted in SM buffer and 5µL volumes were spotted on each of the hosts. Plates were incubated at 25°C (phi6) and 37°C (MS2/phiX174) and examined the following day. A control of the phage-inoculated milk that was not exposed to any temperature (left at room temperature) was also included to establish the original titre.



X/Y	15 sec	1 min	5 min
72°C	✓	✓	
63°C	✓	✓	✓
55°C	✓	✓	✓

Figure 3. Overview of the thermotolerance assay method for phi6, MS2 and phiX174 in reconstituted milk. All three phages were inoculated into the same milk sample as they do not share the same host. Aliquots of this mixture were treated at different time-temperature combinations as displayed in the table. These temperatures were used to replicate common dairy pasteurisation processes and a previous thermotolerance study on coronavirus (Wang *et al*. 2004). Created with BioRender.com.

6.4 Results

6.4.1 ART24 activity

It was observed that both ART24 and surfactin completely inactivated phi6 whereas fengycin had no obvious impact on its titre (Figure 4). In comparison, none of the three compounds affected the titres of either MS2 or phiX174. A significant main difference between these phages is that phi6 is enveloped whereas both MS2 and phiX174 are non-enveloped.

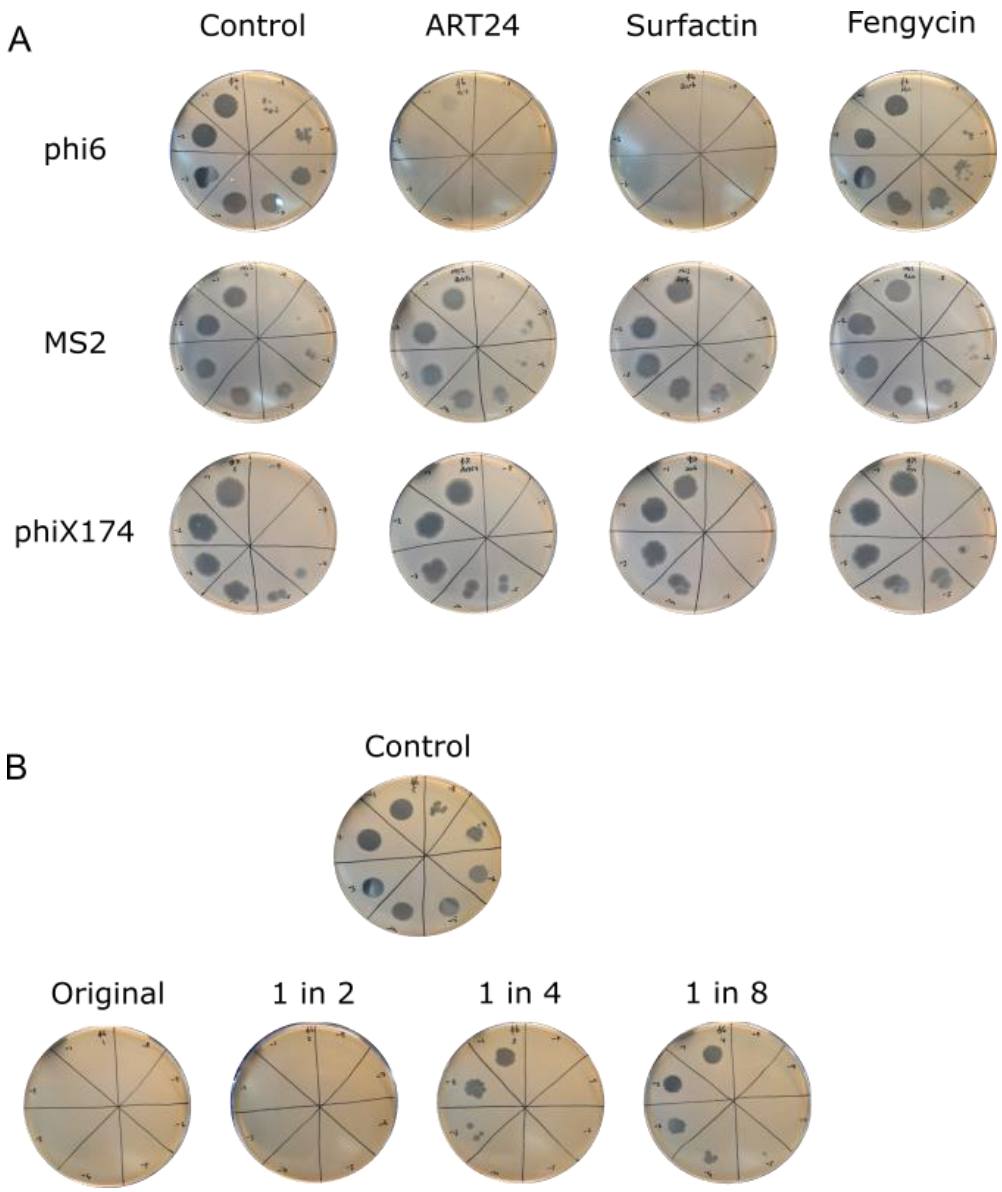


Figure 4. **The effect of ART24 (lipopeptide) exposure on the infectivity of phi6. (A)** Comparison of impact of ART24**,** surfactin and fengycin on the infectivity of the different phages. This highlights the impact that ART24 and surfactin have on enveloped viruses using phi6 as a model virus. The ART24 efficacy can be linked to its surfactin component rather than the fengycin or amylocyclicin fractions. **(B)** Minimum inhibitory concentration (MIC) analysis of ART24 on phi6 infectiveness demonstrates that a 1 in 8 dilution results in a 50% reduction of titre.

It was proposed that ART24 interferes with the phi6 envelope through the action of surfactin and prevents successful host-virus binding and ultimately abolishes infection. This indicates that ART24 could be an exceptionally effective anti-SARS-CoV-2 treatment based on the phi6 model. A recent review by Simon *et al* (2021) details the possible interactions between surfactants and viruses, and how these exchanges can be exploited in the war against coronavirus and other enveloped viruses (Simon *et al*. 2021). The surfactin profile of ART24, with potent surfactant properties, could allow for it to be used either as a carrier component of drug delivery systems, or directly in disinfection. The results from this work indicate that direct application of ART24 to enveloped viruses of high titres results in complete inactivation and prevents any detectable successful infection. Given that ART24 is already being tested in humans for its anti-*C. difficile* properties, utilizing its potential as a COVID-19 treatment could be a viable proposition.

Firstly, phi6 was concentrated by three rounds of ultracentrifugation and then filtered through a 100kDa filter (Microcon). The retained phi6 was then analysed using SDS-PAGE which revealed that virion proteins are intact and largely identifiable. However, concentrated phi6 treated with either ART24 or surfactin prior to filtration showed significant differences from the untreated phage (Figure 5A &5B). The ART24 treated phi6 showed partial destruction of P1 (major inner protein of the procapsid), P2 (RNA-directed RNA polymerase of phi6 stored in the procapsid), and P4 (packaging enzyme embedded in the procapsid), along with complete removal of P3 (spike protein found on the envelope), P5 (peptidoglycan hydrolase located on the nucleocapsid), and P8 (major protein of the nucleocapsid). A similar protein profile was determined for phi6 exposed to surfactin, other than it had little or no impact on P3.

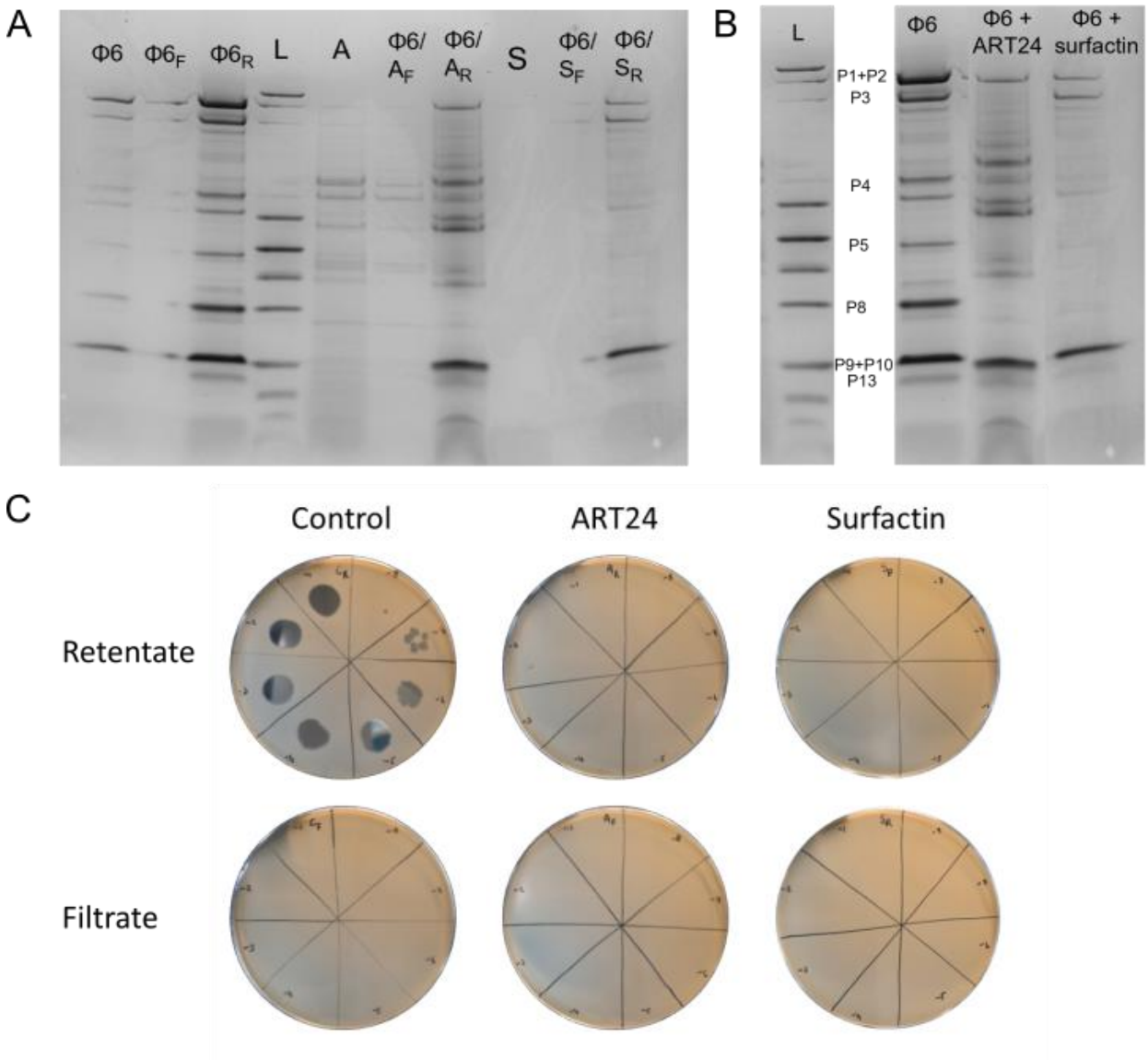


Figure 5. Protein profiles of phi6 on SDS-PAGE gel and spot assays. Following on from fractionation of the treated and untreated phi6 into retentate and filtrate, the **(A)** full gel of the protein profiles, **(B)** a concatonated version of this gel showing retentate profiles of interest, and **(C)** infectivity of phi6 on *P. syringae* were examined.

