

Dynamics of gut bacteriophages in critically ill patients

By

Siu Fung Ho

A thesis submitted to
the University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

Institute of Microbiology and Infection
College of Medical and Dental Sciences
University of Birmingham
August 2023

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

ABSTRACT

Phages are vital members of the gut microbiome and play a significant role in shaping their bacterial host community via predation, lysogeny, and transduction. While previous studies have identified that certain alterations in gut phage composition are associated with several diverse health conditions, the specific dynamics governing phage-host interactions remain largely unknown. Metagenomics and viromics have allowed for the direct identification and analysis of gut phages, regardless of their host's cultivability. However, this process remains complex with significant challenges the scientific community has yet to definitively overcome.

In this study, we benchmarked ten commonly used metagenomic viral identification tools, and provided valuable metrics and insights for other researchers to use and build on. By using a mixture of artificial datasets and mock communities we were able to estimate the recall, precision, and biases of these tools. Overall, we concluded that Kraken2 should be considered when researchers are trying to identify previously characterised phages. When novel phage detection is required we suggest that Kraken2 is used in combination with machine learning based tools such as VirSorter2 and DeepVirFinder.

Subsequently, we identified viral operational taxonomic units (vOTUs)

from a metagenomic sequencing set obtained by a previous study that investigated the gut microbiome of critically ill patients. We were able to define a straightforward yet robust criteria for microbial domination in the gut microbiota and use this to identify the dominant microbe in each patient, including one that had been discovered since the publication of the original study. The relative abundance of dominant microbes and vOTUs predicted to infect them had a near-perfect correlation. This correlation was not affected by the introduction of a time lag, leading us to hypothesise that the majority of our vOTUs originated from integrated prophages. We were also able to detect VFs and ARGs encoded by our vOTUs, although we were unable to determine whether these vOTUs were actively replicating or dormant.

We then built on this by experimentally removing bacterial hosts from stool samples collected from critically ill patients from Queen Elizabeth Hospital Birmingham. in order to long-read sequence the free viral fraction alone. In contrast with the metagenomic analyses, the dominating microbial taxa did not overlap with the most abundant hosts we predicted for our vOTUs, This observation lends support to the "piggyback-the-winner" model of gut phage dynamics. We identified two vOTUs that encoded *cfxA* beta-lactamases, including one that contained an ancestral form of the Tn*4555* transposon, while the second vOTU likely stemmed from a misassembly.

The insights gleaned from this work provide a framework for evaluating the performance of viral identification tools, as well as the advantages and limitations of utilising metagenomic versus viromic sequencing to explore the dynamics of the gut phageome.

For Kobe

ACKNOWLEDGMENTS

First and foremost, I would like to thank my supervisor Professor Willem van Schaik, whose support, mentorship, and genuine investment in my research has been instrumental in helping me complete this thesis. I would also like to thank the Wellcome Trust for funding my project, without which none of this work would have been possible.

I would also like to thank the members of the van Schaik and McNally lab groups. I have learned so much from every one of you and I am grateful for the friendships we have made. I couldn't ask for a better group of people to work with and I hope that we will continue to support one another. In particular, I'd like to give a special mention Greg and Rob, as we joined the lab at a similar time, and your company made me feel so welcome. We spent many hours together working on Contamination Club, a project that I'm very proud to be a part of.



Thank you to my friends and family, especially my parents who have always believed in me and supported all aspects of my life. Finally I want to thank my amazing partner Sylvia, who has been with me every step of the way. Her passion for science and unwavering support has kept me motivated through difficult times, and I couldn't imagine completing this journey without them.

I thought it was going to be bad,

but not this bad.

- Nicole Wheeler

DECLARATION OF AUTHORSHIP

Siu Fung Ho was the first author and main contributor of the publication:
Ho, S.F.S., Wheeler, N.E., Millard, A.D. et al. Gauge your phage: benchmarking of bacteriophage identification tools in metagenomic sequencing data. Microbiome 11, 84 (2023). https://doi.org/10.1186/s40168-023-01533-x.

Results from this publication are presented in Chapter 2 of this thesis.

Contents

Page

1 Introduction **1**

1.1 Bacteriophages 2

1.1.1 Bacteriophage Classification 2

1.1.2 Bacteriophage Life Cycles 5

1.1.3 Horizontal Gene Transfer 8

1.1.4 Bacteriophage Transduction 11

1.2 The Gut Microbiota 15

1.3 The Gut Virome 17

1.3.1 Contribution of Phages to Gut Bacterial Virulence 20

1.3.2 Host-Phage Population Dynamics in the Gut 22
1.4 Phages Beyond the Human Gut 25
1.4.1 Aquatic Phages 25
1.4.2 Soil Phages 27
1.5 Methods to Study the Human Gut Metagenome and Virome 30
1.5.1 Inadequacy of Viral Sequence Databases 38
1.5.2 The Discovery of CrAss-like Phages 40

2 Gauge your phage: Benchmarking of bacteriophage identification tools in metagenomic sequencing data 47
2.1 Background 48
2.2 Aims and Objectives 54
2.3 Methods 55
2.3.1 Benchmarking with RefSeq Dataset 55
2.3.2 Benchmarking with Randomly Shuffled Sequences 57
2.3.3 Benchmarking with Mock Community Metagenomes 58
2.3.4 Benchmarking with Simulated Mock Virome Communities . 58
2.4 Results 60
2.4.1 Benchmarking with RefSeq Phage and Non-Viral Artificial Contigs 60
2.4.2 Benchmarking Tools with Randomly Shuffled Sequences . . . 74
2.4.3 Benchmarking Tools with Mock Community Shotgun Metagenomes 74
2.4.4 Impact of Tool Prediction on Diversity Metric Estimation . . 77
2.4.5 Runtime and Computational Load of Each Tool 80

2.5 Discussion 84
2.5.1 Limitations 88
2.6 Conclusion 90

3 Investigating Bacteriophage Dynamics in the Intensive Care Unit using Metagenomic Sequencing Methods 91

3.1 Background 92
3.2 Aims and Objectives 95
3.3 Methods 96
3.3.1 Sequence Retrieval and Quality Control 96
3.3.2 Taxonomic Identification of Metagenomic Reads 97
3.3.3 Viral Sequence Identification 97
3.3.4 vOTU Taxonomic Identification and Host Prediction 98
3.3.5 Dominant Host-Phage Correlation Analysis 100
3.4 Results 101
3.4.1 Updated Taxonomic Profiles of ICU Metagenomes 101
3.4.2 Defining Microbial Domination 102
3.4.3 Identifying vOTUs from ICU patient gut metagenomes . . . 105
3.4.4 Dominant Host-Phage Dynamics in the ICU Gut 109
3.5 Discussion 120
3.5.1 Limitations 125
3.6 Conclusion 131

4 Investigating Bacteriophage Dynamics in the Gut Microbiome of Intensive Care Unit Patients using Viral Nanopore Metagenomics 132

4.1 Background 134
4.2 Aims and Objectives 135
4.3 Methods 135
4.3.1 Q33 Phage Propagation and Rnumeration 135
4.3.2 Stool Sample Preparation 137
4.3.3 Adaptation of VLP Nucleic Acid Extraction from Stool Protocol for ONT Long-Read sequencing 137
4.3.4 Quantification of DNA Extractions 141
4.3.5 ONT Library Preparation and Sequencing 141
4.3.6 Whole Community DNA Extraction from Stool 144
4.3.7 Whole Community Metagenomic Sequencing 145
4.3.8 Quality control and assembly of long-read ONT VLP reads . 146
4.3.9 Quality Control and Human Sequence Removal from Metagenomes 148
4.3.10 Taxonomic Identification of Whole Community Metagenomic Reads 148
4.3.11 Taxonomic and Functional Profiling of vOTUs 148
4.3.12 Sequence Level Analysis of ARG Encoding vOTUs 149
4.3.13 Calculation of vOTU Absolute Abundances 149
4.3.14 Linear Regression and Correlation of Dominant Host-Phage Pairs 150
4.4 Results 150
4.4.1 Adaptation of VLP Nucleic Acid Extraction from Stool Protocol for ONT Long-Read Sequencing 150

4.4.2 ICU Patient Selection and Implementation of VLP ONT Sequencing Protocol . . . 161
4.4.3 Metagenomic Sequencing of ICU Patient Stool Samples . . . 166
4.4.4 Intestinal Microbiome Domination in the ICU Patients . . . 166
4.4.5 Assessment of Bacterial DNA Contamination in Viromes . . 168
4.4.6 Decontamination of Viromes . . . 169
4.4.7 Taxonomic and Functional Profiling of vOTUs . . . 171
4.4.8 Exploring the Genomic Context of ARG Encoding vOTUs . 177
4.4.9 Calculating Absolute Abundance of vOTUs using Q33 Phage Spike-in . . . 181
4.4.10 Dynamics of Dominant Host-Phage Pairs in the ICU Gut . . 185
4.5 Discussion . . . 190
4.5.1 Limitations . . . 199
4.6 Conclusion . . . 202

5 General Discussion **204**
5.1 Conclusion . . . 214

References **215**

Appendix **301**
5.2 Custom scripts . . . 301
5.2.1 removesmalls.pl . . . 301
5.2.2 Time lagged cross-correlation python script . . . 302
5.3 Supplementary Tables . . . 307
5.4 Supplementary figures . . . 323

List of Figures

1.1 The classical classification of viral taxa infecting bacteria and archaea 4

1.2 Comparison of the old (1991–2017) and new (2019-present) ICTV taxonomic rank hierarchies 6

1.3 The lytic and lysogenic life cycles of bacteriophages 9

1.4 Mechanisms of horizontal gene transfer 12

1.5 Mechanisms of bacteriophage transduction 42

1.6 Abundance of the main taxonomic groups of the human gut microbiota at domain/kingdom level 43

1.7 Bacteriophage distribution in the human gut 44

1.8 Comparison between amplicon sequencing and metagenomic sequencing. 45

1.9 Schematic of virus-like particle extraction and sequencing 46

2.1 Overview of RefSeq Artificial Contig Benchmarking Workflow . . . 63

2.2 Optimisation of detection thresholds (continuous) of viral identification tools 64

2.3 Benchmarking of viral identification tools on artificial RefSeq contigs 65

2.4 Optimisation of detection thresholds (categorical) of viral identification tools 67

2.5 Taxonomy of each tool's viral prediction on RefSeq phage fragments 68

2.6 Comparison of artificial RefSeq contig NZ_CP059254.1 fragment 385 to Streptomyces phiC31 phage reference genome (NC_001978.3) 70

2.7 Comparison of artificial RefSeq contig NZ_CP077715.1 fragment 358 to Sulfolobus virus 1 (provirus) reference genome (NC_001338.1) 71

2.8 Taxonomy of false positive viral prediction on RefSeq chromosomal fragments 72

2.9 Taxonomy of false positive viral prediction on RefSeq plasmid fragments 73

2.10 Comparison of viral identification tools on uneven mock community samples 79

2.11 Estimation of diversity metrics of tool predicted virome populations 82

2.12 Comparison of tool runtimes on the positive RefSeq artificial contig set 84

3.1 Comparison of taxonomic compositions between analyses from Ravi et al. and this study using a representative sample from Patient 35 103
3.2 Genus level taxonomic composition of patients across their stay in the intensive care unit 104
3.3 Network based taxonomic classification 110
3.4 Host taxonomy of vOTUs within each patient across their stay in the intensive care unit 112
3.5 Relative abundance of vOTUs in each patient by lifestyle 116
3.6 Time Lagged Cross-Correlation between Dominant Bacterial Genus and their Phages 117

4.1 Timeline of sequenced stool samples for each intensive care unit patient 165
4.2 Species level taxonomy of metagenomic reads for each patient across their stay in the intensive care unit 170
4.3 Family and genus level taxonomy of vOTUs 173
4.4 Network based taxonomic classification 175
4.5 Relative abundance of vOTUs in each patient by lifestyle 176
4.6 Global alignment of *cfxa* encoding vOTUs with metagenomic contigs and Tn*4555* transposon 178
4.7 Absolute and relative abundance of vOTUs stratified by host taxonomy 183
4.8 Linear regression of vOTU abundance or vOTU alpha diversity and length of stay in the intensive care unit 186
4.9 Dynamics of dominant host genus and corresponding vOTUs 188

5.1 Linear regression between vOTU abundance and length of stay of intensive care unit patients 324

5.2 Linear regression between Inverse Simpson alpha diversity of vOTUs and length of stay of intensive care unit patients 325

List of Tables

2.1 Overview of tools to identify and predict phage sequences in microbial ecosystems 51

2.2 Number of sequences at each stage of the RefSeq benchmarking workflow. 66

2.3 Performance of tools on randomly shuffled artificial phage contigs . 75

4.1 PCR program for amplification and barcoding in SQK-RPB004 . . 142

4.2 Proposed changes to VLP extraction protocol and their outcomes . 155

4.3 Clinical details of ICU patients 161

4.4 Per sample Nanopore sequencing statistics 162

5.1 List of vOTUs from Chapter 3 that encode antibiotic resistance genes307

5.2 List of vOTUs from Chapter 3 that encode virulence factor genes . 310

List of Abbreviations

This list describes abbreviations used more than one in the body of this thesis.

Symbols

α	Lowercase alpha
$\bar{X}$	Sample mean
β	Lowercase beta
$\geq$	Greater than or equal to
$\leq$	Less than or equal to

μ	Population mean
ϕ	Lowercase phi
$\pm$	Plus or minus
φ	Lowercase phi
$<$	Less than
$>$	Greater than
%	Percentage
®	Registered
™	Trademark

Units

°	Degree
°C	Degree Celsius
μg	Microgram
μL	Microlitre
μM	Micromolar
v/v	Volume/volume
w/v	Weight/volume
bp	Base pairs

fmol	Femtomole
g	gram
GB	Gigabytes
h	Hour
Kb	Kilobase pairs
M	Molar
Mb	Megabase pairs
mg	Milligram
min	Minute
mL	Millilitre
ng	Nanogram
nm	Nanometre
nt	Nucleotide
rpm	Rotations per minute
RT	Room temperature
s	Second
spp.	Species
U	Unit

x g	Times gravity

Other Abbreviations

$^{-}1$	Per
AAI	Average amino acid identity
ANI	Average nucleotide identity
ANOSIM	Analysis of similarity
ARG	Antibiotic resistance gene
BLAST	Basic Local Alignment Search Tool
CARD	Comprehensive Antibiotic Resistance Database
CD	Chrohn's disease
CDS	Coding sequence
CPM	Counts per million
CRISPR	Clustered regularly interspaced short palindromic repeats
CTAB	Cetyltrimethylammonium bromide
DNA	Deoxyribonucleic acid
dsDNA	Double stranded DNA
DTR	Direct terminal repeats
ENA	European Nucleotide Archive

FPH	False positive hit
GI	Gastrointestinal
GPD	Gut Phage Database
GTA	Gene transfer agents
HGT	Horizontal gene transfer
HMM	Hidden markov model
ICTV	International Committee on Taxonomy of Viruses
ICU	Intensive care unit
IS	Insertion sequence
ITR	Inverted terminal repeats
IV	Intravenous
KTW	Kill the winner
LSTM	Long short-term memory
LSU	Large subunit
MAG	Metagenome assembled genome
MCP	Major capsid Protein
MDA	Multiple displacement amplification
MIUVIG	Minimum Information about Uncultivated Virus Genome

ML	Machine learning
MOI	Multiplicity of infection
NCBI	National Centre for Biotechnology Information
NGS	Next generation sequencing
NMDS	Non-metric multidimensional scaling
OD	Optical density
ONT	Oxford Nanopore Technologies
ORF	Open reading frame
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PEG	Polyethylene glycol
PES	Polyethersulfone
PFU	Plaque forming units
PHROG	Prokaryotic Virus Remote Homologous Groups
PI	Percent identity
PTW	Piggyback the winner
pVOG	Prokaryotic virus ortholog groups
QC	Query cover

RBPs	Receptor binding proteins
RefSeq	Reference Sequence Database
RNA	Ribonucleic acid
RQD	Red queen dynamics
rRNA	Ribosomal RNA
SDS	Sodium dodecyl sulfate
SMRT	Single molecule real time
ssDNA	Single stranded DNA
ssRNA	Single stranded RNA
SSU	Small subunit
TerL	Terminase large subunit
TFF	Tangential flow filtration
TLCC	Time lagged cross-correlation
tmRNA	Transfer-messenger RNA
tRNA	Transfer RNA
UC	Ulcerative colitis
v	Version
VC	Viral cluster

VF	Virulence factor
VFDB	Virulence Factors Database
VLP	Virus-like particle
VLP	Virus-like particles
vMAG	Viral MAG
vOTU	Viral operational taxonomic unit
vOTU-PPG	vOTU particles per gram of stool
WGA	Whole genome amplification

chapter 1

Introduction

1.1 Bacteriophages

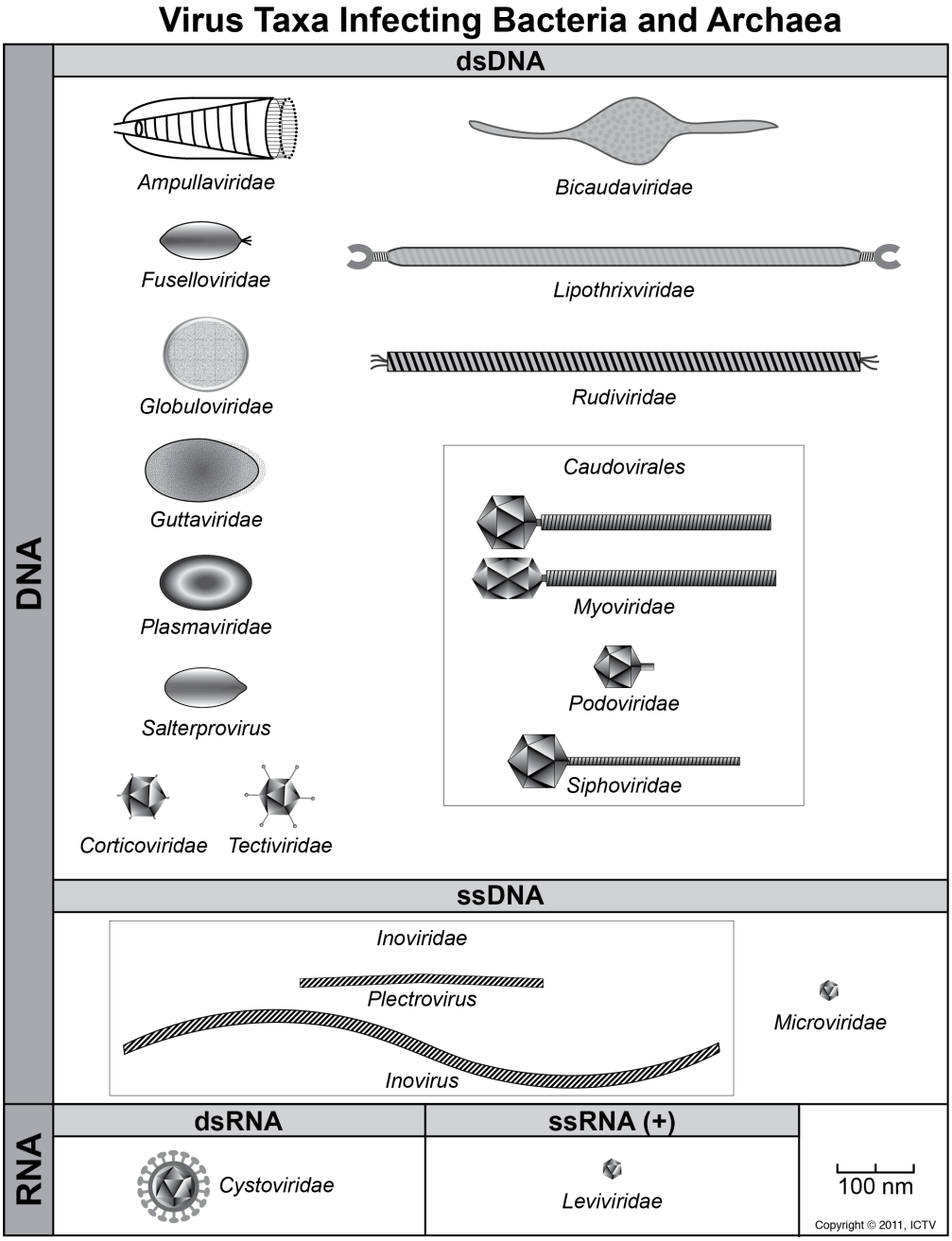
Bacteriophages (phages) were first independently identified in 1915 by Frederick Twort when they observed the clearing of bacterial lawns through plaque formation (Twort, 1915). In 1917 d'Hérelle coined the term "bacteriophage"which is derived from the English term "bacterium" and the Greek "phaegin" meaning "to eat" (d'Herelle, 1917). Phages are viruses that infect and replicate within bacterial cells (Ellis & Delbrück, 1939; Sulakvelidze et al., 2001). They are the most abundant and diverse biological entities on the planet, with an estimated population of 10^{31}, which is more than every other organism combined (Sulakvelidze et al., 2001).

1.1.1 Bacteriophage Classification

Phages are historically classified by the International Committee of Taxonomy of Viruses based on a number of characteristics. These consist of any feature of the phage including the nucleic acid structure; the viral capsid structure; the presence of a tail; their host range; pathogenicity; and sequence comparisons (Lefkowitz et al., 2018). Traditionally, phage identification and classification were based on visualisations of virion morphology. By 2007, at least 5568 phages had been examined under an electron microscope. The vast majority (96%) of these belonged to the *Caudovirales* order of tailed phages, which all encode double-stranded DNA (dsDNA) (Figure 1.1) (H. W. Ackermann, 2007). The order of *Caudovirales* was then split into three families depending on their tail morphology: *Myoviridae* have long contractile tails; *Podoviridae* has short

non-contractile tails; and *Siphoviridae* has long non-contractile tails. Phage tails are responsible for host receptor recognition, and nucleic acid ejection (H. W. Ackermann, 2003). The combination of these features allowed a hierarchical relationship between viral groups to be formed. For example, phages from the family *Inoviridae* have non-enveloped flexible filamentous virions, their genomes are made up of supercoiled, positive-sense, circular single-stranded DNA (ssDNA), and they only infect Gram-negative bacteria (Figure 1.1).

The advent of Next Generation Sequencing (NGS) has led to a rapid expansion in the number of sequenced phage genomes with multiple recent studies identifying thousands of previously unknown phages at a time (Benler et al., 2021; Callanan et al., 2020). Pressure on the traditional International Committee on Taxonomy of Viruses (ICTV) classification system had been mounting with genome sequencing of phage genomes revealing higher than expected genomic diversity. This led ICTV to replace the traditional five-rank taxonomic structure with a 15-rank hierarchical system that can accommodate any level of viral divergence, including the basal evolutionary connection between distantly related viruses (Figure 1.2) (Gorbalenya et al., 2020). Within these taxonomic ranks, it was clear that the traditional classification criteria were at odds with the genomic data. This is exemplified by the tailed phages order *Caudovirales*, where the three families were shown to be polyphyletic (Lavigne et al., 2009; Lavigne et al., 2008; Rohwer & Edwards, 2002). This led to the creation of new genomic-based families within the order by the ICTV. Within this framework, myoviruses have been split into three families (*Ackermannviridae* (Adriaenssens et al., 2018b), *Chaseviridae* (Adriaenssens et al., 2020),



.

Figure 1.1: **The classical classification of viral taxa infecting bacteria and archaea.**

Bacteriophages and archaeal viruses are displayed according to nucleic acid structure and morphology. Reproduced from the ICTV 9th Report (King et al., 2012)

Herelleviridae (Barylski et al., 2020); siphoviruses into two families (*Demerecviridae* and *Drexlerviridae* (Adriaenssens et al., 2020); and podoviruses into one family (*Autographiviridae*) (Adriaenssens et al., 2020). This culminated in a proposal to entirely abolish the order *Caudovirales* and the morphology-based phage families *Myoviridae*, *Podoviridae*, and *Siphoviridae* being ratified by ICTV members in 2022. All the underlying taxonomic ranks were absorbed into the class *Caudoviricetes* (Turner et al., 2021; Turner et al., 2023; P. J. Walker et al., 2022). Outside of the tailed phages, four subfamilies have been proposed to be added to the *Microviridae* ssDNA family. The previously mentioned filamentous *Inoviridae* has been split into *Inoviridae* and *Plectroviridae* (Adriaenssens et al., 2020), with a further division into five new families proposed (Roux et al., 2019b). These old taxonomic groups, especially those of the order *Caudovirales*, will still be referred to occasionally in this thesis due to their relevance to previous studies.

1.1.2 Bacteriophage Life Cycles

Phages are obligate intracellular parasites that require bacterial host machinery to replicate. Phages are either lytic, temperate (lysogenic) or pseudolysogenic; lytic phages only follow the lytic life cycle, whereas temperate phages have the ability to switch between the lytic or lysogenic cycle (Figure 1.3) (Kutter & Sulakvelidze, 2006). In the lytic cell cycle, the host's DNA is degraded and the injected phage DNA is replicated and transcribed to synthesise new viral progeny. In the lysogenic life cycle, phage genomes are integrated into the bacterial chromosome. Phages in

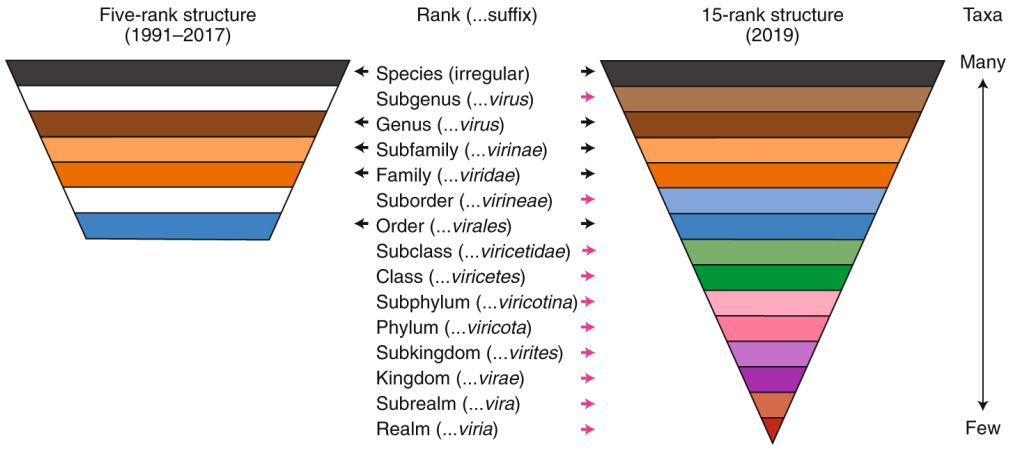


Figure 1.2: **Comparison of the old (1991–2017) and new (2019-present) ICTV taxonomic rank hierarchies.**

The 15-rank structure closely follows the Linnaean taxonomic system and represents the entire genetic diversity of viruses. Black arrows indicate the ranks common to both the 5 and 15-rank structure; pink arrows indicate the new ranks that were introduced in the 15-rank structure. Adapted from Gorbalenya et al., 2020.

this state are known as a prophage and the bacterium which contains the prophage is known as a lysogen. Prophages remain dormant in the genome through cell replications until they are induced into the lytic life cycle by environmental stress. Phage induction can occur by exposure of the host cell to UV irradiation, or DNA damaging agents such as mitomycin C, or antibiotics which trigger the host's SOS response. Lysogeny in temperate phages is preferred when the infected cell is starved or when the multiplicity of infection (MOI) is high (Rokney et al., 2008). Pseudolysogeny occurs in some filamentous and temperate phages, as well as in archaeal viruses (Hay & Lithgow, 2019; Pfeifer et al., 2021; Ripp & Miller, 1997). In this interaction, the phage genome is maintained in the host in an idle, episomal state. In contrast with the lysogenic cycle, replication of the phage genome is not

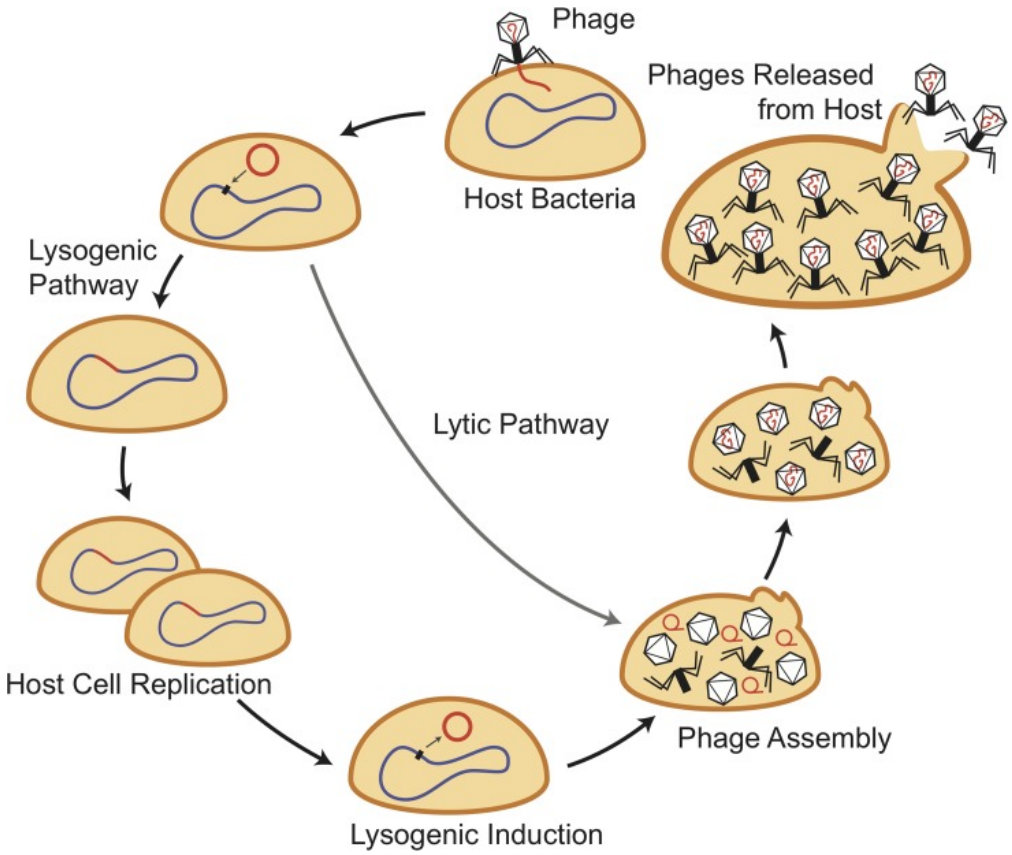
synchronous with the host chromosome, resulting in asymmetrical inheritance of the phage DNA during cell division (Miller & Day, 2008; Ripp & Miller, 1997).

The first step in all phage life cycles is the adsorption of the phage onto the bacterial cell surface. This contact occurs randomly via Brownian motion with phage receptor binding proteins (RBPs) interacting with host cell surface receptors. Some phages, such as *Caulobacter crescentus* phage ϕCb13 and ϕCbK, employ a two-step attachment system, with an initial reversible binding to the host flagellum via a variable length phage head filament, and a second irreversible binding to pilus portal. This two step process increases the likelihood of successful phage absorption by concentrating phage particles within close proximity to its host cell surface receptor (Guerrero-Ferreira et al., 2011). Successful absorption induces a conformational change in the phage particle leading to injection of the viral nucleic acid into the host cytoplasm. It is at this point where temperate phages decide on whether they lysogenise the host or proceed with the lytic cycle. The mechanisms involved in this decision have been most studied in coliphage λ (Ptashne et al., 1980). The decision is regulated by a circuit of viral genes, most important of which are *cI*,*cII*, and *cro* (Ptashne et al., 1980; Svenningsen et al., 2005). This regulatory system integrates a wide range of deterministic factors in the decision making process including: MOI (Oppenheim et al., 2005; Zeng et al., 2010); the nutritional state of the host bacterium (Kourilsky, 1973); host cell size (St-Pierre & Endy, 2008); and the location of absorption on the cell surface (Edgar et al., 2008). Whilst λ phage's lysis-lysogeny decision making process is often treated as a model system (Weitz et al., 2008), it is clear that a multitude of other important strategies have evolved in other phage families. In contrast to

λ, P1 phage, which also infects *Escherichia coli*, pseudolysogenises as a low copy number plasmid, with this decision being entirely independent of MOI (Bertani & Nice, 1954; Rosner, 1972; K. Zhang et al., 2021). *Bacillus subtillis* SPβ phages have been shown to communicate using a quorum-sensing system called the arbitrium system to coordinate their lysis-lysogeny decisions. This system was first described in phi3T and uses the six amino acid peptide AimP as a signalling molecule. In the initial phase of infection, the number of actively infecting phages is low, so AimP is absent in the surrounding medium, promoting the lytic cycle. As the phages continue to replicate, AimP is released from the host cell and accumulates in the intercellular medium. This promotes lysogeny so that the host population is not completely eradicated by the phage (Erez et al., 2017).

1.1.3 Horizontal Gene Transfer

Genetic material is typically transferred between bacteria vertically, that is from parent to offspring. Other movement of genetic information is known as horizontal gene transfer (HGT) and it has been found to be the dominant source of genetic variation and evolution in bacteria (Soucy et al., 2015). This process allows innovations in one bacterium to be shared across the population. HGT occurs through three main mechanisms; transformation, conjugation, and transduction (Soucy et al., 2015) (Figure 1.4). Transformation is where a recipient bacterium uptakes exogenous free genetic material. Integration of this DNA is enabled by homologous recombination (Johnson et al., 2005). Conjugation is the transfer of genetic material from a donor bacterium to a



.

Figure 1.3: **The lytic and lysogenic life cycles of bacteriophages.** In the lytic cycle, the phages infect and lyse their host, releasing phage progeny. In the lysogenic cycle, phage genomes (red) integrate into the host genome as a prophage. Reproduced from Chiang et al., 2019

recipient bacterium through direct cell-to-cell contact, mediated by a thin tube structure, called a pilus, or by direct adhesion. The genetic material is typically in the form of a conjugative plasmid although conjugative transposons and integrative conjugative elements (ICEs) have also been identified (Norman et al., 2009). Transduction is a form of HGT that is mediated by phage particles and

will be discussed below in section 1.1.4. Other non-canonical HGT strategies include: gene transfer agents (GTAs) (Lang et al., 2012), cell fusion (Naor & Gophna, 2013; Schleper et al., 1995), membrane vesicles (Dorward et al., 1989; Stentz et al., 2015), and nanotubes (Dubey & Ben-Yehuda, 2011) (Figure 1.4). GTAs are phage-like particles that randomly package DNA from their host's genome, with the total DNA packaged being less than needed to produce the GTA itself. Of the recognised GTAs, they all resemble tailed phages with smaller heads (Lang et al., 2017).

Besides plasmids and phages, a number of other mobile genetic elements have been described to play important roles in bacterial evolution, including transposable elements (TE), and integrons. TEs are sequences of DNA that can move position within a DNA molecule or move to other DNA molecules (Blot, 1994). In bacteria, TEs can be divided into three categories: insertion sequences (IS), non-composite transposons, and composite transposons. ISs are the smallest TEs ($<$ 2.5 Kb), and typically only contain one or two open reading frames (ORFs), including the transposase. They also contain short inverted repeat sequences at their ends, where the transposase binds and initiates transposition (Galas et al., 1989). ISs have been found in plasmids, phages, and composite transposons, and have been shown to play key roles in the transmission of antibiotic resistance genes via modulating expression of neighbouring genes, and as a part of composite transposons (Mahillon & Chandler, 1998). For example, IS*26* is associated with genes conferring resistance to many different antibiotic classes and their dissemination in gram-negative bacteria (Harmer et al., 2014), including aminoglycosides (Wrighton & Strike,

1987), tetracyclines (Allard et al., 1993), and quinolones (Hordijk et al., 2011). Composite transposons are flanked by two ISs with a section of genomic DNA in the middle which can be mobilised (Harmer et al., 2014). There are also non-composite transposons such as the Tn*3* family which are flanked by inverted repeats. These transposons are not dependent on IS elements but instead are responsible for their own excision and transfer. Transposition of these transposons occurs in a "paste and copy" mechanism, where replications occur during integration into the target sequence (Grindley, 2002).

Integrons are genetic elements that can incorporate exogenous ORFs into gene cassettes by site-specific recombination (Boucher et al., 2007). Integrons always encode an integrase (Messier & Roy, 2001), a primary recombination site that is recognised by the integrase (Partridge et al., 2000), and a promotor which is responsible for driving the transcription of the gene cassette (Lévesque et al., 1994).

1.1.4 Bacteriophage Transduction

Transduction is where genetic material is transferred via an infecting phage that has exogenous host DNA integrated into its genome. There are three types of transduction: generalised, specialised, and lateral transduction (Schneider, 2021) (Figure 1.5). Generalised transduction was the first type of phage-mediated HGT to be identified (Zinder & Lederberg, 1952). In this process, any bacterial DNA can be transferred between bacteria. Generalised transduction occurs when phage terminases erroneously identifies *pac* site homologs in the host

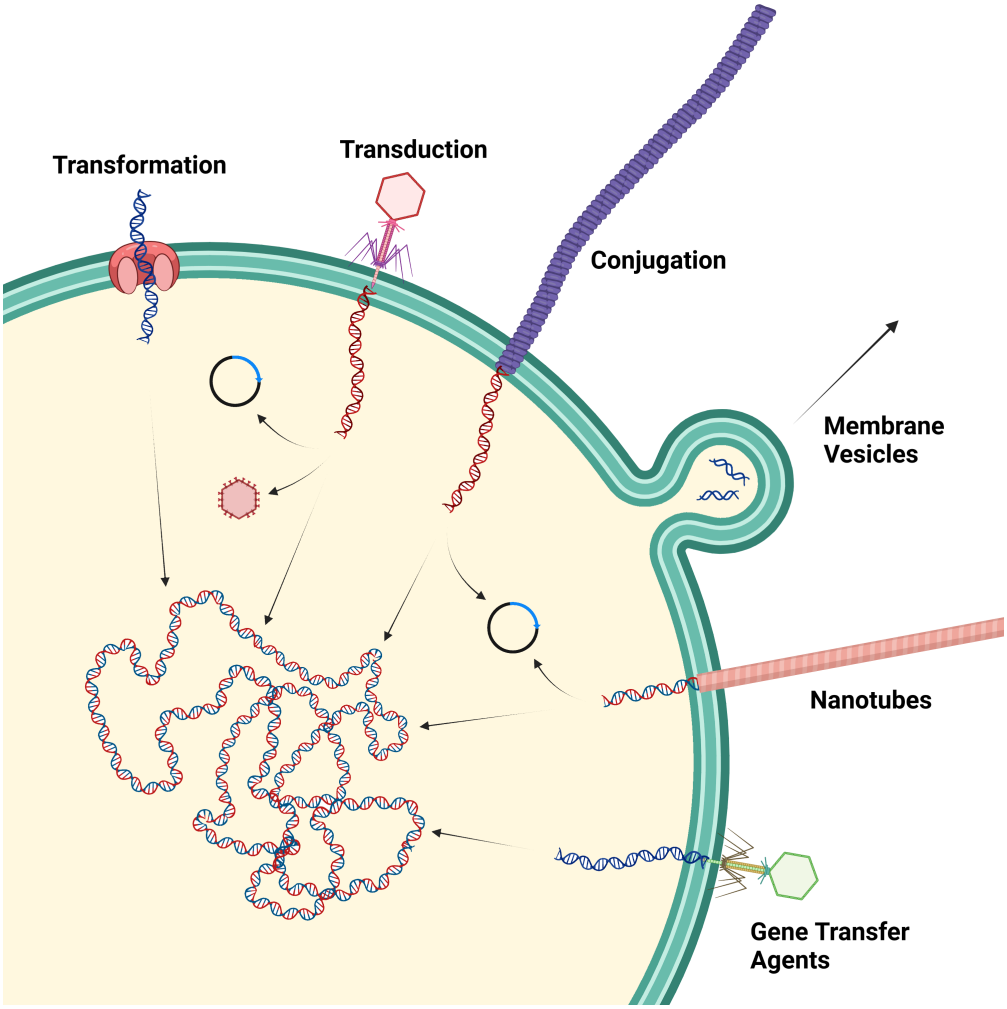


Figure 1.4: **Mechanisms of horizontal gene transfer.**

Horizontal gene transfer (HGT) is defined as the transfer of genetic material from one bacterium to another. There are three classic mechanisms for genetic material to cross the recipient's cell membrane: conjugation, transduction, and transformation. Transformation is where a recipient bacterium uptakes exogenous free genetic material. Conjugation requires both a donor and a recipient bacterium, and is mediated by direct cell-to-cell contact. Transduction is a phage-mediated mechanism of HGT. In transduction, the phage DNA can be replicated and packaged into new virions or integrate into the host chromosome. There are also a number of non-canonical HGT mechanisms including membrane vesicles, nanotubes, and gene transfer agents.

chromosome or plasmid, and cleaves the DNA ready for phage head packaging. When the capsid is at "headful" capacity, the *pac* terminase makes a second non-specific cut. This "headful" typically contains 102% - 110% of the phage genome size and upon infection, the direct terminal repeats of the ends of the sequence undergo homologous recombination to generate a circular genome (Casjens & Gilcrease, 2009). This results in successive packaging events moving the ends of the genome along the sequence by the length of the terminal repeat. The variability of start and end locations of the packaged sequence is further complicated by the process of determining when a capsid is at headful capacity being imprecise, with ranges of ±1000 bp having been observed. The process of initiating cleavage at a *pac* site is also imprecise in the majority of headful phages analysed to date, with cleavage of DNA occurring over 2 Kb from a *pac* site having been shown in phage Sf6 (Leavitt et al., 2013). This transduction is referred to as "generalised" as the *pac* site homologs occur throughout bacterial genomes, so it is thought that any part of the bacterial genome can be packaged and transferred (Schmieger, 1982; Sternberg & Weisberg, 1975). Phage T4 is similar to the *pac* phages with its headful packaging strategy except it was thought that DNA packaging was initiated at a nonspecific site. However, there is evidence that the small terminase in T4 phages, gp16, recognises and binds to a putative *pac* sequence towards the 3' end of gene 16 and fundamentally utilises the same mechanisms as other headful packaging phages (Rao & Black, 2013). Some phages use *cos* packaging sites instead of *pac* sites, which require a second *cos* site to cleave and package the molecule. *Cos* site homologs also exist in the host DNA but the probability of having two *cos* site homologs within "headful" packaging distance is very low, so it is thought that generalised transduction only

occurs in *pac*-type phages (Chiang et al., 2019).