Given that most of these proteins are encapsulated within the envelope and despite the major envelope protein (P9) being detectable in comparable amounts in all three, these findings suggest that ART24 and surfactin are effective at disrupting the protective envelope of phi6. Presumably most of the proteins released by the disrupted virion passed through the filter and

were not retained. The additional activity of ART24 on P3 may be due to a variety of factors, including the fact that the ART24 preparation includes proteases.

Corresponding spot assays revealed that phi6 was rendered inactive in both retentate and filtrate of the ART24 and surfactin treated phage, and it was only the retentate of the untreated control that retained its activity against the *P. syringae* host (Figure 5C).

6.4.2 Thermotolerance

Exposure of phi6 (10^8 pfu/ml) to temperatures above 55°C for a period of more than 15 seconds resulted no detectable plaques ($<10^3$ pfu/ml) (Figure 6). In comparison, applying the same temperature-time conditions to MS2 and phiX174 resulted in no more than two logs decrease in titre.

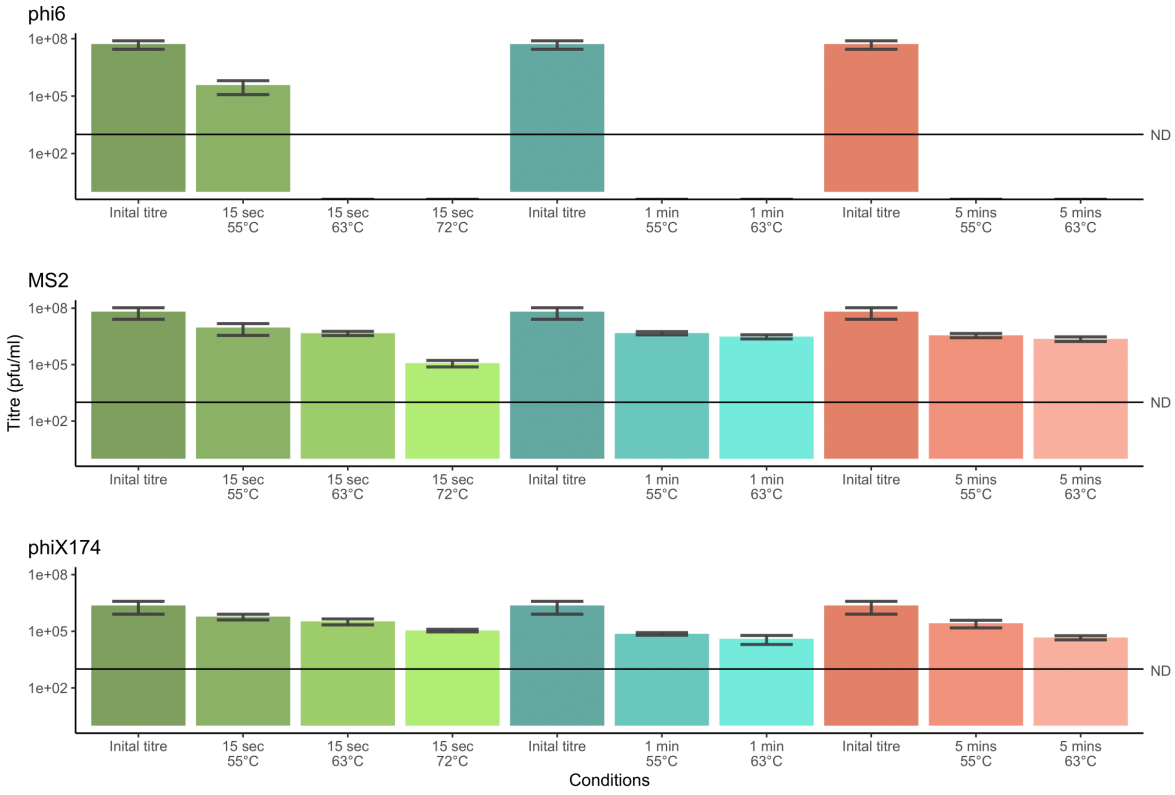


Figure 6. Effect of different time-temperature conditions on the phages phi6, MS2 and phiX174. The titre of phi6 dropped by two logs following 55°C at 15 seconds and was undetectable at every other time-temperature combination. In comparison, MS2 and phiX174 titres dropped in a time- and temperature-dependent manner – as the time and temperature increased, the titre of the phage decreased. The error bars indicate ± standard deviation. (ND = not detectable; level of detection was $>10^3$pfu/ml).

This analysis demonstrates that pasteurisation treatments and exposure to temperatures greater than 55°C for 15 seconds are effective at eliminating any potential SARS-CoV-2 contamination, which is in congruence with previous research in this area. The work carried out by Wang *et al* (2004), reported that at 35°C the nucleocapsid protein of coronavirus begins to unfold and at 55°C it is completely denatured, resulting in inactivation of the virus (Wang *et al*. 2004). It can be proposed that a similar mechanism affects critical proteins of phi6,

rendering it unable to infect, even in typically protective high-lipid environment of milk. It should be noted that these experiments were completed in smaller, representative volumes of milk and not in the typical large-scale volumes of milk. Therefore, these results serve as an insight into the thermotolerance assays on phi6 in the dairy environments and bigger experiments are required to confirm these findings.

6.5 Conclusion

This work suggests that phi6 can be used as a convenient surrogate to generate a preliminary evaluation of the effectiveness of different anti-SARS-CoV-2 therapies and treatments. Two such cases are that of the ART24 compound, which shows promise in preventing the infection of cells with the virus, and different milk heat treatments that are commonly used in dairy industries across the world. Further work investigating the impact of low temperature used in milk storage prior to thermo-treatments may provide value insights into any potential protection offered to the viruses. Future studies may also benefit to examine the implications on the nucleocapsid following human consumption given the body temperature is 37°C where work has found detrimental changes begin at 35°C. Obviously, final tests would have to be performed on SARS-CoV-2 to validate the findings generated with phi6.

Nonetheless, it is evident that there are many advantages to using phi6 as a safe surrogate for screening compounds and treatments that could be used in the fight against SARS-CoV-2. It allows for a large range and number of compounds to be tested as it can be used in BSL-1 laboratories. As the results can be directly applied from phi6 to coronavirus, it restricts any potential mutations in the pathogen. The use of phi6 as a model for enveloped viruses, in particular SARS-CoV-2, offers researchers across the globe a more accessible way to overcome the current pandemic.

6.6 References

Altmeyer, Ralf. 2004. "Virus Attachment and Entry Offer Numerous Targets for Antiviral Therapy." *Current Pharmaceutical Design* 10 (30): 3701–12. https://doi.org/10.2174/1381612043382729.

Barros, Joana, Maria Pia Ferraz, and Fernando Jorge Monteiro. 2021. "Bacteriophage Phi 6 as Surrogate and Human-Harmless Viruses to Study Anti-SARS-CoV-2 Approaches," 3.

Brian, D. A., and R. S. Baric. 2005. "Coronavirus Genome Structure and Replication." *Coronavirus Replication and Reverse Genetics* 287: 1–30. https://doi.org/10.1007/3-540-26765-4_1.

Bukasov, Rostislav, Dina Dossym, and Olena Filchakova. 2021. "Detection of RNA Viruses from Influenza and HIV to Ebola and SARS-CoV-2: A Review." *Analytical Methods* 13 (1): 34–55. https://doi.org/10.1039/D0AY01886D.

Callanan, Julie, Stephen R. Stockdale, Evelien Adriaenssens, Jens H. Kuhn, Mark J. Pallen, Janis Rumnieks, Andrey N. Shkoporov, Lorraine A. Draper, R. Ross, and Colin Hill. 2021. Rename One Class (*Leviviricetes* - Formerly *Allassoviricetes*), Rename One Order (*Norzivirales* - Formerly *Levivirales*), Create One New Order (*Timlovirales*), and Expand the Class to a Total of Six Families, 420 Genera and 883 Species. https://doi.org/10.13140/RG.2.2.25363.40481.

Chowdhury, Trinath, Piyush Baindara, and Santi M. Mandal. 2021. "LPD-12: A Promising Lipopeptide to Control COVID-19." *International Journal of Antimicrobial Agents* 57 (1): 106218. https://doi.org/10.1016/j.ijantimicag.2020.106218.

Doremalen, Neeltje van, Trenton Bushmaker, William B. Karesh, and Vincent J. Munster. 2014. "Stability of Middle East Respiratory Syndrome Coronavirus in Milk." *Emerging Infectious Diseases* 20 (7): 1263–64. https://doi.org/10.3201/eid2007.140500.