Specialised transduction was first discovered in coliphage λ, and unlike generalised transduction it is limited to the transfer of specific areas of the host genome (Morse et al., 1956). It occurs when an aberrant excision event causes the excised molecule to be made up of viral DNA and the adjacent segment of host DNA. The virion then replicates as usual, and if a packaging site is retained, the concatameric DNA will also be packaged into new capsids and transferred to other host cells. This anomalous excision is a rare event and coupled with the limited host DNA that can be transferred, specialised transduction is regarded to only be a small contributor of phage-mediated HGT (Kleiner et al., 2020).

The traditional varieties of transduction (generalised and specialised) are a result of mistakes made by phages during the DNA packaging of their capsids. Lateral transduction was discovered in 2018, in temperate phages of *Staphylococcus aureus* (J. Chen et al., 2018). In contrast to generalised and specialised transduction, lateral transduction appears to be part of the natural phage life cycle. This is primarily a result of staphylococcal prophage's atypical conversion to lytic phase. This mode of transduction has been shown to have a greater than 1000-fold transfer frequency than previously observed with other transduction mechanisms. The assumed pathway of prophage induction follows a non-standard order where excision of the DNA occurs prior to replication and packaging. J. Chen et al., 2018 found that *S. aureus* phages delay excision until later in their life cycle, resulting in the phage genome being replicated whilst it is still attached to the host genome. Concatemers of viral DNA and the adjacent host DNA is then packaged into at least seven viral capsid headfuls. Hundreds of

kilobases of bacterial chromosome have been shown to be transfered, creating "hypermobile" regions of the bacterial genome.

1.2 The Gut Microbiota

The gastrointestinal (GI) tract hosts the largest population of microorganisms in the human body. Collectively known as the gut microbiota, it comprises a diverse array of bacteria, fungi, archaea, and viruses (Figure 1.6). The colonisation of the gut begins at birth, with microbes originating from faeces (Makino et al., 2013) and the vagina (Dominguez-Bello et al., 2010), when the baby is delivered via the birth canal. In contrast, birth by cesarean section is less exposed to microbes during birth (C. J. Hill et al., 2017), with initial GI colonisation typically coming from skin-to-skin contact (Akagawa et al., 2019) or directly from the environment (Dominguez-Bello et al., 2010). Other factors such as weight, breastfeeding, maternal age, ethnicity, and parental smoking are known to affect the developing microbiota (Bokulich et al., 2016; Borre et al., 2014). As they age, the microbiota matures into a comparatively stable form (Arumugam et al., 2011). In healthy adults 99% of the gene catalogue and biomass of the gut microbiota is accounted for by bacteria and archaea (Qin et al., 2010; Sender et al., 2016), with an estimated 3.8 x 10^{13} cells in the average male gut, approximately the same as the total number of human cells (Sender et al., 2016). Unsurprisingly, bacteria and archaea have received the majority of the attention of human microbiota studies, although recent investigations have elucidated the important roles of fungi (Huseyin et al., 2017), protozoa (Laforest-Lapointe &

Arrieta, 2018), and viruses (Shkoporov et al., 2019). The commensal gut microbiota is dominated by the Firmicutes and Bacteroidetes bacterial phyla, representing more than 90% of the community (Arumugam et al., 2011). Actinobacteria, Proteobacteria, and Verrucomicrobia are also common phyla although they are typically less diverse (Huttenhower et al., 2012).

Research into the gut microbiota has grown exponentially over the last decade with it now considered a "super-organ" responsible for critical roles in human health including: directly inhibiting pathogens via competition, nutrient depletion (Momose et al., 2008), production of bacteriocins (Hammami et al., 2013); induction and regulation of the innate and adaptive immune system (Hasegawa et al., 2010; Satoh-Takayama et al., 2008); maintenance of the intestinal epithelium (Gordon & Bruckner-Kardoss, 1961); metabolism of undigested carbohydrates, and production of short chain fatty acids via anaerobic fermentation (Macfarlane & Macfarlane, 2003); synthesis of the vitamins biotin and folate (Entcheva et al., 2002; Santos et al., 2008); and interaction through the gut-brain axis which regulates brain chemistry and modifies neuro-endrocine systems via signalling molecules (Rutsch et al., 2020; Shah et al., 2014). Despite these wide-ranging but specific roles, the composition of the gut is highly variable between individuals and even within individuals over time. Notable factors that alter the community include age (Bartosch et al., 2004; Han et al., 2017), ethnicity (Brooks et al., 2018; Gupta et al., 2017), geography (Gupta et al., 2017), lifestyle (Huxley et al., 2009; Lutgendorff et al., 2008), and diet (Huxley et al., 2009). Dysbiosis in the gut is defined as an imbalance or alteration of the gut microbial composition that is associated with a disease or disorder. Chronic

disorders where dysbiosis has been shown to be a factor are extensive (Vijay & Valdes, 2022) and include: rheumatoid arthritis (Wells et al., 2020; Zhang et al., 2015); type 1 (Kostic et al., 2015) and type 2 diabetes (Larsen et al., 2010); atopic eczema (Ta et al., 2020); asthma (Barcik et al., 2020); non-alcoholic fatty liver disease (De Gottardi & McCoy, 2011); chronic kidney disease (Jazani et al., 2019); irritable bowel syndrome (IBS) (Ghoshal et al., 2012); inflammatory bowel disease (IBD) - which includes ulcerative colitis (UC) (Lepage et al., 2011) and Crohn's disease (CRD) (Mottawea et al., 2016); Celiac disease (Girbovan et al., 2017); colorectal cancer (Y. Yang et al., 2021); and a number of mental health disorders such as autism spectrum disorder (ASD) (Wang et al., 2012) and attention deficit hyperactivity disorder (ADHD) (Prehn-Kristensen et al., 2018).

1.3 The Gut Virome

The gut virome refers to the community of viruses that live in the digestive tract. This is composed of eukaryotic viruses and prokaryotic viruses (phages) with the latter making up over 95% of the gut virome (Gregory et al., 2020). In contrast to other environments where phages dominate the population, gut phages are outnumbered by, or at least exist in similar numbers to, their bacterial hosts, with an estimated 10^9 - 10^{10} virus-like particles (VLPs) g^{-1} (Hoyles et al., 2014) (Figure 1.6). Microscopic and metagenomic studies have shown that the gut virome is largely made up of the class Caudoviricetes (Nayfach et al., 2021). As the majority of colonised bacteria in the distal gut are difficult to culture (e.g. the families of *Prevotellaceae* and *Lachnospiraceae*), culture based methods are mainly utilised to

isolate phages from a limited number of model and clinically relevant hosts such as *E. coli* (Dhillon et al., 1976; Viazis et al., 2011), *Enterococcus faecalis* (Bonilla et al., 2010), *Staphylococcus aureus* (Synnott et al., 2009), and *Clostridioides difficile* (Ramesh et al., 1999). These phages play a crucial role in regulating the bacterial population in the gut microbiota via HGT (Touchon et al., 2017), lysogeny (Lwoff, 1953), and predation (Rodriguez-Valera et al., 2009).

Colonisation of the gut with bacteria and viruses occurs within hours of birth (Lim et al., 2015; Maqsood et al., 2019). It is hypothesised that the initial viral colonisation in the neonatal gut comes from prophage induction within pioneer bacteria, followed by the introduction of viruses that infect the human host, which is mainly driven by environmental factors such as breastfeeding (Mohandas & Pannaraj, 2020; Walters et al., 2023). A recent study on the gut of 53 infants over the course of 3 years found that as the infants aged, their phageomes shifted from being dominated by *Lactococcus* phage to *Microviridae* species, poophages, and gokushovirsues, which resembled the materal gut phageome (Walters et al., 2023). As the diet of the infant changes from liquid to solid food the gut microbiota undergoes a shift towards that of an adult with the most common bacteria found belonging to *Bacteriodes* and *Firmicutes*. This shift subsequently affects the composition of the virome (Minot et al., 2013). Gregory et al. found that viral richness was the highest in children (0 to 3 years) and adults aged 18 to 65 years, with significant decreases between these ages and with elderly individuals. This trend however was mostly driven by *Siphoviridae* phages with eukaryotic viruses peaking at infancy, decreasing in childhood and remaining low throughout life. CrAss-like phages (whose discovery and

importance are explored in more detail in Section 1.5.2) increased from infancy to old age, with the biggest increase occurring from childhood to adulthood. This is hypothesised to be initially driven by the acquisition of crAss-like phages during childhood and early adulthood before they expand to take up a larger proportion of the gut virome (Gregory et al., 2020).

Multiple studies have found significant inter-individual variation in the adult gut phageome, with these differences being the main source of variance at the population level (Minot et al., 2011; Shkoporov et al., 2019). Despite this, the concept of a core phageome, a common group of phages found in 20-50% of healthy people, has been touted (Manrique et al., 2016; Shkoporov et al., 2019). Manrique et al., 2016 identified 23 phages which were found in more than half of their data set (n=63), mostly from the Caudovirales and Microviridae families. Significantly similar phageomes have been found between monozygotic twins and their mothers, and healthy members of a household. In contrast to the high inter-individual variation, the composition of an adult individual gut phageome appears to be stable over time, with 80-95% of viral genomes persisting over a study period of 1-2.5 years (Minot et al., 2013; Reyes et al., 2010).

Specific and persistent alterations in gut phageome composition have been linked to several diverse health conditions. For example, in ulcerative colitis (UC) patients, Zuo et al., 2019 found that their mucosa is enriched in phages compared to healthy controls although their diversity, richness and evenness were decreased. This altered virome correlated with intestinal inflammation and UC patients had more dissimilar viromes to each other than controls. Functional alterations of the mucosal virome were also found in UC patients including those associated

with host bacteria pathogenicity. This led the authors to suggest that the virome dysbiosis may contribute to UC pathogenesis, and that therapeutic interventions to reduce the abundance of mucosal phages should be explored. However, it is unclear whether this finding is a result or cause of UC (Zuo et al., 2019).

1.3.1 Contribution of Phages to Gut Bacterial Virulence

Phages have a significant impact on bacterial evolution through predation, lysogeny, and HGT. Some phages encode toxin genes that play a crucial role in the pathogenicity of the bacterium. A well studied example of this is in shiga-toxin producing *E. coli* (STEC) (Gyles, 2007). STEC is a clinically important foodborne pathogen that causes serious complications such as haemolytic uremic syndrome which can lead to death. This is due to the presence of shiga-toxin genes encoded by lambdoid prophages, which target ribosomal RNA, inhibiting protein synthesis, leading to apoptosis (Krüger & Lucchesi, 2015). Similar toxin genes are encoded by the prophages of other bacteria such as *S. aureus* (toxic shock syndrome toxin (*tst*)) (Ruzin et al., 2001), *Vibrio cholera* (accessory cholera exotoxin (*ace*)) (Trucksis et al., 1993), and *Corynebacterium diphtheriae* (diphtheriae toxin) (Gill et al., 1972). Aside from toxin systems, phages can also contribute to bacterial colonisation. For example, the phage *Sop*φ encodes the SopE effector protein in *Salmonella* spp. which promotes bacterial invasion through actin cytoskeleton rearrangements (M. W. Wood et al., 1996). Filamentous Pf phages in *Pseudomonas aeruginosa* are known to be an important part of their biofilm lifecycle. Pf phages induce

cell death causing the release of bacteria DNA which is a major component of the extracellular biofilm matrix. It has also been shown that Pf phages also directly interact with polymers to form a biofilm. Similar biofilm building capacity has been demonstrated in the related Fd phages of *E. coli* (Secor et al., 2015). Bacteria living within biofilms are significantly more resistant to antibiotics and desiccation, allowing them to survive for long periods on host surfaces and medical devices (Costerton et al., 1999). Phages can also contribute to virulence by interfering with host immunity. *S. aureus* phages *sakϕC* and *sak42D* encode staphylokinase (*sak*) which can neutralise antimicrobial peptides (Nguyen & Vogel, 2016), whilst *φSaeq1* encodes a staphylococcal complement inhibitor (SCIN), which stops the activation of the human complement system (de Jong et al., 2018). Another mechanism by which phages can increase the virulence of enteric bacteria is by transferring antibiotic resistance genes (ARGs). For example, Colomer-Lluch et al. found phages in urban sewage encoding β-lactamase resistance genes bla_{CTX-M} and bla_{TEM}. These were found to be viable when transduced into two *E. coli* recipients (Colomer-Lluch et al., 2011). Transduction of ARGs *in vitro* has been demonstrated in *S. aureus* (Varga et al., 2012), the Enterobacteriaceae family (Frosini et al., 2020; Rodríguez-Rubio et al., 2020), *Enterococcus* (Mazaheri Nezhad Fard et al., 2011), amongst others. Despite this, sequencing of complete phage genomes encoding ARGs is rare, and experimental evidence for ARG transduction is uncommon compared to the number of studies finding ARGs within viromes. Due to this, there is controversy regarding the extent and importance of phage-mediated ARG transfer in natural environments, even under heavy antibiotic pressure (Enault et al., 2017; K. Kang et al., 2021). A caveat to this is that in generalised and lateral transduction, only

bacterial DNA is packaged into the viral progeny (in the case of lateral transduction, the majority of virions). This would explain why ARGs are so rarely sequenced in complete phage genomes. No natural phage encoding ARG(s) has been shown to be able to confer resistance by lysogenic conversion - that is, lyse its original host, infect another host cell, and then continue the cycle (Pfeifer et al., 2022). Pfeifer et al. showed that ARGs were much more common in phage-plasmids genomes than phages genomes. These ARGs were also often encoded with integrons, which had not previously been seen in phages. The researchers were able to induce 1/6 of their phage-plasmids with mitomycin C and their lysogens were resistant to broad-spectrum cephalosporins. This lysogenic conversion was demonstrated for two full cycles. This high frequency of ARGs in phage-plasmids may be explained by their larger size; the genome of phage lambda is 49 Kb, whereas the average P1-like element is 96 Kb. Their large genome size allows the integration of larger elements containing the ARG and the integrons. The genetic homology between phage-plasmid and plasmids may also promote recombination between the two.

1.3.2 Host-Phage Population Dynamics in the Gut

It is well known that phages in the human gut influence the dominance and diversity of the microbial population through predation and feedback loops. There are three main dynamics models where phages can alter the gut microbiota: Red Queen dynamics (RQD; co-evolution of resistant hosts and infective phages; an evolutionary arms race) (Thingstad et al., 2014),

kill-the-winner (KTW; lysis of the most common hosts, which prevents domination of the gut) (Thingstad, 2000), and piggyback-the-winner (PTW; lysogeny allowing phages infecting the dominant host to coexist and thrive) (Knowles et al., 2016; Silveira & Rohwer, 2016). Despite the dense community of bacteria in the human gut, we do not frequently see bursts of phage populations predating the dominant host species under the KTW model. The cyclic change in phage and host bacteria abundances that we see in ocean environment KTW dynamics have not been observed in the adult human gut microbiome (Minot et al., 2011; Reyes et al., 2010). Knowles et al., 2016 analysed data from a wide range of ecosystems, with a focus on coral reefs, and found that viral counts are lower than expected when host abundances are high, due to an increased relative abundance of temperate phages. This comparatively high abundance of lysogenic phages in the gut microbiome compared to marine environments may explain the dynamics observed. KTW dynamics have been observed in the infant's gut, with Lim et al., 2015 detecting an inverse correlation between phage and bacterial populations, although only a single cycle occurs (Lim et al., 2015). In contrast to the KTW model, lysogenic infection is favoured not only at low concentrations of hosts but also at high host titers. Under this model (PTW), as the host and viral densities increase, the energetic cost of generating resistance to infection by carrying a prophage is less than through mutation. An extension to this model was proposed by Silveria and Rohwer, describing the population dynamics across the mucosal surfaces of the gut. Within the top layer of the mucin, where bacterial colonisation is at its highest, PTW dynamics are favoured, which provides an evolutionary advantage for commensals against the invasion of their niche. Deeper into the mucin layers, where the bacterial abundance is lower,

KTW dynamics are favoured which potentially aids in the elimination of pathogens (Silveira & Rohwer, 2016) (Figure 1.7). The stability of the gut phageome as discussed earlier further agrees with the predictions of the PTW model. As for RQD, despite the diversity of the gut, there is no evidence for its existence, as it would otherwise result in continuous co-selection of multi-resistant hosts and generalist phages. Instead, it appears that the metabolic cost of resistance slows bacterial growth, while the generalist phages' reduced efficiency towards specific hosts prevents them from outcompeting specialists (Howard-Varona et al., 2018).

The PTW model provides an intuitive explanation for why phage counts are so low despite the high bacterial load in the gut. However, this model falls short of explaining all the observed phenomena. For instance, strictly virulent *Microviridae* phages have been shown to persist for extended periods in the gut (Minot et al., 2013). In a study by Reyes et al., a set of strictly virulent phages were engrafted into mice which were colonised with 15 bacterial strains. The phages were able to persist stably without any significant changes in bacterial composition (Reyes et al., 2013). As strictly virulent phages are common members of the gut microbiota and do not cause significant perturbations to the composition, other models are required in addition to PTW to explain the full dynamics of the human-bacterial-phage system. These models would perhaps include abiotic factors such as pH, intestinal transit time (ITT), and nutrient variety/density, which have been shown to affect the outcome of virulent phage infections (Reyes et al., 2013).

1.4 Phages Beyond the Human Gut

Phages have been found in every biome on Earth including in aquatic environments (Bergh et al., 1989; Rohwer & Thurber, 2009; Warwick-Dugdale et al., 2019), soil (Ashelford et al., 2003), deserts (Prestel et al., 2008), and in the deep biosphere (Daly et al., 2019).

1.4.1 Aquatic Phages

Previous studies on the ecological significance of phages on microbial communities have mainly focused on marine environments. Marine phages have crucial roles in modulating bacterial populations through HGT, redirection of microbial metabolism through auxiliary metabolic genes (Mann et al., 2003; Thompson et al., 2011), and predation. Analysis by Wigington et al., 2016 found that 95% viral abundance from 22 marine surveys ranged between 10^5 and 10^7 VLPs per mL. Compared to the human gut, the average virus-to-microbial cell ratio is much higher at approximately 10:1, although there is substantial variation between samples, with ratios of up to 160:1 having been observed (Wigington et al., 2016).

It is thought that phages lyse 20-40% of ocean bacteria every day (C. A. Suttle, 2007). The lysis of marine bacteria results in the release of particulate and dissolved nutrients, making them available for use by heterotrophs. This process is known as the "viral shunt" (Middelboe & Lyck, 2002; Miki et al., 2008;

S. W. Suttle, 1999). It is estimated that this process results in the release of up to 145 gigatons of carbon per annum (Lara et al., 2017). Evidence of these KTW dynamics is clearly seen in flavophages whose abundance blooms with their *Flavobateria* hosts (Bartlau et al., 2022). Evidence for PTW interactions has also been shown in microcosm experiments of coastal water and suggests that KTW and PTW dynamics are not mutually exclusive in this niche (X. Chen et al., 2019).

Morphological analysis of *Tara* Ocean samples with quantitative transmission electron microscopy found that non-tailed viruses dominated the upper oceans with them comprising 51-92% of the VLPs observed (Brum et al., 2013). The largest marine virome study to date was conducted on samples from the *Tara* Oceans expedition (Brum et al., 2015) and was compiled into the Global Ocean Viromes 2.0 (GOV 2.0) dataset (Gregory et al., 2019). This dataset combines sequencing data from 145 samples collected across the world's oceans, supplemented by an additional 41 samples taken from the Arctic Ocean. Challenges in taxonomic identification were evident, as over 90% of the viral population from this study could not be taxonomically identified to a specific viral family, but the other 10% were mostly assigned to dsDNA phages. The authors found that their viral communities could be robustly partitioned into five ecological zones in the global ocean, and these were potentially driven by latitude, nutrient availability and temperature. The study also revealed that epipelagic waters and the Arctic Ocean are hotspots for viral biodiversity. This finding contrasts with the expected latitudinal diversity gradient pattern observed in marine organisms, where species richness typically increases from the poles to the tropics (Willig et al., 2003).

RNA viruses have also been shown to be abundant in oceans (Hillary et al., 2022), but assessing their relationship to bacteria is not trivial, as very few RNA viruses (3.4%) are phages (Dominguez-Huerta et al., 2022). As RNA phages do not lysogenise their hosts or participate in HGT, their fundamental ecological interactions differ from DNA phages (Brown et al., 2022).

1.4.2 Soil Phages

Unlike marine environments, soil communities are intrinsically more diverse, partly stemming from the wide variety in composition among different soil types and their spatial heterogeneity in environmental factors such as pH, temperature, particle sizes and moisture (Dion et al., 2020). The presence of plants is also known to affect the viral community, as shown by a study comparing the viromes of rhizospheres and bulk soils (Bi et al., 2021). Meta-analysis of 57 soil samples from nine studies found that viral abundance is highly variable, but correlated with soil samples, ranging from 10^3 VLPs per gram of dry weight in Saudi Arabian desert sands to 10^9 in eastern Virginian forest soil (Williamson et al., 2017). Viral abundance was also shown to be significantly positively correlated with bacterial abundance. Like their counterparts in aquatic systems, soil viruses are also involved in the "viral shunt", and therefore contribute to biogeochemical cycling within the soil (Emerson et al., 2018; Trubl et al., 2018; Williamson et al., 2017). This cycling has been directly shown in one study where they used stable-isotope probing and tracked the flow of carbon-13 from plants to bacterial genomes and then into phage genomes (Starr et al., 2021). Phage lysis of their hosts may release and redistribute

plant-derived carbon and other cellular material into the soil environment (viral shunt).

At the regional scale, sequencing data shows that the majority of soil viral populations are only detected at a single location suggesting high levels of variation across sites and habitats. On a global level, 4% of vOTUs were found to be shared across geographically distributed peatlands (ter Horst et al., 2021), suggesting the existence of a specific niche for these viruses and long distance transport and exchange of these viruses via aerolisation (Reche et al., 2018).

Despite the high richness observed in many soil viral communities, it is perhaps surprising that there has not been a rapid increase in the number of viral sequences derived from whole community soil metagenomes (Roux & Emerson, 2022), as has been the case for other biomes, such as the human gut (Camarillo-Guerrero et al., 2021; Nayfach et al., 2021). This reflects the technical difficulties associated with producing soil metagenomes, such as its microheterogeneity, matrix complexity, and presence of organic inhibitors, resulting in limited recovery of viral genomes (Göller et al., 2020; Pratama & van Elsas, 2018). Some studies have overcome this by using sequencing viral fractions from soil which have been shown to have a greater ability to recover viral populations in terms of both richness and diversity (Santos-Medellin et al., 2021). This process is covered in more detail in Section 1.5 in relation to human gut samples.

A comparison of VLP morphologies from six different soils reported that the majority of viruses found in soil are tailed phages, although one sample of

agricultural soil was dominated by spherical and non-tailed particles (56%) (Williamson et al., 2005). Another study which focused on agricultural soil also found that samples were dominated by spherical, non-trailed particles (70%), with only 5% being tailed (Swanson et al., 2009). Since these morphological studies, metagenomics/viromics have been used to explore the composition of viruses in soil. In one study of a permafrost thaw gradient, the authors combined assembles from 190 metagenomes and 7 viromes and recovered 1,907 vOTUs. Only 17% of these could be assigned taxonomy via protein clustering to NCBI RefSeq database, and the majority (95%) of these were assigned to the tail phage order *Caudovirales* (now absorbed into the class *Caudoviricetes* (Williamson et al., 2017). Another study which leveraged Hi-C to capture phage-host interactions in bulk soil found that they could only assign taxonomy to just over half of vOTUs, which were all in the class *Caudoviricetes* (R. Wu et al., 2023). A recent review of soil viral diversity screened dsDNA viruses from the IMG/VR v3 database and compiled a set of high-quality viral contigs (Jansson & Wu, 2023). Over a third of these were unclassified, with most of the remaining sequences classified to families within the *Caudoviricetes* class including non-tailed phages belonging to the families *Autographivirdae* and *Herelleviridae*.

Compared to aquatic environments, the rate of infection by phages in soil systems is higher due to an increased interaction probability between phages and their hosts (Kuzyakov & Mason-Jones, 2018). There is also evidence that in soil with increased resource availability, bacterial resistance to phage infection is also increased without a concurrent increase in phage infectivity (Gómez et al., 2015). Temperate phages are abundant in many soil types with up to 40% of soil

bacteria carrying inducible prophages (Huang et al., 2021; Williamson et al., 2007). Lysogeny is thought to be more common for phages in soil compared to those in aquatic environments (Ghosh et al., 2009; Williamson et al., 2007). Phage-bacteria interactions vary depending on soil conditions, with low nutrient environments typically promoting PTW dynamics in order to facilitate host survival (Knowles et al., 2016; Silveira & Rohwer, 2016; Thingstad, 2000). Recent studies also suggest that lysogeny is favoured in drier soils, whereas phages in wet soil conditions tend to become lytic (Schimel, 2018; R. Wu et al., 2021; R. Wu et al., 2023). This is partly caused by free soil phages irreversibly binding to soil particles as soil dries, resulting in their inactivation. This is supported by evidence that host-associated phages in post-desiccation soil had reduced transcriptional activity, an indication of lysogeny (R. Wu et al., 2023). These transitions between lifestyles have a significant impact on host populations and soil ecology (Jansson & Wu, 2023).

1.5 Methods to Study the Human Gut Metagenome and Virome

Historically, research into microbiomes relied on *in vitro* culture in order to isolate a single organism which could then be analysed (Tang et al., 2020). As most phages and their hosts were unculturable, the diversity of the microbiota was significantly underestimated. Despite this, culture-based techniques have provided valuable insights to our basic understanding of phage biology. Our current understanding of

phage-host interaction, functional diversity, and ecology is largely derived from *in vitro* studies. These techniques will remain important in the future to complement sequencing efforts.

The first attempts to analyse the metagenome rather than individual organisms were carried out by isolating DNA from environmental samples, and then cloning the DNA fragments into bacterial artificial chromosomes (BACs) or plasmids which were maintained in model organisms such as *E. coli*. The transformants were then screened for their biological activity and production of novel products, with later studies sequencing the cloned fragments (Handelsman et al., 1998). Since then, the field has consistently improved with three major methods being used for microbiome profiling today: rRNA amplicon sequencing, shotgun sequencing, and long-read sequencing. In amplicon sequencing, PCR is used to amplify the hypervariable regions of the bacterial 16S rRNA gene. Variable region 4 (V4) of 16S rRNA has been shown to outperform the other variable regions in terms of discriminatory power and is the most commonly used marker for amplicon sequencing (D'Amore et al., 2016; de Muinck et al., 2017). The amplicons are then barcoded, pooled, and sequenced. This technology relies on the highly conserved nature of the 16S rRNA gene, and comparisons of the genes between organisms can be used to determine evolutionary distance. These amplicons are typically analysed at the phyla or genera level as it is less accurate at the species level. Importantly, this method is also unable to analyse viruses, fungi, or protists due to their lack of the 16S gene (Clarridge, 2004; Van de Peer et al., 1996).

In contrast to 16S rRNA sequencing, shotgun sequencing (aka short read

sequencing) sequences all the DNA in the sample (Figure 1.8). Illumina is the main platform for generating shotgun libraries and this works by fragmenting the DNA into millions of segments which are then sequenced. This sequencing process uses fluorescently labeled nucleotides which are incorporated into the DNA strand based on complementarity. The emissions wavelength and intensity are then recorded to identify what nucleotide base was incorporated. This process is then repeated until the entire strand is sequenced. As shotgun sequencing analyses all the DNA in the sample it leads to more accurate and cross-domain taxonomic identification, as well as functional profiling (Qin et al., 2010; Schmieder & Edwards, 2012). However, assembly of shotgun sequences is often difficult, especially in repetitive genomic regions where short reads cannot bridge across the repeated region. This results in fragmented genome assemblies, especially in phage genomes due to the abundance of CRISPR arrays and insertion sequences (Kingsford et al., 2010), despite their comparatively short length. These challenges can be overcome with single-virus sequencing, which is particularly suitable for low abundance phages or those with high strain variation (Labonté et al., 2015); or with metagenomic fosmid libraries, but these approaches are technically difficult and expensive (Warwick-Dugdale et al., 2019).

Long-read sequencing is currently dominated by two technologies: Oxford Nanopore Technologies (ONT) nanopore sequencing and Pacific Biosciences (PacBio) single-molecule real-time (SMRT) sequencing. Nanopore sequencing works by measuring the fluctuations in ionic current when a single-stranded nucleic acid passes through a nanopore. Different nucleotides produce specific perturbations to the current; therefore, the variations in the current density can

be "basecalled" to determine the DNA or RNA sequence in real-time. SMRT sequencing uses a sequencing-by-synthesis technique that is similar to Illumina sequencing. A single-stranded DNA template is fixed to the bottom of a well. As nucleotides are incorporated into the complementary strand, a fluorescent signal is recorded by the detector and a base call is made accordingly. Recent improvements in these long-read sequencing technologies have positioned them as viable alternatives to the Illumina platform. Previous studies utilising the Nanopore sequencing have consistently produced metagenomic N50s of $>$ 5000 bp (Che et al., 2019; Nicholls et al., 2018), with megabase long reads from isolates having been achieved when DNA extraction and library preparation methods were optimised (Jain et al., 2018). Reads of this length would be sufficient to enclose complete genomes of phages, whose sizes range from 10 to 735 Kb in length (Al-Shayeb et al., 2020; Mahmoudabadi & Phillips, 2018) or at least span repetitive regions which cause fragmentation during assembly. Long-read sequencing approaches have higher error rates than their short-read counterparts ($<$5% for nanopore sequencers, $<$ 1% for SMRT sequencers) (Oxford Nanopore Technologies, 2020; Weirather et al., 2017) but these are continuously being improved with new algorithms and sequencing chemistries. Error rates can also be remedied by combining short and long-read data to produce hybrid assemblies.

A viromic Nanopore sequencing workflow named VirION has previously been developed by Warwick-Dugdale et al., 2019 with a focus on aquatic samples. They showed that this approach increased the number of recovered complete viral genomes, and maximise the capture of viral diversity and genomic islands, whilst

reducing the impact of high error rates. A follow up to this workflow was then developed to further improve long-read viral metagenomics (Zablocki et al., 2021). VirION2 incorporates an optimised PCR amplification on DNA libraries which are used as input for the genomic DNA by ligation kit (SQK-LSK109). This update was able to lower the high DNA requirements from VirION from $\tilde{1}$00 ng extracted from 20 L of seawater to 1 ng, whilst achieving reads that were 76% longer (median length: 6,965 bp).

Studies can then either choose to sequence the entire metagenome and then computationally separate viral sequences (Minot et al., 2011) or attempt to physically separate the virus-like particles (VLP) before preparing the library to produce a metavirome (Shkoporov et al., 2019) (Figure 1.9). As the genomes of viruses are relatively small and therefore make up a small proportion of the sample, the latter approach allows lower abundance phages to be sequenced. These would be missed if the entire microbiota was sequenced, unless the sequencing depth was magnitudes higher. However, VLP sequencing will likely exclude many phages from the sequencing entirely due to their association with the cellular fraction. This is due to the majority of gut phages being integrated into the host genome as prophages, or attached to the host surface (Labonté et al., 2015), or in a pseudolysogenic state due to a lack of resources (Ripp & Miller, 1997). This can be seen as a positive by some researchers as it distinguishes prophages that are not actively producing virions (dormant or defective) from phages that are associated with virions. Current protocols for VLP extraction from faecal samples use a variation of the following steps: (1) homogenisation, (2) centrifugation and filtration, (3) concentration and

purification of VLPs, (4) nucleic acid extraction (Castro-Mejía et al., 2015; d'Humières et al., 2019; Kleiner et al., 2015; Shkoporov et al., 2018b) (Figure 1.9). Despite these commonalities, available protocols differ at every step: centrifuge forces and times; filter sizes; purification and concentration technique. A number of studies have attempted to optimise and standardise this process. Castro-Mejia optimised two VLP extraction methods based on tangential-flow filtration (TFF) or polyethylene glycol (PEG) precipitation for the purification step. However, both of their protocols include a final caesium chloride gradient (CsCl) step which is unsuitable for large scale studies as it is labour intensive, expensive, and has poor reproducibility - issues which are shared by TFF (Castro-Mejía et al., 2015). The simplicity, low cost, reproducibility and diversity of phage genomes recovered make PEG precipitation the choice of many studies, although it can be hampered by the formation of a high molecular weight viscous precipitate which prevents the removal of PEG. This can be solved with a chloroform extraction which has the added benefit of removing any remaining bacterial cells, but results in the loss of lipid enveloped virus groups as chloroform is polar and has a high dielectric constant which allows it to break the hydrophobic interactions between the lipids in the membrane(H.-W. Ackermann et al., 1978; Kleiner et al., 2015; Shkoporov et al., 2018b). Some non-enveloped phages such as those belonging to families *Corticoviridae*, *Inoviridae*, and *Plasmaviridae* have also been shown to be susceptible to chloroform (Conceição-Neto et al., 2015; Thurber et al., 2009). In the case of the filamentous phages, chloroform sensitivity is hypothesised to be due to chloroform inducing morphological changes which resulted in the contraction of the rods into spherical particles (Oh et al., 1999; L. M. Roberts & Dunker, 1993).

Even with these extensive purification steps, contamination is currently inevitable (even when using stricter purification methods like CsCl) (Castro-Mejía et al., 2015; d'Humières et al., 2019). Contamination levels are often calculated by mapping reads to a 16S rRNA gene database. However, most bacterial genomes have multiple copies of 16S rRNA (mean of 4.2; maximum of fifteen (Kembel et al., 2012)) This along with the ribosomes being often co-purified with viral particles means that analysis using 16S rRNA can lead to disproportionately high contamination levels. A novel approach is using the *chaperonin-60* (*cpn60*) gene which is a highly conserved housekeeping gene that only occurs once per genome (Shkoporov et al., 2018b). Median levels of contamination using Shkoporov et al., 2018b PEG-based VLP extraction method was 5.5% using *cpn60* and 8% when using 16S rRNA. Consequently, *in silico* methods to predict viral sequences have to be employed to remove the remaining bacterial sequences (Shkoporov et al., 2018b). These computational methods will be discussed at length in Chapter 2. At every step of VLP extraction, viral particles are unavoidably lost: virions may be absorbed to the surface of larger particles, damaged by nucleases, or excluded by the filtering protocol - filamentous viruses such as *Lipothrixviridae* are often over 2μm in length and giant viruses can reach up to 750 nm (Callanan et al., 2021; Vestergaard et al., 2008). Protocols therefore need to be carefully chosen and optimised to balance the risk of contamination and detection of as many viral types as possible.

After successful VLP extraction from faeces, a number of steps need to be carried out to isolate nucleic acids of sufficient purity and concentration for downstream library prep and sequencing. The addition of DNase and RNase to

the sample to remove extracellular non-viral contaminating DNA and RNA fragments is an important and common step (Kleiner et al., 2015). Recent work, however, has shown that RNase treatment directly affects intact RNA viruses. This is due to the nucleocapsid of some RNA viruses containing an RNA component (Adriaenssens et al., 2018a). If one intends to isolate the RNA fraction of a virome, the reduction or omission of an RNase step will allow the capture of wider RNA viral diversity, at the risk of increased contamination levels, although this can be mitigated with stricter bioinformatic filtering. Extraction of the nucleic acids is often done with commercial kits or using a phenol-chloroform protocol. Formamide/cetyltrimethylammonium bromide (CTAB), which is typically used in nucleic acid extractions from plant tissue, has been used in conjunction with phenol-chloroform by some studies in order to remove contaminating polysaccharides (Verbanic et al., 2019).

When sequencing the gut phageome, whole-genome amplification (WGA) methods are often used as DNA yields from faecal samples are typically low (4 ng - 500 ng per gram of faeces), whilst library preparation kits often require high concentrations. A common WGA technique is multiple displacement amplification (MDA) with ϕ29 DNA polymerase (Blanco et al., 1989). MDA is able to amplify DNA without the use of primers, unlike conventional PCR techniques, whilst producing amplicons greater than 70 Kb with low error rates. A reverse transcription step can also be added to this, which allows the additional sequencing of single-stranded cDNA and RNA genomes. However, the use of WGA leads to significant distortion of the taxonomic composition, especially when the DNA concentration is low. MDA with ϕ29 DNA polymerase

is known to preferentially amplify short single-stranded sequences which leads to over-representation of small circular ssDNA genomes and plasmids. This biased amplification also leads to reduced diversity and makes it difficult to detect low abundance viral groups (Yilmaz et al., 2010). Newer Illumina library preparation kits such as the Swift Biosciences 1S Plus and Nextera XT kits are designed for low abundance input, which reduces amplification steps and should be used as standard in phageome research. ONT also have low abundance input methods such as the SQK-RPB004 kit which uses transposase-mediated cleaving to attach primer binding sites to the DNA for PCR amplification, which has been shown to produce equivalent microbial compositions to a PCR-free ONT method (SQK-LSK109) when sequencing the ZymoBIOMICS Microbial Community Standard (Marquet et al., 2022).

1.5.1 Inadequacy of Viral Sequence Databases

The severe disparity between the demonstrable diversity of gut phages and the number of gut-associated phages in public databases (4569 phage genomes in NCBI RefSeq as of 08/01/2023) is perhaps the largest roadblock to using metagenomics to study the human gut phageome. On study found that $\tilde{8}1\%$ of reads from viral metagenomics studies of the human gut do not align to any viral genomes (Aggarwala et al., 2017) and viral genomes do not contain a universal marker gene that could assist in assigning them a taxonomy. In contrast, $> 90\%$ of the microbial diversity of the human gut is accounted for at the species level due to studies such as the Human Microbiome Project (Aagaard et al., 2013). As

the majority of reads can not be aligned to a reference database, an alternative method is to assemble the reads into contigs which are then classified and annotated using *de novo* and reference based techniques. Reads are then mapped back to the contigs to build count matrices, which can then be analysed to produce α- and β- diversities, without taxonomically identifying any of the contigs. However, producing a high quality assembly from phage populations is particularly difficult due to the modular structure of their genomes, their high levels of diversity especially at the strain level, the high abundance of repeat and hypervariable regions, and their wide variation in abundance (García-López et al., 2015; Sutton et al., 2019). These issues can be remedied by using single virus genomics or long-read sequencing which was discussed earlier in Section 1.5. Assigning taxonomic ranks to these sequences is very challenging due to the lack of a universal phylogenetic marker (such as 16S rRNA gene or *cpn60* in bacteria). Some studies have used the large terminase subunit (TerL) and major capsid protein (MCP) as marker genes as they are highly conserved, however, they are not encoded in all phage genomes (Benler et al., 2021; Cook et al., 2021). A 2019 publication outlined "Minimum Information about Uncultivated Virus Genome" (MIUVIG) guidelines which attempt to provide a robust framework for reporting viral genomes into public databases. This includes the origin of the virus, genome quality, genome annotation, taxonomic classification (based on ICTV criteria), and *in silico* host prediction. Widespread adoption of this framework will allow for systematic reporting of novel phage genomes and enable studies from different research groups to be comparable (Roux et al., 2019a).

1.5.2 The Discovery of CrAss-like Phages

Metagenomic studies have allowed researchers to uncover the complexity of the human gut virome. The initial gut virome studies found that 60%-95% of phages were entirely novel and could not be assigned a taxonomy or linked to a host (Shkoporov & Hill, 2019). The phrase "viral dark matter" is used to describe this current knowledge gap regarding the taxonomic composition and population dynamics of the gut phageome. Perhaps the best argument for the importance of metagenomic studies in uncovering this viral dark matter was the discovery of crAssphage; a previously unknown 97 Kb dsDNA phage genome that was found by cross-assembling reads in human gut metagenomes. It was shown to be highly abundant, comprising up to 90% of reads in gut viromes and 22% in whole community gut metagenomes, with widespread presence across individuals (>50% of Western population colonised) (Dutilh et al., 2014). Prevalence of crAssphage has been associated with an industrialised lifestyle and diet, but most studies have found no significant correlation with any health or disease states (Edwards et al., 2019). In contrast to this, a recent study showed that colorectal cancer patients had a reduced abundance of crAssphage than healthy controls (Gao et al., 2021), whilst another showed that obesity was associated with reduced and differential populations of crAssphage (Cervantes-Echeverría et al., 2023). The host phylum of crAssphage was predicted to be Bacteroidota based on co-abundance analysis, CRISPR-spacer matches, and the presence of BACON (Bacteriodetes Associated Carbohydrate-binding Often N-terminal) domains (Dutilh et al., 2014). This was partially confirmed in 2018 when

ΦcrAss001, which infects a specific strain of *Bacteroides intestinalis*, was isolated in pure culture (Shkoporov et al., 2018a).