Fedorenko, Aliza, Maor Grinberg, Tomer Orevi, and Nadav Kashtan. 2020. “Survival of the Enveloped Bacteriophage Phi6 (a Surrogate for SARS-CoV-2) in Evaporated Saliva Microdroplets Deposited on Glass Surfaces.” *Scientific Reports* 10 (1): 22419. https://doi.org/10.1038/s41598-020-79625-z.

Gendron, Louis, Daniel Verreault, Marc Veillette, Sylvain Moineau, and Caroline Duchaine. 2010. “Evaluation of Filters for the Sampling and Quantification of RNA Phage Aerosols.” *Aerosol Science and Technology* 44 (10): 893–901. https://doi.org/10.1080/02786826.2010.501351.

Jiang, Yuqian, Han Zhang, Jose A. Wippold, Jyotsana Gupta, Jing Dai, Paul de Figueiredo, Julian L. Leibowitz, and Arum Han. 2021. “Sub-Second Heat Inactivation of Coronavirus Using a Betacoronavirus Model.” *Biotechnology and Bioengineering* 118 (5): 2067–75. https://doi.org/10.1002/bit.27720.

O’Donnell, Michelle, Brian Healy, Colin Hill, R Paul Ross, Mary C Rea, Ronald Farquhar, and Laurent Chesnel. 2020. “ART24, a Novel Live Biotherapeutic Product in Development for the Prevention of CDI, Is Active against a Broad Range of C. Difficile Ribotypes in Vitro,” 1.

O’Donnell, Michelle, James W. Hegarty, Brian Healy, Sarah Schulz, Calum J. Walsh, Colin Hill, R Paul Ross, Mary C Rea, Ronald Farquhar, and Laurent Chesnel. Manuscript submitted for publication. “Identification of ART24: A Newly Characterized Strain of Bacillus Velezensis with Direct Clostridiodes Difficile Killing and Toxin Degradation Bio-Activities.” *Scientific Reports*.

Rockey, Nicole, Peter J. Arts, Lucinda Li, Katherine R. Harrison, Kathryn Langenfeld, William J. Fitzsimmons, Adam S. Lauring, *et al*. 2020. “Humidity and Deposition Solution Play a Critical Role in Virus Inactivation by Heat Treatment of N95 Respirators.” *MSphere* 5 (5): e00588-20. https://doi.org/10.1128/mSphere.00588-20.

Silverman, Andrea I., and Alexandria B. Boehm. 2020. “Systematic Review and Meta-Analysis of the Persistence and Disinfection of Human Coronaviruses and Their Viral Surrogates in Water and Wastewate.” *Environmental Science & Technology Letters*, May. https://europepmc.org/articles/PMC7294895.

Simon, Miriam, Michael Veit, Klaus Osterrieder, and Michael Gradzielski. 2021. “Surfactants – Compounds for Inactivation of SARS-CoV-2 and Other Enveloped Viruses.” *Current Opinion in Colloid & Interface Science* 55 (October): 101479. https://doi.org/10.1016/j.cocis.2021.101479.

Sur, Sreyoshi, Tod D. Romo, and Alan Grossfield. 2018. “Selectivity and Mechanism of Fengycin, an Antimicrobial Lipopeptide from Molecular Dynamics.” *The Journal of Physical Chemistry. B* 122 (8): 2219–26. https://doi.org/10.1021/acs.jpcb.7b11889.

Vatter, Petra, Katharina Hoenes, and Martin Hessling. 2021. “Photoinactivation of the Coronavirus Surrogate Phi6 by Visible Light.” *Photochemistry and Photobiology* 97 (1): 122–25. https://doi.org/10.1111/php.13352.

Vries, Rory D. de, Katharina S. Schmitz, Francesca T. Bovier, Camilla Predella, Jonathan Khao, Danny Noack, Bart L. Haagmans, *et al*. 2021. “Intranasal Fusion Inhibitory Lipopeptide Prevents Direct-Contact SARS-CoV-2 Transmission in Ferrets.” *Science* 371 (6536): 1379–82. https://doi.org/10.1126/science.abf4896.

Wang, Yulong, Xiaoyu Wu, Yihua Wang, Bing Li, Hao Zhou, Guiyong Yuan, Yan Fu, and Yongzhang Luo. 2004. “Low Stability of Nucleocapsid Protein in SARS Virus.” *Biochemistry* 43 (34): 11103–8. https://doi.org/10.1021/bi049194b.

Wei, Xiping, Julie M. Decker, Hongmei Liu, Zee Zhang, Ramin B. Arani, J. Michael Kilby, Michael S. Saag, Xiaoyun Wu, George M. Shaw, and John C. Kappes. 2002. “Emergence of Resistant Human Immunodeficiency Virus Type 1 in Patients Receiving

Fusion Inhibitor (T-20) Monotherapy." *Antimicrobial Agents and Chemotherapy* 46 (6): 1896–1905. https://doi.org/10.1128/AAC.46.6.1896-1905.2002.

Yuan, Lvfeng, Shuai Zhang, Yongheng Wang, Yuchen Li, Xiaoqing Wang, and Qian Yang. 2018. "Surfactin Inhibits Membrane Fusion during Invasion of Epithelial Cells by Enveloped Viruses." *Journal of Virology* 92 (21): e00809-18. https://doi.org/10.1128/JVI.00809-18.

Thesis summary and future work

RNA phages have played a vital role in molecular biology, from determining the genetic code to deciphering translation and replication, but they are an understudied and underrepresented portion of the global phage community. During this PhD and detailed in this thesis, it has become apparent that when examining both human and environmental phageomes, the diversity and abundance of +ssRNA phages requires careful consideration. It is also evident that these phages, in particular dsRNA phages, offer an exciting range of applications, especially during this pandemic era.

Chapter I reviews our current understanding of these entities. It details the problems encountered in the detection of RNA phages in comparison to their DNA counterparts and illustrates how few studies enrich for these viruses. It is a timely compilation of the existing and relevant information about RNA phages which provides insights into their key biological features. It details how this can be manipulated to enable the sequence-based discovery and subsequent annotation of RNA phage genomes.

Chapter II provides an evaluation of the methodologies used in human gut virome analysis and their associated biases in relation to RNA virus recovery, especially in terms of RNA phage yields. It highlights how different extraction protocols, including those used in highly influential human virome papers, introduced biases that may affect the conclusions drawn from these experiments. It also discusses the possibility of generating a standardised and unbiased method for examining the human virome.

Chapter III reveals the expansion of +ssRNA phages from a range of environmental and aquatic metatranscriptomic samples, across America, Austria, Japan, and Singapore, using a specialised HMM detection tool. The search yielded 15,611 new partial-genomes and 1,015 near-complete +ssRNA phages. This expansion enabled comparisons of these viruses to reveal

two highly distinct lineages that share a conserved genome architecture and display no evidence of homologous recombination or genome mosaicism.

Chapter IV details the generation of a comprehensive taxonomy for +ssRNA phages that was accepted by the International Committee on Taxonomy of Viruses (ICTV) in March 2021. This work built upon the previous chapter where a novel, flexible taxonomic scheme was proposed to incorporate the diverse range of +ssRNA phages. It describes the classification and taxon criteria based on a combination of HMM cluster profiles and pairwise amino acid identity (PAAI) of different core proteins.

Chapter V determines the effectiveness of different methods at recovering spike-in controls of +ssRNA phages, MS2 and Qbeta. These include a typical PEG Method, a Filtration Method, and an innovative Biopsy Method, each with an RNA-enriched and a DNA-enriched portion. Following the extraction method, the recovery of MS2 and Qbeta was examined, along with the richness and diversity associated with each method. The previously described HMM search tool was applied to these samples and those associated with an in-house study to identify novel and known +ssRNA phages. This work suggests that the Biopsy Method, a high-throughput, cost-effective protocol, would offer researchers an efficient and robust means to comprehensively study the human phageome.

Chapter VI assesses a range of viable treatment options for coronavirus using dsRNA phage phi6 as a safe surrogate model. It was found that exposure to lipopeptide ART24 and a variety of important temperatures rendered phi6 inactive while having no or limited effect on the alternative phage controls. This chapter demonstrates the myriad potential of RNA phages have as model systems in not only molecular biology but in food and medicine.

The development of search tools for specific viruses, especially phages, could prove pivotal to expanding this area of research. With the creation of the HMM search tool for +ssRNA phages in Chapter III comes a wide range of possible applications, including using

this framework to build similar search tools for other phages, such as crAssphage. It could also be applied in the search for viruses and other microbes of interest from different samples and niches, including identifying pathogenic bacteria from sequenced clinical samples. This, coupled with the ever-evolving bioinformatics methodologies, should enable us to gain a better and more detailed understanding of the difficult to isolate members of the human and global phageome.

Another change that could be implemented in future studies relating to RNA phages, is examining their abundance and diversity in different sections of the GIT, in of the human GIT. Recent work by Shkoporov *et al* (2021) found differences in the virome and phageome composition between samples acquired along the GIT in mammals (Shkoporov *et al*. 2021). It seems plausible that the use of faecal samples to solely represent the entire virus community fails to capture the entire composition and complexity of the human phageome, especially the RNA fraction.

Future applications of the RNA phages and their products are expansive, including nanotechnology, vaccinology, and evolutionary and environmental studies. As seen in Chapter VI, these phages offer themselves as surrogates in assessing treatments for human and animal pathogenic viruses such as influenza, HIV, and hepatitis. Given the global implications of one virus, SARS-CoV-2, having access to safe and robust model systems will enable a more rapid response to potential virus outbreaks of the future. RNA phages could also lend themselves to advancing vaccine science as they are being studied as optimal vaccine delivery vehicles. Both MS2 and Qbeta have been manipulated for vaccines against HIV infection (Sun *et al*. 2011; Peabody *et al*. 2008) and drug dependency (McCluskie *et al*. 2015) by fusing different antigens to their coat proteins, which could provide a vast amount of vaccine varieties.

This work signifies that the richness and assortment of RNA phages is only beginning to be explored with vast amounts expected to be recorded in coming years. Understanding the

biology of these entities is crucial for uncovering the true diversity of the human and global microbiome, as well as the foundations of exciting future applications.