Since the discovery of crAssphage, hundreds of crAss-like phage genomes have been discovered *in silico* all around the world, including divergent strains in great apes, Old World monkeys, and New World monkeys (Edwards et al., 2019). This provides evidence that crAssphage may be tens of millions of years old and have co-evolved with the human lineage as a core element of the normal gut virome. In 2022, ICTV ratified a proposal to create a new order named *Crassvirales*, which was split into four families, ten subfamilies, 42 genera, and 73 species (P. J. Walker et al., 2022). This was primarily due to their significant genomic differences to the other members of the order of *Caudovirales*, which itself has been abolished (as described in section 1.1.1), as well as the increasing diversity of the crAss-like phages themselves. This is exemplified by the fact that the original crAssphage had no nucleotide sequence similarity to the first cultured crAssphage. As a point of interest, the original crAssphage strain is now classified under the species *Carjivirus communis* (Turner et al., 2023)

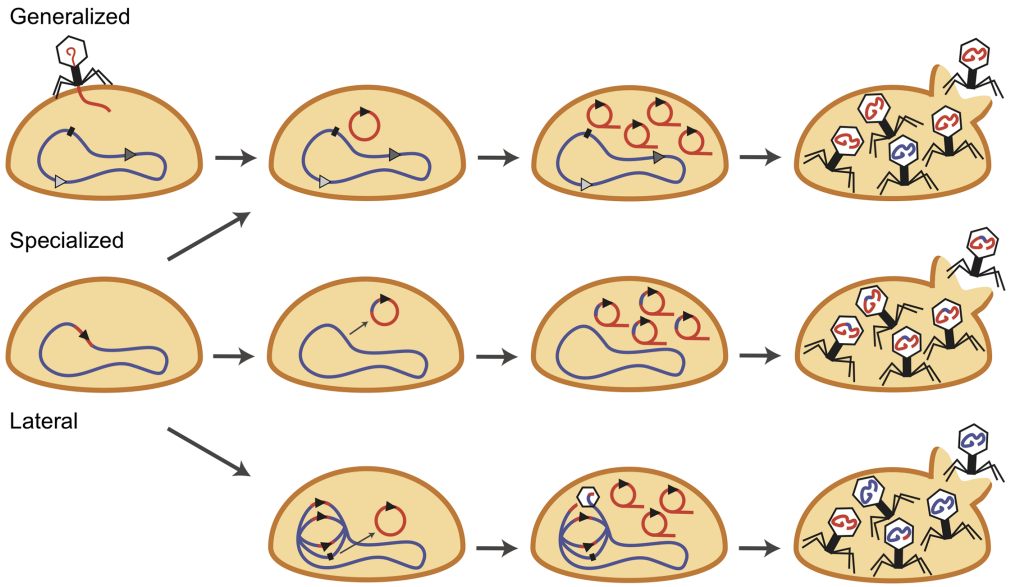
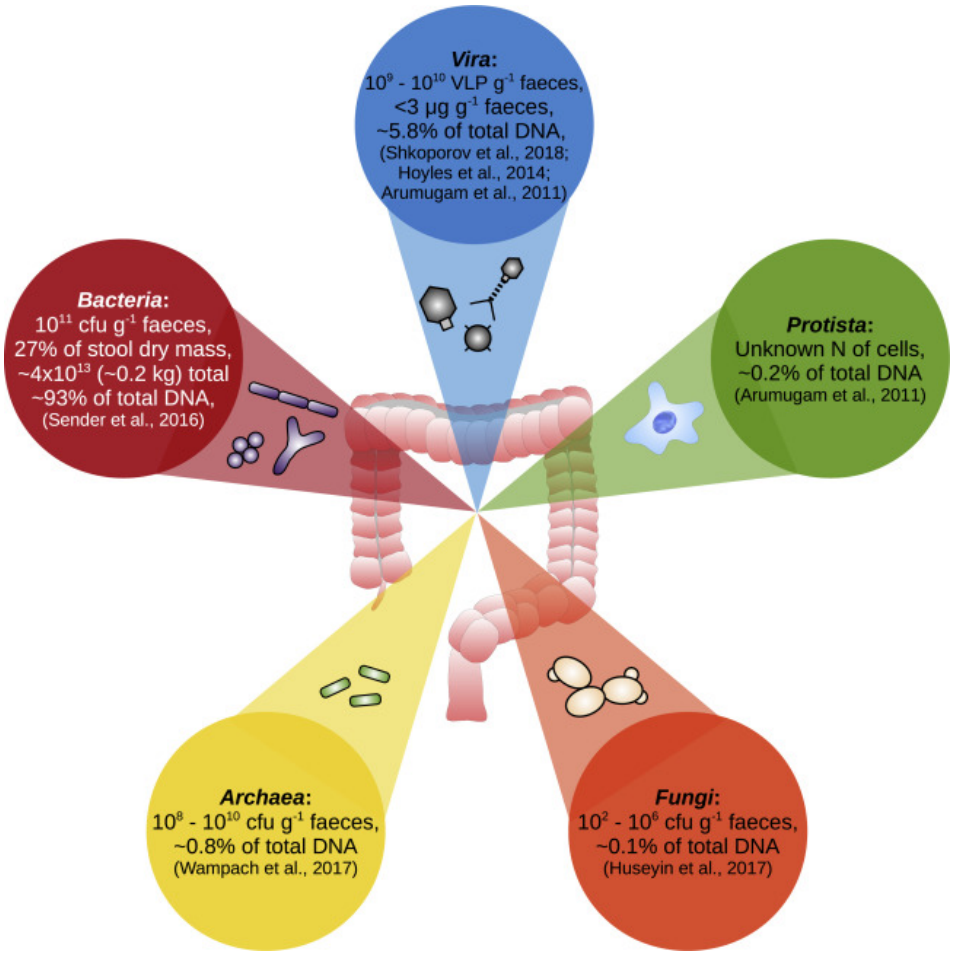


Figure 1.5: **Mechanisms of bacteriophage transduction.**
Generalised transduction occurs when bacteriophage terminases incorrectly identifies *pac* site homologs (grey triangles) in the host DNA (in blue) resulting in host DNA being packaged into the capsid head. Specialised transduction is the result of an aberrant excision event which causes the host DNA adjacent to the viral genome (in red) to also be packaged. Lateral transduction is a mode of horizontal gene transfer that has so far only been identified in *Staphylococcus aureus* phages and appears to be a natural part of their life cycle. In these phages, replication of the phage genome happens whilst it is still integrated in the host genome which results in downstream host DNA also being packaged until the viral capsid is full. Reproduced from Chiang et al., 2019



.

Figure 1.6: **Abundance of the main taxonomic groups of the human gut microbiota at domain/kingdom level.**

Reproduced from Shkoporov and Hill, 2019

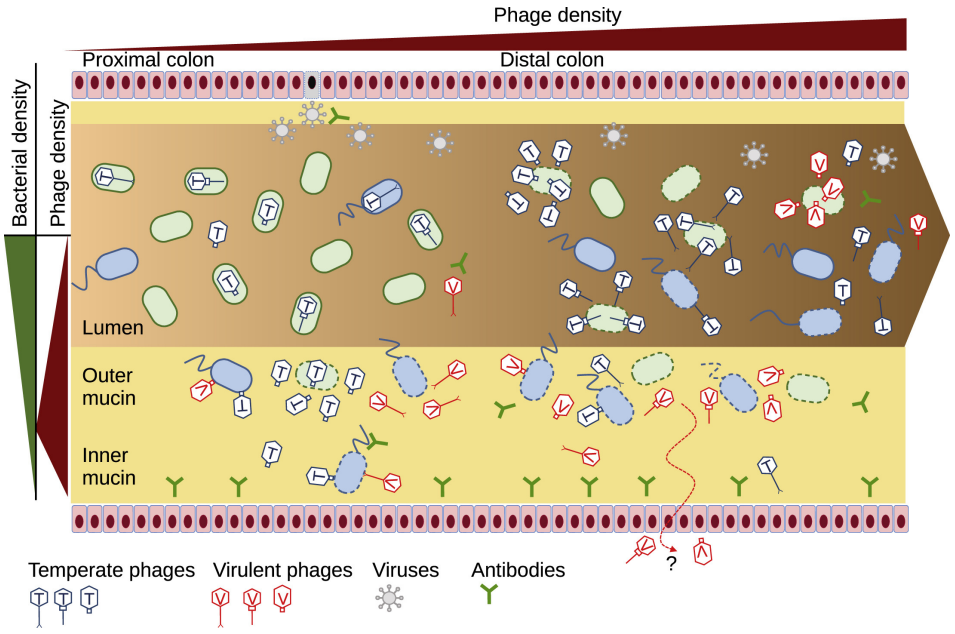
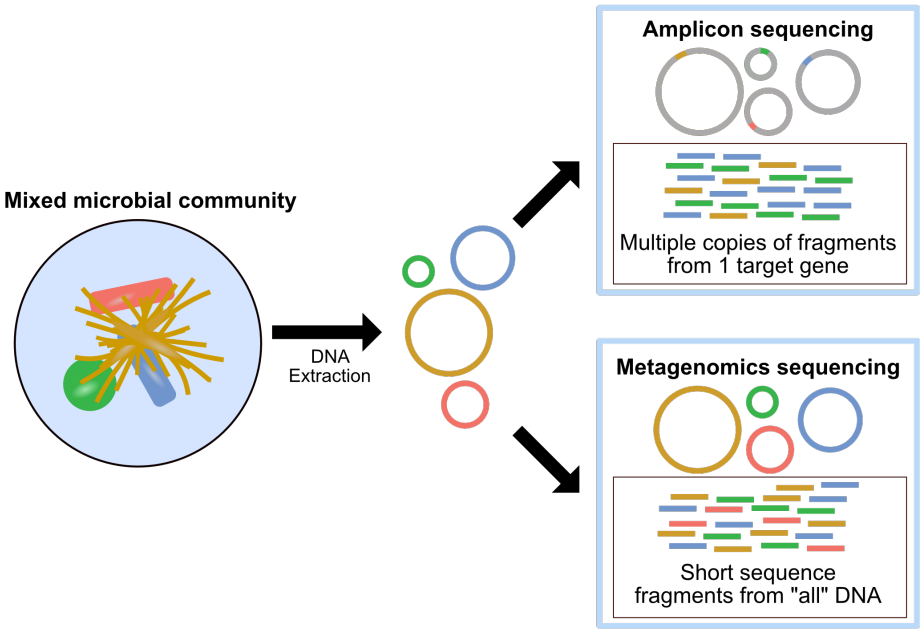


Figure 1.7: **Bacteriophage distribution in the human gut.** Commensal bacteria are more abundant towards the top of the lumen where mucin concentration is low. The high abundance of bacteria favours lysogeny over the lytic cycle (piggyback-the-winner model). In the mucin layer, bacterial density is relatively low, which favours a lytic switch in temperate bacteriophages (kill-the-winner model). This switch also occurs towards the distal colon as nutrient levels drop. A pathogenic bacterium would encounter two levels of phage-mediated immunity: competition with commensal lysogens in the lumen, and lytic infection in the mucin layers. Reproduced from (Shkoporov & Hill, 2019)



.

Figure 1.8: **Comparison between amplicon sequencing and metagenomic sequencing.** Reproduced from https://astrobiomike.github.io/misc/amplicon_and_metagen

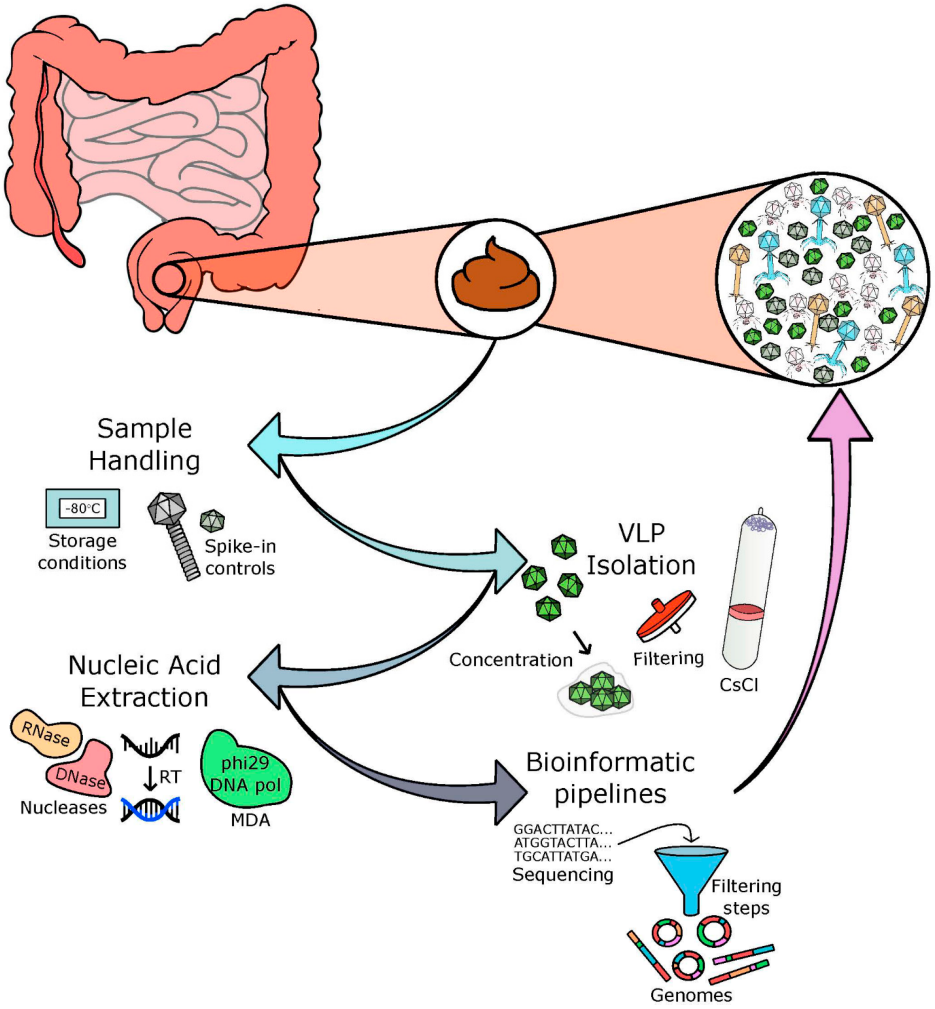


Figure 1.9: **Schematic of virus-like particle extraction and sequencing** faecal samples are first collected as a proxy for the human gut and stored at -°80. The faeces is then homogenised, centrifuged and filtered to remove debris and large cells. VLPs are then purified using polyethylene glycol precipitation, tangential flow filtration, or using a caesium chloride gradient. Nucleic acids can then be extracted and sequenced for downstream bioinformatic pipelines. Reproduced from Callanan et al., 2021.

chapter 2

Gauge your phage: Benchmarking of bacteriophage identification tools in metagenomic sequencing data

2.1 Background

Phage identification and characterisation traditionally relied on isolation and culturing techniques, which are not only time consuming and often require a significant level of expertise. These methods are also often impractical as many hosts, and their corresponding phages, cannot be cultured under laboratory conditions (A. W. Walker et al., 2014). The advent of high-throughput next generation sequencing has revolutionised the field of phage research by enabling the routine generation of metagenomic data from various environments, and providing a direct means for the identification and analysis of all genetic material in a sample, regardless of cultivability (Loman et al., 2013).

Over the past ten years, numerous tools have been developed for identifying viral sequences from mixed metagenomic and virome assemblies (Table 2.1). While previous tools focused on prophage prediction (PhiSpy (Akhter et al., 2012), Phage_Finder (Fouts, 2006), PHAST/PHASTER (Arndt et al., 2017), ProPhinder (Lima-Mendez et al., 2008)) or virome analysis (MetaVir2 (Roux et al., 2014), VIROME (Wommack et al., 2012)), VirSorter (Roux et al., 2015) was one of the first tools to identify phage sequences in metagenomes by detecting viral hallmark genes that have homology to reference databases, and by building probabilistic models based on different metrics (viral-like genes, PFAM genes, uncharacterised genes, short genes, and strand switching) which measure the confidence of each prediction. Since VirSorter's release, other gene based homology based tools such as VIBRANT (Kieft et al., 2020), and VirSorter2 (Guo et al., 2021) have been developed. VIBRANT uses a

multi-layer perceptron neural network based on protein annotation from Hidden Markov Model (HMM) hits to several databases to recover a diverse array of phages infecting bacteria and archaea, including integrated prophages. In addition to this, it characterises auxiliary metabolic genes and pathways after identification. VirSorter2 builds on its predecessor by incorporating five distinct random forest classifiers for five different viral groups into one algorithm to improve the diversity of viruses that it can detect accurately. MetaPhinder, unlike the tools above, uses BLAST based homology hits to a custom database to calculate average nucleotide identity and the likelihood that a sequence is of viral origin.

VirFinder was the first machine learning viral identification tool to utilise k-mer signatures (Ren et al., 2017). VirFinder was shown to have considerably better rates of recovery of viral sequences than VirSorter, especially on shorter sequences (<5 Kb), but had issues of variable performance in different environments, perhaps due to biases introduced by the reference data used for training the machine learning model (Ponsero & Hurwitz, 2019). DeepVirFinder (Ren et al., 2020) improves on VirFinder by applying a convolutional neural network that was trained on an enlarged dataset containing viral sequences from environmental metavirome sequencing data. DeepVirFinder boasts increased viral identification at all contig lengths over its predecessor VirFinder, whilst mitigating the latter's bias towards phages that are easily cultivable in the laboratory. Kraken2 is a k-mer based metagenomic taxonomic classifier (D. E. Wood et al., 2019) that can be used for viral detection (Yutin et al., 2021). It queries k-mers to a database which associates it to their lowest common

ancestor taxa which is then used to assign the taxonomic label. PPR Meta uses three convolutional neural networks to identify if a sequence is of phage, plasmid, or chromosomal origin (Fang et al., 2019). Sequence features are extracted by the network directly, instead of using pre-selected features such as k-mer signatures or genes. The three networks are also trained on three groups of different sequence lengths to improve their performance on shorter fragments, which some gene-based tools struggle with due to the low number of full-length genes available for analysis. Seeker also uses a neural network, in this case, a Long Short-Term Memory (LSTM) model, which is not based on pre-selected features but instead reads through the entire DNA sequence (Auslander et al., 2020). MetaviralSPAdes (Antipov et al., 2020) uses an entirely different approach by leveraging variations in depth between viral and bacterial chromosomes in assembly graphs. The tool is split into three separate modules: a specialised assembler based on metaSPAdes (viralAssembly); a viral identification module that classifies contigs as viral/bacterial/uncertain using a Naive Bayesian classifier (viralVerify); and a module which calculates the similarity of a constructed viral contig to known viruses (viralComplete).

It is important to note that machine learning tools have the potential to identify novel species, which is especially important with the enormous diversity of phages that is theorised to still be unknown (Shkoporov & Hill, 2019). With the development of so many tools using a variety of approaches, it is essential to comprehensively compare and benchmark them to evaluate which tools are most applicable to researchers. The performance of each method can vary based on sample content, assembly method, sequence length, classification thresholds

and other custom parameters. Therefore, to address these issues and provide an objective comparison, in this chapter, I have benchmarked ten metagenomic viral identification tools using artificial contigs, mock communities and real samples. This study aims to provide researchers with a reliable and accurate evaluation of the performance of various metagenomic viral identification tools, allowing them to choose the most appropriate tool for their specific research needs.

Table 2.1: **Overview of tools to identify and predict phage sequences in microbial ecosystems.** The method utilised by each tool is described. Tools in italics were not included in this study as they were either not relevant to this study or technical difficulties were encountered during their use.

Tool	Description	Reference
DeepVirFinder	Predicts viral sequences via a k-mer based deep learning method using convolutional neural networks (CNN). Successor to VirFinder.	(Ren et al., 2020)
Kraken2	A taxonomic classifier that uses exact k-mer matches to match a query sequence to it's lowest common ancestor.	(D. E. Wood et al., 2019)
MARVEL	Machine learning tool for predicting phage sequences in metagenomic bins.	(Amgarten et al., 2018)
MetaPhinder	Integrates BLAST hits to multiple phage genomes in a database to identify phage sequences in assembled contigs.	(Jurtz et al., 2016)

Tool	Description	Reference
viralVerify (metaviral-SPAdes)	viralVerify is a module of metaviralSPAdes which classifies contigs with a Naïve Bayes classifier based on Hidden Markov Models protein hits.	(Antipov et al., 2020)
PhaMers	Identifies phage sequences using a supervised machine learning model based on k-mer frequencies.	(Deaton et al., 2017)
PPR-Meta	Bi-path CNN approach that simultaneously identifies both phages and plasmids	(Fang et al., 2019)
Seeker	Deep learning framework that uses Long Short -Term Memory models (LSTM) which does not depend on sequence motifs.	(Auslander et al., 2020)
VIBRANT	Deep learning neural network based on protein signatures which also highlights auxiliary metabolic genes and pathways.	(Kieft et al., 2020)
ViraMiner	Extension of DeepVirFinder that is trained to identify any virus that may colonise human samples.	(Tampuu et al., 2019)
VirFinder	K-mer based machine learning method for identification of viral contigs.	(Ren et al., 2017)

Tool	Description	Reference
virMine	Iterative pipeline that relies on the abundance of non-viral sequences in databases to strictly filter out unwanted contigs. Pipeline accepts both reads or assembled contigs.	(Garretto et al., 2019)
VirMiner	Web-based pipeline that handles genome assembly, functional annotation using a variety of databases and identification of phage contigs via a random forest algorithm	(Zheng et al., 2019)
VirNet	Deep learning neural network using an attentional neural model trained on nucleotide viral fragments.	(Abdelkareem et al., 2018)
VIROME	Web-based pipeline that classifies viral sequences based on homology to databases and functional annotates them. No local version.	(Wommack et al., 2012)
VirSorter	Uses referenced-based and reference-free approaches in unison relying on probabilistic similarity models and referenced based protein homology searches to increase novel virus detection.	(Roux et al., 2015)
VirSorter2	Builds on VirSorter by applying machine learning to evaluate "viralness" using genomic features. Works with a wider variety of viral groups than its predecessor.	(Guo et al., 2021)

Tool	Description	Reference
VirusSeeker	Made up of two BLAST-based pipelines – Virome and Discovery. Virome aligns reads to a curated database to identify viral sequences and compute their abundance in the sample. Discovery focuses on contig-based analysis to aid novel virus discovery.	(Zhao et al., 2017)

2.2 Aims and Objectives

1. To determine the optimal parameters and thresholds for each tool.

2. To benchmark each tool using artificial contigs.

3. To benchmark each tool using mock community data.

4. To determine the diversity metric bias of each tool.

5. To produce recommendations on which tool(s) to use for future studies.

2.3 Methods

2.3.1 Benchmarking with RefSeq Dataset

Complete bacterial and archaeal chromosome and plasmid sequences, and phage genomes deposited in RefSeq (Na et al., 2016) since 1 January 2020 (inclusive) were downloaded on 12 August 2021 to construct a benchmarking set. Sequences in this set with ≥95% identity to the pre-2020 RefSeq sequences and training datasets for DeepVirFinder, PPR Meta, Seeker, VIBRANT, VirFinder, and VirSorter were removed with dedupe.sh (`minidentity=95`) (Bushnell et al., 2017) to reduce any potential overfitting. Chromosome and plasmid sequences were then randomly down-sampled by a factor of 14.3, using reformat.sh (from BBtools suite) (Bushnell et al., 2017), to produce a host:phage ratio of approximately 19:1. Phigaro (v2.3.0) (`-e tsv bed -not-open`) (Starikova et al., 2020) was run on the chromosomal and plasmid sequences to detect prophage sequences. Prophage sequences were then removed using BEDtools `complement` and `getfasta` (Quinlan & Hall, 2010). The resulting multifastas were then split into individual sequences using faidx `-split-files` from SAMtools (H. Li et al., 2009). PhageBoost (v0.1.7) (Sirén et al., 2021), was run on these split sequences to identify any remaining prophage sequences. This process was parallelised with GNU parallel (v20201122) (`-j 8`). The identified prophage sequences were then again removed with BEDtools `complement` and `getfasta`. Host sequences were searched for HMM hits to the pVOG database (Grazziotin et al., 2017) with HMMER3 (v3.3.2). Host sequences that had ≥30% open reading frames (ORFs)

with HMM hits to the pVOG database were removed as contamination. All sequences were then uniformly fragmented to between 1 Kb and 15 Kb, using a custom Python script (available at https://github.com/sxh1136/Phage_tools/blob/master/fragment_genomes.py), to create artificial contigs. Each viral prediction tool was then run on the three sets of contigs (chromosome, plasmid, and phage) with default settings except for VIBRANT and VirSorter where the phage-derived contig set was additionally run using their virome decontamination modes, due to their authors noting that there may be improved performance in datasets consisting of mainly viral fragments when using this mode. Commands used to run each tool and their version numbers can be found at https://github.com/sxh1136/Phage_tools/blob/master/manuscript_tools_script.md. RefSeq benchmarking datasets are available for download and use at https://figshare.com/articles/dataset/RefSeq_Datasets_for_benchmarking_-_Ho_et_al_/19739884.

For tools where score/probability thresholds can be manually adjusted (DeepVirFinder, MetaPhinder, PPR Meta, Seeker and VirFinder), F1 curves were plotted (100 data points) with ggplot2 (v3.3.2) (Wickham, 2016) and optimal thresholds were determined by maximal F1 score. For viralVerify and VirSorter, which have categorical thresholds for phage identification, the category sets with the highest F1-score was plotted; for VirSorter two category sets were compared, category 1,2,4,5 and all categories, as these are commonly used (Nayfach et al., 2021; Roux et al., 2015). ViralVerify was also additionally run using both Pfam-A 33.0 and a provided database of virus/chromosome

specific HMMs as these are listed on the tool's GitHub usage guide. Kraken2 was run with the pre-built kraken2-microbial database, available at https://lomanlab.github.io/mockcommunity/mc_databases.html as this is listed on Kraken2's downloads page under "pre built kraken2 databases". The reference genomes of the two chromosomal fragments that Kraken2 identified as viral were downloaded from NCBI RefSeq (NC_CP059254.1, and NZ_CP07771.1). The fragments and the reference genomes were annotated with pharokka (Bouras et al., 2023) and globally aligned and visualised with clinker (Gilchrist & Chooi, 2021). Taxonomy of tool true positive and false positive viral predictions were retrieved using the R package Taxonomizr (v0.10.3) (Sherrill-Mix, 2023).

The run time of each tool on this dataset, utilising 16 threads, was recorded using a Linux virtual machine provided by Cloud Infrastructure for Big Data Microbial Bioinformatics (CLIMB-BIG-DATA) (Connor et al., 2016), with the following configuration: CPU: Intel®Xeon®E3-12xx v2 (Ivy Bridge, IBRS) (16 VCPU); GPU: Cirrus Logic GD 5446; Memory: 108 GB Multi-Bit ECC.

2.3.2 Benchmarking with Randomly Shuffled Sequences

All artificial phage contigs created in the previous benchmark were randomly shuffled whilst preserving dinucleotide distribution using esl-shuffle from the HMMER3 suite (v3.3.2, -d) (Mistry et al., 2013) Each identification tool was then run on the randomly shuffled sequences using the optimised thresholds that were determined in the RefSeq benchmark, and false positives were recorded.

2.3.3 Benchmarking with Mock Community Metagenomes

Three shotgun metagenomic sequencing replicates of an uneven mock community (Kleiner et al., 2017) were retrieved from the European Nucleotide Archive (BioProject PRJEB19901). These communities contain five phage strains: ES18 (H1), F0, F2, M13, and P22 (HT105). The quality of the data was checked using FASTQC (v0.11.8, default settings) (Andrews et al., 2019) and overrepresented sequences were removed with Cutadapt (v2.10, –max-n 0) (M. Martin, 2011). Cleaned paired-end reads were then assembled with MetaSPAdes (v3.14.1, default settings) (Nurk et al., 2017) and contigs <1 Kb were removed. Prophages were then removed from the contigs by sequentially running Phigaro and then PhageBoost, as with the RefSeq chromosomal and plasmid datasets. Each tool was then run on the three sets of contigs using optimal parameters as determined previously with the exception of viralVerify and VirSorter where all categorical thresholds were re-evaluated. MetaQUAST (v5.0.2, default settings) (Mikheenko et al., 2016) was used to map contigs to reference phage genomes and calculate coverage.

2.3.4 Benchmarking with Simulated Mock Virome Communities

Four mock communities (samples 13, 2, 7, and 9) containing between 500 and 1000 viral genomes created by Roux et al. were selected for analysis (Roux et al., 2017). These samples belonged to four different beta diversity groups and did

not share any of their 50 most abundant viruses. Each simulation of 10 million paired-end reads was quality controlled with Trimmomatic (Bolger et al., 2014) and assembled with MetaSPAdes by Roux et al. (Roux et al., 2017). The contigs were then downloaded for benchmarking. As before, contigs with length ≤1 Kb were removed, and then inputted into each viral identification program. Positive viral contig sets for each tool were then extracted and reads were mapped to these with BBMap (Bushnell, 2014) with ambiguous mapped reads assigned to contigs at random (`ambiguous=random`), as in Roux et al. (Roux et al., 2017). Primary mapped reads with pairs mapping to the same contig were then extracted with SAMtools (v1.16) (`options -F 0×2 0×904`) (H. Li et al., 2009) and mapped to a pool of non-redundant population contigs. This pool was created by clustering all four samples with nucmer (v3.1) (Marçais et al., 2018), at ≥95% ANI (average nucleotide identity) across ≥80% of their lengths. Abundance matrices for each tool were calculated by normalising read counts by contig length and total library size to produce counts per million (CPM). These abundance matrices were then used to calculate Shannon, Simpson, and Bray-Curtis dissimilarity indices using the vegan package (v2.5.7) (Oksanen et al., 2020). Non-metric multidimensional scaling (NMDS) and PERMANOVA (adonis2) using Bray-Curtis dissimilarity were also conducted with vegan. The independent variable for adonis2 was set as 'tool' and permutations were constrained by sample. Pairwise PERMANOVA p-values were corrected with the Benjamini–Hochberg method (Benjamini & Hochberg, 1995). Pairwise comparisons were carried out with the `pairwise_adonis` wrapper from https://rdrr.io/github/gauravsk/ranacapa/man/pairwise_adonis.html. Seed and permutations were set to 123 and 9999, respectively, where possible. All plots were generated with ggplot2 (v3.3.2) (Wickham, 2016) and arranged with

ggarrange from ggpubr (v0.4.0) (Kassambara, 2020).

2.4 Results

2.4.1 Benchmarking with RefSeq Phage and Non-Viral Artificial Contigs

Ten commonly used tools for viral sequence identification in metagenomes were selected for evaluation: DeepVirFinder, Kraken2, MetaPhinder, PPR Meta, Seeker, VIBRANT, viralVerify, VirFinder, VirSorter, and VirSorter2. All of these tools can be run locally without relying on a web server, accept metagenomic contigs as input, and have been published in the past decade (as of 21/01/2020).

We first evaluated all the programs on the same uniform datasets. All complete phage genomes deposited in RefSeq between 1 January 2020 and 12 August 2021 were downloaded, quality controlled, and fragmented to create a true positive set of artificial contigs. A negative set was constructed from all RefSeq bacterial and archaeal chromosomes and plasmids, submitted in the same time period. Multiple steps were taken to ensure these datasets did not include false positives or false negatives. First, all bacterial, archaeal, and phage RefSeq genomes deposited prior to 2020 and the training sets of each machine learning tool were used to dereplicate the datasets used in this study to remove any similar species that may cause overfitting of some tools. In addition, chromosome and plasmid sequences with $\geq$30% of their open reading frames having hits to the

pVOG database were removed to exclude any remaining viral sequences. As the negative dataset was considerably larger than the positive dataset, we subsampled the negative set by a factor of 14.3, resulting in 253 host chromosomes and 309 host plasmids (Table 2). This sampling rate was chosen to produce a phage:host ratio (∼1:19) that was similar to what is found in human gut microbiomes (Shkoporov & Hill, 2019). Finally, we removed chromosomes and plasmid sequences that contained integrated prophages using two state-of-the-art prophage detection tools, Phigaro (Starikova et al., 2020) and PhageBoost (Sirén et al., 2021), to prevent their erroneous identification as viral contigs (Figure 2.1). In total 2088 prophages were removed from the chromosome set and 91 from the plasmid set. The cleaned artificial contigs were then run on the benchmarking tools (Table 2.2).

All evaluated programs, except Kraken2, produce thresholds or confidence ranges for viral identification. For tools (DeepVirFinder, MetaPhinder, PPR Meta, Seeker, VirFinder, and VirSorter2) that assign a continuous threshold (score, identity, or probability), a F1 curve was plotted, and an optimal threshold was determined (Figure 2.2). For VirSorter and viralVerify, the categories that returned the highest F1 score were used (Figure 2.4) In most tools there was a trade-off between precision and recall. This is likely due to relaxed thresholds allowing for more viral and non-viral sequences to be detected, increasing recall, and decreasing precision simultaneously. For VIBRANT and VirSorter, the positive dataset was additionally run in virome mode and virome decontamination mode respectively, as this improves viral recovery in samples composed mainly of viral sequences by adjusting the sensitivity of the tool (Kieft et al., 2020; Roux et al., 2015).

The tools we benchmarked on this dataset had highly variable performance in terms of F1-score (0.44 – 0.93), precision (0.47 – 1.00), and recall (0.46 – 0.96) (Figure 2). VirSorter2, VIBRANT, and PPR Meta achieved the highest F1-scores of 0.93, 0.93 and 0.92 respectively. VirSorter2 achieved this with high precision (0.92) and well as high recall (0.93), VIBRANT had a higher precision (0.97) and lower recall (0.89), and PPR Meta had a slightly lower precision at 0.88 but higher recall at 0.96. The majority of the remaining tools performed well, with six tools (DeepVirFinder, Kraken2, MetaPhinder, VirFinder, and VirSorter) having F1-scores of over 0.83 (Figure 2.3). Kraken2 had a precision score of almost 1, with only 2 chromosomal fragments and no plasmid fragments flagged as viral, whilst still correctly identifying over 5000 phage fragments. These chromosomal fragments were originally part of the complete chromosomes of *Streptomyces albidoflavus* strain J1074/R2 and *Saccharolobus shibatae* strain BEU9 and Kraken2 identified them as *Streptomyces* phiC31 phage and *Sulfolobus* virus 1 (provirus) respectively. Global alignment of the fragments to their Kraken2 identified reference revealed limited homology between them and suggests that these are true false positive predictions (Figure 2.6 and 2.7). ViralVerify had a low precision score of 0.55 whilst its recall score of 0.88 was comparable to the other tools. Seeker had poor performance in both precision (0.48) and recall (0.41) compared to the other tools. Generally, k-mer based tools performed better than reference similarity/gene-based tools, although this trend was primarily influenced by the relatively low F1-scores of Seeker and ViralVerify, and the sample sizes of the investigated tools was too small to draw statistically significant conclusions. Across our benchmark, only 0.06% (4/6665) of phage contigs from our positive

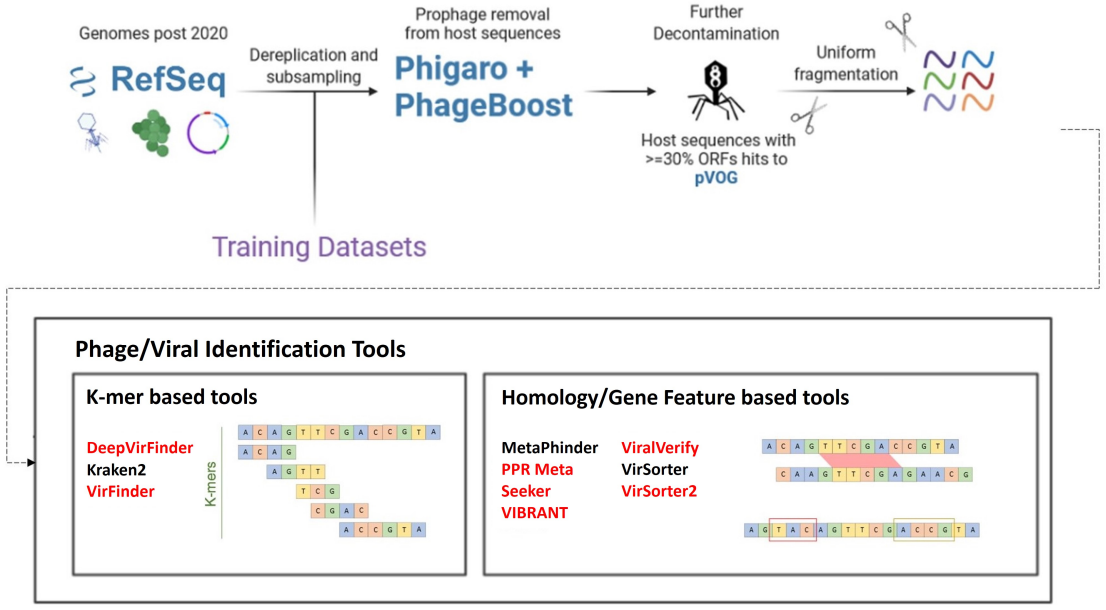


Figure 2.1: **Overview of RefSeq Artificial Contig Benchmarking Workflow.** All bacterial and archaeal chromosomes and plasmids and phage genomes that were deposited in the NCBI Reference Sequence Database (RefSeq) database between 1 January 2020 and 12 August 2021 inclusive were downloaded. The phage genomes were used to create a positive test set and the chromosomes and plasmids for a negative set. The sequences were dereplicated with the training sets for each machine/deep learning tool that was benchmarked (highlighted in red), as well as any RefSeq sequences deposited prior to 2020. The negative set was down sampled to produce a positive:negative ratio of approximately 1:19 to replicate a typical gut microbiome. Prophages were identified and removed with Phigaro and PhageBoost. Any host sequences with $\geq$30% of their open read frames having hits to the Prokaryotic Virus Orthologous Groups Database were then removed. All sequences were then uniformly fragmented into artificial contigs with lengths between 1 and 15 Kb. All identification tools were then run on the artificial contig sets.

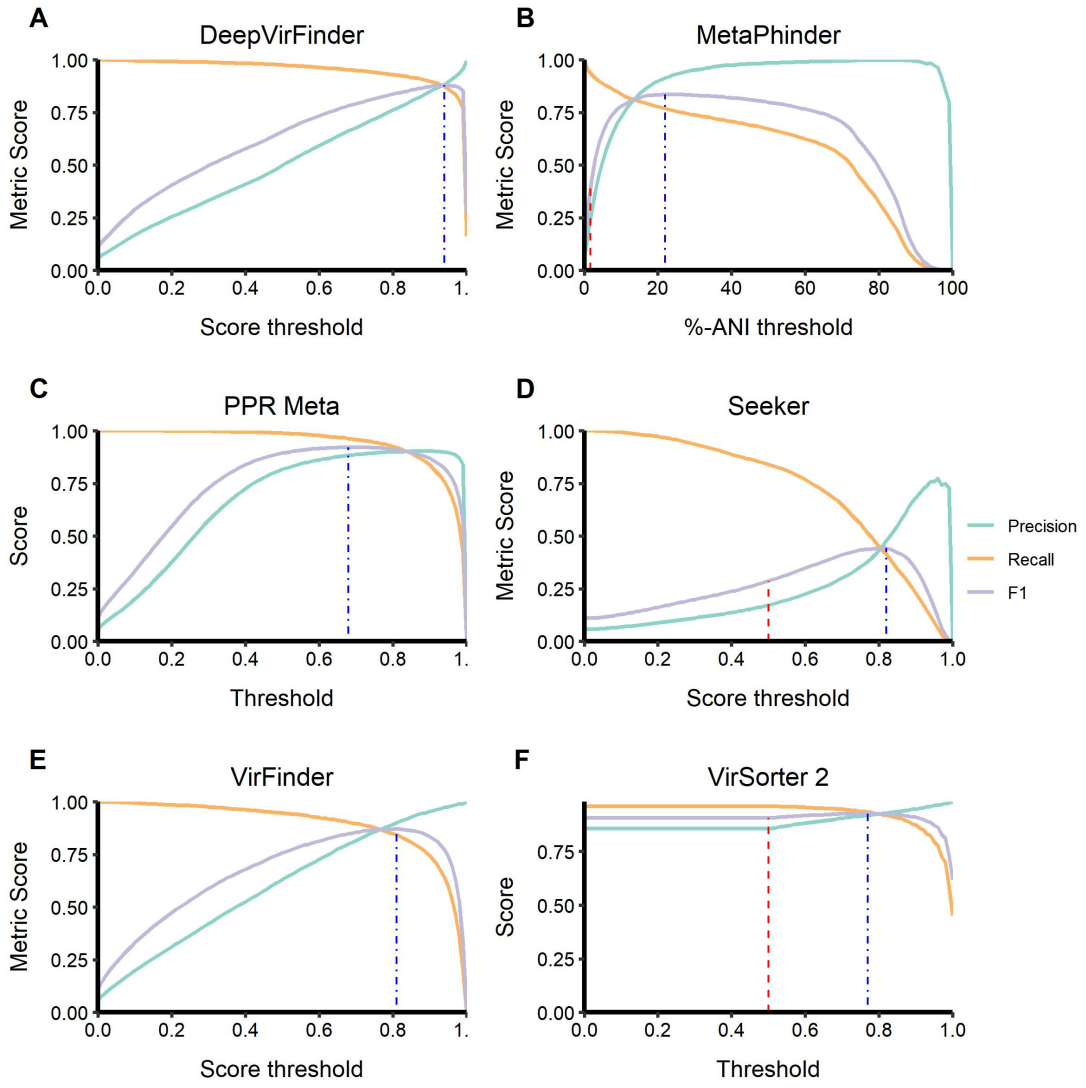


Figure 2.2: **Optimisation of detection thresholds (continuous) of viral identification tools.** DeepVirFinder, MetaPhinder, PPR Meta, Seeker, VirFinder, and VirSorter2 assign a continuous threshold (score, identity, or probability) for viral identification. We plotted precision, recall and F1-score against this threshold for 100 data points. The default threshold for MetaPhinder, Seeker, and VirSorter2 are indicated by the red dashed line. The peak F1-score is indicated by the dashed blue line.

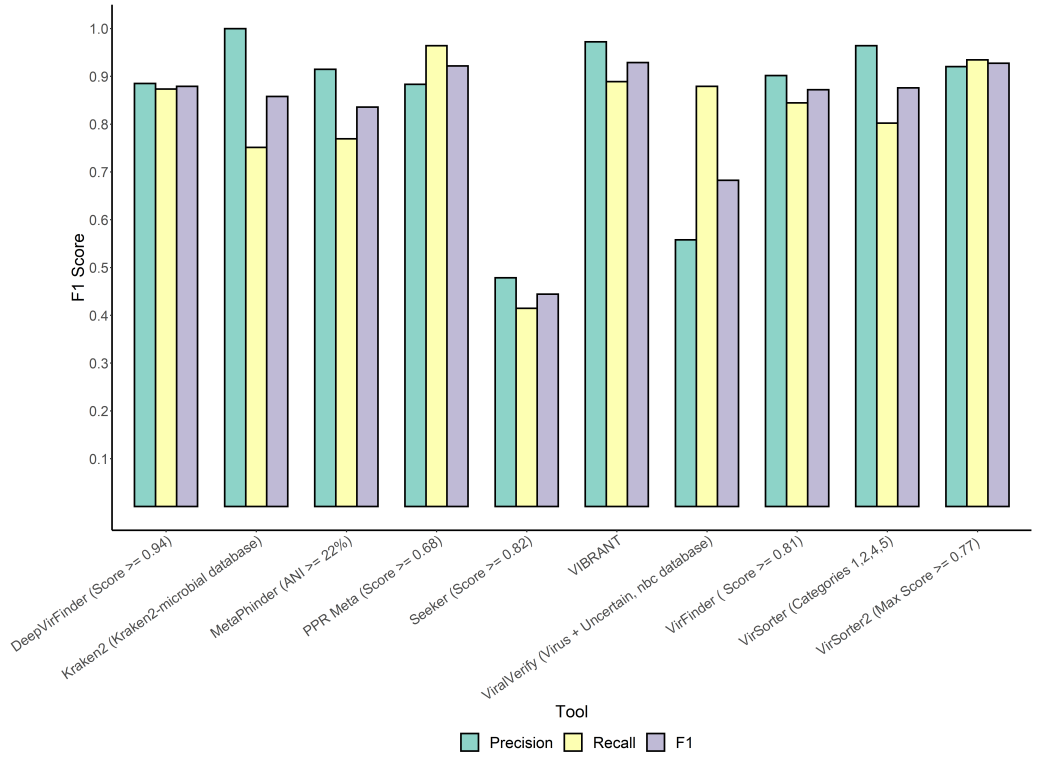


Figure 2.3: **Benchmarking of viral identification tools on artificial RefSeq contigs.** Contigs were generated by randomly fragmenting complete bacterial/archaeal/phage genomes and plasmids deposited in the NCBI Reference Sequence Database (RefSeq) between 1 January 2020 and 12 August 2021, to a uniform distribution. Each tool was then separately run on the true positive (phage genome fragments) and negative (bacterial/archaeal chromosome and plasmid fragments) datasets. For tools whose score/probability threshold or categories could be manually adjusted, values/categories were selected based on optimal F1-scores.

Table 2.2: **Number of sequences at each stage of the RefSeq benchmarking workflow.** Columns labelled with (n) contain the number of sequences at each step and columns with (bp) indicate the number of base pairs at each step. Sequences that were deposited between 1 January 2020 and 12 August 2021 were downloaded from NCBI Reference Sequence Database (RefSeq). They were then dereplicated with RefSeq sequences deposited before 1st January 2020 and with the training sets for DeepVirFinder, Seeker, VIBRANT, VirFinder, VirSorter2. Chromosome and plasmid sequences were subsampled by a factor of 14.3. Prophages were removed from chromosome and plasmid sequences using Phigaro and PhageBoost. The number of prophages removed are indicated in parentheses. Chromosomes and plasmid sequences with ≥30% of their open reading frames having hits to the pVOG database were removed. All sequences were then uniformly fragmented to between 1 and 15 Kb.

Dataset	Chromosome (n)	Plasmid (n)	Phage (n)	Chromosome (bp)	Plasmid (bp)	Phage (bp)
Post-2020	7,400	9,960	1,849	30,034,515,475	1,009,997,498	128,660,045
Post-2020 Dereplicated	3,546	4,453	901	14,845,391,115	480,796,703	62,387,054
Subsampled	253	309	901	1,011,740,231	28,782,822	62,387,054
Prophage Removed	2,307 (2,088)	400 (91)	901	965,959,872	27,323,488	62,387,054
pVOG Removed	2,065	313	901	848,361,160	24,433,971	62,387,054
Fragmented Artificial Contigs	104,003	2,754	6,664	830,889,456	21,783,942	53,426,665

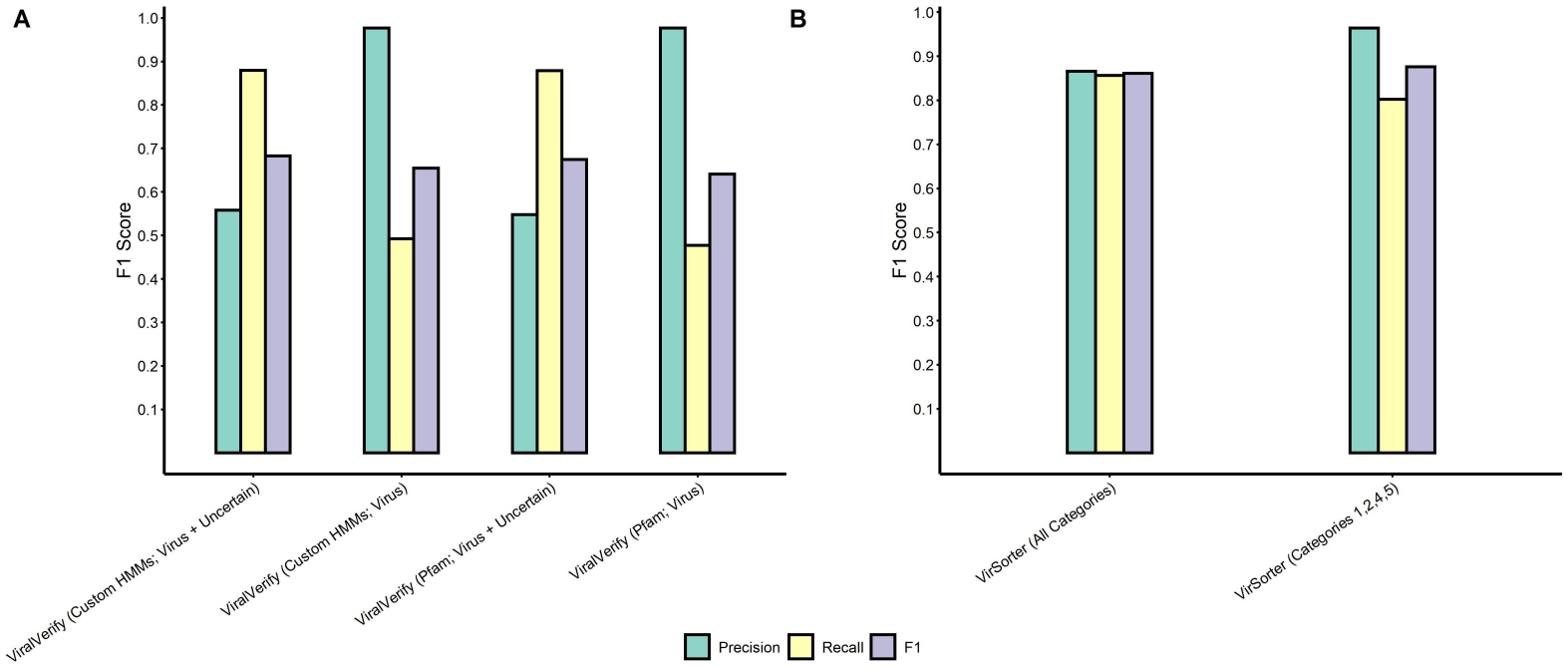


Figure 2.4: **Optimisation of detection thresholds (categorical) of viral identification tools.** ViralVerify and VirSorter have categorical settings that adjust their confidence for viral classification. We plotted the precision, recall, and F1-score for these categories. **(A.)** ViralVerify is packaged with two different databases on their GitHub page - their custom HMMs, and Pfam A 33.0. ViralVerify also has two categories for viral identification - Virus and Uncertain. We plotted the evaluation metrics for both databases with just the Virus category or for both the Virus and Uncertain category. **(B.)** VirSorter has six different categories for viral identification. Categories 1 - 3 are for viral fragments and 4 - 6 are for prophages. We split our optimisation of VirSorter into two sets; the first set contained any contig that was classified into any of the six categories, and the second set only included the top two categories for viral fragments and prophages (categories 1,2,4,5).

set were not detected by any tool, with 89.7% being identified by over half the tools, and 11.6% (775/6665) found by all 10 tools.

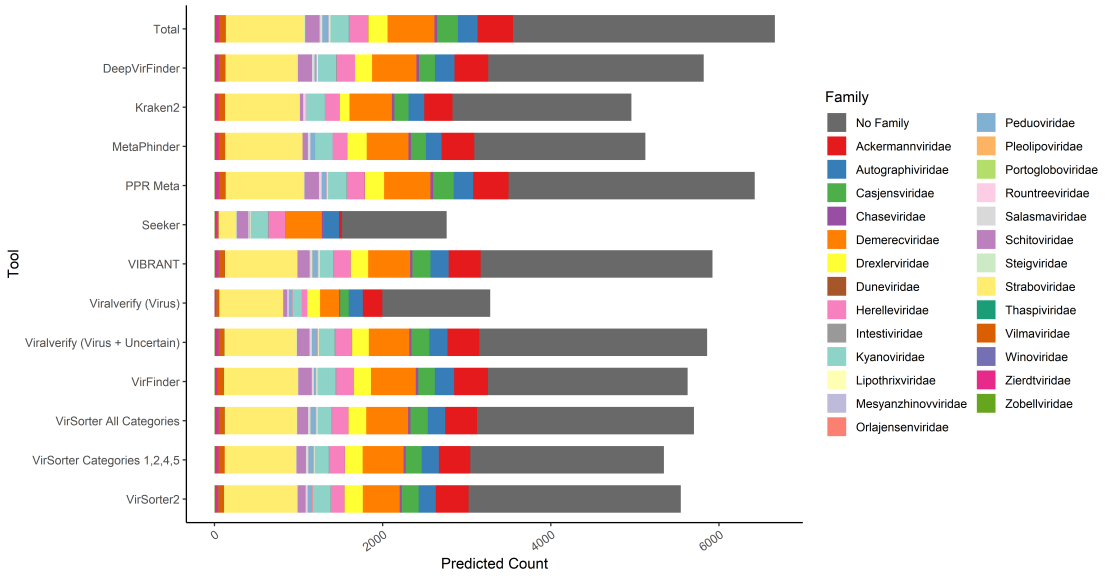


Figure 2.5: **Taxonomy of each tool's viral prediction on RefSeq phage fragments.** Viral predictions of each tool on the RefSeq true positive phage genome fragments were computed and the associated RefSeq accessions were extracted. The taxonomy of these accessions was obtained using the R package Taxonomizr. Family level taxonomy was then visualised facetted by tool predictions and compared to the taxonomy of the original fragments (shown in Total). Sequences with no family level taxonomy assigned were labelled as "No Family".

No significant taxonomic bias was found for any viral family or between each tool. The only exception is with Seeker which classifies fewer Ackermannviridae fragments than the other tools (Figure 2.5). The taxonomy of false positive viral predictions did however vary between tools and between the chromosomal and plasmid datasets. Nearly half (44.7%) of DeepVirFinder's chromosomal false positive hits (FPHs) belonged to Clostridia despite the class

only making up 1.27% of the dataset. False positive predictions by Seeker on the chromosomal dataset were biased towards Bacilli (30.0% of FPHs versus 13.8% of the chromosomal dataset) and Actinomyces (29.2% of FPHs versus 16.1% in the chromosomal dataset). For all other tools, except for Kraken2, the chromosomal false positive taxonomic profiles roughly matched with the overall distribution in the dataset (Figure 2.8). Seekers' false positive hits on the plasmid dataset were dominated by fragments belonging to the Halobacteria class (75.6% of FPHs versus 3.63% of the plasmid dataset. VirFinder also had a bias towards Halobacteria plasmid fragments with it making up 20.0% of its FPHs. Plasmid fragments of the Mollicutes class were overrepresented in the false positive hits of every tool except Kraken2 and viralVerify (Virus), which had zero and two false positive hits respectively. This overrepresentation was particularly pronounced in PPR Meta (10.5% of its FPHs); VIBRANT (5.66% of its FPHs); VirSorter All Categories (11.4% of its FPHs); VirSorter Categories 1,2,4,5 (14.2% of its FPHs); and VirSorter2 (8.2% of its FPHs), whereas Mollicutes made up only 0.36% of the total plasmid fragments (Figure 2.9). These percentages must be taken in the context of each tool's overall performance as the total number of FPHs varies greatly between the tools - ranging from 2 to 4331 in the chromosomal dataset and between 0 and 313 in the plasmid dataset.

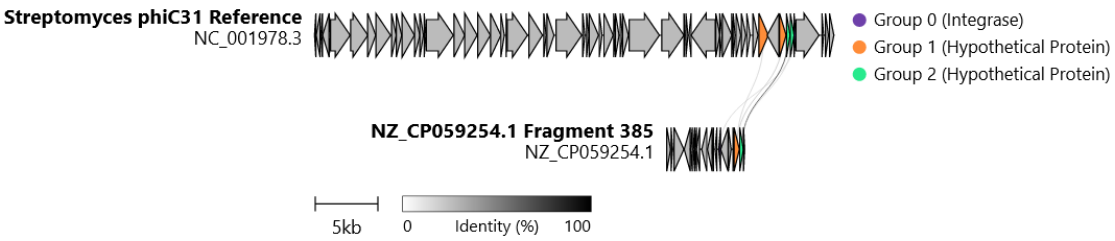


Figure 2.6: **Comparison of artificial RefSeq contig NZ_CP059254.1 fragment 385 to Streptomyces phiC31 phage reference genome (NC_001978.3).** The 6217 bp contig NZ_CP059254.1 fragment 385 was identified as Streptomyces virus phiC31 by Kraken2. The contig originally belonged to Streptomyces albidoflavus strain J1074/R2 complete chromosome. The contig and Streptomyces virus phiC31 reference genomes were annotated with pharokka and the genes were globally aligned and visualised with clinker. Homologous genes between the two sequences are coloured by homology and annotated with their predicted function from pharokka.

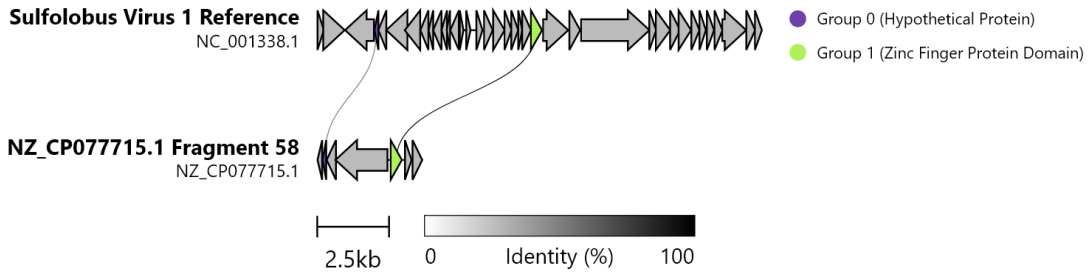


Figure 2.7: **Comparison of artificial RefSeq contig NZ_CP077715.1 fragment 358 to Sulfolobus virus 1 (provirus) reference genome (NC_001338.1).** The 3647 bp contig NZ_CP077715.1 fragment 58 was identified as Sulfolobus virus 1 by Kraken2. The contig originally belonged to Saccharolobus shibatae strain BEU9 complete chromosome. The contig and Sulfolobus virus 1 reference genome were annotated with pharokka and the genes were globally aligned and visualised with clinker. Homologous genes between the two sequences are coloured by homology and annotated with their predicted function from pharokka.

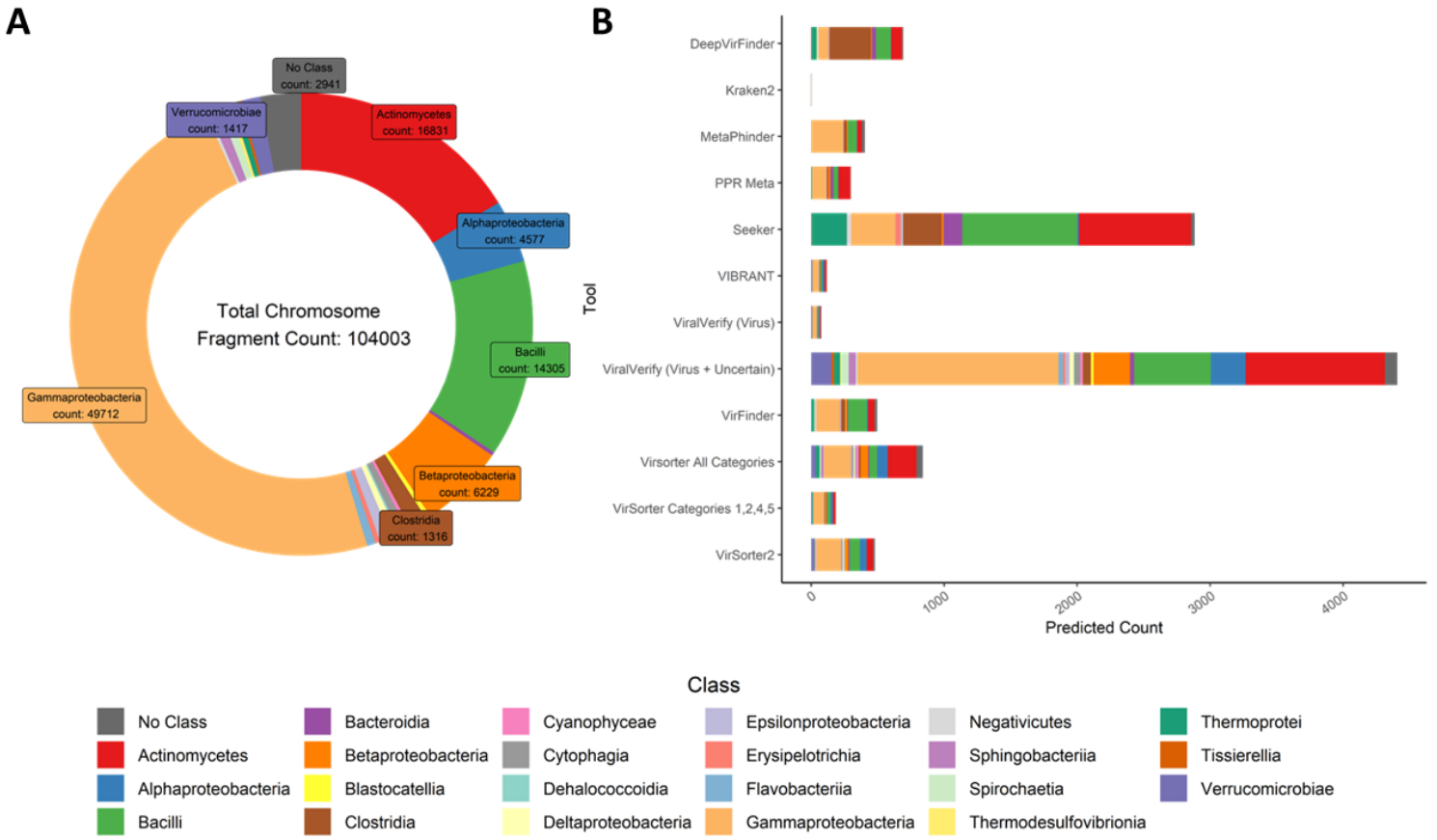


Figure 2.8: **Taxonomy of false positive viral prediction on RefSeq chromosomal fragments.** Accessions of NCBI Reference Sequence Database (RefSeq) chromosomal fragments were converted into class level taxonomy with Taxonomizr. Sequences with no class level annotation were labelled as "No Class". Legends and colours are shared by parts A and B. **(A)** The taxonomic proportion of bacterial classes in the RefSeq chromosomal fragment dataset. Classes with over 1000 fragments were labelled for clarity. **(B)** Taxonomy of false positive viral predictions on the RefSeq chromosomal dataset by each tool.

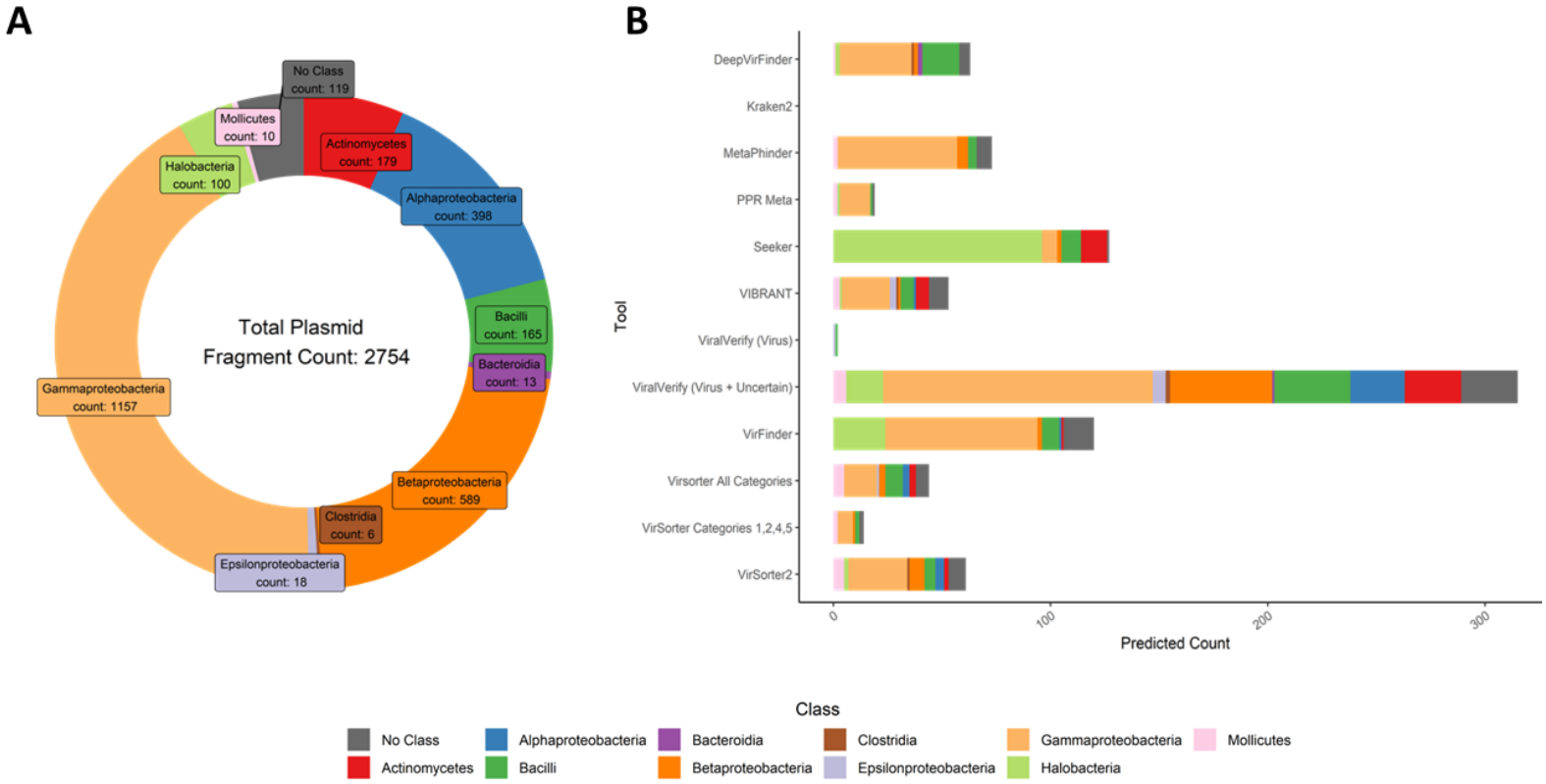


Figure 2.9: **Taxonomy of false positive viral prediction on RefSeq plasmid fragments.** Accessions of NCBI Reference Sequence Database (RefSeq) plasmid fragments were converted into class level taxonomy with Taxonomizr. Sequences with no class level annotation were labelled as "No Class". Legends and colours are shared by parts A and B. **(A)** The taxonomic proportion of bacterial classes in the RefSeq plasmid fragment dataset. All classes are labelled on the doughnut chart. **(B)** Taxonomy of false positive viral predictions on the RefSeq plasmid dataset by each tool.

2.4.2 Benchmarking Tools with Randomly Shuffled Sequences

To serve as a further negative control, the positive RefSeq benchmark contigs were randomly shuffled at a nucleotide level to produce sequences that should not be identified as viral by any of the tools. Of the tools tested, four identified zero shuffled contigs (MetaPhinder, viralVerify, VirSorter, and VirSorter2), Kraken2 classified three contigs, DeepVirFinder and VirFinder detecting 742 and 1070 respectively, with the rest of the tools identifying over 2500 shuffled contigs as viral, including PPR Meta which incorrectly classified 99.2% (6608/6664) of all the shuffled contigs as viral (Table 2.3).

2.4.3 Benchmarking Tools with Mock Community Shotgun Metagenomes

We next sought to compare these tools on real community shotgun metagenomic contigs. Thus, we obtained sequencing data of an uneven mock community created by Kleiner et al. (Kleiner et al., 2017), containing 32 species from across the tree of life, including five bacteriophages, at a large range of cell abundances (0.25%-21.25%). Integrated prophages were removed with Phigaro and PhageBoost to prevent them from being flagged as viral. No contigs belonging to the five bacteriophages were removed during this process. This allowed us to assess the performance of our tools on real data whilst retaining knowledge of the

Table 2.3: **Performance of tools on randomly shuffled artificial phage contigs.** Artificial NCBI Reference Sequence Database (RefSeq) phage contigs generated in the previous benchmark were randomly shuffled whilst preserving the dinucleotide distribution using esl-shuffle from the HMMER3 suite. Phage detection tools were then run on the shuffled contigs, and any positive hits were recorded. Tools shaded in grey indicate methods utilising machine/deep learning models.

Tool	False positives
DeepVirFinder	11.13%
MetaPhinder	0.00%
VIBRANT	58.76%
ViralVerify	0.00%
VirSorter	0.00%
VirSorter2	0.00%
Seeker	38.99%
PPR Meta	99.16%
VirFinder	16.06%
Kraken2	0.05%

ground truth (sample composition) and determine each tool's detection limit on low abundance species. The optimised parameters found in the RefSeq benchmark were used for each tool, with the exception of viralVerify and VirSorter. These tools have categorical thresholds which drastically change the profile of identified viral contigs, and the F1-scores between the thresholds were

very close, so the parameters were further analysed in this benchmark. In general, the tools' F1-scores were considerably lower on this dataset than on the RefSeq artificial contigs, with F1-scores dropping by an average of 40.6%, compared to the RefSeq benchmark (Figure 2.10). Kraken2 outperformed all other tools with a F1-score of 0.86, 0.3 higher than DeepVirFinder in second place. This is due to its high precision of 0.99 whilst identifying 76% of the phage contigs. DeepVirFinder had a high recall rate (0.80), but unlike in the RefSeq benchmark, had a lower precision of 0.42. Several other tools had similar results, with PPR Meta, MetaPhinder, and VirFinder all achieving high recalls of 0.91, 0.98, and 0.83 respectively, but having comparatively low precision scores (0.24, 0.28, and 0.35 respectively). VirSorter (categories 1, 2, 4, and 5) and Seeker had comparable performances to the tools above, with them both having F1-scores of 0.44. Seeker, along with Kraken2, were the only tools to attain similar F1-scores in this benchmark to the RefSeq benchmark. VirSorter2, which performed best in the RefSeq benchmark, had a lower F1-score (0.36) than its predecessor mainly due to its high number of false positive hits. VIBRANT, which also performed well in the previous benchmark, again had poor precision and middling recall resulting in a F1-score of 0.32. ViralVerify (NBC database, virus only) had both low precision (0.24) and recall (0.22) resulting in the lowest F1-score (0.22) in this test. K-mer tools on average had a higher F1-score than the reference similarity/gene based tools but this difference was not statistically significant due to the small sample size of the number of tools in this study ($p = 0.094$, one-tailed Welch's t-test).

Out of the four DNA phage species found in the assemblies (Phage F2 was

not sequenced due to being a ssRNA virus), only MetaPhinder was able to detect M13, whilst F0, the most abundant phage species, was detected by all benchmarked tools in all samples.

MetaPhinder identified contigs belonging to all four phage species in two samples and three in the other sample. PPR Meta, VIBRANT, VirSorter, and VirSorter2 were able to identify contigs belonging to three species in all three samples. ViralVerify and VirFinder were able to identify the three phage species in two out of three samples, missing out on contigs belonging to phage ES18. DeepVirFinder and Kraken2 classified viral contigs belonging to three phage species in one out of three samples, and detected two species in the other samples. Seeker was only able to identify contigs belonging to the most abundant phage F0. No correlation was found between F1-score and the number of phage strains detected (Rs = -0.371, p = 0.29) but a positive, but not statistically significant, correlation, was observed between tools that identified more contigs of viral origin (true positives + false positives), and the number of phage strains identified (Rs = 0.604, p = 0.06).

2.4.4 Impact of Tool Prediction on Diversity Metric Estimation

To test the impact of these tools on diversity estimations, four simulated mock community metaviromes containing an average of 719 viral genomes were retrieved from (Roux et al., 2017). Reads were mapped to contigs ($\geq$1 Kb) that were identified as viral by each tool, and these mapped reads were then mapped

to a set of population contigs to estimate their abundance in each sample. Reads were also directly mapped to the population contigs as a control. Read counts were then normalised by their length and sequencing depth, which Roux et al., 2017 found to be a reliable normalisation method. Diversity estimation metrics were then calculated using the normalised population counts. All tools returned fewer genomes per sample compared to the initial population, although there was significant variation between tools. PPR Meta, MetaPhinder, and Kraken2 retrieved the greatest percentage of genomes with 86.8%, 89.1%, and 83.7% respectively (Figure 2.11A). All other tools were able to retrieve more than 50% of the genomes with the exception of Seeker and viralVerify, which were only able to recover 32.4% and 41.3%, respectively. All Shannon's alpha diversities calculated from the count matrices of each tool were within 10% of the default population with the exception of Seeker, whose *H'* score was on average 27.0% lower (Figure 2.11B). Simpson alpha diversity indices showed similar performance, with all tools having a diversity score within 1% of the initial population, with the exceptions of DeepVirFinder and Seeker, who were 1.4% and 5.1% divergent, respectively (Figure 2.11C). PPR Meta was the only tool to estimate a comparatively higher alpha diversity than the default population. In order to compare the diversities between tools and samples, we calculated the beta diversity using pairwise Bray-Curtis dissimilarities. This was then plotted on a non-metric multidimensional scaling plot (NMDS) with two dimensions (stress = 0.11). Bray-Curtis dissimilarities were chosen as they are able to handle null values which are prevalent in these abundance matrices. Using NMDS, we observed that communities predicted by Seeker did not cluster with the default population or those predicted by the other tools (Figure 2.11D). Subsequent

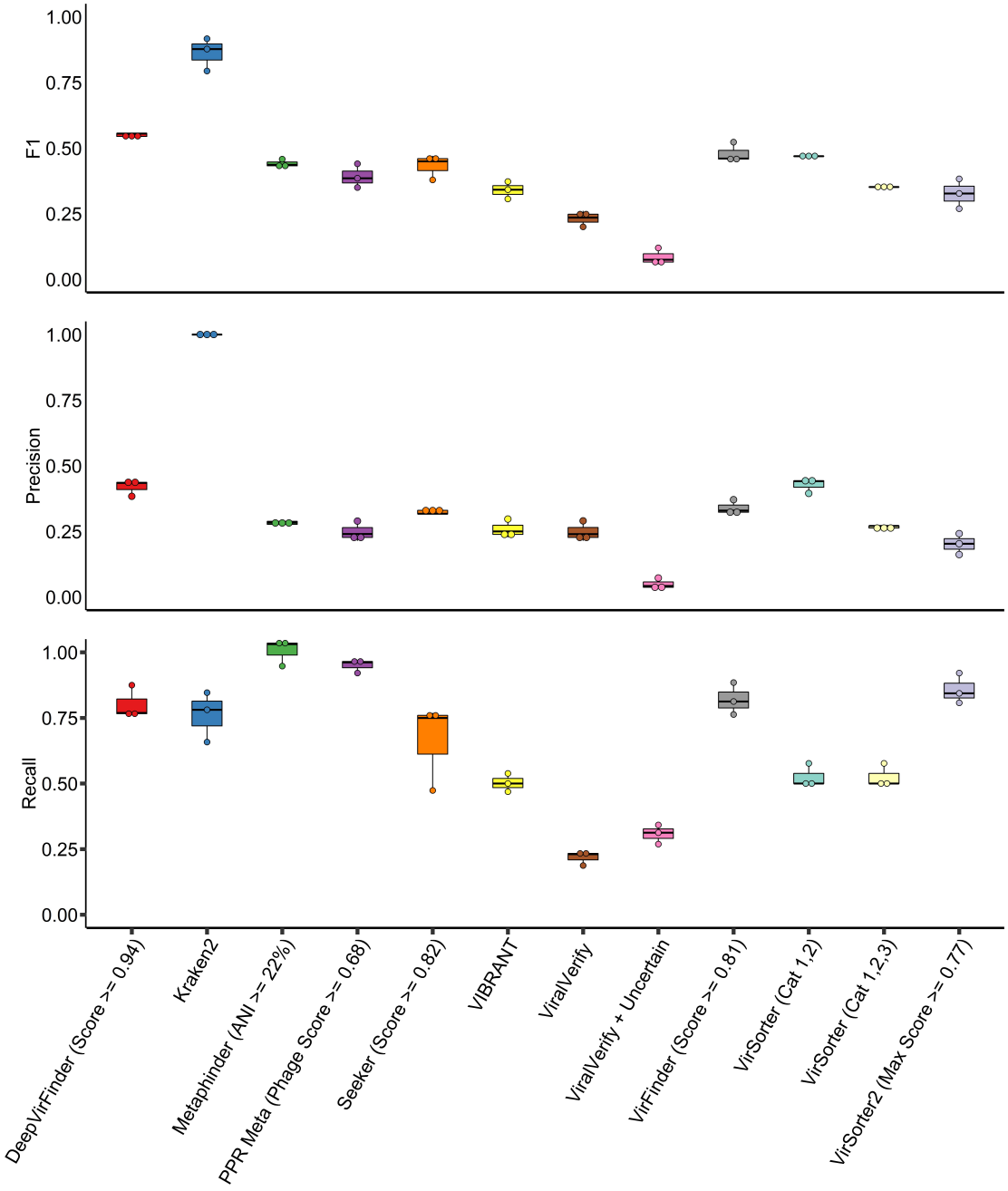


Figure 2.10: Caption on next page.

Figure 2.10: **Comparison of viral identification tools on uneven mock community samples.** Mock community reads were retrieved from a previous study (Kleiner et al., 2017), and assembled with metaSPAdes. Prophages were detected and removed with Phigaro and PhageBoost before running each identification tool using optimal thresholds based on previous benchmarks, except for viralVerify and VirSorter where both sets of categories are shown. F1-score, Precision, and Recall metrics are displayed as separate panels. Each sample is plotted as a single point for each tool, with a boxplot indicating the interquartile ranges, extremes and mean of all three samples.

PERMANOVA analysis indicated that tool selection had a significant effect on community composition, though the explained variation was relatively small ($p = 0.001$, $R^2 = 0.07$). The modest impact on variation may be due to Seeker uniquely introducing a substantial amount of variation compared to other tools. To test this hypothesis we conducted pairwise PERMANOVAs as a post-hoc test. However, these did not reveal any statistically significant differences, likely due to our small sample size of 4, although the explained variation for pairwise comparisons involving Seeker were much higher than those which did not (6.7% vs 0.07% respectively).

2.4.5 Runtime and Computational Load of Each Tool

We also recorded the running times of each tool on the RefSeq positive dataset on a high-performance cluster (16 VCPU, 108 Gb RAM) (Figure 2.12). Kraken2 and VirFinder were the fastest tools finishing in under 5 minutes. DeepVirFinder, MetaPhinder, PPR Meta, and Seeker finished in under half an hour with viralVerify

finishing just over that mark. VirSorter and VirSorter2 took the longest time to run on this dataset (2.9 hours and 3.8 hours to complete, respectively).

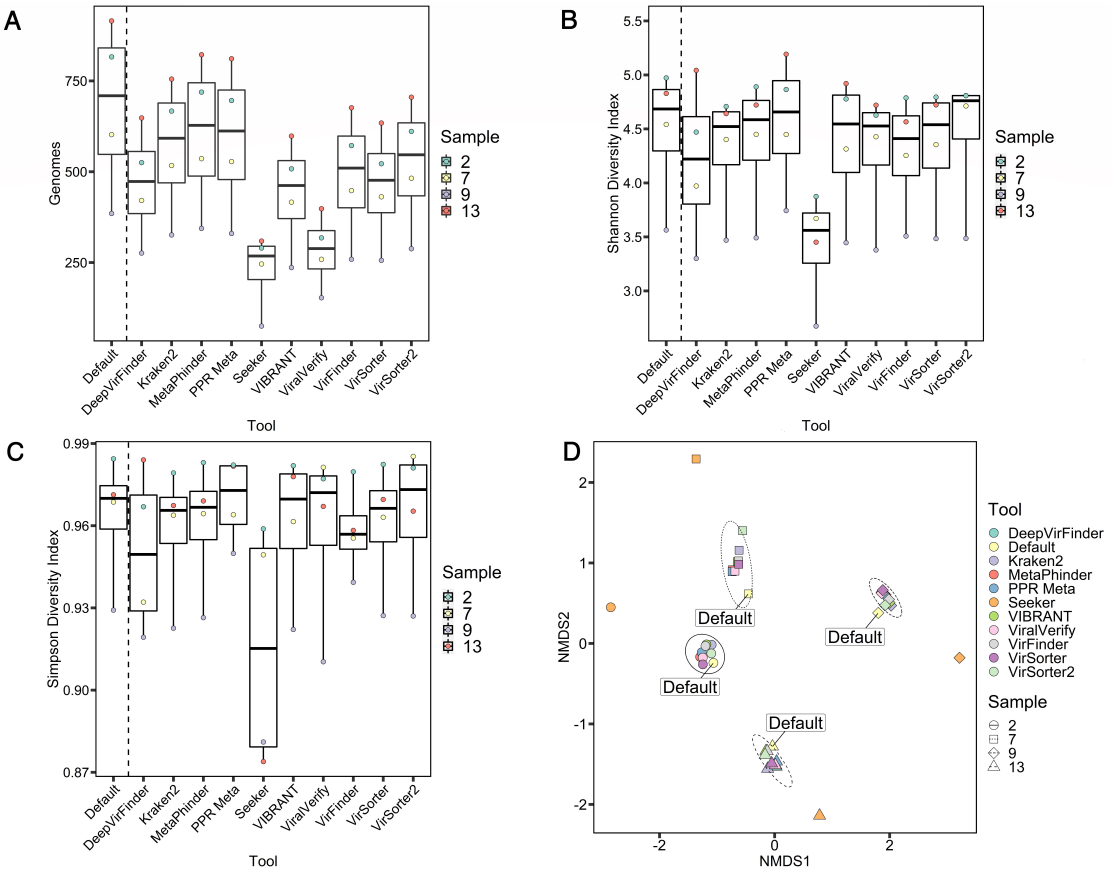


Figure 2.11: (Caption on the next page.)

Figure 2.11: **Estimation of diversity metrics of tool predicted virome populations.** To assess the impact of each tool on population diversity, four simulated virome assemblies from Roux et al. were downloaded. Each program was then run to determine the subset of predicted viral contigs. Reads were mapped to these contig subsets and mapped reads were then subsequently mapped to a pool of population contigs. All diversity metrics were computed by the R package "vegan". 'Default' in each plot indicates each sample's original assembly. **(A.)** Number of genomes observed from read mapping to predicted viral contig populations for each tool. **(B.)** Comparison of estimated Shannon diversity indices from each tool's virome subset. Estimations are based on read counts that were normalised by contig size and sequencing depth of the virome. **(C.)** Comparison of Simpson diversity indices from each tool's virome subset. **(D.)** Non-metric multidimensional scaling (NMDS) ordination plot of Bray-Curtis dissimilarity of virome subsets predicted by each viral identification tool. The stress for this plot is 0.11. Ellipses indicate the 95% confidence interval for each sample cluster's centroid. Samples are represented by the same symbol and ellipse line type; tools are denoted by colour.