References

McCluskie, Michael J., Jennifer Thorn, David P. Gervais, David R. Stead, Ningli Zhang, Michelle Benoit, Janna Cartier, *et al*. 2015. "Anti-Nicotine Vaccines: Comparison of Adjuvanted CRM197 and Qb-VLP Conjugate Formulations for Immunogenicity and Function in Non-Human Primates." *International Immunopharmacology* 29 (2): 663–71. https://doi.org/10.1016/j.intimp.2015.09.012.

Peabody, David S., Brett Manifold-Wheeler, Alexander Medford, Sheldon K. Jordan, Jerri do Carmo Caldeira, and Bryce Chackerian. 2008. "Immunogenic Display of Diverse Peptides on Virus-like Particles of RNA Phage MS2." *Journal of Molecular Biology* 380 (1): 252–63. https://doi.org/10.1016/j.jmb.2008.04.049.

Shkoporov, Andrey, Stephen Stockdale, Aonghus Lavelle, Ivanela Kondova, Cara Hueston, Aditya Upadrasta, Ekaterina Khokhlova, *et al*. 2021. "Viral Biogeography of Gastrointestinal Tract and Parenchymal Organs in Two Representative Species of Mammals." https://doi.org/10.21203/rs.3.rs-803286/v1.

Sun, Shipeng, Wenli Li, Yu Sun, Yang Pan, and Jinming Li. 2011. "A New RNA Vaccine Platform Based on MS2 Virus-like Particles Produced in Saccharomyces Cerevisiae." *Biochemical and Biophysical Research Communications* 407 (1): 124–28. https://doi.org/10.1016/j.bbrc.2011.02.122.

Appendix I

An Insight into the Elusive RNA Bacteriophages

This piece was published as an article in Capsid & Tail.

https://phage.directory/capsid/rna-phage-expansion

The delight of finding Wally amongst the crowd of people at the funfair is similar to that when you detect a rare RNA bacteriophage in metagenomic or metatranscriptomic samples. These phages were originally identified in 1961 by Loeb and Zinder when they isolated a phage which had an RNA genome as opposed to the typical DNA. Since their initial discovery, RNA phages have served as important molecular models for understanding some of biology's most intricate molecular pathways such as gene regulation, transcription, and translation. They have also been central to many molecular milestones such as the first gene (the coat protein gene) to be sequenced in 1967 and the first entire genome to be fully sequenced in 1976.

A little insight into these little phages

Despite their historical significance, there is very little known about RNA phages, with only two families described in the literature, the *Cystoviridae*, with dsRNA genomes, and the *Leviviridae*, which have ssRNA genomes. *Cystoviridae* have tri-segmented genomes, which generally range from 12.7 to 15.0 kbp in length, enclosed in a protein envelope. The three segments are organised to encode different functional units. In the latest ICTV report, there is only one recognised genus, *Cystovirus*, with a single species. Additionally, within the latest ICTV report, there are six more phages belonging to *Cystoviridae*, with several more being associated with this family.

Leviviridae have a positive-sense, single-stranded genome of approximately 4,000 bp. This family is currently separated into two genera, *Levivirus* and *Allolevivirus*, primarily based on whether they have three or four genes. There are 25 complete and partial ssRNA phage genomes belonging to these two genera in the latest ICTV report, and 32 more sequences recognised as potential *Leviviridae*.

Importance of RNA phages

General interest in all things phage has exploded in recent years, as studies have uncovered the important role they play in shaping and structuring bacterial communities. Another important feature is their potential application as therapeutics in a post-antibiotic era. RNA phages have been found to target a variety of bacteria, including those listed by the World Health Organisation (WHO) as some of the deadliest pathogens such as *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Klebsiella pneumoniae*.

Problems with trying to find them

However, RNA phages have remained somewhat enigmatic in phageome studies, where extraction and isolation methods may be biased in favour of DNA phages. Another problem associated with these phages is the delicate nature of RNA and the ubiquitous presence of RNases in environments. It may also be that bioinformatics protocols during downstream processing do not capture RNA phage sequences. These issues have limited the expansion in the knowledge and diversity of these interesting groups of phages. The development of an optimised protocol for the isolation and characterisation of RNA phages should significantly improve our attempts in this field.

Just the beginning

A recent study by Krishnamurthy *et al*. in 2016 (Krishnamurthy *et al*. 2016) showed that RNA phages may be more abundant than previously believed as they noted the partial genomes of five cystoviruses and 138 leviviruses through the mining of metatranscriptomic datasets. This expansion also revealed novel hosts including a Gram-positive bacterium. This was one of the first papers I read as I began my PhD journey, and it was where my interest in these unusual and understudied phages peaked. How can we progress with our understanding of the human

phageome without taking RNA phages into account? To address the potential issue bioinformatics protocols, we built a specific ssRNA-phage search tool. Application of this method to a variety of metatranscriptomic samples, we expanded the available sequences by 60-fold, as well as enabling us to examine their genome structure and phylogenetic relationships (Callanan *et al*. 2020).

Given that eukaryotic RNA viruses make up the largest portion of the total human virome, it seems unlikely that there would be so few RNA phages associated with us and our microbiome. The search for these phages and their roles is just beginning in an RNA phage-based renaissance.

References

Callanan, Julie, Stephen R. Stockdale, Andrey N. Shkoporov, Lorraine A. Draper, R. Paul Ross, and Colin Hill. 2020. "Expansion of Known SsRNA Phage Genomes: From Tens to over a Thousand." *Science Advances* 6 (6): eaay5981. https://doi.org/10.1126/sciadv.aay5981.

Krishnamurthy, Siddharth R., Andrew B. Janowski, Guoyan Zhao, Dan Barouch, and David Wang. 2016. "Hyperexpansion of RNA Bacteriophage Diversity." *PLOS Biology* 14 (3): e1002409. https://doi.org/10.1371/journal.pbio.1002409.

Appendix II

Rename one class (*Leviviricetes* - formerly *Allassoviricetes*), rename one order (*Norzivirales* - formerly *Levivirales*), create one new order (*Timlovirales*), and expand the class to a total of six families, 420 genera and 883 species

The total body of work is available as a technical report on ResearchGate and on the ICTV website, I have selected the main text of the proposal to be included here.

Callanan, Julie, Stephen R. Stockdale, Evelien M. Adriaenssens, Jens H. Kuhn, Mark Pallen, Janis Rumnieks, Andrey N. Shkoporov, Lorraine A. Draper, R. Paul Ross, and Colin Hill. "Rename one class (*Leviviricetes*-formerly *Allassoviricetes*), rename one order (*Norzivirales*-formerly *Levivirales*), create one new order (*Timlovirales*), and expand the class to a total of six families, 420 genera and 883 species." (2021).

http://dx.doi.org/10.13140/RG.2.2.25363.40481

Part 3: TAXONOMIC PROPOSAL

Name of accompanying Excel module

2020.095B.R.Leviviricetes.xlsx

Abstract

The relatively simple genome architecture of all bacterial positive-sense single-stranded (+ssRNA) viruses identified to date contain three core genes; a maturation protein (MP), a coat protein (CP), and an RNA-directed RNA polymerase (RdRP), in that order. We present the characterization of 1,868 near-complete bacterial +ssRNA virus genomes, defined as sequences encoding a MP with a minimum length of 350 amino acids and an RdRP greater than 500 amino acids.

As nucleotide sequences are poorly conserved between bacterial +ssRNA, following the existing demarcation criteria for viruses classified in family *Leviviridae* (*Allassoviricetes*: *Levivirales*): pairwise amino-acid comparisons of the RdRP for species and genus demarcation were determined as 80% and 50% identity, respectively (Callanan *et al*. 2020). Profile hidden Markov models (HMMs) were used to detect more distant relationships between core bacterial +ssRNA virus proteins.

Phylogenetic relationships between bacterial +ssRNA virus RdRPs are in agreement with protein clustering, resulting in a proposed taxonomic structure of two orders, six families, 420 genera, and 883 species, encompassed within a single class.

Text of proposal

Species demarcation criteria

We have chosen 80% pairwise amino-acid identity of the viral core-encoded RdRP protein as the criterion for establishing species (Figure 1). This cutoff was applied in a bottom-up approach to 1,868 +ssRNA viruses that met specific criteria, including a minimum-length maturation protein (350 amino acids) and RdRP (500 amino acids). The 1,868 sequences originated from NCBI available sequences, and the studies of Callanan *et al.*, Starr *et al.*, Shi *et al.*, and Krishnamurthy *et al.* (Callanan *et al*. 2020; Starr *et al*. 2019; Shi *et al*. 2016; Krishnamurthy *et al*. 2016). This yielded 883 species, with all sequences assigned a species membership contained within a distinct genus.

Genus demarcation criteria

We determined that the currently classified bacterial +ssRNA viruses, which are presently classified in genera *Levivirus* and *Allolevivirus*, share 50% amino-acid identity in their RdRPs. Applying this criterion, the 1,868 bacterial +ssRNA viruses clustered into 420 genera. All sequences classified with the 420 genera are contained within a distinct family taxon.

New higher taxa and naming origins

Order and family names, which are derived from scientists that studied bacterial +ssRNA viruses, were arbitrarily assigned to groups. No scientist's name was deliberately associated with a particular group of viruses, in order to prevent author bias towards interpreting the merits or achievements of any individual.

Class

***Leviviricetes* (formerly named *Allassoviricetes*)**: The class is based on the current highest taxonomic rank encompassing all bacteria-infecting +ssRNA viruses that share the same genome architecture of their three core genes. Previous analysis of +ssRNA viruses suggested that bacterial-specific +ssRNA viruses form two distinct groups (Figure 2) (Wolf *et al*. 2018).

Orders

***Norzivirales* (formerly named *Levivirales*):** This order is based on the phylogeny and clustering of bacterial +ssRNA virus RdRP protein sequences. It is named after Norton Zinder (1928-2012), who isolated the first bacterial virus that contained RNA as its genetic material and continued to make crucial findings about these entities.