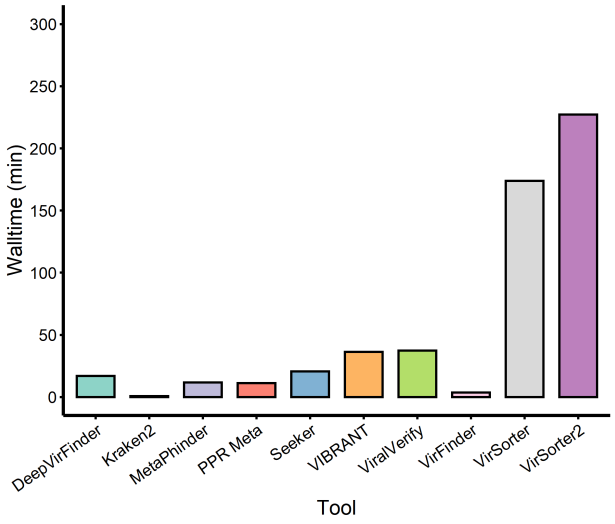


Figure 2.12: **Comparison of tool runtimes on the positive RefSeq artificial contig set.** Wall runtime of each tool on mock community samples was recorded on a 16 VCPU, 108GB RAM, Linux high-performance cluster without GPU acceleration. Tools were run with 16 threads where it could be set as a parameter (all tools except MetaPhinder, PPR Meta and Seeker). The RefSeq positive set contains approximately 53.4 million bp and 6664 contigs.

2.5 Discussion

Bacteriophages are crucial members of microbial communities in nearly every ecosystem on Earth, responsible for controlling host population size as well as having wider impacts on community functions. Tools designed to recover viral sequences from mixed community metagenomic and virome samples are fundamental to studying the role of bacteriophages in the wider context of their environment. Advancements in this field have produced an extensive suite of

viral identification tools that each claim to improve on the performance of similar tools. Selecting which tool among these is ideal for a dataset is thus not straightforward, especially as each novel tool typically only benchmarks against two or three other existing tools. Most tools developed for this purpose, especially those released in recent years have utilised machine/deep learning to classify sequences, whereas others rely on categorising sequences based on their similarity to sequences in databases. Both these approaches have the potential to improve over time with newly discovered viral genomes being added to training datasets and databases.

Here we compare ten methods for identifying viral sequences from metagenomes across three datasets. We first benchmarked the tools on positive and negative datasets to evaluate their performance on an ideal set of contigs (size between 1 and 15 Kb, without misassemblies), and determine approximate optimal thresholds. Most tools performed well here, detecting the majority of phage sequences, whilst keeping false positives low. PPR Meta, VIBRANT, and VirSorter2, which all use different machine learning methods, had the best performance across the tools. Whilst the optimal thresholds that we determined may not necessarily be ideal for all other datasets, we believe they can be used as a basis for further usage of these tools as in each case they produced considerably better results than the default parameters. We therefore encourage researchers to apply these thresholds and parameters within the context of their prospective dataset. Phage taxonomy did not affect viral prediction of the tools whilst chromosomal and plasmid taxonomy did reveal biases in several tools. Therefore, when analysing metagenomes it is important to consider what viral hosts you

expect to encounter in your sample, and whether the bias of the selected viral prediction tool will affect your downstream analysis.

Contrasting results between machine learning and more traditional tools were found when analysing their identification of randomly shuffled phage sequences. Of the six tools that utilise machine/deep learning methods, five identified a significant proportion of the sequences as viral, with only VirSorter2 as the only exception, probably due to its classifier being trained on a range of sequence and gene features. Three of the other four tools returned zero false positives and Kraken2 only returned three. This highlights that whilst these machine/deep learning algorithms have the capability to detect novel phages, their performance may be unpredictable when exposed to novel data with features that differ from those in our training sets.

When tested on real metagenomic data, most tools performed significantly worse than in the RefSeq benchmark, with the exceptions of Kraken2 and Seeker. Generally, k-mer tools had a smaller drop in F1-score from their RefSeq benchmark compared to reference similarity/gene-based tools. This is probably due to the comparatively shorter phage sequences that were assembled from the metagenomic reads, which provide less genetic context, thereby negatively impacting the algorithms of the reference similarity and gene-based tools. MetaPhinder was the only tool which was able to detect the one contig of phage M13 that was assembled in each sample. Unfortunately, this is likely due to MetaPhinder's low precision in this benchmark resulting in many false positive calls. Most other tools were able to identify contigs belonging to the other three phages, with the exception of Seeker which could only identify the most

abundant phage F0 in each sample. This suggests that when analysing metagenomic datasets where phage species are likely to be in low abundance, k-mer based tools such as Kraken2 and DeepVirFinder are the best choices, the former being the favoured option when precision is of particular importance, whilst the latter's use is appropriate if the discovery of novel phages is of interest due to its deep learning algorithm.

We also gauged any potential biases and impacts these tools may have on the diversity of its predicted viral population. Most tools performed well with alpha diversity indices within 10% of the default population with the exception of Seeker which returned a considerably lower value due to the very low number of viral population genomes Seeker originally predicted. Some tools such as PPR Meta predicted higher alpha diversity than the default population. This is due to the tools missing some high abundance genomes from their predictions, resulting in a more even diversity distribution. When evaluating beta diversity, Seeker was the only tool that produced results that had significant dissimilarity from the other tools and did not cluster with the other programs, again as a result of the low proportion of genomes it recovered in this dataset. Hence beta diversity trends of the tools examined here, with the exception of Seeker, are accurate to the original population, even when only half the genomes are recovered.

Runtime and computational load are also important factors to examine since these can become practical limitations if large samples take many hours or days to be analysed. Most tools were reasonably fast, although VirSorter, and its successor VirSorter2, took multiple hours to complete their runs. It is important to note that VIBRANT, VirSorter, and VirSorter2 annotate the identified viral

genomes and predict prophages which comes at the expense of runtime, although these can be useful for some applications. Kraken2 was by far the fastest tool, taking less than a minute to run on our dataset. However, Kraken2 requires very high RAM usage compared to the other benchmarked tools so it may not be feasible for researchers with limited computational resources.

2.5.1 Limitations

Although these benchmarks comprehensively compared the performance of state-of-the-art tools, there are a number of limitations to our study. Firstly, we used the default database(s) or the original trained model(s) that was provided with each tool. Whilst providing each tool the same database, or dataset to be trained on, may have been a fairer comparison of the underlying algorithms, this was beyond the scope of our study. We note that most users are unlikely to retrain these tools prior to their use or rigorously benchmark the results.

Whilst our RefSeq benchmarking set included all types of phage genomes including dsDNA, ssDNA, RNA, and archaeal viruses, we did not investigate the performance of each tool on each group. While we have approached phages as a single group for benchmarking purposes, there are many features which are specific to each viral group as they are governed by different ecological and evolutionary factors. As the vast majority of phages studied and sequenced to date are dsDNA phages from the *Caudoviricetes* class, some of the tools we benchmarked may struggle with the less researched viruses. An exception to this may be VirSorter2 which includes separate classifiers for five different virus groups including ssDNA

and RNA phages. Further work should identify how each tool performs with the ever-growing number of viral groups, so researchers can tailor their bioinformatic pipelines to their viruses of interest.

In addition, whilst we noted that tools have varying performance in each of the benchmarks we did not investigate the features of each method that led to those results. Further research into this would be especially useful with the randomly shuffled phage contigs, as some tools like PPR Meta had extremely high false positive rates, and knowledge of what exactly is causing those is crucial for future tool development.

We also did not assess the performance of combining multiple tools, which could provide meaningful insights that would be missed when only one single tool is used, as in Marquet et al. where the authors combined multiple tools into a single workflow (Marquet et al., 2022). However, this approach does not favour any particular prediction tool, resulting in the user having to manually select which contigs they believe to most likely be from phage genomes. An alternative would be to automatically integrate the tools using the benchmarking results in this chapter, in the same manner as how phage host prediction methods have been combined into a single framework by Roux et al., 2023.

Many of the tools also have additional functionalities, which we did not benchmark here but may nevertheless impact a researcher's choice of tool such as prophage prediction (VIBRANT, VirSorter, VirSorter2); plasmid prediction (PPR-Meta); taxonomic identification (Kraken2); and functional annotation (VIBRANT).

Finally, a few recently developed tools we found during our study were not included in our benchmarking either due to (1) requiring the use of its own web server and therefore not being scalable (VIROME, VirMiner), (2) lack of clear installation/running instructions (ViraMiner), or (3) errors when attempting to use the tool, which we were unable to resolve (PhaMers, VirNet, VirMine).

2.6 Conclusion

Our comparative analysis of ten currently available metagenomic virus/phage identification tools provides valuable metrics and insights for other investigators to use and build on. Using mock communities and artificial datasets, precision, recall and biases of these tools could be calculated. By adjusting the filtering thresholds for viral identification for each tool and comparing F1 scores, we were able to optimise performance in every case. In the artificial contig benchmark, most tools performed well, with PPR Meta, VIBRANT, and VirSorter2 having the highest F1-scores. In the mock-uneven community dataset, tools generally performed worse with the exception of Kraken2, whose performance included almost perfect precision with above average recall scores. All tools except Seeker were able to produce a diversity profile with similar indices to the original virome population, and are suitable for phage ecology studies. We suggest that of the currently available metagenomic phage identification tools, Kraken2 should be considered when researchers are trying to identify previously characterised phages. When novel phage detection is required Kraken2 should be used in combination with tools such as VirSorter2 and DeepVirFinder.

chapter 3

Investigating Bacteriophage Dynamics in the Intensive Care Unit using Metagenomic Sequencing Methods

3.1 Background

Patients admitted into an intensive care unit (ICU) are subjected to various stresses such as antibiotic exposure (Iapichino et al., 2008; Lankelma et al., 2017), non-antibiotic drug interventions (Moron et al., 2019), nutritional changes (Krezalek et al., 2017), sepsis (Alverdy & Krezalek, 2017), and surgery (Krezalek et al., 2016). These factors are all associated with dysbiosis in the gut flora, which in turn increases patients' susceptibility to disease and infection (Kitsios et al., 2017). Antibiotic therapy in particular is related to significant alterations to the commensal microbiota, as the regimens often target large parts of the community indiscriminately (Modi et al., 2013). Furthermore, the use of antibiotics in the ICU is ubiquitous, with 71% of all patients in the United States receiving at least one course of antibiotics, with 35% of these being estimated as unnecessary (Davison & Wischmeyer, 2019; Moron et al., 2019).

The observation that critical illness alters the human microbiota was first made in 1969 by Johanson et al., decades before the advent of high-throughput sequencing (Johanson et al., 1969). In this study, they noted that the healthy oral microbiota was replaced by gram-negative aerobes such as those from the phylum Proteobacteria in critically ill patients. Notably, this change was not seen in hospital workers but depended on the severity of their illness. This is likely due to fewer food-associated microbes being introduced, as well as a decreased supply of nutrition for the commensal gut microbiota (David et al., 2014).

The main route of microbial elimination in healthy individuals is via

transit through the GI tract. A healthy adult expels approximately 10^{14} bacterial cells a day through defecation (Simon & Gorbach, 1984). This transit time is significantly slowed down in critically ill patients, due to a variety of therapeutic (catecholamines (Mittal et al., 2017), opiates (Poulsen et al., 2016), and sedatives (Martel et al., 2018)) and pathophysiological (hyperglycemia and electrolyte imbalance) factors (Müller et al., 2018). Transit time through the stomach also slows, as the acidic environment is neutralised by treatments such as proton pump inhibitors and antacid (Lombardo et al., 2010). Production of IgA (H. B. Coutinho et al., 1997), and bile salts (de Vree et al., 1999), which are also responsible for microbial elimination, are lowered (Dickson, 2016). Growth conditions in the intestinal tract are also altered in critical illness, which changes the growth rates of intestinal microbes. Hypoperfusion and reperfusion of the intestinal lining result in extensive mucosal inflammation (de Haan et al., 2009). This leads to an altered mucosal oxygen gradient (Albenberg et al., 2014), and increased nitrate concentrations (Winter et al., 2013), which promotes the growth of bacteria in the phylum Proteobacteria, which includes clinically important gram-negative rods, such as *E. coli* and *P. aeruginosa*, and members of the phylum Firmicutes, such as *Enterococcus* spp. and *Staphylococcus aureus* (Grootjans et al., 2010). Crucially, in many of these patients, the mucosal layer is damaged, leading to hyperpermeability and loss of hydrophobicity (Bardou et al., 2015; Jung & Bae, 2020). Under normal circumstances, this mucosal layer plays a crucial role as a physical barrier, preventing the movement of bacteria, water soluble toxic molecules, and digestive enzymes, across the epithelium and into the mucosal surface.

The net result of these alterations is an unstable gut microbiota with extremely low diversity. The stomach and the adjacent small intestine, which are normally sparsely populated, become overgrown by a small number of species, such as *E. coli* and *Enteroccocus* spp (Agudelo-Ochoa et al., 2020; Alverdy & Krezalek, 2017; Evans et al., 2023; Ravi et al., 2019). The GI tract, which in healthy individuals contains hundreds of microbial species, rapidly loses diversity, with a few, and in some cases one, species dominating the community ($\geq$50% relative abundance). Dominant bacterial species include *Enterococcus* spp., *S. aureus*, *Clostridioides difficile*, and members of the *Enterobacteriaceae* family such as *Klebsiella* spp. and *E. coli*. Additionally, *Candida* spp., which are typically rare in the gut, have been shown to bloom and thrive (Ravi et al., 2019; Romanowski et al., 2012). A study of 14 ICU patients reported that four patients had ultra low diversity gut communities of 1 to 4 bacterial taxa (Zaborin et al., 2014). This low microbial diversity and dominance of certain pathogens is a marker of severe disease and predictive of adverse outcomes such as multiple organ dysfunction syndrome (MODS) (Alverdy & Krezalek, 2017).

In a typical gut microbiota, viruses comprise around 10% of the community (Shkoporov & Hill, 2019). A small proportion of these are eukaryotic viruses ($<$10%) which include human enteric viruses such as coronaviruses (Liang et al., 2020), hepatitis E (Lampejo et al., 2022), and rotaviruses (Ganime et al., 2012), all of which have been implicated in nosocomial transmission events in the ICU and excess mortality. These viruses may have been latent in the patient's gut microbiome prior to ICU admission and immunosuppression therapy and may remain in the gut after clearance of the infection (Z. Cao et al., 2022; Zuo et al.,

2021). Phages account for >90% of a typical gut virome (Shkoporov & Hill, 2019), but to the best of our knowledge, no previous study has studied the effects of an ICU stay on this population. Data on this area is desperately needed with ICU beds making up less than 10% of beds within hospitals but accounting for greater than 20% of inpatient care cost (J. Martin et al., 2008; Pastores et al., 2012).

Therefore, in this chapter I aimed to investigate the role of phages in the gut microbiota of 24 critically ill patients during their stay in the ICU, using a publicly available shotgun metagenomic dataset.

3.2 Aims and Objectives

1. To define criteria for microbial domination and identify the dominant microbe in each patient.

2. To identify viral operational taxonomic units and predict their taxonomy and host.

3. To investigate the dynamics of the dominant host-phage pairs.

4. To identify actively replicating phage sequences and investigate their potential for horizontal gene transfer of virulence factors and antimicrobial resistance genes.

3.3 Methods

3.3.1 Sequence Retrieval and Quality Control

Metagenomic sequencing reads of 233 samples were retrieved from European Nucleotide Archive (ENA) project PRJNA533528 with a further four missing samples (13-6929532, 13-6929609, 136929647, and 13-6929647) downloaded from the following Dropbox link after personal correspondence with the authors: https://www.dropbox.com/sh/tzgwrdf571k3p5s/AADxy8aVGLEF_ZavbUiW9SCya?dl=0 (Ravi et al., 2019). This Dropbox link is now expired and the missing samples have not been added to the ENA project so we have made them available at: https://figshare.com/articles/dataset/Missing_reads_from_Ravi_et_al_2019/24805785. The FASTQ files were then renamed from their ENA sample accession to their sample ID, e.g. SRR8926088_1.fastq to 13-6929817_1.fastq.

Quality profiling, low-quality read filtering, adaptor trimming, and exact deduplication were carried out with fastp (v23.1) (S. Chen et al., 2018). Samples 13-6929545, 13-6929633, 13-6929634 and 13-6929899 were omitted from further analysis as they contained less than 100,000 reads. Human sequences were detected by aligning the reads to GRCh38.p14 reference genome from NCBI (https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.40) with Bowtie2 (v2.4.2) (Langmead & Salzberg, 2012). Unmapped reads were extracted and sorted using SAMtools (v1.13) (H. Li et al., 2009) (`view -f 4 | sort -n`). This was then

converted back into a FASTQ file using BEDtools (v2.30) (`bamtofastq`) (Quinlan & Hall, 2010).

3.3.2 Taxonomic Identification of Metagenomic Reads

Microbial hosts in each sample were identified using MetaPhlAn (v4.1.0) (`-unknown-estimation`) (Blanco-Miguez et al., 2022). The resulting abundance tables were then merged with `merge_metaphlan_tables.py` and genus level classifications were extracted (`grep -E "(g__)|(^ID)"`). The data was then imported into RStudio (RStudio Team, 2020) and visualised with ggplot2.

3.3.3 Viral Sequence Identification

To identify viral operational taxonomic units (vOTUs), samples were individually assembled using SPAdes (v3.15.4) (`-meta`) (Nurk et al., 2017) and the scaffolds were then concatenated. Scaffolds shorter than 10 Kb were then removed with a custom perl script (`removesmalls.pl`) (section 5.2.1) unless then were predicted as circular by ccfind (v1.4.5) (Nishimura et al., 2017). Scaffolds which fulfilled at least one of the following criteria were taken forward for analysis: (1) classified as viral by Kraken2 (D. E. Wood et al., 2019) using the standard database from https://benlangmead.github.io/aws-indexes/k2 (accessed 23rd April 2021), which is built from the following RefSeq databases: archaea, bacteria, human, plasmid, UniVec_Core, and viral. Sequences classified as viral by Kraken2 were separated with VirKraken; (2) predicted as a prophage by Phigaro (v2.2.6) (`-e`

`bed -d`) (Starikova et al., 2020); or (3) had a MASH distance (Ondov et al., 2016) of less than 0.05 to any sequence in the Gut Phage Database (GPD) (Camarillo-Guerrero et al., 2021). Scaffolds that passed criteria 1 and 3 were then concatenated and dereplicated using CD-HIT-EST (v4.8.1) (`-c 0.95 -aS 0.85 -n 9 -d 0 -p 1 -t 4 -g 1`) (W. Li & Godzik, 2006). Scaffolds passing criteria 2 were dereplicated separately to avoid absorption of prophages into their original host sequences. Host regions flanking the prophage regions as predicted by Phigaro were removed using the sub-commands of BEDtools `complement` and `getfasta`. Quality, completeness, and contamination of vOTUs were estimated with CheckV (v1.01) (`end_to_end`) (Nayfach et al., 2020). Filtered reads were then mapped to vOTUs using Bowtie2. Read counts were complied with SAMtools `flagstat`, and relative abundances were calculated with CoverM (`genome -m relative_abundance`) (https://github.com/wwood/CoverM).

3.3.4 vOTU Taxonomic Identification and Host Prediction

vOTUs were annotated using pharokka (`-meta`) (v1.3.2 (Bouras et al., 2023)) using phanotate as the gene predictor (McNair et al., 2019). The E-value for MMseq2 (Steinegger & Söding, 2017) search of PHROGs (Terzian et al., 2021), CARD (Jia et al., 2017), and VFDB (L. Chen et al., 2005) and PyHMMER (Larralde & Zeller, 2023) annotation with PHROGs was set at 1E-05. A gene-to-genome mapping file was then created from the annotated proteins using the `vcontact2_gene2genome` (`-s Prodigal-FAA`) script from vConTACT2 (Jang et al., 2019). The mapping file and annotated proteins were then

concatenated with equivalent files from the INPHARED dataset (INfrastructure for a PHAge REference Database): `2Nov2022_vConTACT2_gene_to_genome.csv` and `2Nov2022_vConTACT2_proteins.faa` respectively (Cook et al., 2021). The resulting proteins were then clustered using vConTACT2 (v0.9.19) (`-db 'None' -pcs-mode MCL -vcs-mode ClusterONE`) and the resulting network was visualised with Cytoscape (v3.8.0) (Shannon et al., 2003) and the family annotation file from INPHARED. Putative hosts for the vOTUs were predicted using IPHOP (v1.2.0) using the default database (`Sept_2021_pub`) (Roux et al., 2023). Phage lifestyle was predicted using DeePhage (v1.0) (S. Wu et al., 2021). The stage of infection of prophage vOTUs was estimated by running Prophage Activity Estimator (PropagAtE) (v1.1.0) (Kieft & Anantharaman, 2022) on vOTUs that were originally identified by Phigaro, before the host flanking regions were removed. vOTUs predicted to infect *Akkermansia* spp. were searched against the NCBI nucleotide collection (nr/nt) using nucleotide BLAST (megablast) (Altschul et al., 1990). We also performed linear regression analysis using the `geom_smooth` function (`stat = "smooth", method = "lm", se = FALSE` from the ggplot2 package (Wickham, 2016). Total vOTU relative abundance or vOTU Inverse Simpson alpha diversity as used as the dependent variable, and the day of ICU stay as the independent variable. Multiple comparisons correction was applied to p-values using the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995).

3.3.5 Dominant Host-Phage Correlation Analysis

For patients with microbial domination during their stay in the ICU, we calculated their most abundant genus by summing up their relative abundance across their samples using the R package dplyr (Wickham et al., 2023). For patients 38 and 53, their dominant microbe was only classified to class and family taxonomic level respectively, so the next most dominant genus was selected instead. Relative abundance of vOTUs which were predicted to infect these dominant hosts were also calculated. The relative abundance of the dominant hosts and their phages were then plotted on the same figures using two `geom_area` objects, with one figure for each patient (Wickham, 2016). Data for patient 38 was not visualised as no phage infecting *Candidatus Gallimonas* were identified. The data from the remaining 11 patients were then exported into separate csv files, again stratified by the patient. These files were imported into Python (v3.10.8) (Van Rossum & Drake, 2009) using the pandas library (v1.5.3) (McKinney, 2010). The Pearson correlation coefficient was calculated between host and phage relative abundance using the `pearsonr` function from `scipy.stats` (v1.10.1) (Virtanen et al., 2020). The time-lagged cross-correlation (TLCC) coefficients between host and phage abundance, with days in the ICU as the lag unit, was then calculated using the `correlate` function from `scipy.signal` and was normalised to between -1 and 1 for visualisation (Harris et al., 2020). The TLCC was then plotted for each patient using the `matplotlib` (v3.7.1) library (Hunter, 2007). A dashed vertical line indicating lag at peak correlation was drawn using the `axvline` function. The full script for calculating the TLCC coefficients and plotting can be found in

section 5.2.2.

3.4 Results

3.4.1 Updated Taxonomic Profiles of ICU Metagenomes

As the taxonomic profiling of the metagenomes from Ravi et al., 2019 were determined with a previous version of MetaPhlAn (v2.7.7), we reanalysed all 233 samples using MetaPhlAn4. This version of MetaPhlAn uses approximately 5.1 million unique clade-specific markers genes obtained from over 1 million genomes, compared to approximately 1 million unique clade-specific marker genes from 16,903 genomes in the MetaPhlAn version used by Ravi et al., which significantly improves the classification resolution. As MetaPhlAn2, unlike MetaPhlAn4, does not have a `unclassified_estimation` function which estimates the percentage of the metagenome that is made up by microbes which are not in its database, we cannot directly compare their performance on this dataset. However, a previous benchmarking paper found that MetaPhlAn2 had a recall of 0.73±0.08 (± indicates standard deviation) when predicting the presence and absence of species in a mock community (Parks et al., 2021). For our dataset the estimated recall (1 - $\bar{X}$, where $\bar{X}$ is the mean unclassified percentage) was 90.8±8.73. This illustrates the improved performance of MetaPhlAn4 and the necessity to reanalyse the metagenomes. We note any major changes in taxonomic composition as a result of this reanalysis in the next section.

3.4.2 Defining Microbial Domination

In the original study, Ravi et al. reported that 75% of their long-stay ICU patients had episodes of intestinal domination, with particularly high relative abundances of the organisms *Enterococcus faecium*, *E. coli*, and *Candida albicans*. However, the authors did not specify what their criteria for microbial domination were, and only 12/24 (50%) of the samples were shown in their domination plot. In this study, we define microbial domination in a patient as: $\geq$30% relative abundance of at least one species at one time point. This is similar to the criteria in Taur et al., 2012 although their calculations were at genus level. Using this definition, we also found that 75% (18/24) of patients displayed microbial domination during their stay in the ICU. This included patients 10, 35, 36, 45, 46, 51, and 52 which were not shown in the author's original figure. Patient 29 which was included in the original figure was not determined to show microbial dominance under this criteria, as the most abundant species found was *Bacteroides fragilis* at 25.5% relative abundance (Figure 3.2). Across all patients classified as having episodes of microbial domination, the taxa in question were consistent between Ravi et al. and our own analysis, with the exception of patient 35. The MetaPhlAn and metagenome assembled genome (MAG) analysis in the original publication found that *Candida albicans* was the only dominating species during patient 35's stay in the ICU, with a maximum relative abundance of 82.4%. In contrast, our MetaPhlAn analysis found two dominating taxa: *C. albicans* and *Candidatus* Borkfalkia ceftriaxoniphila, with maximum relative abundances of 46.6% and 83.2%

respectively (Figure 3.1). Within one sample taken on day 9 of the patient's stay, the estimated relative abundances of *C. albicans* were 70.84% and 17.9% in the Ravi et al. and this study respectively. Due to the recency of the discovery and description of *Candidatus* Borkfalkia ceftriaxoniphila, it would not have had representative marker genes in databases used in Ravi et al.'s analyses, and therefore would not be detected (Hildebrand et al., 2019). In the following analyses, we focused on samples from the 18 patients with microbial domination.

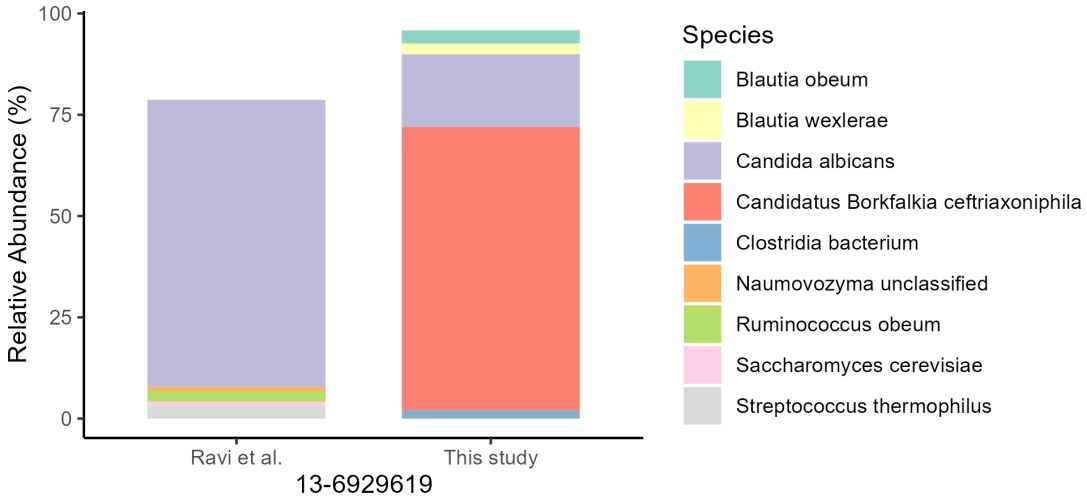


Figure 3.1: **Comparison of taxonomic compositions between analyses from Ravi et al. and this study using a representative sample from Patient 35.** Taxonomy and relative abundance from the Ravi et al. study was downloaded and plotted against the equivalent data from our MetaPhlAn4 analysis. Sample 13-6929619 represents day 9 of patient 35's stay in the ICU.

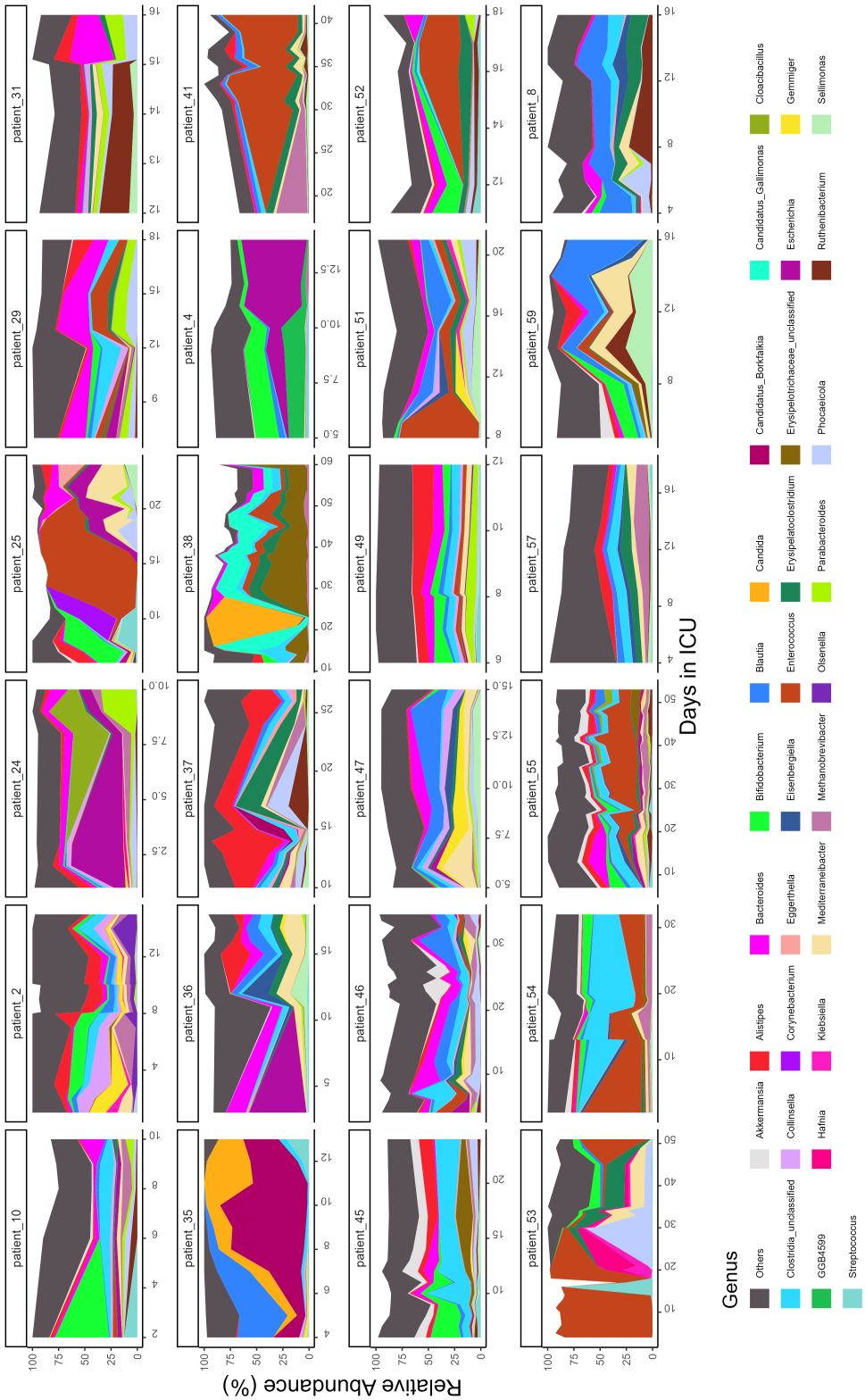


Figure 3.2: (Caption next page.)

Figure 3.2: **Genus level taxonomic composition of patients across their stay in the intensive care unit.** Taxonomy and relative abundance for each sample was assigned using MetaPhlAn4. Only genera that have a relative abundance of ≥20% in at least one sample are displayed. All other genera were collapsed into the "other" category.

3.4.3 Identifying vOTUs from ICU patient gut metagenomes

Metagenomic reads were assembled for all samples individually using MetaSPAdes. The resulting scaffolds were identified as viral if they passed one of the following criteria: Classified as viral by Kraken2; had a MASH distance of ≤0.05 to any genome in the GPD; or classified as a prophage by Phigaro. Kraken2 was selected as it had the highest precision across our benchmarks in Chapter 2. We utilised a newer version of the database we benchmarked on, as it includes approximately two years more of RefSeq sequence data, corresponding to a 47.1% increase in total taxonomy nodes. Scaffolds that closely resemble genomes in the GPD were also kept to increase the number of gut phage genomes which were detected as the RefSeq database that Kraken2 uses is not representative of the diversity of complete phage genomes. This is due to only exemplar phage genomes, as defined by the ICTV, being added to RefSeq. Due to this, RefSeq is always outdated compared to databases like GPD and INPHARED.

Originally, we had planned to identify novel phage sequences using a machine learning (ML) based tool, as these are able to identify novel phage genomes that are not in databases. However, when we used DeepVirFinder, which was one of the better performing ML tools in Chapter 2, during manual

curation of its output we noticed a considerable number of false positives. An example of this was a scaffold that was predicted as viral by DeepVirFinder, but when it was aligned with megablast, the top six results all returned with it belonging to *C. albicans* mitochrondrion sequences with an identity and coverage of over 99%. This is perhaps unsurprising, with DeepVirFinder having a fair low recall when benchmarked on mock community data (0.42) (Section 2.4.3). This may also be impacted by our the threshold being set by score instead of p-value resulting in no multiple comparisons correction being applied. In any case, we decided on the more conservative method of filtering by MASH distances to a database of highly curated phage genomes derived from gut metagenomes (GPD). This database contains 142,809 species-like vOTUs mined from 28,060 publically available human gut metagenomes which is significantly more than the Gut Virome Database (GVD) which was published the year before, where the authors mined 33,242 species-like vOTUs from 2,697 human gut metagenomes, or the 54,118 species-like vOTUs recovered from 11,810 human gut metagenomes by the authors of the Metagenomic Gut Virus (MGV) catalogue.

A number of studies have shown that the gut phageome is dominated by temperate phages (Knowles et al., 2016; Silveira & Rohwer, 2016). In order to capture prophages which may have been missed by the previous two methods due to the presence of flanking host sequences we used Phigaro, which we also used in Chapter 2 to remove prophages from host sequences. A benchmarking study published in 2022 found that Phigaro had very high accuracy (0.980), the highest precision (0.748) out of all the ten included tools except for PhiSpy (0.772), which was published as a part of the benchmarking study, and an average recall score

(0.566; 6th out of 10 tools) (Roach et al., 2022). All the metrics in this study were calculated at the CDS level using manually annotated bacterial genomes as the ground truth. For example, the accuracy score is the ratio between all the correctly labelled CDS by each tool and the total number of CDS. PhageBoost which we also used in Chapter 2 was shown to have poor precision ($\bar{X} = 0.399$) by the benchmarking study so it was not used here.

Several studies have shown that if all dereplicated scaffolds are considered as distinct vOTUs, then the total population count will be significantly over-estimated, with an average of two to three vOTUs per actual genome (Aziz et al., 2015; García-López et al., 2015). Therefore, we removed all scaffolds that were <10 Kb, unless they were predicted to be circular. The remaining scaffolds were then dereplicated by clustering the scaffolds at 95% identity and 85% coverage as suggested by the Minimum Information about an Uncultivated Virus Genome (MIUViG) standards (Roux et al., 2019a). This resulted in a total of 3085 vOTUs with a median length of 18.6 Kb (Quartile 1 (Q1) = 12.9 Kb, Quartile 3 (Q3) = 33.9 Kb) We estimated the completeness and quality of these vOTUs using CheckV (Nayfach et al., 2020). This tool works by clustering sequences against a viral genome database using average amino acid identity (AAI) to calculate the expected genome length. It then predicts complete closed genomes by searching for direct terminal repeats (DTRs), prophage-host boundaries, and inverted terminal repeats (ITRs). Note that the prophage-host search does not work on our vOTUs as host regions have already been removed. In total, we found that 56 (1.82%) vOTUs were complete, 132 (4.28%) were high quality (≥90% completeness), 377 (12.2%) medium quality, 1612 (52.3%) low

quality, and 908 (29.3%) were not determined. The median completeness, excluding vOTUs that were not determined, was estimated to be 30.8% (Q1 = 18.5%, Q3 = 49.5%), and 89.4% of all vOTUs contained no host contamination. Across each sample, the average relative abundance of all vOTUs was 7.89%±3.85.

To investigate the composition of the vOTUs at higher taxonomic levels, we clustered them with 19,720 complete (or near complete) reference genomes from INPHARED (Figure 3.3A). Only 291 (9.43%) of vOTUs clustered with a reference genome, indicating that they were approximately in the same genus. Notably, 11 vOTUs clustered with crAssphage cr127_1 (accession MT774393), an uncultured member of the Crevaviridae family (Figure 3.3B). 665 (21.6%) vOTUs were classified as outliers as they are likely to be related to the viral clusters (VCs) but not at the genus level (family or sub-family level). The other 2129 vOTUs either only clustered with other vOTUs (1944; 63.0%) or were labelled as singletons (185; 6.00%), indicating 893 putative novel genera.

We functionally annotated the vOTUs with pharokka, which uses PHANOTATE, a gene calling tool specifically designed for phage genomes (McNair et al., 2019). Functional annotations are then assigned by searching the CARD (Alcock et al., 2023), PHROGs (Terzian et al., 2021), and VFDB (L. Chen et al., 2005) databases. All three databases are searched with MMseq2 and PHROGs is additionally annotated using HMM profiles. Pharokka also predicts CRISPR, tRNA, and tmRNA sequences. A median of 28 (Q1 = 19; Q3 = 50) coding sequences (CDS) were found in our vOTUs. The vast majority of CDS did not have a known function (82.6%). 785 vOTUs encoded at least one

tRNA. A total of 137 virulence factors genes (VF) encoded by 38 vOTUs, and 62 antibiotic resistance genes (ARGs) encoded by 54 vOTUs were identified. This included ARGs which can confer resistance to tetracyclines (*tetQ*), macrolides (*msr(D)*, *ermB*, *mel*), and beta-lactams (*cfiA2*, *cfxA2*, *cfxA4*). *cfxA2* was identified in 8 different vOTUs, *cfxA4* in 3 vOTUs, and *tetQ* in 9 vOTUs (Supplementary Table 5.1). VFs that were detected included pore forming exotoxin hemolysin E (*hlyE*), the Type III secretion system effector *espL1*, and the endocarditis specific antigen *efaA* (Supplementary Table 5.2). Twenty six vOTUs encoded at least one ARG and VF gene, with vOTU_1943 containing 3 ARGs (*KpnH*, *emrA*, and *emrR* and 1 VF (ATP-dependent chaperone ClpB).

3.4.4 Dominant Host-Phage Dynamics in the ICU Gut

To study bacterial-host phage dynamics we predicted hosts for each vOTU with a framework called iPHoP, which integrates five host-based methods: BLASTn to host genomes; BLASTn to a CRISPR spacer database; WIsH - a k-mer composition based tool (Galiez et al., 2017); VirHostMatcher - a network based tool (Ahlgren et al., 2016); and Prokaryotic virus Host Predictor - a tool that uses a Gaussian model based on k-mer frequencies (Lu et al., 2021). Additionally, iPHop integrates the phage-based tool RaFAH, which predicts hosts based on phage protein clusters with the known host(s), instead of relying on phage-host genome similarities (F. H. Coutinho et al., 2021). The classifications from each method and their evolutionary distance are collected as input for random forest classifiers and a composite score is calculated for each virus-candidate host pair. This allowed us to predict hosts

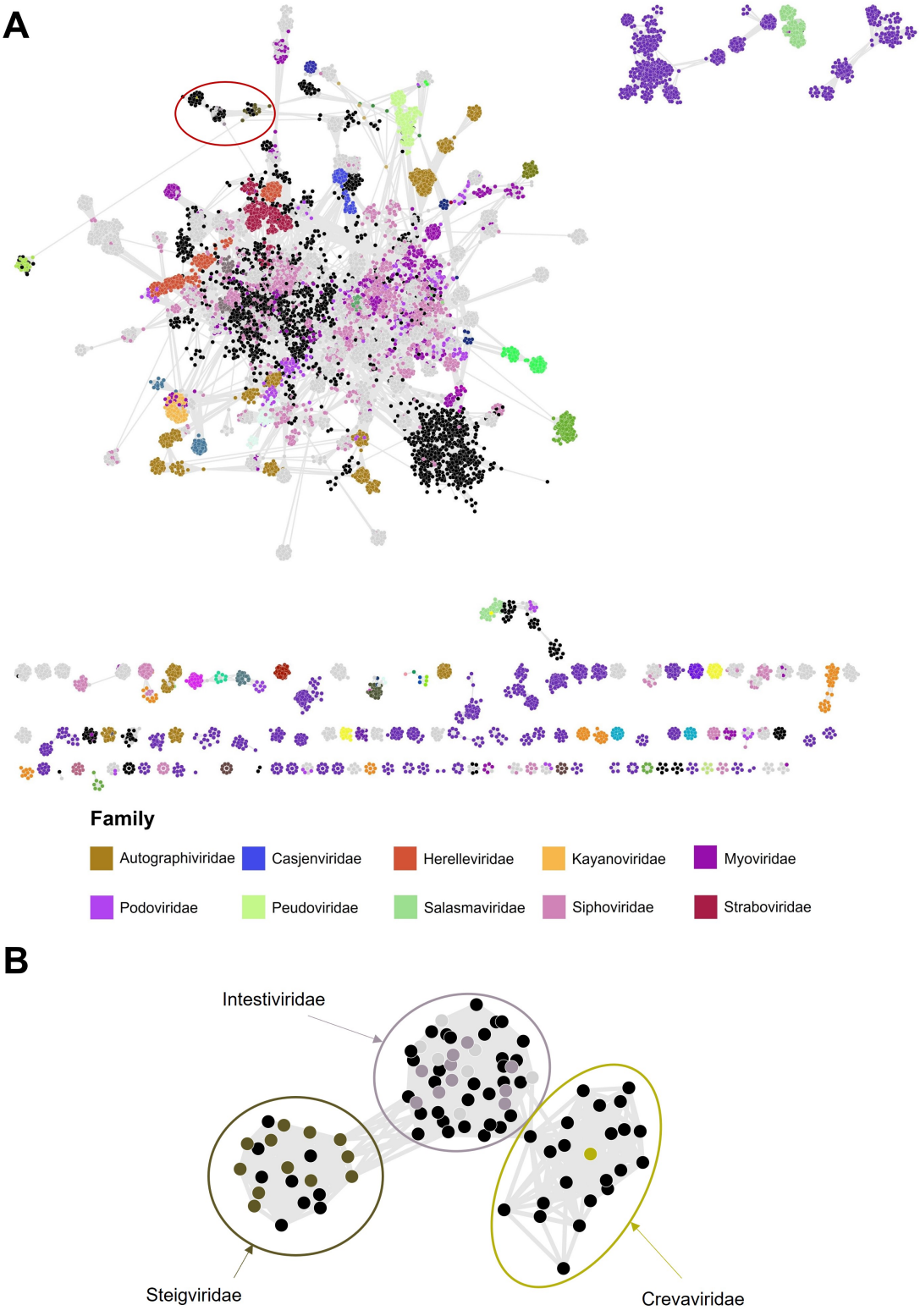


Figure 3.3: (Caption on next page.)