Timlovirales: This order is based on the phylogeny and clustering of bacterial +ssRNA virus RdRP protein sequences. It is named after Timothy Loeb (1935-2016) who, with Norton Zinder, isolated the first bacterial +ssRNA virus.

Families

Familial taxonomic groups were based on distinct phylogeny of bacterial +ssRNA virus RdRP protein sequences, which is supported by coat protein (CP) clustering using OrthoMCL. There are nine instances (out of 883 bacterial +ssRNA viruses) for which the predicted CP cluster did not confidently match its predicted corresponding RdRP cluster (difference in RdRP E-values < 1E-10). Therefore, no order or familial taxonomic rank is designated for these bacterial +ssRNA viruses until they are further investigated (see example AVE006, Figure 4).

Atkinsviridae: named after John Atkins (1944-present) for his discovery of the lysin protein from Escherichia virus MS2.

Blumeviridae: named after Thomas Blumenthal (1943-present) for his findings on the replication of bacterial ssRNA viruses, in particular the structure and function of the replicase.

Duinviridae: named after Jan van Duin (1937-2017) for his discoveries related to novel bacterial ssRNA viruses and RNA folding within bacterial ssRNA viruses.

***Fiersviridae* (formerly named *Leviviridae*)**: named after Walter Fiers (1931-2019), who sequenced the first gene and genome of any organism, *Escherichia* virus MS2.

Solspiviridae: named after Sol Spiegelman (1914-1983) who discovered an RNA chain of only 218 nucleotides that could be reproduced by an RdRP.

Steitzviridae: named after Joan Argetsinger Steitz (1941-present) for her determination of an initiation sequence that is central to modern-day ribosome profiling.

Genus and species name generation

Genera

Establishing a nomenclature for the 420 proposed genera was conducted as follows: A bacterial +ssRNA virus was chosen to represent the genus if (1) it was previously described and available in the ICTV archives, (2) its sequence had been deposited in GenBank, (3) or it was the longest contig sequence of all remaining available.

Some genera names for isolated phages were manually designed based on their current type species exemplar isolate names, including their phonetics: *Emesvirus* for Escherichia virus MS2, *Qubevirus* for Escherichia virus Qbeta, *Pepevirus* for Pseudomonas virus PP7, *Cunavirus* for virus C-1, *Empivirus* for the M-pili dependent virus, *Hagavirus* for enterobacteria virus Hgal1, *Perrunavirus* for Pseudomonas virus

PRR1, *Apeevirus* for Acinetobacter virus AP205, and *Cebevirus* for Caulobacter virus Cb5.

While several additional unique genera names were generated manually, others were manipulated using three different scripts, written to subtly alter the spelling of chosen terms. For bacterial +ssRNA virus sequences from the Starr *et al*. study (Starr *et al*. 2019), random grass names were chosen from a list of world grasses, as this was the plant-soil interaction study. All metagenomically assembled bacterial +ssRNA virus sequences identified in the Callanan *et al*. study include within their strain name the accession code for the raw sequence reads (i.e., SRR1234567) (Callanan *et al*. 2020). This code also enables the tracking of each sequence to its original study location. Unique names were therefore derived from the sequence's original study location by modifying the anglicized names of cities, towns, or villages.

The name-modifying scripts altered terms as follows. Each letter of the term to be mutated was randomly assigned number 1, 2, or 3. Characters were then passed through a three-step script to create the following changes:

- *a1→ah, b1→p, c1→k, d1→t, e1→eh, g1→j, i1→ih, k1→c, o1→oh, t1→d, u1→uh*
- *a2→e, e2→i, i2→o, o2→u, u2→a*
- *a3→i, e3→o, i3→u, o3→a, u3→e, j3→g, k3→c, p3→b, t3→d*

To maximize the likelihood that the mutated term was pronounceable, only the first occurrence of a repeating letter was kept. All long terms were truncated after the seventh character and shortened further to the last occurring vowel, if needed (to prevent a hard consonant before the genus level suffix "-*virus*"). Each mutated name needed to be a minimum of five characters in length and contain two consonants and two vowels.

All names were checked against the ICTV Species Master List 2019.v1 to ensure the uniqueness of taxon name word stems (Realm→Species).

As an example, a unique genus name was derived from a representative phage that was isolated from a metagenomic study of Japanese environmental samples. A randomly chosen Japanese city, Kakunodate, was modified to ultimately generate the proposed genus name *Kecuhnavirus*.

Species

Binomial species names were generated by combining the genus name with a Latin species epithet based on a characteristic of the exemplar isolate or characteristic of the sample it was found in. The etymology of the species epithets is indicated in the comments section of the Excel module. The species naming was inspired by the preprint on Latin binomials for bacteria and archaea (Pallen, Telatin, and Oren 2021).

* The authors would like to acknowledge and thank Aharon Oren for corrections of the Latin grammar of the proposed species epithets.

Taxonomy assigning profile Hidden Markov Models

Profile hidden Markov models (HMMs designed to detect bacterial +ssRNA viral proteins were first presented in the Callanan *et al.* study (Callanan *et al*. 2020). Updated HMMs are now available at *https://figshare.com/articles/dataset/Bacterial_ssRNA_virus_Hidden_Markov_Models/12745394*. These HMMs are designed to aid researchers in finding bacterial +ssRNA viruses and inferring higher taxonomic assignments. Concatenated HMM profiles for bacterial +ssRNA virus proteins detect two RdRP protein clusters, three maturation-protein clusters, and nine CP clusters.

HMM profile searches return order and family taxonomic information for RdRP and CP hits, respectively. By curating the HMM search output to determine the best hit, a scheme for rapidly advancing bacterial +ssRNA phage taxonomy is available: RdRP_A→*Norzivirales*; RdRP_B→*Timlovirales*; CP_A, CP_B, and CP_H→*Fiersviridae*; CP_C→*Atkinsviridae*; CP_D and CP_F→*Steitzviridae;* CP_E→*Blumeviridae;* CP_G→*Solspiviridae*; and CP_AP205-like→*Duinviridae*.

The phylogeny of bacterial +ssRNA viral RdRP proteins agree with the current RdRP and CP clusters used to generate the taxonomy assigning HMM profiles (Figure 3).

Supporting evidence

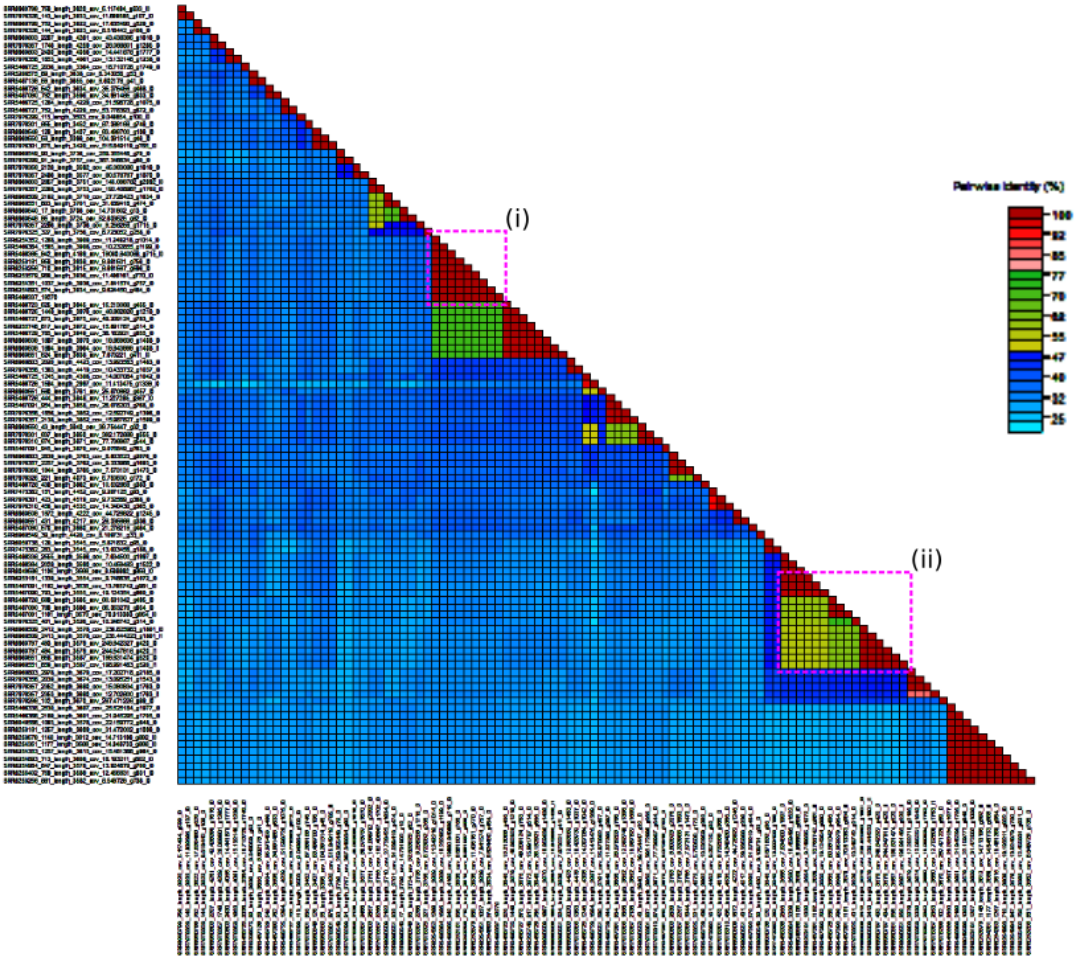


Figure 1. Example of species and genus demarcation cutoffs of 80% and 50%, respectively, applied to pairwise RNA-directed RNA polymerase (RdRP) amino-acid comparisons. The pairwise amino-acid comparisons of the RdRP protein sequences for the members of the proposed *Atkinsviridae*. The image inset dotted-box **(i)** shows a distinct species clustering (red-colored boxes), whereas the dotted-box **(ii)** shows three species represented by multiple sequences and a species representing a single sequence, clustered into a genus (yellow-green-colored boxes). Pairwise comparisons in shades of blue do not meet the species or genera clustering criteria.