Figure 3.3: **Network based taxonomic classification. A.** vConTACT2 network analysis of all viral operational taxonomic units (vOTUs) from this study and a database of phage genomes from INPHARED. Coding sequences in each vOTU were predicted with phanotate and clustered reference genomes before visualisation with Cytoscape. vOTUs nodes are coloured black. Reference nodes are coloured by family. The ten most common families are displayed in the legend. Nodes with a degree of less than 5 are not shown. The red circle indicates the location of clusters shown in part B. **B**. Three distinct network clusters with all reference nodes (coloured) belonging to the order Crassvirales. Each cluster contains reference genomes belonging to only one family, allowing each vOTU node (black) to inherit its taxonomy.

for 2403/3085 (77.9%) of vOTUs at the genus level, with a minimum confidence score of 90.

The most abundant predicted host genera by average relative abundance were *Bacteroides* (μ = 1.92%), *Escherichia* (μ = 0.65%), and *Enterococcus* (μ = 0.60%) (Figure 3.4). When comparing this to the most abundant hosts estimated by MetaPhlAn across our entire dataset, *Bacteroides* was the fifth most abundant genus (μ = 4.1%), *Enterococcus* was the most abundant (μ = 13.4%), and *Escherichia* the 9th. Conversely, an unclassified member of the class Clostridia was the second most abundant genus as estimated by MetaPhlAn (μ = 5.48%), and *Blautia* the third (μ = 4.73%) (Figure 3.2). This unclassified Clostridia could correspond to many of the host genera that were assigned to the vOTUs including *Blautia*, *Ruminococcus*, *Enterocloster*, *Gemmiger*, *Faecalibacterium*, and *Eubacterium*. These host genera are the 12th, 11th, 15th, 14th, 19th, and 18th most abundant respectively.

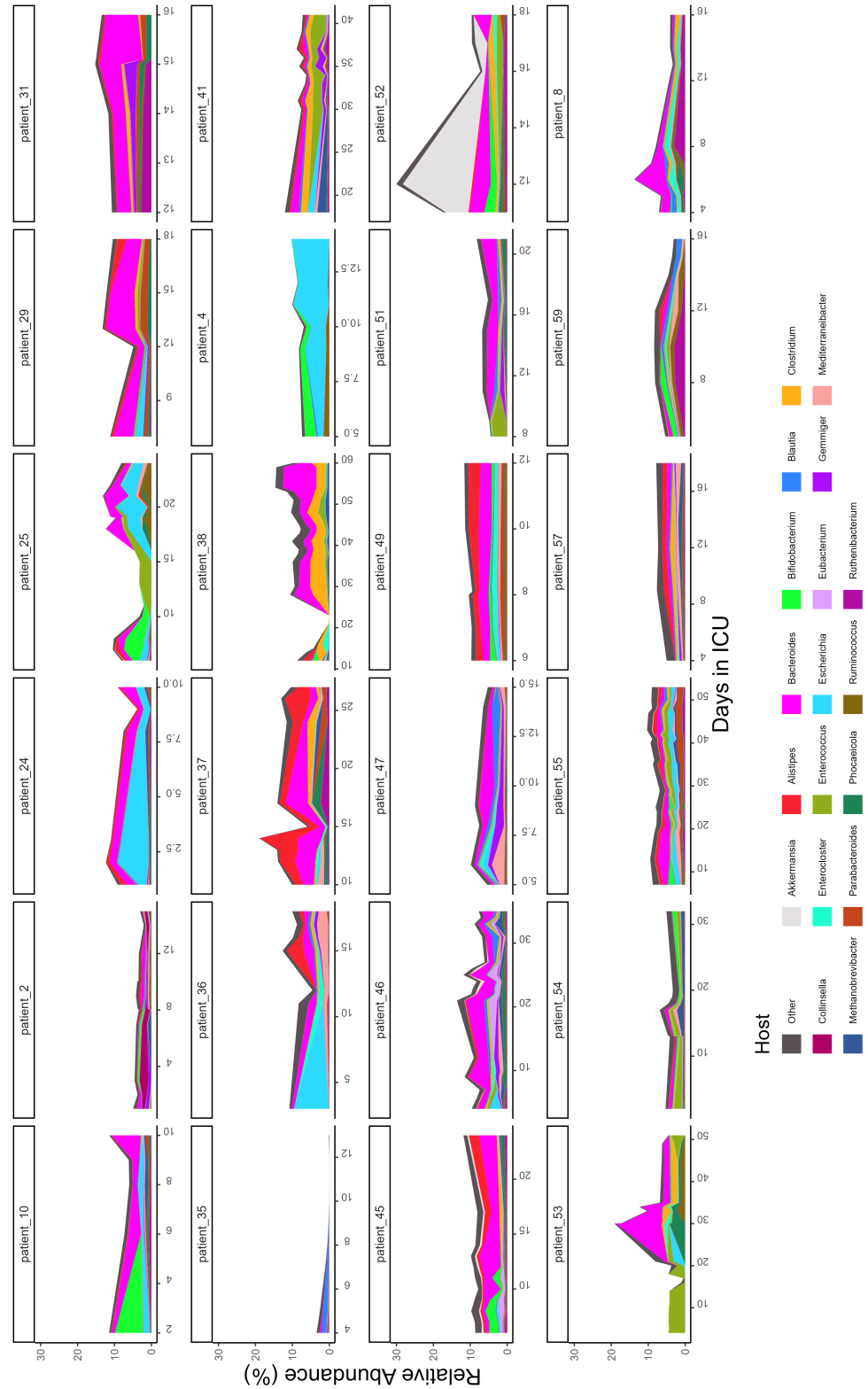


Figure 3.4: (Caption on next page.)

Figure 3.4: **Relative abundance of vOTUs by host taxonomy within each patient across their stay in the intensive care unit.** Host taxonomy of each vOTU was assigned using iPhoP. All quality controlled reads from each sample were mapped to vOTUs using Bowtie2 and relative abundances were calculated with CoverM. Only the top 20 genera with the highest peak abundance are shown. All other genera were collapsed into the "other" category.

The abundance of vOTU host genus was consistent with the abundance of the corresponding host genus in each sample, except for samples from patient 52. This patient had a bloom of vOTUs with *Akkermansia* as their predicted host genus, with a peak relative abundance of 18.7% on day 12. *Akkermansia* spp. itself only reached a peak of 2.78% during patient 52's ICU stay. The abundance of *Akkermansia* vOTUs was mainly made up of three scaffolds: vOTU_1698, vOTU_1612, and vOTU_1819. To confirm that these scaffolds were not misidentified as phages, we conducted a sequence similarity search against the NCBI nr/nt using the megablast algorithm. The closest match for all the scaffolds was Bacteriophage sp. isolate 2960_50739, which is a partial genome phage MAG (Accession: OP072475.1) from a human gut virome study in Japan (Accession: PRJNA862966). Query cover (QC) for these matches were 94%, 95%, and 95% respectively, and percent identity (PI) were 97.6%, 97.6%, and 98.07% respectively. All three vOTUs encoded two tail proteins (phrog_376 and phrog_394), one exonuclease (phrog_69), one endolysin (phrog_435), one RuvC-like Holliday junction resolvase (phrog_159), three PHROG proteins with unknown function (phrog_3742, phrog_3657, and phrog_26724), and either 55 (vOTU_1612) or 56 (vOTU_1698 and vOTU_1819) hypothetical proteins.

Linear regression analysis was conducted to assess the relationship between the Inverse Simpson alpha diversity of vOTUs or the total relative abundance of vOTUs (dependent variable) and the day of the patient's stay in the ICU (independent variable). The results revealed that no patients exhibited a significant linear trend for alpha diversity or vOTU relative abundance over time (Supplementary Figure 5.1 and 5.2).

We then analysed the relationship between the dominant genus in each patient's gut and the vOTUs which are predicted to infect that genus. First, we calculated the dominant genus by summing the relative abundance of each genus for each patient. In some cases, there were multiple dominating genera during a patient's ICU stay. However, for simplicity, we only retained the genus with the highest total relative abundance. We also extracted the abundances of vOTUs infecting those genera. Patients 35 and 38 were dominated by *Candidatus Borkfalkia* and *Candidatus Gallimonas* respectively. These genera belong to the class Clostridia and are currently uncultured, and iPhoP did not predict any vOTUs that could infect them. This may be due to no vOTUs who infect these genera being present in the samples, or the vOTUs which infect these genera not having any host assigned to them by iPhoP due to low confidence score (22.1% of vOTUs had not prediction at genus level). Among the other 16 patients, seven were dominated by *Enterococcus*, two by *Escherichia*, two by *Bifidobacterium*, two by *Bacteroides*, one by *Blautia*, and one by *Alistipes*.

We observed a strong and significant correlation between the dominant genus and its corresponding vOTU abundances in each patient (mean Pearson's r = 0.95±0.05). This correlation, which is close to 1, suggests that the host and

phage abundances are synchronized, rather than one driving the other. To confirm this, we conducted a time-lagged cross-correlation analysis for each dominant host-vOTU pair in each patient (Figure 3.6). This analysis calculated the similarities between the abundance time series at different time lags. For all 16 patients, the highest degree of synchrony was observed at 0 days, providing further evidence that the host genus and vOTU abundances are not mutually driving each other, but instead moving in sync.

As this synchrony could be attributed to the detection of prophages, we predicted the lifestyles of all of our vOTUs with DeePhage. While only 18.3% (566) of vOTUs were originally detected as prophages from Phigaro, the other vOTUs may have also been derived from prophages. Across all samples, the average proportion of temperate phages was 77.0%±12.0 (Figure 3.5). This dominance of temperate phages is normal in the human gut and may explain why the abundance series are so synchronous. However, many of these temperate phages may have been induced, a distinction that DeePhage does not determine. Indeed, it is estimated that temperate phages make up 20-50% of free phages (Reyes et al., 2010; Sausset et al., 2020). On the other hand, it is difficult to distinguish viable temperate phages from defective ones, i.e. phages that have lost the ability to enter the lytic cycle due to certain mutations.

In order to distinguish actively replicating prophage vOTUs we employed the tool PropagAtE, which estimates whether a prophage is in its lytic or lysogenic cycle based on prophage-to-host read coverage ratios. As this tool only works on prophage sequences flanked by the host, we used the prophage vOTUs that were predicted by Phigaro. 115 (20.3%) of the prophage vOTUs were found to be active

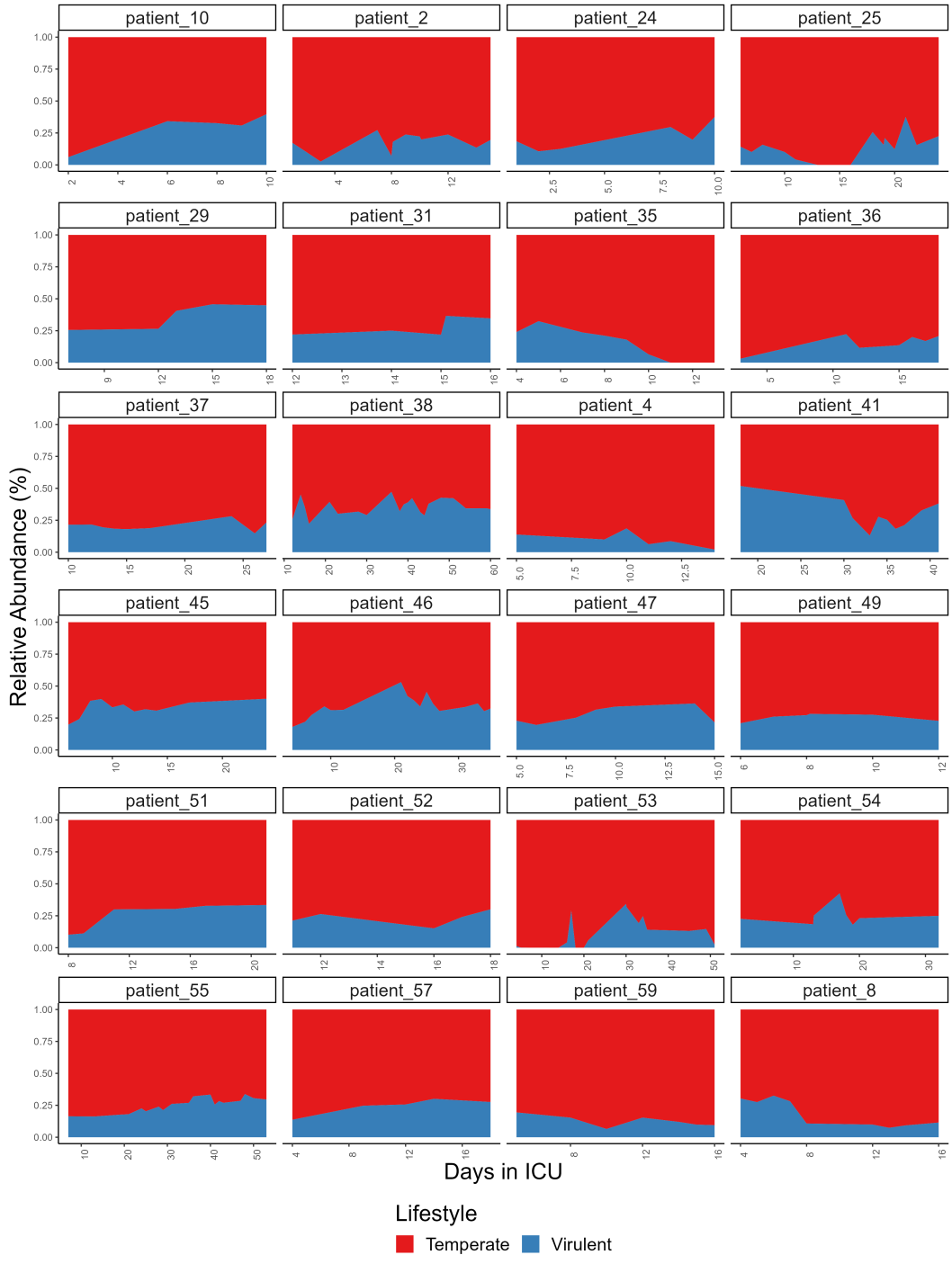


Figure 3.5: **Relative abundance of vOTUs in each patient by lifestyle.** Lifestyle for vOTUs were predicted using DeePhage. Each subplot represents a single patient across their stay in the ICU. Total relative abundance of vOTUs was normalised to 100% per sample.

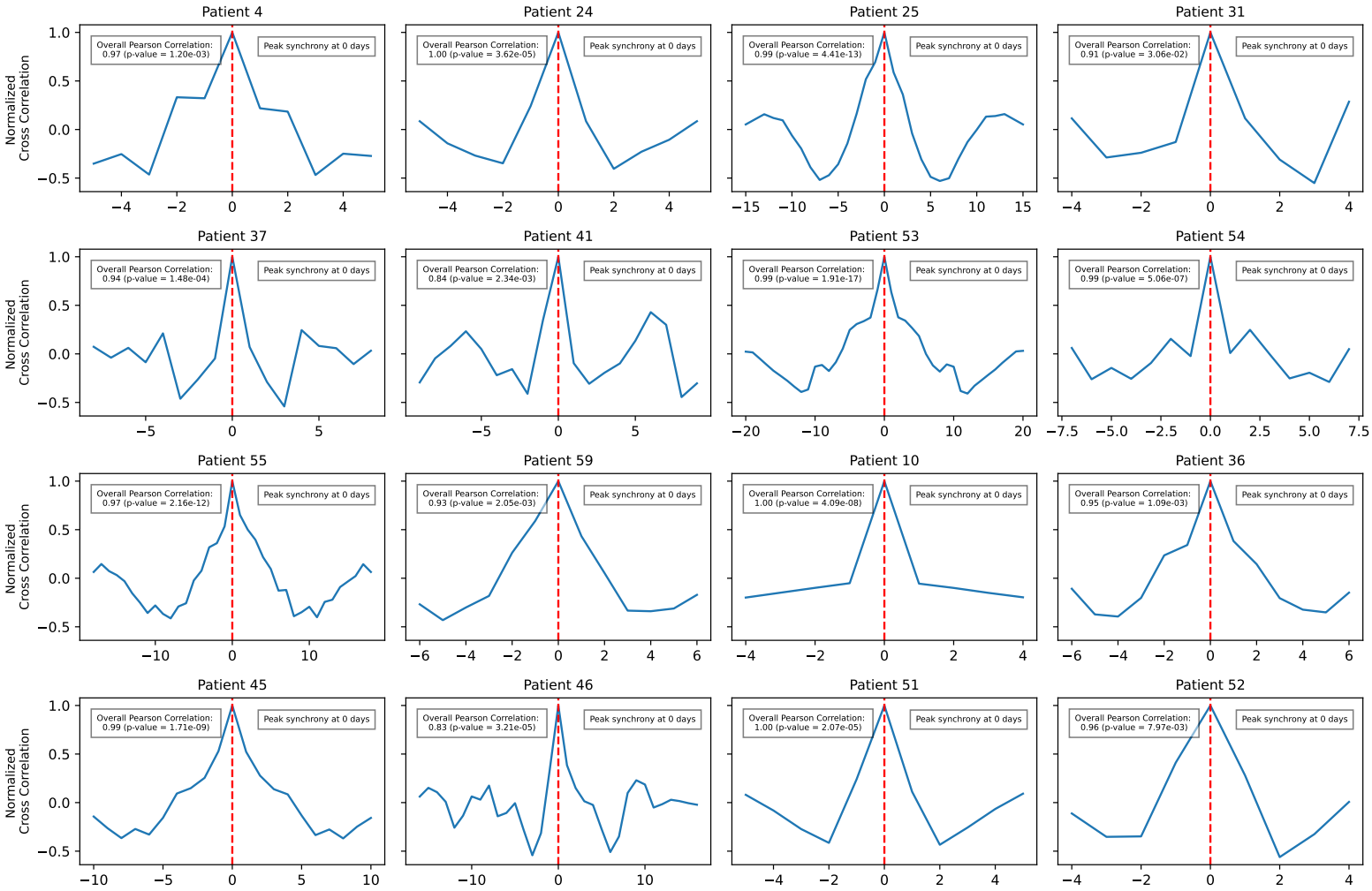


Figure 3.6: (Caption on next page.)

Figure 3.6: **Time lagged cross correlation between dominant bacterial genus and their phages.** Relative abundance of dominant hosts and their predicted infecting vOTUs were calculated for the 18 patients who exhibited microbial domination during their stay in the ICU. Data for patients 35 and 38 were not visualised as no phages infecting their dominant microbes were identified. Pearson correlation and time-lagged cross-correlation (TLCC) coefficients were calculated between host and phage relative abundance, with the latter using days in the ICU as the lag unit. Coefficients were normalised to between -1 and 1 for easy visualisation. The TLCC was then plotted for each patient, with a red dashed vertical line indicating the lag at peak correlation.

in at least one sample with a median of 2 active prophage vOTUs per sample (Q1 = 1, Q3 = 4).

No CDS within the active prophage vOTUs were annotated with any virulence factors or ARGs. When we separated ARG and VF encoding vOTUs by lifestyle predicted by DeePhage, 59/62 of ARGs and all VFs hits were found within the temperate vOTUs. vOTU_156, one of two vOTUs containing an alignment to CARD that was predicted to have a virulent lifestyle, had hits to two ARGs: *mel* (accession ARO:3000616), a msr-type ABC-F subfamily protein that is linked to macrolide resistance at a nucleotide sequence identity of 0.879; and *ermG* (accession ARO:3000522), a Erm 23S ribosomal RNA methyltransferase which confers resistance to macrolides, lincosamides, and streptogramin B, at a sequence identity of 0.948. *mel* was also found in the other virulent vOTU, vOTU_1512 at a sequence identity of 0.878, and does not appear in any other vOTU, regardless of lifestyle. When we aligned the protein sequences of the two *mel* homologs they had a QC of 99% and a PI of 99.01%.

Alignment to the NCBI nr database revealed that this *mel* homolog is likely a macrolide efflux MFS transport *mefA* gene, which is typically expressed on the same operon as *mel* (Gay & Stephens, 2001). The CARD database does not have a separate entry for *mefA* even though the other MFS-type efflux protein, mefE, which also confers macrolide resistance and is expressed on the same operon, has its own entry (https://card.mcmaster.ca/ontology/36908). The *mefE/mel* opeon is carried within the Macrolide Efflux Genetic Assembly (Mega) gene element (Gay & Stephens, 2001). This led to the pharokka annotating the two vOTUs with the incorrect ARG. vOTU_156 and vOTU_1512 are also related with both of them being predicted by iPhoP to infect *Bacteroides faecis*. A 40.7Kb section in the middle of vOTU_156, which is 54.5Kb long, also aligns with the middle of vOTU_1512 (14,854 - 60,772bp), which is 61.7Kb long (QC = 69.9%, PI = 99.9%). Homologs of *mefA* are encoded by *Streptococcus* species such as *S. pneumoniae* and *S. pyogenes* (M. C. Roberts et al., 1999), as well as *mefC* in *Photobacterium* (Nonaka et al., 2012), *mefI* in *S. pseudopneumoniae* and *S. pyogenes* (Blackman Northwood et al., 2009), and *mefO* in *S. dysgalatiae* (Fyfe et al., 2016). Whilst both virulent and temperate phages are involved in transduction, from this data alone, we are not able to predict whether these vOTUs are actively replicating or whether the ARGs are capable of conferring resistance.

A recent study (Pfeifer et al., 2022) showed that phage-plasmids, unlike traditional phages, are much more likely to encode ARGs. Therefore, we sought to identify any vOTUs that were actually phage-plasmids using a random forest classifier. Four vOTUs were classified as plasmids: vOTU_174, vOTU_511,

vOTU_2517, and vOTU_2912. Again, no VFs or ARGs were found to be encoded by these vOTUs.

3.5 Discussion

The gut of critically ill patients often exhibits significant dysbiosis and extremely low diversity, with one or two taxa dominating the space. To the best of our knowledge, no prior studies have investigated the role and dynamics of the gut phageome in critically ill patients. Previous studies have instead focused on the role of bacteria and fungi. Given that phages are essential members of the healthy gut community, and play a vital role in regulating their bacterial hosts, further research is needed to explore their role in disease states.

The aim of this chapter was to mine phage sequences from metagenomic sequencing data produced in a previous study (Ravi et al., 2019), and identify their significance in shaping the critically ill gut "pathobiome". The previous study found that 75% of their patients had episodes of microbial domination, whilst two-thirds of patients experienced a dramatic drop in microbial diversity at some stage during their stay in the ICU. This loss of diversity was significantly correlated with meropenem use but not the administration of other antibiotics. Since the original study was conducted, two major versions of MetaPhlAn have been released, which include marker genes from approximately 60 times more reference genomes. Therefore we sought to reanalyse the taxonomic composition of these samples. Although we could not directly compare the classified

proportion of the samples between the two versions because the older version lacked the "unclassified_estimation" function, our estimated recall was significantly higher than those obtained by benchmarking studies using MetaPhlAn2 (Parks et al., 2021; Ye et al., 2019). While the overall taxonomic compositions remained largely consistent between the two analyses, there was a notable exception in the case of patient 35. In this instance, the original authors had overestimated the bloom of *Candida albicans* while completely missing the bloom of a second microbe, *Candidatus* Borkdalkia ceftriaxoniphila. This was due to the fact that the latter had only been discovered in 2019, as a low abundance commensal in at least one-third of adult guts (Hildebrand et al., 2019). *Candidatus* Borkfalkia ceftriaxoniphila has previously been shown to opportunistically bloom in response to antibiotic intervention, as was observed in this patient, during the administration of piperacillin/tazobactam. This underscores the critical importance of utilising updated databases for taxonomic classification, given the continuous identification of new taxa. On average, approximately 9.8% of our samples, remained unclassified with MetaPhlAn4. As databases expand, reanalysis of this data with future versions of MetaPhlAn is likely to further diminish this unclassified proportion, enhancing profiling accuracy. However, despite this ongoing refinement, the majority of overall compositions remained consistent across analyses, including the number of patients who experienced microbial domination in their gut. Consequently, it is unlikely that future reanalysis will result in drastic shifts in our conclusions.

The original study by Ravi et al. did not specify the criteria for identifying an episode of microbial domination in a patient. Therefore, we defined it as when

the relative abundance of at least one species is ≥30% at one time point. This threshold was chosen due to the fact that it preserved the number of patients that exhibited microbial domination as reported by the original authors. Similar criteria has also been used in other studies (Taur et al., 2012). However, it is unclear whether the patients determined to have episodes of microbial domination under this criteria were the same as those reported by Ravi et al. as only 50% of them are displayed in Figure 1 of the original publication. Of those that were included in that figure, patient 29 was determined to not have microbial domination under our criteria. The maximum abundance of *Bacteroides fragilis* in that patient was close to the threshold (25.5%), which suggests our threshold may be too stringent.

The results from this chapter showed that it is possible to isolate phage derived sequences from this data, including a small number of complete genomes. However, the majority of the identified vOTUs were classified as genome fragments by CheckV in accordance with MIUVIG guidelines (Roux et al., 2019a). The reason for this classification is likely due to the fact that most vOTUs represent fragments of larger viral genomes, as it is extremely difficult to assemble complete genomes from metagenomic data, especially with viral genomes which are mosaic in nature. Additionally, novel genomes would also be classified as genome fragments due to having no database representatives. This is the case for over half of our vOTUs (69%), as they did not cluster with reference genomes in our gene sharing network analysis.

Compared to previous studies on healthy gut phageomes, we found an elevated relative abundance of phage sequences across our samples. This was on average 36% higher than what was reported by Arumugam et al., 2011. It is

possible that the relative abundances of these vOTUs are inflated by the presence of dormant lysogenic phages. However, extremes of up to 22% phage relative abundance in the gut have been previously reported (Dutilh et al., 2014). The abundance of vOTUs categorised by host genera matched the abundance of corresponding host genera in each sample, with the notable exception of samples from patient 52. This patient had a bloom of vOTUs infecting the genus of *Akkermansia*, reaching a peak relative abundance of 18.7%. *Akkermansia* spp. itself only reached a peak of 2.78% during patient 52's ICU stay. The abundance of *Akkermansia* vOTUs was mainly made up of three scaffolds: `vOTU_1698`, `vOTU_1612`, and `vOTU_1819`. When these scaffolds were searched against the NCBI nr/nt with megablast, their closest references all returned as MAGs of bacteriophage isolates, suggesting that it is unlikely that these are falsely identified vOTUs. It is unclear why such a large bloom of *Akkermansia* vOTUs is presented in this patient in the absence of a bloom from its host.

The vOTUs we identified in this chapter had on average 2.4x fewer CDS than the median genome in INPHARED as a result of most of them being genome fragments. Despite this, we were still able to identify tRNAs, VFs and ARGs encoded by the vOTUs. Interestingly, we observed that a quarter of vOTUs contained tRNA CDS, which is slightly lower than the number reported by Cook et al., 2021 (32%). While we identified over double the number of VFs than ARGs, these were concentrated in fewer vOTUs, which is expected given that VFs tend to be host species-specific and are unevenly distributed among phage genomes. Many of the VFs and ARGs that were identified have clinical relevance, such as *cfxA2* and *cfxA4*, which are class A beta-lactamases, which

have been found in clinical multidrug-resistant *B. fragilis* strains (Sydenham et al., 2019). While *cfxA* beta-lactamases have been found in chromosomal and plasmid sequences (Boiten et al., 2023; Handal et al., 2005), we believe this is the first identification of it on a phage genome. VFs that were identified by pharokka included the *E. coli* exotoxin haemolysin E gene *hlyE*. This pore-forming toxin can lyse erythrocytes, induce apoptosis in macrophages, and is cytotoxic to mammalian cells in culture. Homologs of HlyE are present in the genomes of *Shigella flexneri* and *Salmonella enterica* serovar Typhi, which can cause dysentery and typhoid fever respectively (Hunt et al., 2010; Wallace et al., 2000). To the best of our knowledge, this is the first time *hlyE* has been found on a phage genome, although bacteriophage-associated enterohaemolysins have been identified in *E. coli* previously (*ehly1* and *ehly2*) (Beutin et al., 1990; Oscarsson et al., 2002), and exotoxins are historically the VFs that are most associated with phages (Abedon & LeJeune, 2007).

When focusing on the dominant microbe in each patient, we discovered a strong (near-perfect) correlation between its abundance and the vOTUs that were predicted to infect it. This suggests that the abundance of the microbe and its corresponding vOTUs are coupled instead of one driving the other. We hypothesised that this coupling was due to the sequencing of uninduced prophages and their respective hosts. Indeed the vast majority of our vOTUs were predicted to be temperate, which is the norm in healthy patients (Lin & Lin, 2019). Despite this, it is not possible to determine whether they were actively replicating within this dataset, with the exception of the prophages originally identified by Phigaro (18.4% of all vOTUs). From there, we found that

approximately a fifth of them were likely to be actively replicating based on their coverage. Notably, none of these prophage vOTUs were found to encode any VFs or ARGs. This result was unexpected, given the well-documented presence of VFs in many phage genomes (Abedon & LeJeune, 2007; Boyd, 2012; Silveira et al., 2020). As for ARGs, there is still debate on how commonly they are encoded within functional phage genomes (Enault et al., 2017; Pfeifer et al., 2022). Our findings support recent studies suggesting that ARG carriage in phage genomes is a rare phenomenon. However, our stringent vOTU filtering, and the large number of actively replicating prophages that we likely missed, point to the need for additional research in this area.

3.5.1 Limitations

The results here show that we are able to mine phage sequences from metagenomic data, allowing us to predict their hosts, and study their dynamics within the gut of ICU patients. However, our initial attempt at using DeepVirFinder as part of our phage sequences prediction method produced a substantial number of false positive hits. This was an issue as a high false positive rate would affect the confidence of our downstream analyses. Therefore, we opted to use a distance-based analysis tool (MASH) to compare our samples against a large gut-derived phage genome database (GPD), resulting in a cleaner set of predictions. As a result of this, our methodology did not have the capacity to detect novel phage genomes. It is also worth noting that the GPD was created using identification tools which also have non-negligible false positive rates, as

shown in Chapter 2 (VirFinder and VirSorter). This database did however have substantial quality control as shown by the author's CheckV analysis (Camarillo-Guerrero et al., 2021), so we believe that there is little contamination within it. We could have also replaced DeepVirFinder with another tool that performed well in our benchmarking analysis such as VIBRANT. However, VIBRANT still had a poor precision score on mock community data and would likely still leave us with a higher number of false positives than our MASH-based method. In order to identify prophage regions in our metagenomic assemblies we opted to use Phigaro which was found to have both high accuracy and high precision by a prophage identification benchmarking study (Roach et al., 2022). However, we failed to note that Nayfach et al., 2020's study had conflicting accuracy results when testing Phigaro on mock prophages with varying host contamination on the fragments, with the tool on average overestimating the amount of contamination, although the IQRs were large. This means that Phigaro is likely to truncate prophage regions with viral genes on its edges being interpreted as host encoded, resulting in shorter vOTUs which are more difficult to classify and contain less genomic context which is important when identifying potential MGEs. Nayfach et al., 2020 also found that Phigaro identified 28.3% of entirely viral fragments as prophages, which implies that real phage regions may be discarded by Phigaro and that contigs may be mistakenly identified as integrated prophages. In future studies, it is advisable to use tools like CheckV, which uses gene annotations and GC content to identify viral-host boundaries, to validate prophage predictions and remove host contamination.

The vOTUs that we identified had a median length of 18.6 Kb, which is

considerably shorter than the median length of high-quality phage genomes reported in the literature, which ranges from 40 to 50 Kb (Camarillo-Guerrero et al., 2021; Pfeifer et al., 2021; Zrelovs et al., 2021). This is mainly due to the limitations of Illumina shotgun sequencing, which in this study, generated 150bp paired-end reads. Assembling such short reads into complete genomes is challenging, as the read length is shorter than the repetitive sections of DNA (such as IS elements) (Minot et al., 2012). This is especially problematic for metagenomic sequences, due to the uneven coverage of different genomes, as coverage information is often used to resolve these low complexity regions in isolate sequencing (Ghurye et al., 2016). The presence of multiple individuals from the same species with high micro-diversity can also cause assemblies to break (Shen & Millard, 2021). Assembly of viral genomes is often further complicated due to the lack of reference genomes or marker genes (Sutton et al., 2019), high mutation rates (Roux et al., 2017), and hypervariable regions that are associated with host interaction (Warwick-Dugdale et al., 2019). As a result, genomes are fragmented into shorter sequences, causing the loss of linkage and contextual information (Smits et al., 2014). This had significant implications for our prophage detection tool Phigaro, which relies on identifying areas of high phage gene density and prophage boundaries using a triangular window function. (Starikova et al., 2020). This further impacts our ability to detect actively replicating vOTUs and any VFs or ARGs that they may encode and transfer via transduction. These issues could have been alleviated by binning the scaffolds in MAGs, so that stronger inferences could be made into the taxonomy and functional potential of the vOTUs. However, viral scaffolds are rarely binned due to the same difficulties that affect assembly that are described above (Kieft et al.,

2022). Most binning algorithms are designed for bacterial and archaeal genomes and often rely on microbial single-copy genes. Several tools aimed at binning viral genomes from metagenomes have been developed since the work in this chapter was conducted. CoCoNet uses an unsupervised neural network to group fragments based on tetranucleotide compositions and coverage information (Arisdakessian et al., 2021). vRhyme, uses a supervised machine learning approach based on a number of sequence features and coverage information (Kieft et al., 2022). An alternative approach to producing high quality phage genomes was taken by the recent tool Phables (Mallawaarachchi et al., 2023). Phables leverages assembly graph information to identify genomic paths, producing sequences that are longer than individual scaffolds produced by assembly software. Any of these tools would likely improve the linkage and contextual information within our vOTUs. However, most binning tools struggle to separate closely related viral variants, and typically only a small fraction of the overall vOTUs can be binned. Long-read sequencing technologies such as Oxford Nanopore or PacBio sequencing platforms are capable of producing contiguous DNA fragments ($>$ 10 Kb) that can span low complexity regions (Amarasinghe et al., 2020), allowing for the assembly of substantially longer contigs. Some studies have even recovered full length, high quality viral genomes without assembly using Nanopore sequencing (Beaulaurier et al., 2020). However, recovery of phage genomes with long-read metagenomics is still difficult, especially when it comes to getting sufficient coverage or isolating high molecular weight DNA from complex and/or low biomass sample types. In addition, there are concerns with the error rates in long-read sequencing compared to their short-read counterparts. For example, the median raw read

accuracy for the R9.4.1 flow cell was less than Q10 in 2019. However this technology has consistently improved, and ONT claim that their Q20+ chemistry and R10.4.1 flow cells can produce raw read accuracies of Q24 (Oxford Nanopore Technologies, 2023a). Despite this, at the time of writing, R10.4.1 flow cells and Q20+ chemistry are not available for all Nanopore workflows, such as the Rapid PCR Barcoding Kit (SQK-RPB004).

One of the aims of this chapter was to detect phages encoding VFs and/or ARGs and investigate whether there was any evidence of HGT of these genes. Although we were able to identify some clinically relevant VFs and ARGs carried by a small number of vOTUs, it remains unclear whether they can confer any resistance or whether they can be transferred via transduction. In addition to this, due to time constraints, we did not manually inspect the genomic context of each ARG encoding vOTU. As a result, it is possible that some non-phage derived sequences remain. Out of the 38 and 54 vOTUs that carried VFs and ARGs respectively, only two of each were originally identified as prophages by Phigaro. To assess their potential for transduction, we carried out prophage:host ratio analysis with PropagAtE, to predict whether they were actively replicating. Neither of these vOTUs were predicted to be active, and hence no evidence for HGT of VFs and ARGs was found. However, it is worth noting that this approach has a low recall rate, for two reasons. Firstly, we could only run this analysis on prophages that assembled as integrated parts of a host scaffold (and then identified by Phigaro). Many of our vOTUs, including those that encoded VFs and ARGs, are likely to be prophages that did not assemble in this way due to the limitations of Illumina sequencing, as discussed previously. Secondly, since

active prophages will lyse their hosts and produce viral progeny, there is a lower probability of sequencing them than there is for integrated prophages. A potential solution to this is to sequence purified VLPs instead of the whole community. This approach could enable us to distinguish actively replicating phages from dormant lysogenic prophages. This would also increase the sequencing coverage and depth of viral DNA, as their genomes are very small when compared to bacteria and host-associated DNA. It is estimated that 5.8% of the DNA in a stool sample is viral (Shkoporov & Hill, 2019), so by only sequencing this fraction, the recovery of lower abundance genomes would be increased. However, depending on the VLP isolation technique, significant biases can be introduced into the composition, especially if a nucleic acid amplification step is included. Whole community metagenomic sequencing also provides valuable information about the bacterial host population, that is lost when VLP sequencing is carried out alone. Other than the prediction of lifestyle and whether the prophages were actively replicating we did not manually inspect the functional annotations of vOTUs that encoded ARGs and/or VFs due to time limitations. This would have been useful as we may have been able to identify genes such as integrases or excisionases which would suggest a specific lifestyle or specific MGEs within the vOTUs such as transposons or integrons that would provide an alternative route for HGT than transduction. We would also be able to compare the vOTUs to any related reference phage genomes which would provide us with greater genomic context, phenotypic information if they have been cultured, and any evolutionary history.

3.6 Conclusion

The results here showed that valuable insights into the composition and role of phages in the gut microbiota of ICU patients could be obtained from whole community metagenomic sequencing data, despite the limitations discussed above. We were able to define simple but clear criteria for microbial domination in the gut microbiota and use this to identify the dominant microbe in each patient, including one that had been discovered since the publication of the original study. The majority of the vOTUs we identified were novel, and their relative abundance was slightly higher than typically reported in healthy guts, albeit stable. The relative abundance of dominant microbes and vOTUs predicted to infect them had a near-perfect correlation. This correlation was not improved when introducing time lag, which led us to hypothesise that it was due to the majority of our vOTUs being derived from integrated prophages. We were also able to detect VFs and ARGs encoded by our vOTUs, although we were unable to determine whether these vOTUs were actively replicating or dormant. Future work will use long-read based VLP sequencing in order to capture actively replicating phages, with improved contextual information.

chapter 4

Investigating Bacteriophage Dynamics in the Gut Microbiome of Intensive Care Unit Patients using Viral Nanopore Metagenomics

4.1 Background

In the previous results chapter, I investigated the dynamics of the phage population in the gut of critically ill patients, using whole community shotgun sequencing data. The findings in Chapter 3 revealed that the viral population was stable over time, especially when compared to their bacterial hosts. Strong positive correlations between the dominant pathogen and their infecting vOTUs in each patient were found, potentially due to the majority of these vOTUs being integrated prophages. VFs and ARGs were identified in a number of vOTUs, but they were not found to be actively replicating with our coverage-ratio based method. These issues highlight the limitations of using whole community shotgun sequencing to investigate the dynamics of the gut virome.

ONT long-read sequencing has recently been applied to metagenomics, as it enables the retrieval of MAGs with higher contiguity when compared to Illumina sequencing platforms which generate relatively short reads. This approach has been applied to various sample origins, such as the human gut microbiome, wastewater, marine, and cow rumens. It has also recently been used to profile the gut virome of healthy volunteers (J. Cao et al., 2020; Lee et al., 2022). In this chapter, I aimed to couple ONT sequencing of VLP extractions with whole community shotgun sequencing in order to capture more complete vOTUs, a more representative set of actively replicating phages, as well as HGT dynamics with their hosts. No previous studies have utilised ONT technology to investigate the gut virome of non-healthy individuals. Therefore, this study also aims to explore the benefits and limitations of ONT sequencing on stool samples

taken from critically ill patients.