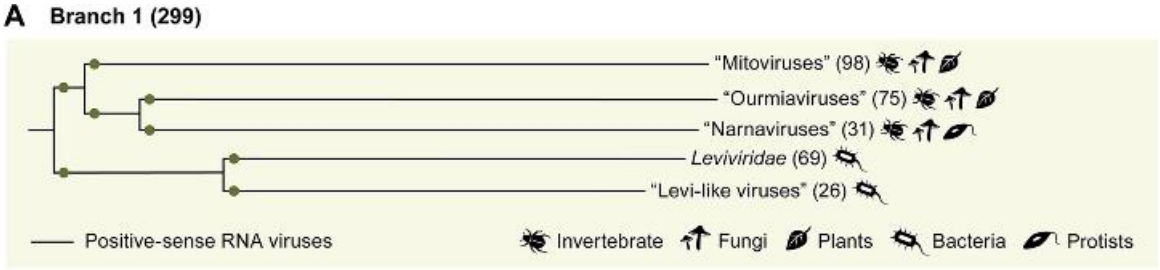


Figure 2. Phylogenetic tree of positive-sense single-stranded (+ssRNA) viral RNA-directed RNA polymerases (RdRPs) of *Leviviricetes*. This information was sourced from Figure 2A of Wolf *et al.* (2018) (Wolf *et al*. 2018). This phylogenetic tree indicates the predicted separation of bacterial +ssRNA viruses into two clades, termed "*Leviviridae*" and "Levi-like viruses". The numbers in parentheses indicate approximately how many distinct virus RdRPs are present in each respective branch. Symbols to the right indicate presumed virus host(s). Olive-green dots indicate that these branches are well-supported (≥0.7).

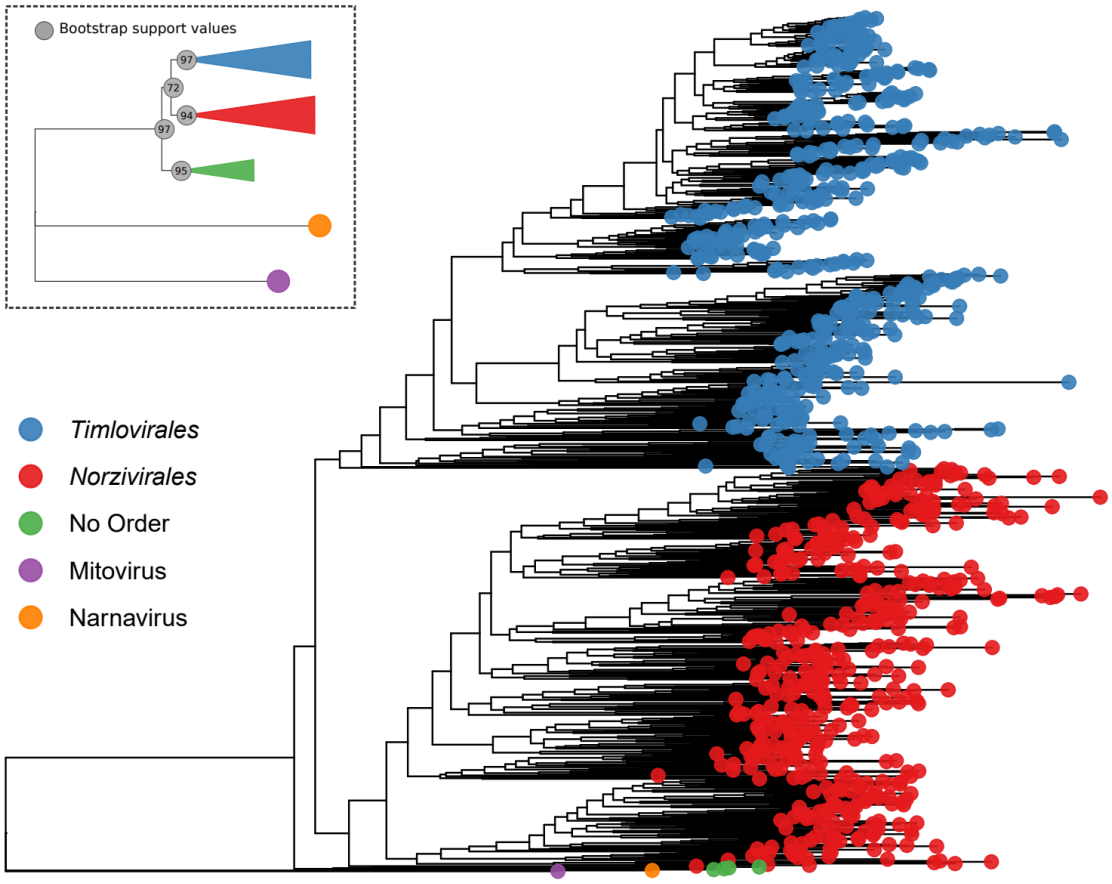


Figure 3. Phylogenetic analysis of bacterial positive-sense single-stranded (+ssRNA) virus RNA-directed RNA polymerase (RdRP) protein sequences. Mitovirids and narnavirids were used to root the bacterial +ssRNA viral RdRP tree, generated using maximum-likelihood-based phylogenetic reconstruction in IQ-TREE with the VT+F+R10 model and 1,000 bootstrap replicates. RdRP sequences used to generate the tree were made non-redundant at 95% BLASTp identity across 95% of coverage length. The image inset, top left, shows a simplified version of the phylogenetic tree with bootstrap support values for the major branches. Sequences without specific corresponding coat protein and RdRP sequences, and which were not assigned order and familial taxonomic ranks (see text), are highlighted as "No Order". RdRP phylogeny is the demarcation criteria proposed for establishing the *Norzivirales* and *Timlovirales* +ssRNA viral orders.

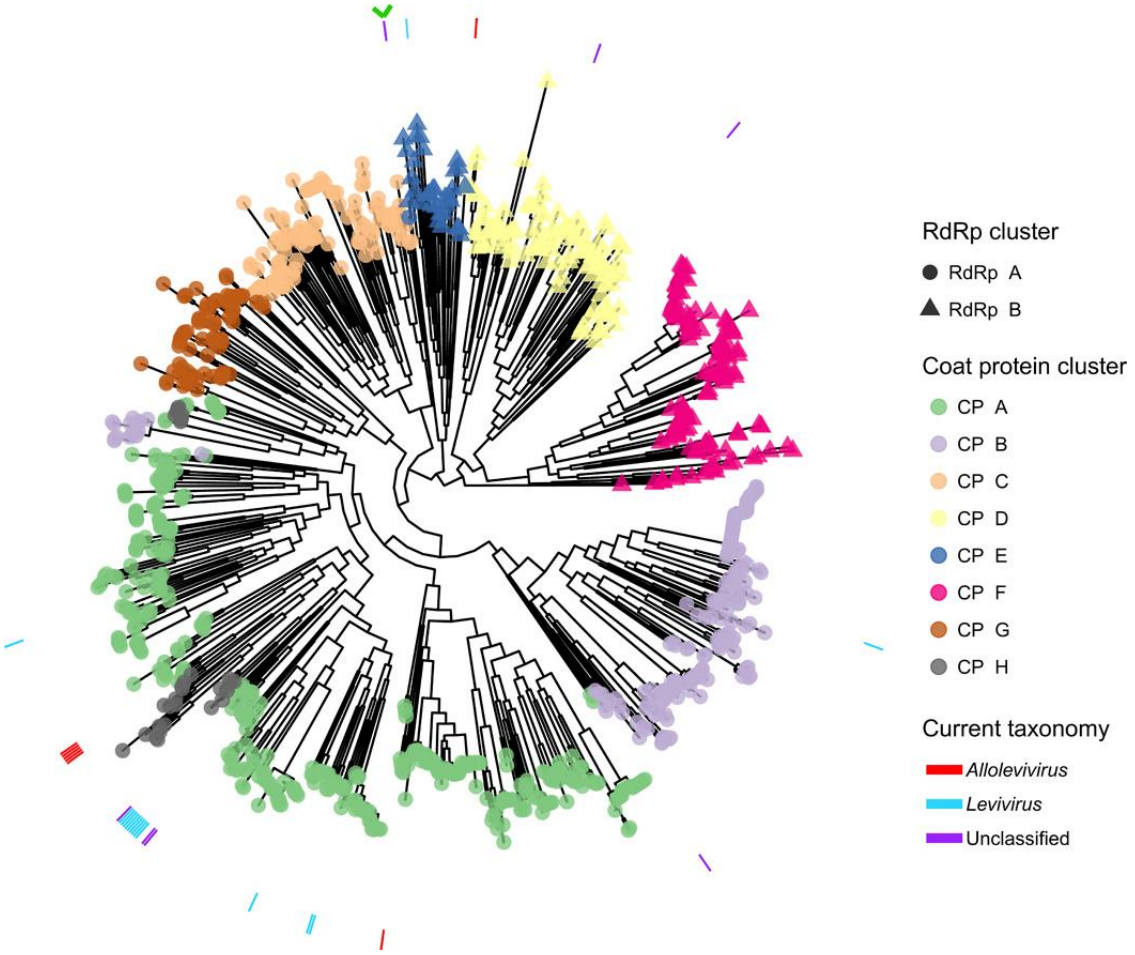


Figure 4. Phylogenetic assessment of bacterial positive-sense single-stranded (+ssRNA) virus core proteins. This information was sourced from Figure 3 of Callanan *et al.* (2020) (Callanan *et al*. 2020). Phylogeny of concatenated bacterial +ssRNA viral maturation protein (MP), coat protein (CP), and RNA-directed RNA polymerase (RdRP) sequences, which closely agree with the phylogeny of RdRP alone. Twenty-nine previously characterized and 1,015 newly identified viruses were included in this core protein phylogenetic analysis. Branch tip shapes indicate the specific RdRP protein cluster: circular = *Norzivirales*, triangular = *Timlovirales*, while branch tip colors indicate CP clusters. The +ssRNA viral CP clusters are used as the family demarcation criteria for *Leviviricetes* viruses. The family *Fiersviridae* is represented by CP clusters CP_A, CP_B, and CP_H, the family *Steitzviridae* by clusters CP_D and CP_F, while all other families are represented by singular coat protein clusters. The encircling annotation ring depicts *Leviviridae* ICTV taxonomy (ICTV Master Species List

2018.v2). A green arrowhead points to virus AVE006, which encodes a unique RdRP and CP association and is therefore not assigned to an order or family within the *Leviviricetes* proposal.

References

Callanan, Julie, Stephen R. Stockdale, Andrey N. Shkoporov, Lorraine A. Draper, R. Paul Ross, and Colin Hill. 2020. "Expansion of Known ssRNA Phage Genomes: From Tens to over a Thousand." *Science Advances* 6 (6): eaay5981. https://doi.org/10.1126/sciadv.aay5981.

Krishnamurthy, Siddharth R., Andrew B. Janowski, Guoyan Zhao, Dan Barouch, and David Wang. 2016. "Hyperexpansion of RNA Bacteriophage Diversity." *PLOS Biology* 14 (3): e1002409. https://doi.org/10.1371/journal.pbio.1002409.

Pallen, Mark J., Andrea Telatin, and Aharon Oren. 2021. "The Next Million Names for Archaea and Bacteria." *Trends in Microbiology* 29 (4): 289–98. https://doi.org/10.1016/j.tim.2020.10.009.

Shi, Mang, Xian-Dan Lin, Jun-Hua Tian, Liang-Jun Chen, Xiao Chen, Ci-Xiu Li, Xin-Cheng Qin, *et al*. 2016. "Redefining the Invertebrate RNA Virosphere." *Nature* 540 (7634): 539–43. https://doi.org/10.1038/nature20167.

Starr, Evan P., Erin E. Nuccio, Jennifer Pett-Ridge, Jillian F. Banfield, and Mary K. Firestone. 2019. "Metatranscriptomic Reconstruction Reveals RNA Viruses with the Potential to Shape Carbon Cycling in Soil." *Proceedings of the National Academy of Sciences* 116 (51): 25900–908. https://doi.org/10.1073/pnas.1908291116.

Wolf, Yuri I., Darius Kazlauskas, Jaime Iranzo, Adriana Lucía-Sanz, Jens H. Kuhn, Mart Krupovic, Valerian V. Dolja, and Eugene V. Koonin. 2018. "Origins and Evolution of the Global RNA Virome." *MBio* 9 (6). https://doi.org/10.1128/mBio.02329-18.

Appendix III

Guidance on Creating Individual and Bulk Latinized Binomial Virus Species and Other Taxon Names.

This manuscript is in preparation for submission. I have included the relevant piece regarding my work. Figures and tables were generated by Dr Stephen Stockdale while I contributed to the text.

Thomas S. Postler, Luisa Rubino, Evelien M. Adriaenssens, Bas E. Dutilh, Balázs Harrach, Sandra Junglen, Andrew Kropinski, Jens H. Kuhn, Arcady Mushegian, Janis Rumnieks, Sead Sabanadzovic, Peter Simmonds, Arvind Varsani, Murilo Zerbini, Julie Callanan, Mark Pallen, Lorraine A. Draper, Colin Hill, and Stephen R. Stockdale

***Leviviricetes* – A case study of bulk name formation**

The recently ratified taxonomic proposal *Leviviricetes* updated and restructured bacterial-infecting positive-sense single-stranded RNA (+ssRNA) viruses (Chapter 4; (Callanan *et al*. Accepted for publication)). The class *Leviviricetes* (formerly *Allassoviricetes*) has been expanded from two genera containing four species to 428 genera containing 882 species. This significant expansion incorporated the vast diversity of viruses identified in recent environmental metagenomic and metatranscriptomic studies (Callanan *et al*. 2020; Starr *et al*. 2019; Shi *et al*. 2016; Krishnamurthy *et al*. 2016).

Latinized binomial species names for all 882 +ssRNA viral species were generated by combining genus names with Latin species epithets. Species epithets were based on a characteristic of the exemplar isolate or characteristic of the sample it was found in. For detailed information on the etymology of the species epithets used, see the comments section of the ICTV taxonomic proposal Excel module. Species naming followed the Latinized binomial species name formation rules outlined above and were inspired by a recent proposal to create additional Latin binomials for bacteria and archaea (Pallen, Telatin, and Oren 2021).

Establishing a nomenclature for the 428 proposed genera was conducted as follows: a bacterial +ssRNA virus was chosen to represent the genus if (1) it was previously described and available in the ICTV archives, (2) its sequence had been deposited in GenBank, (3) or it was the longest contig sequence of all remaining available. Genera names for many isolated phages were manually designed based on their current type species exemplar isolate names, including their phonetics: *Emesvirus* for *Escherichia* virus MS2, *Qubevirus* for *Escherichia* virus Qbeta, *Pepevirus* for *Pseudomonas* virus PP7, *Cunavirus* for virus C-1, *Empivirus* for the M-pili dependent virus, *Hagavirus* for enterobacteria virus Hgal1, *Perrunavirus* for *Pseudomonas* virus PRR1, *Apeevirus* for *Acinetobacter* virus AP205, and *Cebevirus* for *Caulobacter* virus Cb5.

Most genus names required for the *Leviviricetes* taxonomic proposal were to represent novel genera. Therefore, to circumvent the challenge of creating hundreds of completely original genus names *de novo*, we devised a relatively simple approach to mutate an assembled list of terms (a.k.a. strings). A simplification of the R code used to mutate terms is provided in Table 1. The logic for mutating terms, rather than using the original words, was to mitigate any potential negative connotations that may arise when viruses were named after a person, place, or thing and obtaining permission was not feasible.

Desired function	**Desired function as R code**
Count the number of characters within terms stored in df$names	nchar(x = df$names)
Randomly generate numbers between 1 and 3 based on the number of characters within terms stored in df$names	stringi::stri_rand_strings(n = nrow(df), length = nchar(df$names), pattern = "[1-3]")
Paste the first and second characters of df$names to the first and second randomly generated numbers stored in df$rnumber, without introducing spaces	paste(substr(x = df$names, start = 1, stop = 1), substr(df$rnumber, 1, 1), substr(df$names, 2, 2), substr(df$rnumber, 2, 2), sep = "")
Change the alphanumeric combination "a1" to "ah" stored in df$mix	gsub(pattern = "a1", replacement = "ah", x = df$mix)
Remove any remaining numbers from df$mix	gsub("[1-3]", "",df$mix)
Remove repeated characters from df$newTerms	gsub("([[:alpha:]])\\1+", "\\1", df$newTerms)
Keep only the first seven characters of df$newTerms	ifelse(nchar(df$newTerms) > 7, substr(df$newTerms, 1, 7), df$newTerms)
End terms within df$newTerms on a vowel	stringr::str_split_fixed(string = df$newTerms, pattern = "[^aeiou]*$", n = 2)
Count the number of vowels within terms stored in df$newTerms	nchar(gsub("[^aeiouy]", "", df$newTerms, ignore.case = TRUE))

Table 1. Examples of R functions to mutate a list of original terms, as was performed during the *Leviviricetes* taxonomic proposal. A working knowledge of the R programming language is required for completing and utilizing the provided code. The prompts indicating the components of functions are written in full during in the first example usage of each function. Abbreviations used: dataframe = df, $ = separating a dataframe and one of its columns.

Lists of terms were sourced from the corresponding publications identifying +ssRNA phage sequences, by taking basic data such as geographical location or study design. Through Google searches, lists of cities, towns, villages, or regions near the study's location (be that a country or state) were extracted from publicly available webpages and pasted into comma

separated value files for importation into R (RStudio Team 2020). Fitting with the "be creative" guidelines above, genera names for +ssRNA phages identified in the Starr *et al.* (2019) study were in fact based on random grass names as this was the plant-soil interaction study (Starr *et al*. 2019).

The list of terms imported into R were mutated using "gsub" functions. A specific seed was set for reproducibility, before each letter of the term to be mutated was randomly assigned number 1, 2, or 3 using the "stringi" package in R (Wickham and RStudio 2019). Alternating between each character and its corresponding number, an alphanumeric combination was pasted together. The letter-number combinations were then passed through a three-step script, where each step uniquely changed specific alphanumeric combinations. The specific mutations used in the *Leviviricetes* proposal are outlined in Table 2. Substitutions of specific alphanumeric combinations to a new letter were designed to minimally change the phonetics of the resultant term. Due to the case-sensitivity of R, the first letter of a term was never mutated. Furthermore, the frequency at which letters are mutated could be refined in future studies by generating alphanumeric combinations using numbers that would not be recognising by a substitution function (e.g. "a5" would not have be recognised by the *Leviviricetes'* scripts). A final "gsub" function was used to remove numbers and return mutated terms based on the starting list.

Script 1		Script 2		Script 3	
Original	**Mutated**	**Original**	**Mutated**	**Original**	**Mutated**
a1	Ah	a2	e	a3	I
b1	P	e2	i	e3	O
c1	K	i2	o	i3	U
d1	T	o2	u	j3	G
e1	Eh	u2	a	k3	C
g1	J			o3	A
i1	Ih			p3	B
k1	C			t3	D
o1	Oh			u3	E
t1	D				
u1	Uh				

Table 2. Outline of the three *Leviviricetes* scripts designed to substitute alphanumeric combinations to generate novel terms from a starting list.