4.2 Aims and Objectives

1. To develop and optimise VLP extraction from stool protocol for ONT long-read sequencing.
2. To investigate the benefits and limitations of ONT long-read sequencing of VLPs.
3. To investigate the host-phage dynamics within a critically ill gut.
4. To investigate the contribution to HGT by phages in the critically ill gut.

4.3 Methods

4.3.1 Q33 Phage Propagation and Rnumeration

Lactococcal phage Q33 and *Lactococcus lactis* SMQ-86 were kindly provided by Colin Hill (APC Microbiome Ireland, University College Cork). In order to propagate Q33 phage, *L. lactis* SMQ-86 was grown overnight in M17 broth (Oxoid) supplemented with 0.5% glucose (GM17) at 30 °C without agitation. The overnight culture was then diluted into 50 mL of GM17 broth to obtain an optical density at a wavelength of 600 nm (OD_{600}) = 0.01. This culture was then incubated at 30 °C without agitation, and grown to an OD_{600} of 0.15, before the

addition of Q33 phage lysate for a multiplicity of infection (MOI) of 1, with subsequent incubation at 37 °C for 5 h without agitation. The phage lysate was then transferred into a 50 mL sterile conical centrifuge tube, and centrifuged at 4,000 x g for 20 min. The supernatant was then twice filter sterilised using 0.22 µm syringe-mounted polyethersulfone (PES) filters. 15 mL of the filtrate was then added to the upper reservoir of an Amicon® Ultra-15 centrifugal filter unit, Ultracel 100 kDa membrane (Millipore). These were then centrifuged at 4000 x g for 5 min. The filtrate was then discarded, and another volume of phage lysate was added and centrifuged until all the lysate had been processed. Once the lysate in the upper reservoir was < 10 mL, 15 mL of SM buffer was added, before centrifuging again at 4,000 x g for 5 min to wash the phage lysate. This wash step was then repeated, and the phage lysate was concentrated to < 10 mL. This was then sterilised using a 0.22 µm syringe-mounted PES filter. The phage lysate was then mixed with an equal volume of glycerol and stored at -80 °C for long term storage.

The titres of the phage lysate were determined using the double-agar method as previously described (Lillehaug, 1997), with several alterations. A serial dilution of phage lysate was made up to 10^{-7} with GM17 broth. 20 µL of these dilutions were then each mixed with 4 mL of melted top agar (0.4% *w/v* agar GM17, supplemented with 10 mM $CaCl_2$, held at 50 °C), and 200 µL of overnight (16 h) *L. lactis* SMQ-86 culture. These molten mixtures were then quickly poured onto GM17 1% agar plates, and the plates were tilted to evenly distribute the agar. Once the top agar had solidified, they were placed in an incubator at 30 °C, lid side up, overnight. The plaques on each plate were then

counted and the titre of the phage lysate in plaque forming units per mL (PFU mL^{-1}) was calculated with the following formula:

$$\text{PFU mL}^{-1} = \text{plaque count} \times \text{volume (20 µL)} \times \text{dilution factor}$$

4.3.2 Stool Sample Preparation

A single stool sample, H1, used for preliminary VLP extraction protocol optimisation was collected and kindly provided by Dr Amanda Rossiter (University of Birmingham). Ethical approval for the study was given by the Bradford Leeds Research Ethics Committee (REC 16/YH/0100).

The other stool samples used in this study were collected from patients staying in the ICU of the Queen Elizabeth Hospital Birmingham. Ethical approval for this study was given by the University Hospitals Birmingham Research Governance Office (REC 20/YH/0067; IRAS Project ID 268838). Stool samples were divided into 1.5 g aliquots and stored at -80 °C until use. This work was carried out by Ann Snaith (University of Birmingham).

4.3.3 Adaptation of VLP Nucleic Acid Extraction from Stool Protocol for ONT Long-Read sequencing

We adopted a stool VLP nucleic extraction protocol from Shkoporov et al., 2018b for ONT long-read sequencing. The following changes to the protocol were trialled for optimisation (changes are listed in order of when they are carried out in the

protocol, not in chronological order).:

1. Increase of input stool to 1 g from 0.5 g.

2. Homogenisation of resuspended stool in SM buffer by incubation on HulaMixer™ (Invitrogen) for 10 min and 1 min of vigorous vortexing instead of vigorous vortexing for 5 min.

3. Centrifugation of resuspended stool for 30 min instead of 10 min.

4. Increase in the weight of PEG-8000 powder added to filtrates from 10% *w/v* to 20% *w/v*.

5. Samples were centrifuged for 1 hr instead of 20 min after the overnight incubation with PEG-8000 and NaCl.

6. Removal of the chloroform extraction step after overnight precipitation.

7. Addition of a chloroform extraction step post-Phenol/Chloroform/Isoamyl Alcohol extraction.

8. Sodium chloride (0.2M final concentration) instead of sodium acetate (0.3M final concentration, pH 5.2) salt for ethanol precipitation.

9. Removal of the final round of purification with the DNeasy Blood & Tissue Kit (Qiagen).

10. Addition of the Zymobiomics DNA Clean and Concentrator Kit after the DNeasy Blood & Tissue Kit.

11. Cetyltrimethylammonium bromide (CTAB) (G-Biosciences 786-564) extraction post Phenol/Chloroform/Isoamyl Alcohol extraction using the manufacturer's protocol default protocol. Phenol/Chloroform product was used instead of a frozen powdered sample.

12. Phage lysis and nucleic acid extraction with only the Phage DNA Isolation Kit (Norgen Biotek).

13. Phage lysis and nucleic acid extraction with only the ZymoBIOMICS DNA Microprep Kit.

Not all of the listed changes were implemented into the final protocol. Details and reasons for whether each change was retained are described in Section 4.4.1. The final VLP nucleic extraction protocol we used is as follows: Aliquots of 500 mg of stool were resuspended in 10 mL of SM buffer 10 mM Tris-HCl pH 7.5, 8 mM $MgSO_4$, 100 mM NaCl) in a 15 mL conical centrifuge tube and homogenised by incubating on a HulaMixer™ for 10 min and 1 min of vigorous vortexing. For each set of extractions, a negative control of 1 mL 1.5 x 10^7 Q33 phage was used instead of 500 mg of stool. HulaMixer™ settings were set as: orbital rotation - 70 rpm, orbital rotation time: 0 s, reciprocal motion time - 30 s, reciprocal motion turning angle - 90°, vibrating motion - 5 s, vibrating motion turning angle - 1°. Tubes were then incubated on ice for 5 min before centrifugation at 5000 x g in a swing bucket rotor for 30 min at 4 °C. Supernatants were transferred to fresh tubes, which were centrifuged again at 5000 x g in a swing bucket rotor for 30 min at 4 °C. The supernatants were transferred to fresh tubes again, which were then spiked with either 50 µL or 100 µL of 7.5 x 10^8 PFU/mL Q33 phage lysate, for

final concentrations of 7.5 x 10^8 PFU/g and 3.75 x 10^8 PFU/g respectively. The resulting solution was then subjected to two rounds of filtration through sterile 0.45 µm pore PES syringe-mounted membrane filters (Fisherbrand). PEG-8000 and NaCl were added to the filtrates at a final concentration of 10% *w/v* and 0.5 M respectively. Tubes were then gently shaken until the powder completely dissolved, before being incubated overnight (16 h) at 4 °C. The next day, the samples were centrifuged at 5000 x g for 1 hr at 4 °C to pellet the precipitate. The supernatant was removed using a micro-pipette, and the tubes were inverted onto paper towels for 5 min. The tubes were then gently tapped onto the paper towels to remove any remaining supernatant. Pellets were then resuspended in SM buffer and transferred to a 1.5 mL microcentrifuge tube. An equal volume of chloroform was then added, and emulsified by gentle inversion. These were then centrifuged at 2500 x g for 5 min using a benchtop centrifuge, and 360 µL of the aqueous phase was aspirated into a clean 1.5 mL microcentrifuge. 40 µL of a solution of 50 mM $MgCl_2$ and 10 mM $CaCl_2$, 8 U of TURBO™ DNase (Thermo Fisher Scientific) and 20 U of RNase A (Qiagen) were then added to the microcentrifuge tubes, and DNA/RNA digestion was carried out at 37 °C for 1 hr, before inactivation of the enzymes at 70 °C for 10 min. Nucleic acid extraction was then carried out with the ZymoBIOMICS DNA Microprep Kit according to the manufacturer's instructions with the following alterations: 600 µL of lysis solution was added to the samples instead of 750 µL; 5% *v/v* of Proteinase K was added to the samples and incubated for 30 min at 55 °C, along with the lysis solution, as suggested by the manufacturer under appendix B; bead beating was omitted, but replaced with 10 minutes of inverting the tubes; the DNA was eluted with 15 µL DNase/RNase free water.

4.3.4 Quantification of DNA Extractions

DNA quality was determined by examining 1 µL of the sample on a NanoDrop™ 2000 Microvolume Spectrophotometer, using the dsDNA option. DNase/RNase free water was used as a blanking solution.

DNA concentration was examined using the Qubit™ dsDNA HS Assay Kit (Invitrogen). A master mix was made by diluting the Qubit™ dsDNA HS Reagent 1:200 in Qubit™ dsDNA HS Buffer. Standards were made by mixing 10 µL dsDNA HS Standard #1 and Standard #2 with 190 µL of the master mix in two separate Qubit™ Assay Tubes. To measure the concentration of each DNA sample, 199 µL of the master mix was combined with 1 µL DNA sample in a Qubit™ Assay Tube. The solutions were vigorously vortexed for 5 s, being careful to not introduce bubbles, and then incubated for 2 min at RT. DNA concentrations were read using a Qubit™ 4 Fluorometer (Invitrogen).

4.3.5 ONT Library Preparation and Sequencing

Following DNA isolation from VLP extractions, the isolated DNA was prepared for long-read sequencing using the ONT PCR Barcoding Kit (SQK-RPB004) (Oxford Nanopore Technologies). Up to 5 ng and 3 µL of DNA was transferred into a 0.2 mL thin-walled PCR tube. For samples that had a DNA concentration of $\geq 1.\overline{6}$ ng/µL, they were made up to 3 µL using DNase/RNase free water. 1 µL of Fragmentation Mix was then added and mixed gently by flicking the tube. The solution was then spun down and placed in a Mastercycler® Pro Thermal Cycler

(Eppendorf) and DNA was tagmented by incubating at 30 °C for 1 min and then at 80 °C for 1 min. Samples were then briefly put on ice to cool down. A PCR reaction mix consisting of 20 µL DNase/RNase free water, 4 µL tagmented DNA, 1 µL 10 mmol RLB primer (01-12A), and 25 µL LongAmp® Taq 2X master mix (New England Biolabs). The PCR tubes were then gently mixed by flicking the tube and then spun down. The PCR was then carried out using the program described in Table 4.1. The number of cycles was increased from 14 to 20 due to low input DNA concentration, as recommended by ONT.

Table 4.1: PCR program for amplification and barcoding in SQK-RPB004

PCR step	Temperature	Time	Number of cycles
Initial denaturation	95 °C	3 min	1
Denaturation	95 °C	15 s	
Annealing	56 °C	15 s	20
Extension	65 °C	6 min	
Final extension	65 °C	6 min	1

Amplified samples were then transferred to clean 1.5 mL DNA LoBind® tubes (Eppendorf) and mixed with 30 µL of thoroughly resuspended AMPure XP Beads (Beckman Coulter), by pipetting. Samples were then incubated for 5 min at RT on a HulaMixer™. Samples were spun down and pelleted on a MagnaRack™ Magnetic Separation Rack (Invitrogen). The supernatant was then removed, and

200 µL 70% fresh ethanol was added to the pellets for 30 s before being removed. This ethanol wash was repeated twice before air drying for approximately 30 s. Tubes were then removed from the magnetic rack and the pellet was resuspended in 10 µL of 10 mM Tris-HCl pH 8.0 with 50 mM NaCl. After 2 min of incubation at RT, the tubes were placed back on the magnetic rack. Once the eluate became clear and colourless, it was transferred into a clean 1.5 mL LoBind tube. Samples were then quantified using a Qubit fluorometer as described in 4.3.4. Equimolar amounts of each barcoded sample were then pooled into a 1.5 mL DNA LoBind tube to a total of 75 fmoles in 10 µL of 10 mM Tris-HCl pH 8.0 with 50 mM NaCl. Adapters were then ligated onto the pooled DNA by adding 1 µL of rapid adapter protein (RAP). This was mixed gently by flicking the tube and spinning down, before incubation for 5 min at RT.

The resulting DNA library was sequenced on an R9.4.1 Flow Cell (Oxford Nanopore Technologies) on a MinION Mk1B (Oxford Nanopore Technologies). Flowcells were first primed by removing any air bubbles from the priming port and loading 800 µL of priming mix (30 µL of Flush Tether mixed with a tube of Flush Buffer) into the priming port. This was then incubated for 5 min at RT. During this incubation, the sequencing library was prepared by mixing 34 µL Sequencing Buffer, 25.5 µL Loading Beads, 4.5 µL DNase/RNase free water, and 11 µL DNA library in a fresh DNA LoBind tube. Flow cell priming was completed by gently lifting the SpotON sample port cover to expose the SpotON sample port, and loading 200 µL of the remaining priming mix into the priming port. The sequencing library was then gently mixed by pipetting and added into the SpotON sample port in a dropwise manner. Sequencing was carried out for 72 hrs or until no

pores were actively sequencing. Active channel selection and reserved pore feature options were both toggled on. Basecalling was set to fast, and barcoding was set to on. Raw reads were saved as FAST5 files with no Qscore filtering. Human free Nanopore reads are available for download under ENA project PRJEB66265.

4.3.6 Whole Community DNA Extraction from Stool

Whole community DNA was extracted from 66 stool samples using the FASTDNA™ Spin Kit for Soil (MP Biomedicals). 200 mg of stool was added to Lysing Matrix E tubes along with 979 µL sodium phosphate buffer and 122 µL MT buffer. In cases where the stool samples took up a large volume, proportionately less of the two buffers were added to prevent overflowing. The samples were then homogenised using a FastPrep-24™ 5G bead beater (MP Biomedicals) for 40 s at a speed of 6.0 m/s before they were centrifuged at 14,000 x g for 10 min to pellet debris. The supernatants were then combined with 250 µL Protein Precipitation Solution in a new microcentrifuge tube, and mixed by inverting the tubes 10 times. Tubes were then centrifuged at 14,000 x g for 5 min to pellet the precipitated proteins, and the supernatants were transferred to 15 mL conical centrifuge tubes. 1 mL of Binding Matrix Solution was then added to the tubes and mixed by inversion for 2 min before they were allowed to settle on a rack for 3 min. 500 µL of the supernatant was then removed, the Binding Matrix was resuspended, and 600 µL of the suspension was added to a SPIN™ filter and centrifuged at 14,000 x g for 1 min. The flow-through was discarded and the step was repeated until all the remaining suspension had been processed

through the filter. The pellet was then resuspended using 500 µL SEWS-M solution and the SPIN™ filter was then centrifuged at 14,000 x g for 1 min. The flow-through was discarded and the centrifugation was repeated for 2 min. The SPIN™ was transferred to a clean 1.5 mL DNA LoBind tube (Eppendorf) and air-dried for 5 min at RT. The Bind Matrix in the SPIN™ filter was then gently resuspended with 50 µL DNase/RNase free water and incubated for 5 min at 55 °C in a heat block. The final eluate was then centrifuged at 14,000 x g for 1 min and stored at -20 °C. The DNA was quantified using the Qubit™ dsDNA HS Assay Kit and NanoDrop™ 2000 Microvolume Spectrophotometer (Section 4.3.4).

4.3.7 Whole Community Metagenomic Sequencing

For whole community metagenomic sequencing, samples extracted in Section 4.3.6 which passed quality control, were sent to NovoGene (Cambridge, UK) for sequencing on an Illumina NovaSeq 6000 2x150 bp paired-end platform (Illumina). Quality control requirements included, OD260/280 between 1.8 and 2.0, DNA concentration $\geq$1 ng/µL, low DNA degradation, and low RNA/Protein contamination. The latter two requirements were analysed using gel electrophoresis by NovoGene. Human free metagenomic reads are available for download under ENA project PRJEB66265.

4.3.8 Quality control and assembly of long-read ONT VLP reads

Raw reads from long-read sequencing were moved onto a basecalling workstation that was equipped with a NVIDIA GeForce RTX 3080 10GB VRAM graphics processing unit (GPU). Basecalling was carried out with Guppy basecaller (v6.0.1) (Oxford Nanopore Technologies) using the Super accuracy (SUP) model (`--chunk_size 3000 --chunks_per_runner 200 --gpu_runners_per_device 4 --min_qscore 7 --trim_barcodes --config dna_r9.4.1_450bps_sup.cfg --recursive`). Chunk size was fixed to 3000, as some community testing showed that PHRED quality score increases with this parameter, and starts to plateau at this value (Manzanilla, 2020). Chunks per runner was set at 200, and GPU runners per device at 4, as these were the highest values where basecalling would run stably on our system. Sequencing and quality statistics were calculated and plotted with NanoPlot. Long-reads were filtered using Filtlong (v0.2.1) `--min_length 1000 --keep_percent 95` (https://github.com/rrwick/Filtlong). Human reads were detected by mapping to a repeat masked version of the GRCh38.p14 reference genome from NCBI (https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.40) with minimap2 (v2.24) (`-ax map-ont`) (H. Li, 2018), non-human reads were extracted with SAMtools (`fastq -f 4`) (H. Li et al., 2009). Human decontaminated reads were then mapped to the *L. lactis* Q33 reference genome (Accession NC_049809.1) and negative control reads with minimap2 and non-mapped reads were extracted as with the non-human reads. Percentage of reads that mapped the *L. lactis* Q33

reference genome was calculated with SAMtools `flagstat` . Virome contamination was estimated with ViromeQC (v1.0) (Zolfo et al., 2019). Contamination was also estimated by mapping to a database of 549–567 bp segments of *cpn60* (J. E. Hill et al., 2004) with minimap2.

The filtered reads were then assembled with Flye (v2.9) (`--nano-hq --meta --scaffold` (Kolmogorov et al., 2019). Assemblies were then polished with two rounds of Racon (v1.4.22) (Vaser et al., 2017), iterated with Raconinator (https://github.com/stevenjdunn/Raconinator with default settings, and one round of Medaka (v1.7.3) `-m r941_min_sup_g507` (https://github.com/nanoporetech/medaka). All polished assemblies were then concatenated and dereplicated with CD-HIT-EST as in section 3.3.3. Filtered reads for each sample were then mapped to the dereplicated scaffolds with minimap2, and a coverage table was created with `jgi_summarize_bam_contig_depths` from MetaBAT2 (v2.15) (D. Kang et al., 2019). The dereplicated scaffolds were then binned using vRhyme (v1.1.0) (–red 5) (Kieft et al., 2022). Scaffolds were linked using vRhyme's `link_bin_sequences.py` script to generate a single scaffold per bin. Unbinned scaffolds less than 10 Kb that were predicted as circular by vRhyme's `flag_circular.py` script, unbinned scaffolds greater than 10 Kb, and the binned sequences were concatenated into a single multifasta. In order to further decontaminate the scaffolds, any scaffold that fulfilled at least one of the viral sequence identification criteria outlined in Chapter 3, section 3.3.3, were retained as vOTUs. Quality, completeness, and contamination of vOTUs were estimated with CheckV (Nayfach et al., 2020). Read mapping percentage was calculated

with SAMtools `flagstat`.

4.3.9 Quality Control and Human Sequence Removal from Metagenomes

Quality control of metagenomic reads and removal of human sequences was carried out as in Chapter 3, section 3.3.1.

4.3.10 Taxonomic Identification of Whole Community Metagenomic Reads

Taxonomic assignment of whole community metagenomic reads was carried out as in Chapter 3, section 3.3.2.

4.3.11 Taxonomic and Functional Profiling of vOTUs

High level taxonomic classification of vOTUs, functional annotation, and lifestyle prediction was performed as in Chapter 3, section 3.3.4. We also carried out family and genus level taxonomic classification with VPF-Class (v0.1.2) (Pons et al., 2021), retaining the classification with the highest membership score for each vOTU as long as it was $\geq$0.3 and its confidence score was $\geq$0.8.

4.3.12 Sequence Level Analysis of ARG Encoding vOTUs

vOTUs encoding ARGs were identified using pharokka (`-meta`) (v1.3.2 (Bouras et al., 2023)) using CARD (Jia et al., 2017). The closest relative to our ARG vOTUs was determined by alignment to the NCBI nr/nt database with the megablast algorithm. Any phage termini were identified using CheckV (v1.01) (`end_to_end`) (Nayfach et al., 2020). Metagenomic contigs were assembled with SPAdes (v3.15.4) (`-meta`) (Nurk et al., 2017). We then aligned our ARG vOTUs against the subsequent concatenated metagenomic assemblies with the BLASTN program (v2.12.0+) (`-outfmt "7 qseqid qlen sseqid slen qstart qend sstart send evalue length qcovs pident"`). We also aligned both of the vOTUs against TN*4555*, whose sequence had been downloaded from NCBI (Accession: U75371.3) . The ARG encoding vOTUs, their closest metagenomic contigs, Tn*4555*, and their closest GPD reference were then globally aligned and visualised with clinker (Gilchrist & Chooi, 2021).

4.3.13 Calculation of vOTU Absolute Abundances

Relative abundance of vOTUs was calculated as in Chapter 3, section 3.3.3, except alignment of reads was performed by minimap2 instead of Bowtie2. Absolute abundance was then calculated from the relative abundances and mapping rate to Q33 spike-in genome that was determined in section 4.3.8 using the following formula:

$$\text{Absolute abundance} = \frac{1000}{\text{Input stool (mg)}} \times \frac{\text{Spike-in PFU mL}^{-1}}{\text{Spike-in mapping \%}} \times \text{Relative abundance}$$

4.3.14 Linear Regression and Correlation of Dominant Host-Phage Pairs

Linear regression and TLCC analysis of dominant host-phage pairs were conducted as in Chapter 3, section 3.3.4 and section 3.3.5 respectively.

4.4 Results

4.4.1 Adaptation of VLP Nucleic Acid Extraction from Stool Protocol for ONT Long-Read Sequencing

A reproducible protocol for isolating the VLP fraction from stool samples, and extracting nucleic acids from it, ready for high-throughput next generation sequencing was published by Shkoporov et al., 2018b. This protocol was selected as it does not require labor intensive techniques such as CsCl density gradient purification (Castro-Mejía et al., 2015; Kleiner et al., 2015), with relatively low bacterial contamination. ONT sequencing typically requires higher amounts of input DNA (400 ng for Rapid library prep (Oxford Nanopore Technologies, 2023c), and 1000 ng for Ligation based library prep (Oxford Nanopore Technologies, 2023b)) compared to Illumina sequencing which can have input

DNA requirements of as low as 1 ng (Illumina, 2023). High molecular weight DNA is also recommended for ONT sequencing as the resulting read lengths are dependent on the input fragment length. This posed issues for using Shkoporov et al.'s protocol *verbatim*, as the nucleic acid yield was as low as 60.5 ng from 0.5 g of stool from a healthy individual. The protocol also used a silica-based spin-column kit for the purification of nucleic acids, which are known to shear DNA during the binding and elution steps (Maghini et al., 2021). Therefore, we attempted to adapt Shkoporov et al.'s protocol to produce higher DNA yields at a higher molecular weight.

We used the same healthy stool sample as the input into each optimisation attempt, as we had a limited quantity of stool samples from ICU patients, and we wanted attempts to be comparable. In the first attempt, we doubled the input amount of stool from 0.5 g to 1 g and compared the effect of increasing the PEG-8000 concentration used for precipitating VLPs from 10% *w/v* to 20% *w/v*, by running two extractions side by side. This resulted in a DNA yield of 124 ng and 78 ng for the 10% *w/v* and 20% *w/v* PEG-8000 extractions respectively. This lower particle yield for higher PEG-8000 concentrations than 10% has been shown for the isolation of extracellular vesicles (Ludwig et al., 2018). We therefore decided to use 10% *w/v* PEG-8000 to concentrate VLPs in the subsequent optimisation attempts. The NanoDrop UV spectrum of the extracted DNA had an A260/280 ratio of 1.60, and an A260/230 ratio of 0.2. These low ratios were characterised by a high peak at 230 nm, and a smaller peak at 270 nm. We hypothesised that the 230 nm peak was caused by contamination by an organic compound, and the peak at 270 nm was caused by phenol contamination

shifting the typical DNA peak of 260 nm up. This hypothesis was corroborated by the DNA concentration reported by the NanoDrop, which was 100x more (240.4 ng/uL) than the concentration quantified using a Qubit™ 4 Fluorometer. In the next attempt, we decided to add a chloroform extraction step after phenol-chloroform extraction in order to remove any remaining phenol, and sodium acetate was replaced with sodium chloride as the salt used during ethanol precipitation, as sodium acetate can cause the sodium dodecyl sulfate (SDS) used in the protein solubilisation step to precipitate. These changes resulted in a slightly higher DNA yield of 148 ng. However, DNA purity issues remained, with an A260/280 ratio of 1.7, and an A260/230 ratio of 0.73. The absorbance spectra profile was similar to before, although the peak at 270 nm had shifted over to 260 nm, and was taller, relative to the peak at 230 nm. The DNA concentration reported by the NanoDrop was 200x higher (581.3 ng/uL) than quantified by the Qubit assay. This suggested that phenol contamination was the culprit behind the 270 nm peak, so the chloroform extraction step was implemented in subsequent optimisation attempts. The large difference in DNA concentration reported by the two machines also suggests that another contaminant is inflating the peak at 260 nm. We hypothesised that the 230 nm peak may been caused by polysaccharide carryover. This is typically an issue when extracting nucleic acid from plant matter, and polysaccharides can be precipitated using the catonic surfactant CTAB. CTAB extraction has previously been used in conjunction with phenol-chloroform in generic stool DNA isolation protocols, as well as for stool samples from highly vegetarian species such as elephants (Doyle & Doyle, 1987; Kouakou et al., 2022; Tamari et al., 2013; B.-W. Zhang et al., 2006). Therefore, we attempted the protocol with a CTAB extraction after the ethanol

precipitation step which yielded 198 ng of DNA. However, no nucleic acid was recovered post-CTAB extraction. We then tried to remove this 230 nm contaminant by using the ZymoBIOMICS DNA Clean and Concentrator Kit after the DNeasy Blood & Tissue kit, as an additional purification step. Despite this, the peak at 230 nm remained, and the DNA yield was reduced to 50 ng. Next, we replaced the entire phage lysis and nucleic acid extraction of the protocol with the Phage DNA Isolation Kit by Norgen Biotek. This kit is specifically designed for purifying DNA from phages, and uses a heat and chemical based lysis process, as well as a silica-based spin-column to bind DNA. The 230 nm peak on the DNA absorbance spectrum remained but was slightly reduced (A260/280 ratio = 1.78, A260/230 ratio = 0.81), whilst DNA yield was quantified at 151 ng by Qubit assay. As all the extraction attempts we had carried out to this point had DNA yields that were significantly lower than the input requirement for ONT's Rapid or Ligation based library preparation methods, we opted for ONT's Rapid PCR Barcoding kit, which has an input requirement of between 1-5 ng. This library preparation method produces read lengths of approximately 2 Kb, and whilst this is significantly shorter than the read lengths produced by ONT's non-PCR methods, which are dependent on input fragment length, it is still considerably longer than those produced by Illumina's sequencing platforms (typically up to 300 bp). As this kit uses PCR to amplify the input DNA, it also has the benefit of diluting any contaminants that were not removed in the extraction process, and therefore improved downstream sequencing results. When we attempted to amplify the DNA extracted with the Phage DNA Isolation Kit, not enough DNA was produced to sequence, even after the number of PCR cycles was increased from 14 to 20 (9.5 fmol in 10uL). This

is likely due to the 230 nm contaminant inhibiting the PCR process. Therefore we went back to modifying our protocol in order to produce a sufficient pure DNA extraction. We decided to replace the Phage DNA Isolation Kit with ZymoBIOMICS DNA Microprep kit as the latter kit is designed for extracting DNA from a wide array of sample types including stool, whereas the former kit is primarily used for the isolation of DNA from liquid propagations of phage. The ZymoBIOMICS kit also includes a proprietary PCR inhibitor removal step which can remove prohibitive impurities such as polyphenols. This modified protocol produced a DNA yield of 76 ng. The UV absorbance spectrum no longer had defined peaks at 230 nm and 260 nm, but instead displayed a smooth decay curve from 220 nm down to ∼290 nm (A260/280 ratio = 1.97, A260/230 ratio = 0.40). However, the estimated DNA concentration by the NanoDrop was much closer to the Qubit quantification than in previous estimates (NanoDrop: 9.9 ng/uL; Qubit: 5.1 ng/uL). After amplification and barcoding with the Rapid PCR Barcoding kit, the nucleic acid peak at 260 nm was recovered and its ratio with 280 nm and 230 nm indicated that the sample had high purity (A260/280 ratio = 1.70, A260/230 ratio = 2.05). The final barcoded sample was quantified at 54.4 ng/uL. As this DNA yield was significantly higher than the input requirement for the ONT Rapid PCR Barcoding kit, and the quantity of stool we were receiving from ICU patients was low (average of 2.1 g per sample) we reduced the mass of input stool back down to 0.5 g.

In addition to the above modifications, we also attempted to remove the chloroform extraction step after PEG precipitation, as phages belonging to families such as *Inoviridae*, *Corticoviridae*, and *Plasmaviridae*, have been shown to be

susceptible to chloroform (Conceição-Neto et al., 2015). However, the omission of this extraction resulted in no DNA being recovered. This may be related to the viscous high molecular weight compound reported by Kleiner et al., 2015 after the addition of PEG to their filtrate. We also increased the centrifugation times of the resuspended stool and PEG precipitate, as the samples were not completely pelleted using the times proposed by Shkoporov et al., 2018b. The lower number of large particles in the stool supernatant resulted in more efficient filtering, as filters were less prone to being blocked. Finally, we combined 10 minutes of HulaMixer™ rotation and vibration, and 1 minute of vigorous vortexing to homogenise the stool samples, instead of 5 minutes of vigorous vortexing alone, as this vastly increased our throughput when processing multiple samples.

Table 4.2: Proposed changes to VLP extraction protocol and their outcomes

Proposed Change	Reasoning	Result	Final Protocol
Increase input stool from 0.5 g to 1 g	Higher DNA yield needed for ONT sequencing	DNA yield increased	No, as lower DNA yield was adequate for PCR Barcoding Kit
Homogenisation of stool with HulaMixer	Higher throughput and less labour intensive	More time efficient protocol	Yes

Proposed Change	Reasoning	Result	Final Protocol
Centrifugation of resuspended stool for 30 min instead of 10 min	Stool particulates were not completely pelleted and blocked filters	Fewer particulates in supernatant and filter	Yes
20% *w/v* PEG-8000 instead of 10% *w/v*	A few protocols use the higher concentration of PEG and report more efficient precipitation of VLPs	DNA yield decreased by 37%	No
Centrifugation of PEG precipitate for 1 hr instead of 20 min	Small amounts of PEG precipitate were still visible in the supernatant	No difference in DNA yield	Yes, as we are theoretically isolating more VLPs, even if no change was seen in our one comparison

Proposed Change	Reasoning	Result	Final Protocol
Removal of chloroform extraction step after PEG precipitation	Excess PEG should be removed by the subsequent steps in the protocol, and lipid enveloped phages are sensitive to chloroform	Almost no DNA could be recovered	No
Addition of a chloroform extraction after the phenol-chloroform extraction step	Poor DNA quality (Low A260/230 and A260/A280) may result from phenol contamination	DNA quality improved slightly but contamination remained	No, as phenol-chloroform extraction was not included in the final protocol

Proposed Change	Reasoning	Result	Final Protocol
Sodium chloride (0.2M final concentration) instead of sodium acetate (0.3M final concentration, ph 5.2) salt for ethanol precipitation	Sodium acetate can cause SDS to precipitate	Increased DNA yield but DNA quality remained poor	No, as phenol-chloroform extraction was not included in the final protocol
Addition of the ZymoBIOMICS DNA Clean and Concentrator Kit at the end of the protocol	To increase DNA quality and concentrate the elution	Contamination was not removed, and DNA yield was low	No

Proposed Change	Reasoning	Result	Final Protocol
CTAB extraction after ethanol precipitation with sodium chloride	Carbohydrate carryover may be the culprit behind low A260/230 and A260/A280 ratios, which often occurs when using plant matter. CTAB is often used to remove this contaminant.	No DNA was recoverable after CTAB extraction	No
Phage lysis and nucleic acid extraction with the Phage DNA Isolation Kit (Norgen Biotek) alone	Using a kit that is designed for phage DNA extraction may solve the issues with poor DNA yield and quality	Poor DNA yield and quality	No

Proposed Change	Reasoning	Result	Final Protocol
Phage lysis and nucleic acid extraction with the ZymoBIOMICS DNA Microprep Kit alone	This kit contains lysate filtration and PCR inhibitor removal steps which may help remove the DNA contaminant(s).	Low DNA yield but the absorbance peak at 220 nm disappeared	Yes, with the contaminant removed, we were able to use a PCR based library preparation method to increase DNA concentration

Briefly, the final optimised protocol included the following changes: (1) resuspended stool was homogenised by 10 min of rotation and vibration with a HulaMixer™ and 1 min of vigorous vortexing, instead of just 5 min of vigorous vortexing, (2) the two rounds of centrifugation to seperate particulate matter were extended to 30 min each, up from 10 min, (3) centrifugation of PEG-precipitated viral particles was extended to 1 hr, up from 20 min, (4) excess supernatant was removed from pelleted PEG precipitate using a micro-pipette, in addition to inverting onto paper towels for 5 min, (5) Lysis, nucleic acid extraction, and purification was carried out using the ZymoBIOMICS DNA Microprep Kit, instead of using a guanidinium isothiocyanate based lysis buffer, phenol-chloroform extraction, and DNeasy Blood and Tissue spin-column kit. (6) Proteinase K digestion was moved to after cell lysis, per the ZymoBIOMICS DNA Microprep Kit manufacturer's instructions for enhanced extraction of DNA viruses. (7) The DNA elution volume was reduced to 15 µL, down from 50 µL.

4.4.2 ICU Patient Selection and Implementation of VLP ONT Sequencing Protocol

In order to study phage dynamics in the gut of critically ill patients using viral ONT metagenomics, we selected five individuals who had extended stays ($>$ 20 days) in the Queen Elizabeth Hospital Birmingham (QEHB) ICU, and provided $>$ 5 stool samples. These patients were all admitted to the ICU for non-infectious disease related reasons (Table 4.3). All patients received at least two different intravenous (IV) antibiotics across their stay (Figure 4.1).

Table 4.3: Clinical details of ICU patients

Patient	Sex	Age	Clinical Details	Length of ICU Stay (Days)
21	F	74	Road traffic collision	51
25	F	61	Post-surgery for malignant neural sheath tumour	24
26	M	51	Boerhaave's syndrome	22
29	M	51	Subarachnoid hemorrhage	30
46	M	75	Fall	77

The optimised VLP nucleic acid extraction protocol had mixed success on the ICU patient stool samples, with only 46.4% (32/69) of samples being

successfully sequenced. However, the majority (70.3%; 26/37) of samples which could not be sequenced were from one patient, patient 21. Only one sample from patient 21 had sufficient DNA yield to be sequenced. We believe this is due to the samples containing naturally low numbers of phage, or some component of this patient's stool interfering with our protocol. In total, 32 VLP extractions were sequenced, producing 42.6 Gb of raw sequence data with a median read length of 2,428 and a median read quality of 13.5. Low quality reads were removed by filtering out reads with an average Phred quality score of less than 7. The bottom 5% of reads by Phred score that were remaining were also removed, along with reads shorter than 1 Kb. The dataset was filtered to remove reads aligning to human and Q33 spike-in genomes. This left us with an average of 953 Mb per sample, with a median read length of 2,428, and a median read quality of 13.7.

Table 4.4: Per sample Nanopore sequencing statistics

Sample (Patient No. - Day in ICU)	Total bases (Gb)	Read length N50	Mean read quality
P21 - 9	1.235	2,931	13.6
P25 - 6	3.884	3,420	13.2
P25 - 7	0.708	3,260	13.2
P25 - 8	3.653	3,341	12.8
P25 - 9	1.367	3,398	13.1
P25 - 10	1.465	3,815	12.8
P25 - 13	1.459	4,291	13.2

Sample (Patient No. - Day in ICU)	Total bases (Gb)	Read length N50	Mean read quality
P25 - 15	0.893	3,898	13.0
P25 - 16	1.000	4,631	12.9
P25 - 18	0.777	4,006	12.8
P26 - 16	1.280	2,435	13.4
P26 - 19	1.293	2,228	13.2
P29 - 12	0.849	2,348	12.7
P29 - 19	1.110	2,653	13.1
P29 - 20	1.145	2,470	13.1
P29 - 21	0.897	3,487	13.2
P29 - 23	1.094	3,088	13.0
P29 - 24	1.147	2,910	13.2
P29 - 26	0.773	4,036	13.3
P29 - 29	0.939	2,777	13.5
P46 - 4	1.157	2,193	13.1
P46 - 9	1.132	2,445	13.4
P46 - 12	1.323	2,190	13.4
P46 - 13	1.388	2,539	13.8
P46 - 14	0.862	2,324	13.8
P46 - 15	1.061	1,988	14.0
P46 - 17	0.326	2,561	14.1
P46 - 18	1.034	2,259	14.0
P46 - 20	0.937	2,711	13.8

Sample (Patient No. - Day in ICU)	Total bases (Gb)	Read length N50	Mean read quality
P46 - 23	1.259	2,498	13.6
P46 - 27	1.361	2,456	14.0
P46 - 30	1.348	2,860	14.0
Q33 only control 1	1.149	2,785	13.4
Q33 only control 2	1.109	2,599	13.6

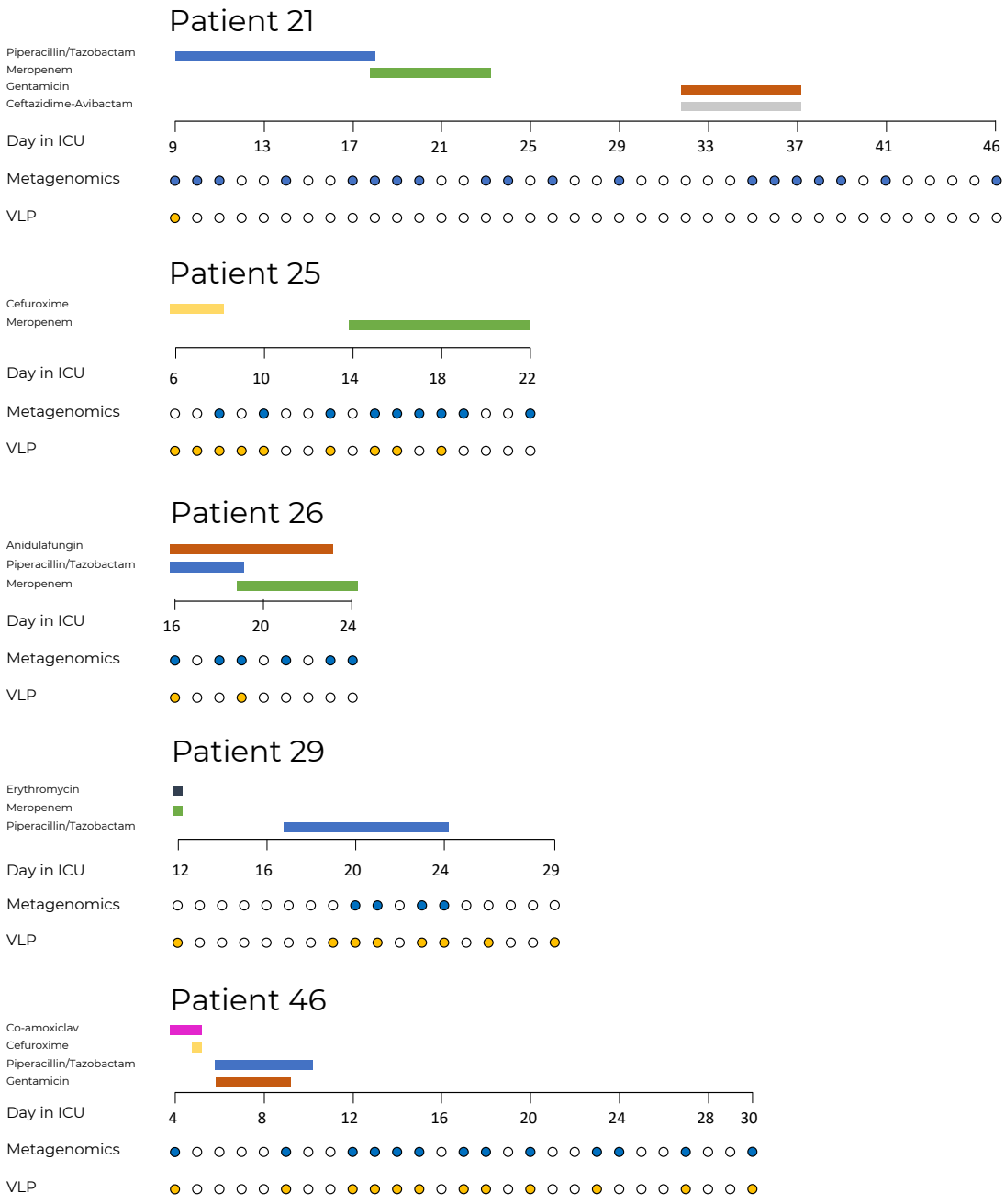


Figure 4.1: (Caption on next page.)