To maximize the likelihood that any mutated term was pronounceable, only the first occurrence of a repeating letter was kept. All long terms were truncated after the seventh character using the "stringr" package in R (Wickham and RStudio 2019). Terms were subsequently shortened further to the last occurring vowel to prevent a hard consonant before the genus level suffix "-*virus*". Each mutated name needed to be a minimum of five characters in length and contain two consonants and two vowels. Due to these specific requirements of new potential genus names, it is recommended that the initial starting list of terms be contain 1.5x to 2.0x times the number of genus names required while avoiding short terms. All names were checked against the ICTV Species Master List 2019.v1 to ensure the uniqueness of taxon name word stems (Realm → Species). Furthermore, potential genera names were Googled, and the top hits assessed to ensure produced terms are novel. A graphical overview of the automated genera name creation strategy with example terms is provided as Figure 1.

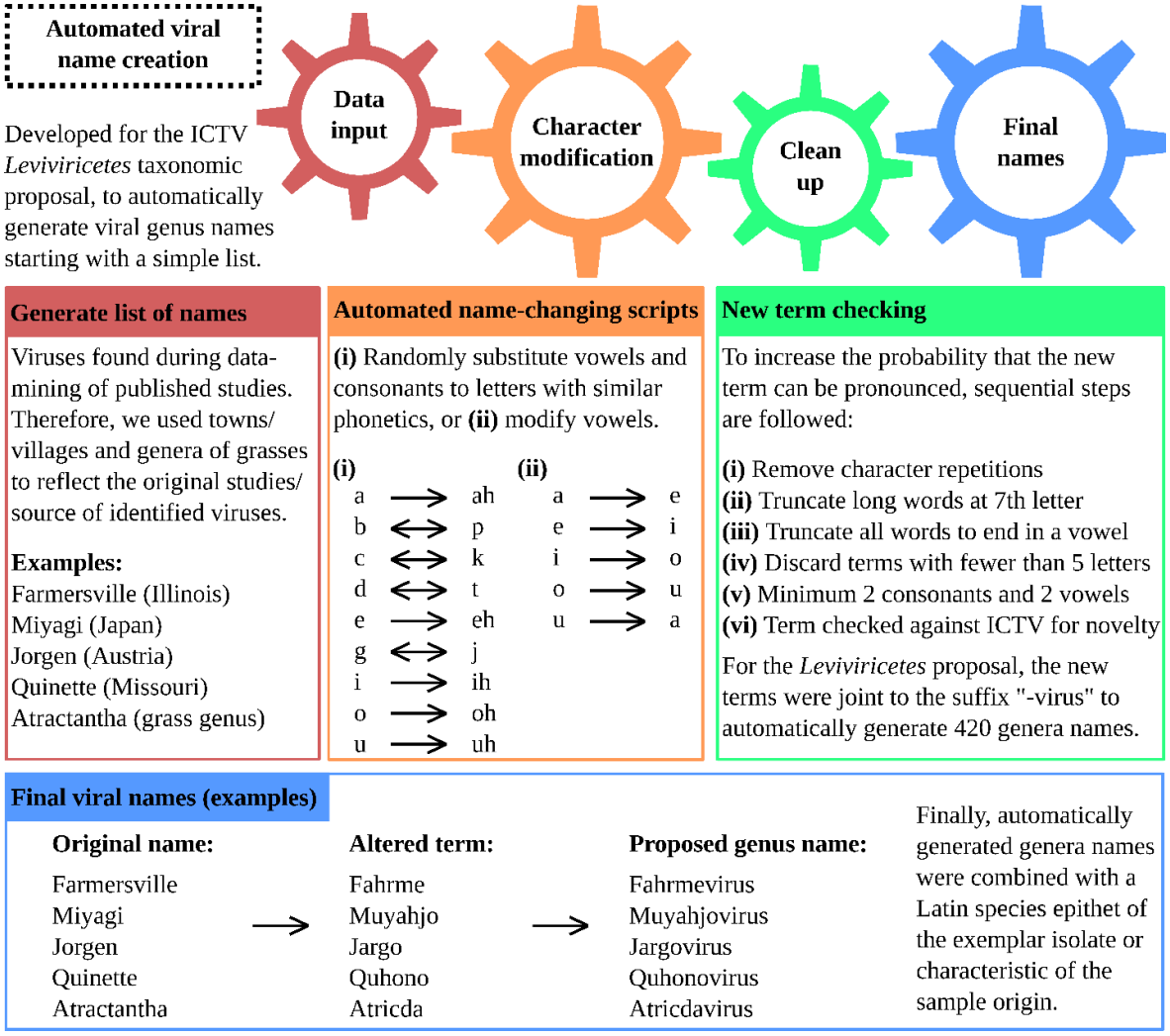


Figure 1. Graphical overview simplifying the automated name creation strategy employed to generate genus names for the *Leviviricetes* taxonomic proposal recently ratified by the ICTV.

References

Callanan, Julie, Stephen R. Stockdale, Evelien M. Adriaenssens, Jens H. Kuhn, Janis Rumnieks, Mark J. Pallen, Andrey N. Shkoporov, Lorraine A. Draper, Paul R. Ross, and Colin Hill. Accepted for publication. "*Leviviricetes*: Expanding and Restructuring the Taxonomy of Bacteria-Infecting Single-Stranded RNA Viruses." *Microbial Genomics*.

Callanan, Julie, Stephen R. Stockdale, Andrey N. Shkoporov, Lorraine A. Draper, R. Paul Ross, and Colin Hill. 2020. "Expansion of Known ssRNA Phage Genomes: From Tens to over a Thousand." *Science Advances* 6 (6): eaay5981. https://doi.org/10.1126/sciadv.aay5981.

Krishnamurthy, Siddharth R., Andrew B. Janowski, Guoyan Zhao, Dan Barouch, and David Wang. 2016. "Hyperexpansion of RNA Bacteriophage Diversity." *PLOS Biology* 14 (3): e1002409. https://doi.org/10.1371/journal.pbio.1002409.

Pallen, Mark J., Andrea Telatin, and Aharon Oren. 2021. "The Next Million Names for Archaea and Bacteria." *Trends in Microbiology* 29 (4): 289–98. https://doi.org/10.1016/j.tim.2020.10.009.

RStudio Team. 2020. "RStudio: Integrated Development for R. RStudio, PBC, Boston, MA." 2020. https://rstudio.com/.

Shi, Mang, Xian-Dan Lin, Jun-Hua Tian, Liang-Jun Chen, Xiao Chen, Ci-Xiu Li, Xin-Cheng Qin, *et al*. 2016. "Redefining the Invertebrate RNA Virosphere." *Nature* 540 (7634): 539–43. https://doi.org/10.1038/nature20167.

Starr, Evan P., Erin E. Nuccio, Jennifer Pett-Ridge, Jillian F. Banfield, and Mary K. Firestone. 2019. "Metatranscriptomic Reconstruction Reveals RNA Viruses with the Potential to Shape Carbon Cycling in Soil." *Proceedings of the National Academy of Sciences* 116 (51): 25900–908. https://doi.org/10.1073/pnas.1908291116.

Wickham, Hadley, and RStudio. 2019. *Stringr: Simple, Consistent Wrappers for Common String Operations* (version 1.4.0). https://CRAN.R-project.org/package=stringr.

Acknowledgements

I have learned that a PhD journey is only ever as good as the people you meet along the way, and I have been extremely fortunate to have met and learned from some of the best.

Firstly, to my supervisors Professor Colin Hill and Professor Paul Ross, thank you for all your guidance, support, and advice that has helped shape me into the researcher I am today – I will be forever grateful.

To everyone across the lab groups, especially those in the Gut Phage Lab, both past and present, thank you so much - every single person has offered invaluable advice, support, and moments of joy all through my years. I really won the PhD lotto.

There are a few individuals that I have met in the lab during my PhD that I want to give a special mention to. Andrey, thank you for everything you have taught me over the years. Colin B, thank you for your unwavering interest and enthusiasm for everything science related. Lorraine, thank you for always being there to help with any issue and concocting some brilliant ideas – you are the best lab manager imaginable. Neda, thank you for helping me towards the final stretch. Stephen, thank you for everything, for being the best bioinformatic sensei, and such a scientific inspiration. Aonghus, Andrei, Ciara D, Ciara T, Imme, Joan, Karen, Katia, Michelle, Muireann, Orla, Pedro, Rory, and Shona thank you all for being extraordinary sources of knowledge and fun.

Ellen, thank you for being so incredibly bright and an all-round amazing human. Emma, thank you for being my mentor at the start of my PhD and for never laughing at my questions. Lauren, thank you for being a constant source of laughter and such a thoughtful person. Olivia, thank you for never letting me forget to appreciate the smaller things in life – I think a Friday rave should be compulsory. I have no doubt our friendships will continue for many years to come.

To all the lads, the gals and everyone I've lent on over the past four years and more, I hope you know how much I value your friendship. To Claire, Dave Lee, Dinah, Katie, Naomi, Nessa, Niamh, and Sofie thank you for being constant sources of inspiration, joy, and fun over the years. To the McCarthys thank you all for your inspiring optimism and strength. To my boys: Barry, Colin, Conor, Darragh C, Darragh H, Luke, and Michael, thanks for listening to me waffle about science and for always making sure I never took anything too seriously.

To Joe and Mary O'Brien, thank you for the constant interest in my work and never tiring of taste-testing my procrastination baking.

To the O'Learys, thank you all for everything over the years. To Andrew, Elaine, Mairéad, and Niamh thank you for making me feel like family. To Dermot, thank you for having such an interest in my science and my food choices.

To my family, I don't think words will ever be enough. Grandad Neil, Grandad Peter, and Granny Teresa, thank you all for keeping an eye on everyone. Granny Sheila, thank you for always making me smile and reminding me to appreciate every moment, no matter how small. Nelson, thank you for being the best boy. James, thank you for your support and motivating determination. Ruth, thank you for always being there to listen, laugh, and inspire me, even from Australia. Mags, thank you for always being a source of motivation with your optimism, enthusiasm, and endless encouragement. Ger, thank you for always being an inspiration with your curious nature and incredible strength. To the both of you, thank you for believing in me and always being there.

And finally, to mo ghrá geal, Gar, thanks for being you. I am exceptionally lucky to have someone like you in my life with your unwavering support, belief, and incredible calmness. I cannot wait for our next set of adventures.

None of this work would have been possible without each and every one of you.