Figure 4.1: **Timeline of sequenced stool samples for each intensive care unit patient.** Stool samples that underwent virus-like-particle and/or metagenomic sequencing are indicated by coloured dots under the day they were collected. The days when patients were administered intravenous antibiotics or antifungal drugs are represented by the colour bars above the timeline. Empty dots indicate samples that either were not successfully sequenced, or no stool was collected on that day.

4.4.3 Metagenomic Sequencing of ICU Patient Stool Samples

Metagenomic sequencing was also carried out on stool samples from the five ICU patients to provide a wider community context to the virome sequences. Whilst we attempted to cover the exact same samples that were processed through the VLP protocol, 2 samples from patient 25 and 4 samples from patient 29 could not be successfully sequenced due to one of two reasons: (1) after VLP isolation and extraction, the remaining quantity of stool was not sufficient for metagenomic extraction; (2) metagenomic sequencing was unsuccessful due to low quality DNA extracted from certain samples. Conversely, for patients 21, 25, 26 and 46 we also shotgun sequenced a number of samples that we were unable to produce viromes for.

4.4.4 Intestinal Microbiome Domination in the ICU Patients

Intestinal microbiome domination by a single species ($\geq$30% relative abundance) was identified in all patients except for patient 26 (Figure 4.2). Patient 21's gut was dominated by five different species during their stay in the ICU. First, their

gut was dominated by *E. coli* which had bloomed from 5.68% relative abundance to 43.7% between day 9 and day 10. This was then momentarily replaced by *Pseudomonas aeruginosa* which had a relative abundance of 65.0% on day 13, before *E. coli* returned on day 16 (65.4%). This *E. coli* dominance lasted till day 19, where it was replaced by *E. faecalis*, which was then itself replaced by *P. aeuginosa* on day 23. No samples were sequenced until day 28, when no single dominating species was identified. This continued till day 34 where a significant bloom of *E. faecalis* occurred which reached a peak of 93.6% relative abundance on day 38. This was finally replaced by *Phocaeicola dorei* on day 40 which had a relative abundance of 51.2%.

Patient 25 was only dominated by *E. faecium* over the period of samples that we sequenced. This domination started between day 12 and day 15 of their ICU stay, where *E. faecium* bloomed up into over half of their gut microbiome (50.2%). This continued till the last day that we sampled (day 21) where *E. faecium* had a relative abundance of 72.6%.

Patient 29 was dominated by a bacterium in the *Clostridia* class. This is likely made up of multiple species belonging to genera such as *Blautia*, *Ruminococcus*, or *Faecalibacterium*, but MetaPhlAn did not have enough information to classify them at lower taxonomic levels. Dominance by this taxon began between day 20 and day 22 of patient 29's ICU stay, with a relative abundance of 48.5%. This rose to 60.7% on day 23 which was the last sample that was sequenced.

The dominating species in patient 46 was *Akkermansia muciniphila* which

grew from 7.23% relative abundance on day 8 to 33.4% on day 11. The relative abundance of *A. muciniphila* peaked on day 13 at 43.5%. Its abundance then dropped below our domination threshold on day 17 (22.6%) before recovering slightly on day 19 (30.1%), before dropping to pre-domination levels (6.15%) by day 23.

Whilst patient 26 did not experience our definition of domination, it is notable that a stable population of *Ruthenibacterium lactatiformans* was found across their samples, with a mean relative abundance of 22.0%.

4.4.5 Assessment of Bacterial DNA Contamination in Viromes

To assess the enrichment rate of our viromes, we used ViromeQC (Zolfo et al., 2019), a tool that maps reads against a database of rRNA sequences and 31 single-copy bacterial marker genes, and estimates an enrichment score compared to the median alignment rate in unenriched human gut metagenomes, which is normalised to an enrichment score of 1x. The median enrichment score in our viromes was 0.4x, which indicated that our sequences had high levels of bacterial contamination. This score is calculated using the highest alignment rate to small subunit (SSU) rRNA, large subunit (LSU) rRNA, or the single copy marker genes, and Zoflo et al. showed that the abundances of the rRNAs were correlated with the marker gene abundances. For all of our samples, the enrichment score was calculated with the SSU rRNA abundance, which was not correlated with the single copy marker abundances (Spearman's rank coefficient = -0.42), which had a median enrichment of 41.6x. Additionally, we assessed bacterial DNA contamination by aligning our

viromes to the *cpn60* gene database (J. E. Hill et al., 2004). *cpn60* genes are present in virtually all bacteria as a single copy, and their presence has previously been used to estimate the level of bacterial contamination in viromes (Shkoporov et al., 2018b), and was not one of the 31 single copy marker genes included in ViromeQC. The median percentage of reads that aligned to the CPN60 database was 0.0024% (Q1 = 0.0000%, Q3 = 0.0044%). Assuming that the average genome size of gut bacteria is 3.9 Mb (Nayfach & Pollard, 2015) and the average *cpn60* gene is 565 bp, this level of *cpn60* gene contamination corresponds to around 16.9% (Q1 = 0.00%, Q3 = 31.1%) of DNA in our samples was of bacterial origin.

4.4.6 Decontamination of Viromes

Assembly of each sample was carried out using only ONT reads, as we did not have corresponding Illumina reads for every time point. Assembled contigs were then polished with two rounds of Racon and one round of Medaka, before being combined and dereplicated, resulting in 4,356 sequences. We then applied vRhyme, a novel binning tool for constructing viral metagenome assembled genomes (vMAGs), on our contigs to produce the most contiguous sequences possible. Due to the non-negligible bacterial DNA contamination in our samples, we employed the same *in silico* viral identification steps as in Chapter 3, section 3.3.3. This left us with 684 vOTUs, of which 3 were vRhyme derived vMAGs, with a median length of 37.7 Kb (Q1 = 19.9 Kb, Q3 = 56.7 Kb). The average percentage of sample reads that could be recruited to the vOTUs was 88.9%. To assess the completeness and quality of our vOTUs we used CheckV. In total, we

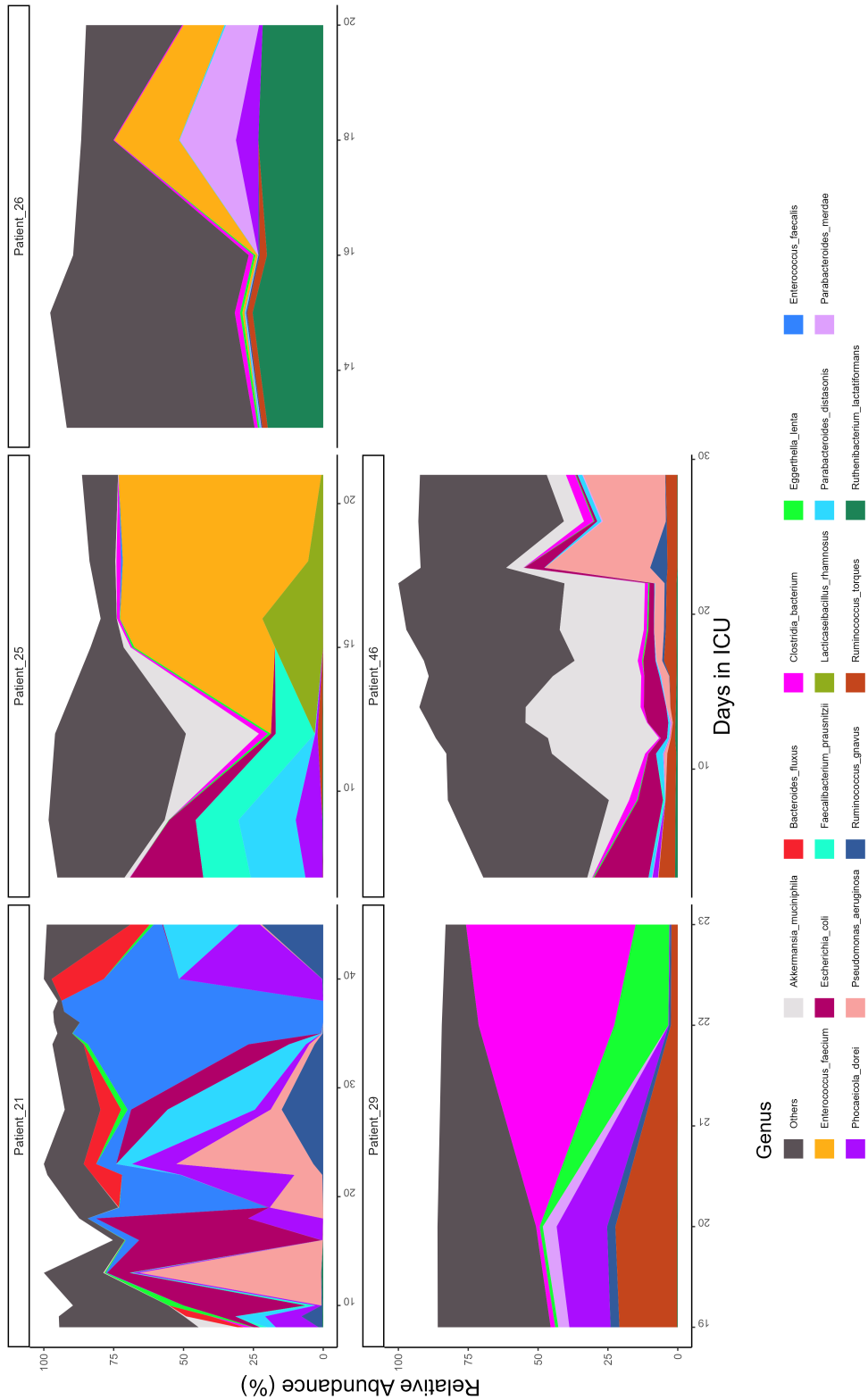


Figure 4.2: (Caption next page.)

Figure 4.2: **Species level taxonomy of metagenomic reads for each patient across their stay in the intensive care unit.** Taxonomy and relative abundance for each sample were assigned using MetaPhlAn4. Only genera that have a relative abundance of ≥15% in at least one sample are displayed. All other genera were collapsed into the "other" category.

found that 6 (0.88%) vOTUs were complete, 336 were high quality (49.1%), 121 (17.7%) were medium quality, 197 (28.8%) were low quality, and 24 (3.5%) were not determined. The vOTUs whose quality was classified as "not-determined" are due to them not having high enough similarity to any CheckV reference genome and no viral HMMs were found. This is either due to these vOTUs being too short, although their median length was 11.7Kb, or they belong to a very novel viral group which is not closely related to any genomes in the CheckV database, or they are false positives that slipped through our decontamination steps. The median completeness, excluding vOTUs which were not determined, was estimated to be 92.3% (Q1 = 43.1%, Q3 = 99.2%), and 98.1% of all vOTUs contained no host contamination.

4.4.7 Taxonomic and Functional Profiling of vOTUs

To investigate the composition of the vOTUs at higher taxonomic levels, we used vConTACT2 to cluster them with the INPHARED database (Figure 4.4). 341 (49.9%) vOTUs clustered with a reference genome, which indicated that they were related at the genus level. 144 (21.1%) vOTUs were included in the global network but did not cluster with other genomes (outliers), as they are likely to be related

to the VCs but not at genus level. The other 196 vOTUs either only clustered with other vOTUs (173; 25.3%) or were labelled as singletons (26; 3.80%), indicating 81 putative novel genera. We taxonomically classified the vOTUs at family and genus level using VPF-Class, which automates this classification by assigning proteins in viral contigs to classified orthologs, which are called Viral Protein Families (VPFs). Using a minimum confidence score of 0.8 and a minimum membership ratio of 0.3, we were able to obtain taxonomic classifications for 609/684 vOTUs at family level. All 609 were placed into the three tailed phage families, 100 in *Myoviridae*, 100 in *Podoviridae*, and 409 in *Siphoviridae*. These families have been abolished from the ICTV taxonomy (see Chapter 1 Section 1.1.1) but are still used in the current version of VPF-Class, although the authors said in October 2022 that they plan to update their classification in a future version (https://github.com/biocom-uib/vpf-tools/issues/31). Using the same parameters, VPF-Class was able to classify 605/684 vOTUs at the genus level, with the lowest membership score being. The most common genus classification was for *Cba41virus* (12.3%), which was renamed to *Lightbulbvirus* by the ICTV in 2019 https://ictv.global/taxonomy/taxondetails?taxnode_id=202200630&taxon_name=Lightbulbvirus%20Cba41. Two species of *Lightbulbvirus* are currently recognised by ICTV: *Lightbulbvirus Cba41* and *Lightbulbvirus Cba172*, both of which known to infect strains of *Cellulophaga baltica* (Holmfeldt et al., 2007; Mihara et al., 2016), a marine associated member of the Flavobacteriaceae family. Phage infecting *Cellulophage* have previously been shown to be enriched in obese patients with type 2 diabetes mellitus when compared to lean controls (K. Yang et al., 2021). DeePhage was then used to predict the lifestyles of our vOTUs. 415 (60.7%) vOTUs were predicted to be temperate. When we factor in the relative abundances of these vOTUs, temperate

lifestyle accounts for only 21.1% across our samples.

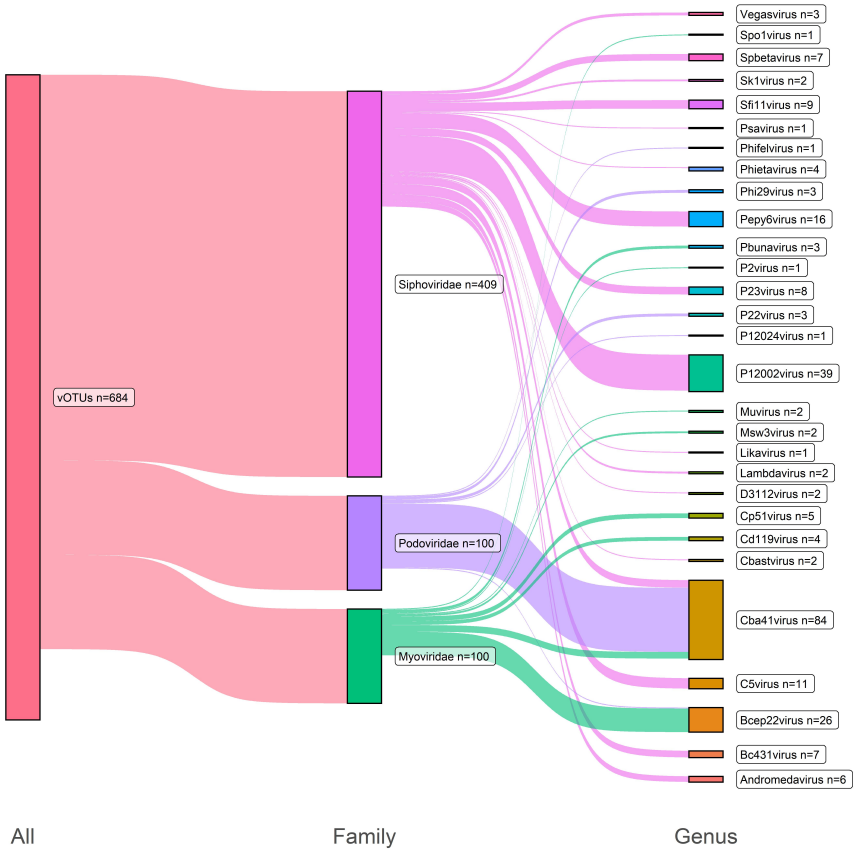


Figure 4.3: **Family and genus level taxonomy of vOTUs.** Family and genus level taxonomy of vOTUs were assigned by VPF-Class. We retained the classification with the highest membership score for each vOTU as long as it was $\geq$0.3 and its confidence score was $\geq$0.8.

We then analysed the functional profiles of our vOTUs with pharokka. A median of 69 (Q1 = 41; Q3 = 97) CDS were found across our vOTUs, with the vast majority having no known function (81.0%). 212 (31.0%) vOTUs contained at least one tRNA. No VF genes were found in any of vOTUs, whilst two beta-lactamase

ARGs (*cfxA2* and *cfxA5*), encoded by two separate vOTUs were identified.

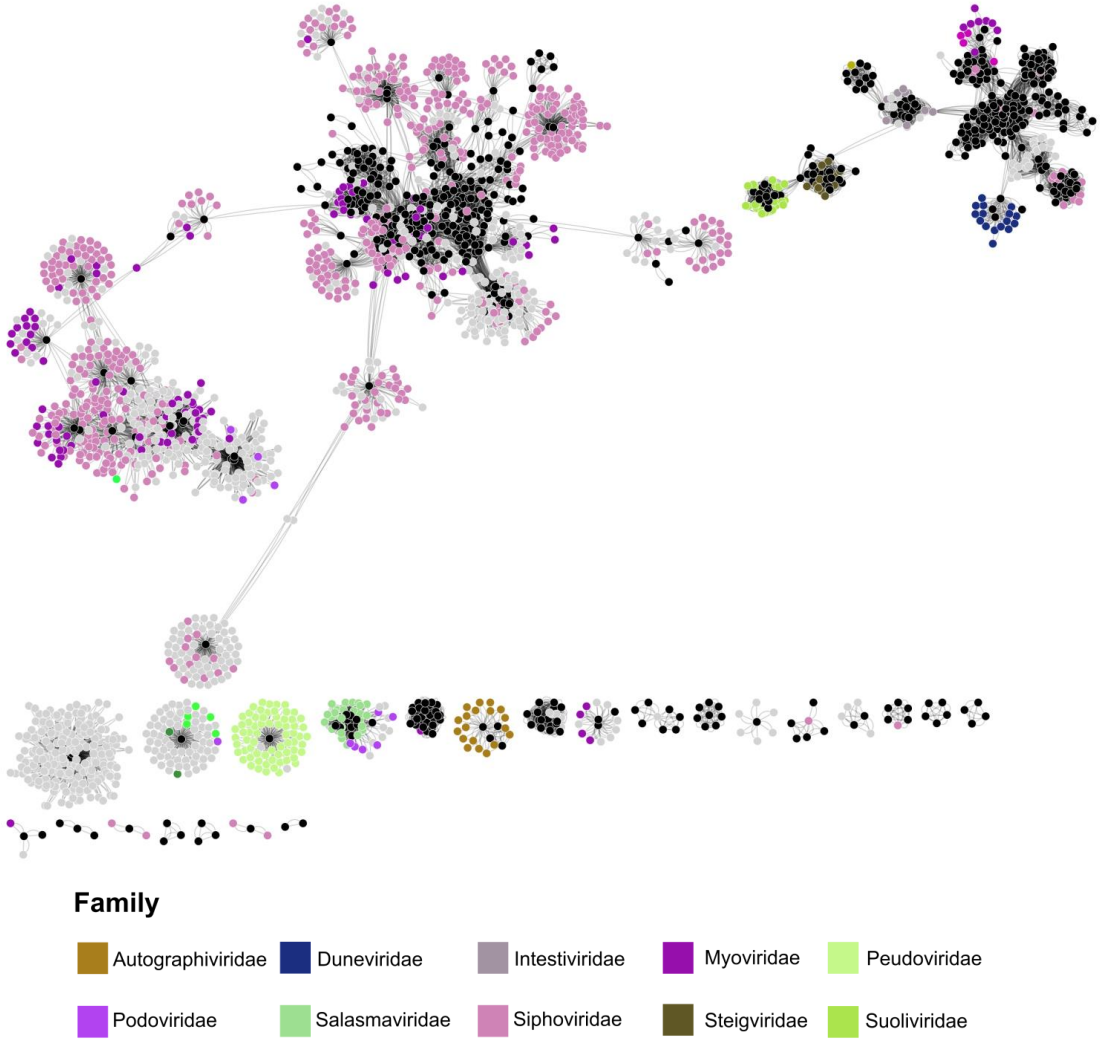


Figure 4.4: **Network based taxonomic classification.** vConTACT2 network analysis of all viral operational taxonomic units (vOTUs) from this study and a database of phage genomes from INPHARED. Coding sequences in each vOTU were predicted with phanotate. vOTUs were then clustered with reference genomes before visualisation with Cytoscape. vOTUs nodes are coloured black. Reference nodes are coloured by family. The ten most common families are displayed in the legend. Only nodes with a connecting edge to a vOTU node were retained in the network.

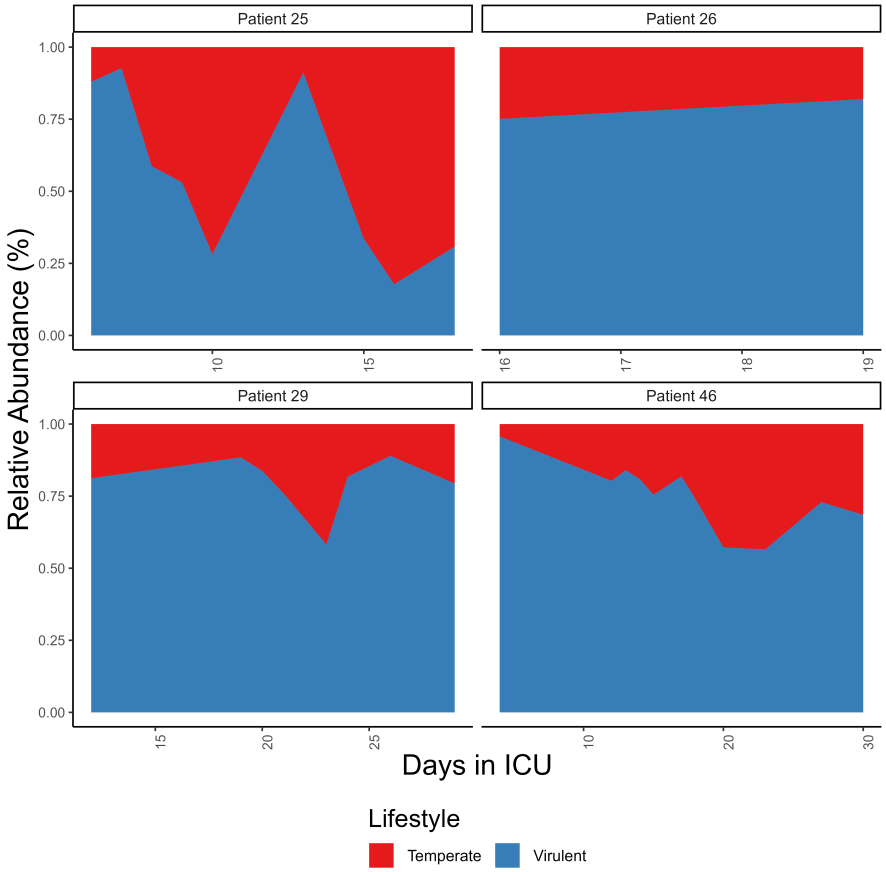


Figure 4.5: **Relative abundance of vOTUs in each patient by lifestyle.** Lifestyle for vOTUs were predicted using DeePhage. Each subplot represents a single patient across their stay in the ICU. Total relative abundance of vOTUs was normalised to 100% per sample.

4.4.8 Exploring the Genomic Context of ARG Encoding vOTUs

cfxA2 encoding vOTU

cfxA2 is encoded by the vOTU B1.5_contig_500, which is a 25.6 Kb sequence containing 41 CDS. Excluding *cfxA2*, only annotated 3 other CDS were not identified as hypothetical proteins; an excisionase (phrog_4691), an integrase (phrog_1), and a DnaB-like replicative helicase (phrog_19). No phage termini were identified in this vOTU. This vOTU was originally identified by MASH alignment to GPD. Alignment to the NCBI nr/nt with the megablast algorithm produces a hit to *Parabacteroides distasonis* strain BFG-592 chromosome (Accession: CP103256.1) with a query coverage (QC) of 98% and percent identity (PI) of 99.76%. To identify the termini of this vOTU, and investigate whether this phage has been contributing to HGT we aligned it against our set of metagenomic contigs with BLASTN. The top alignment was to a 81.0 Kb contig (NODE_190_length_80962_cov_164.008232) with a QC of 100% and PI of 99.8% (Figure 4.6A). Megablast of this metagenomic contig also had a significant alignment to the *P. distasonis* strain BFG-592 chromosome (QC = 99%, PI = 99.8%). No phage boundaries were identified in this contig either.

As *cfxA* genes in *Bacteroides* species have been shown to be associated with the mobilisable transposon Tn*4555*, we aligned B1.5_contig_500 to said transposon (Accession: U75371.3). This produced a near-perfect alignment (QC = 84%, PI = 99.69%), except for a 1,907 bp section from *Tn4555* (between 9,227 bp - 11,168 bp) being missing in B1.5_contig_500 (Figure 4.6A). When we

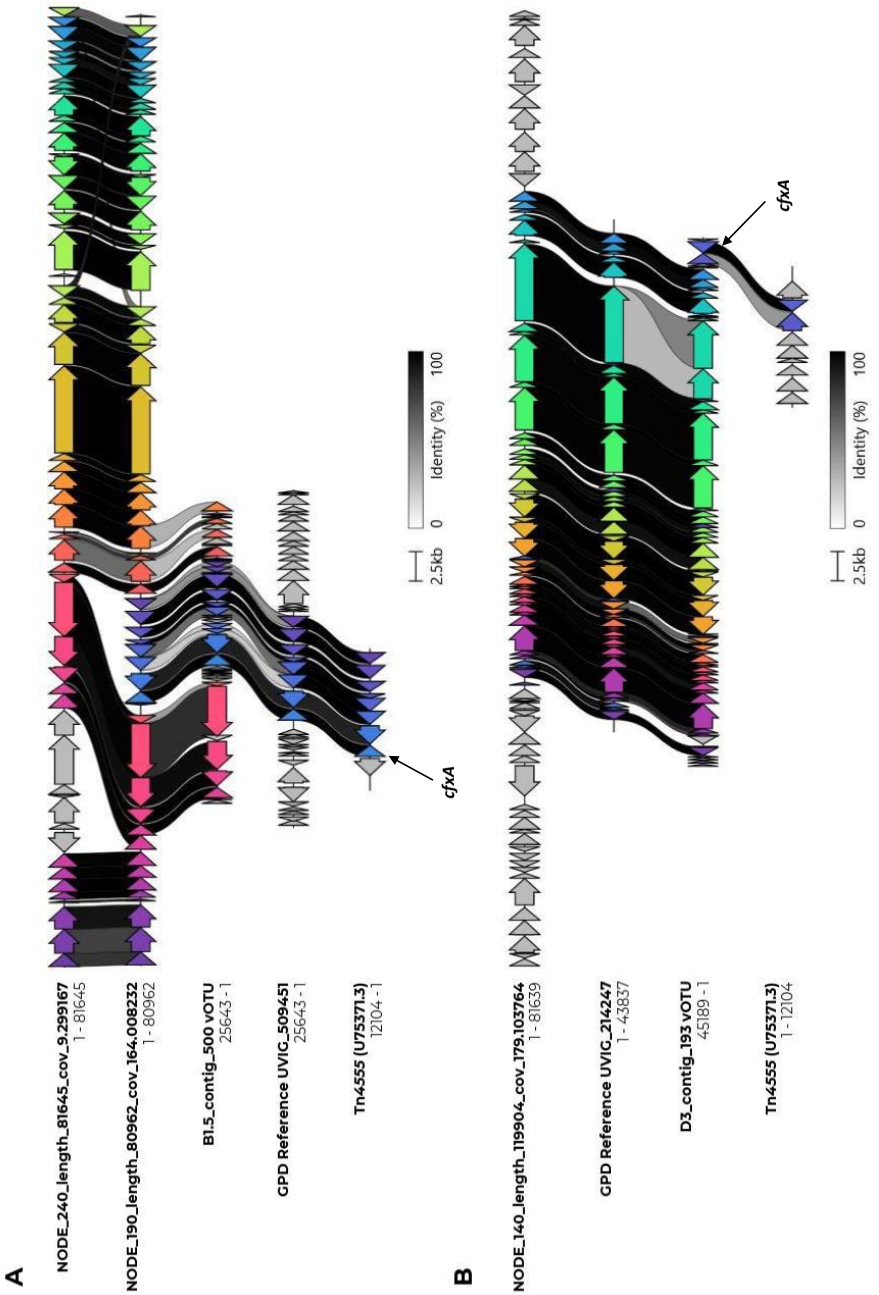


Figure 4.6: (Caption next page.)

Figure 4.6: **Global alignment of *cfxa* encoding vOTUs with metagenomic contigs and Tn*4555* transposon.** vOTUs were annotated with pharokka, and metagenomic contigs were annotated with prokka. Top hits to our metagenomic assembly set were identified with BLASTn. Sequences were then globally aligned with clinker. Open reading frames (ORFs) are colour coded by homology across sequences. **A.** Global alignment of vOTU B1.5_contig_500 to its top two megablast hits as well as the transposon Tn*4555*. Sequences of B1.5_contig_500 and Tn*4555* are reversed. **B.** Global alignment of vOTU D3_contig_193 to its top megablast hit, as well as the transposon Tn*4555*. The sequence of D3_contig_193 is reversed. The sequence of NODE_140_length_119904_cov_179.103764 is cut off at 81,639 bp.

search for insertion sequences (IS) in Tn*4555* we get a perfect alignment to IS*Bf8*, a 1,663 bp sequence from the IS4 family, between 9,260 bp and 10,922 bp. Towards the gamma terminus of the transposon, there is a further 226 bp section that is missing in our vOTU compared to Tn*4555*. The absence of IS*Bf8* suggests that the transposon that inserted itself into B1.5_contig_500 is an ancestral form of Tn*4555*. By taking the alpha and gamma terminus features that were described by Tribble et al., 1997, we were able to identify the boundaries of our Tn*4555*-like transposon to be 4,701 bp and 14,907 bp, resulting in 10,207 bp molecule. We also aligned the GPD reference, and the only homology shared between this vOTU and the reference was the Tn*4555* region (Figure 4.6A).

We also investigated another metagenomic contig, NODE_240_length_81645_cov_9.299167, which had 57% QC and 97.6% PI to B1.5_contig_500. This contig was very similar to NODE_190_length_80962_cov_164.008232 (84% QC; 98.4% PI), except it was

missing the Tn*4555*-like transposon and instead had a 6,108 bp section inserted 8,652 bp upstream (Figure 4.6A). This section contains 7 CDS, all of which are annotated as hypothetical proteins, and no IS elements were detected.

cfxa5 encoding vOTU

cfxa5 is encoded by the vOTU D3_contig_193, which is a 45.2 Kb sequence containing 62 CDS. The majority (47/62) did not have functional annotations and were marked as hypothetical proteins. Amongst the 15 annotated CDS, five encode phage specific structural or packaging proteins: a tail protein (phrog_6305), a head-tail adaptor (phrog_195), a terminase large subunit (phrog_38), a portal protein (phrog_113), and a head maturation protease (phrog_2976). A further three CDS encode proteins related to nucleotide metabolism: a reverse transcriptase (phrog_1423), a DNA binding protein (phrog_4151), and a nucleoside 2-deoxyribosyltransferase (phrog_355). Two CDS were annotated as transposon related: a DNA transposition protein (phrog_296), and a transposase (phrog_310). The remaining CDS had miscellaneous functions and included an amidase (phrog_822) for host cell wall lysis, a CI-like repressor (phrog_6201) which functions as a transcriptional regulator for lysogeny-lytic switching, an ATPase (phrog_6891), and a DNA methyltransferase (phrog_56). This vOTU was originally identified by MASH alignment to GPD. No phage termini were identified in D3_contig_193, although CheckV analysis identified this vOTU to be 93.36% complete based on average amino acid identity (AAI). Alignment of D3_contig_193 to NCBI nr/nt with megablast results in hits to a number of different genomes including:

Bacteroides zhangwenhongii strain HF-5287 (QC: 99%; PI = 99.16%; Accession: CP103098.1), *Phocaeicola vulgatus* strain BFG-191 chromosome (QC = 94%; PI = 99.44%; Accession: CP081912.1), and Caudoviricetes sp. isolate ctNK21 vMAG (QC = 94%; PI = 98.71%; Accession: BK047744.1). The latter two hits do not cover the first 2,512 bp of the D3_contig_193, and whilst the first hit does cover this section, the alignment is split between two different parts of the genome, which are approximately 460 Kb apart. When we aligned D3_contig_193 to our metagenomic contigs, we again found that all of the top alignments started at 2,512 bp. This was also true when we aligned D3_contig_193 to its GPD reference (Figure 4.6B)

As with *cfxA2* in B1.5_contig_500, we aligned D3_contig_193 to Tn*4555*. This revealed a 2,156 bp alignment of a middle section of Tn*4555* (7,047 bp – 9,211 bp) to the beginning of D3_contig_193 (350 bp - 2,514 bp) (Figure 4.6B). The first 350 bp is accounted for by a fragment of IS*942*, but the junction between Tn*4555* and the rest of the contig occurs within the middle of *mobA* and no repeat sequences could be identified nearby.

4.4.9 Calculating Absolute Abundance of vOTUs using Q33 Phage Spike-in

As we had spiked in a known quantity of Q33 phage into each sample before VLP extraction and sequencing, we were able to calculate the absolute abundance of each vOTU using their respective read mapping rate. The median alignment rate of reads to the Q33 genome was 0.67% (Q1 = 0.17%; Q3 = 3.72%), although this

percentage varied a lot throughout our samples, with the last two samples from patient 29 reaching 15.2% and 32.3% respectively. This resulted in an average of 6.86 x 10^8 vOTU particles per gram of stool (vOTU-PPG) although this number varied by patient, with patients 21 and 26 having significantly lower counts of 1.33 x 10^7 and 2.21 x 10^7 vOTU-PPG respectively. Patients 25 and 29 had very similar counts on average with 1.91 x 10^8 vOTU-PPG. On average, patient 46 had the highest count, with 6.40 x 10^8 vOTU-PPG.

When we compared the dynamics of our vOTUs as revealed by relative and absolute abundance, we noticed a number of discrepancies (Figure 4.7). Across all four of our patients (patient 21 was omitted due to only having one virome sample), the trends in the overall relative abundance of vOTUs were not replicated in their absolute abundance. In patient 25, two peaks of vOTUs are found in their absolute abundance, which is mostly masked by the relative abundance. In patient 29, the relative abundance of vOTUs suggests that the viral population was relatively stable across the patient's ICU stay, whereas the absolute abundance reveals a sharp drop in vOTUs around day 19. When looking at the dynamics of vOTU relative abundance of patient 46, we find a steady rise, with small peaks at days 20 and 30. In contrast, peaks in vOTU absolute abundance are seen between days 12 and 18, whilst the steady rise is noticeably shallower. No visible differences between relative and absolute differences are seen in patient 26, although only two samples underwent virome sequencing for this patient.

We predicted hosts for each of our vOTUs using the iPHoP framework. In total 501/684 (73.2%) vOTUs were assigned genus level host predictions. The most abundant host genus was *Parabacteroides* with an average of 5.24 x 10^7 vOTU-

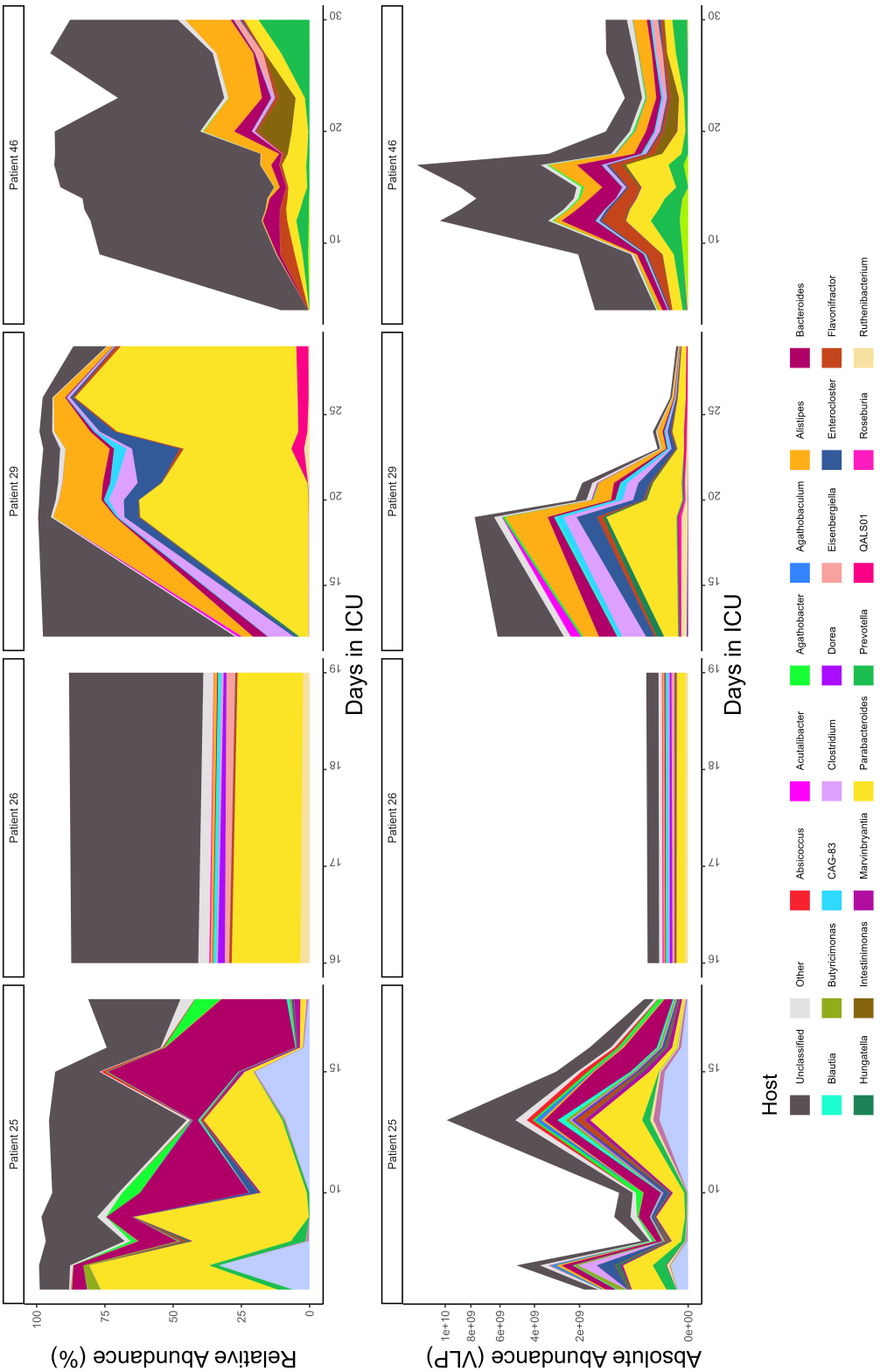


Figure 4.7: (Caption next page.)

Figure 4.7: **Absolute and relative abundance of vOTUs stratified by host taxonomy.** Patient 21 was omitted from this analysis as only one sample had its virome successfully sequenced. Host taxonomy was predicted using iPHoP, and clade specific information was removed. All quality controlled reads from each sample were mapped to vOTUs with minimap2 and relative abundances were calculated with coverM. Absolute abundance was calculated from relative abundances using the Q33 phage spike-in mapping rate. The top 20 host genera by absolute and relative abundance were retained, with all other taxa being collapsed into the "other" category. Y axis for absolute abundance has been square root transformed.

PPG followed by *Bacteroides* at 2.02 x 10^7, *Alistipes* at 1.33 x 10^7, *Prevotella* at 7.70 x 10^6, and *Flavonifractor* at 4.96 x 10^6 (Figure 4.7). When we compare this to the top five bacterial genera in our metagenomic samples by relative abundance, no overlap was found. They are as follows: *Enterococcus* (μ = 17.9%), *Akkermansia* ($\mu = 7.56\%$), *Escherichia* ($\mu = 6.87\%$), *Phocaeicola* ($\mu = 5.92\%$), and *Pseudomonas* (μ = 5.53%). *Parabacteroides* was the 7th most abundant genus as estimated by MetaPhAln (μ = 4.04%), *Bacteroides* was 8th at (μ = 3.43%), *Alistipes* was 10th (μ = 2.72%), *Prevotella* was 150th (0.00184%), and *Flavonifractor* was 44th (μ = 0.247%).

Linear regression analysis was then conducted to assess the relationship between the day the patient's stay in the ICU (independent variable) and the total absolute abundance of vOTUs or the Inverse Simpson alpha diversity of vOTUs (dependent variable) (Figure 4.8). Patients 21 and 26 were omitted from this analysis as we only had virome sequences for 1 and 2 samples respectively. Whilst we did not find any significant linear trend for vOTU absolute abundance over the duration of a patient's stay, patient 29 appears to fit a linear negative trend (slope = -3.91 x 10^7), and its adjusted p-value of 0.078 is only marginally above our α level of 0.05. When we attempted to fit linear regression lines to alpha diversity of vOTUs over a patient's stay, no significant trend could be found for patient 25 or 29, but a significant positive regression line could be fitted for patient 46 (slope = 0.859, $p.adj = 0.002$).

4.4.10 Dynamics of Dominant Host-Phage Pairs in the ICU Gut

We then attempted to correlate the dominant host genus in each patient's gut and the vOTUs that are predicted to infect that genus. Patients 21 and 26 were again omitted from this analysis. For patient 25, as the dominant taxon predicted by MetaPhlAn was the class *Clostridia*, we focused on their second most abundant taxon which was the genus *Mediterraneibacter*. *Mediterraneibacter* does not classify as a dominant genus, although it did reach a peak of 25.3% relative abundance (Figure 4.9A). A significant correlation was found between the abundances of the dominant pathogen and its corresponding

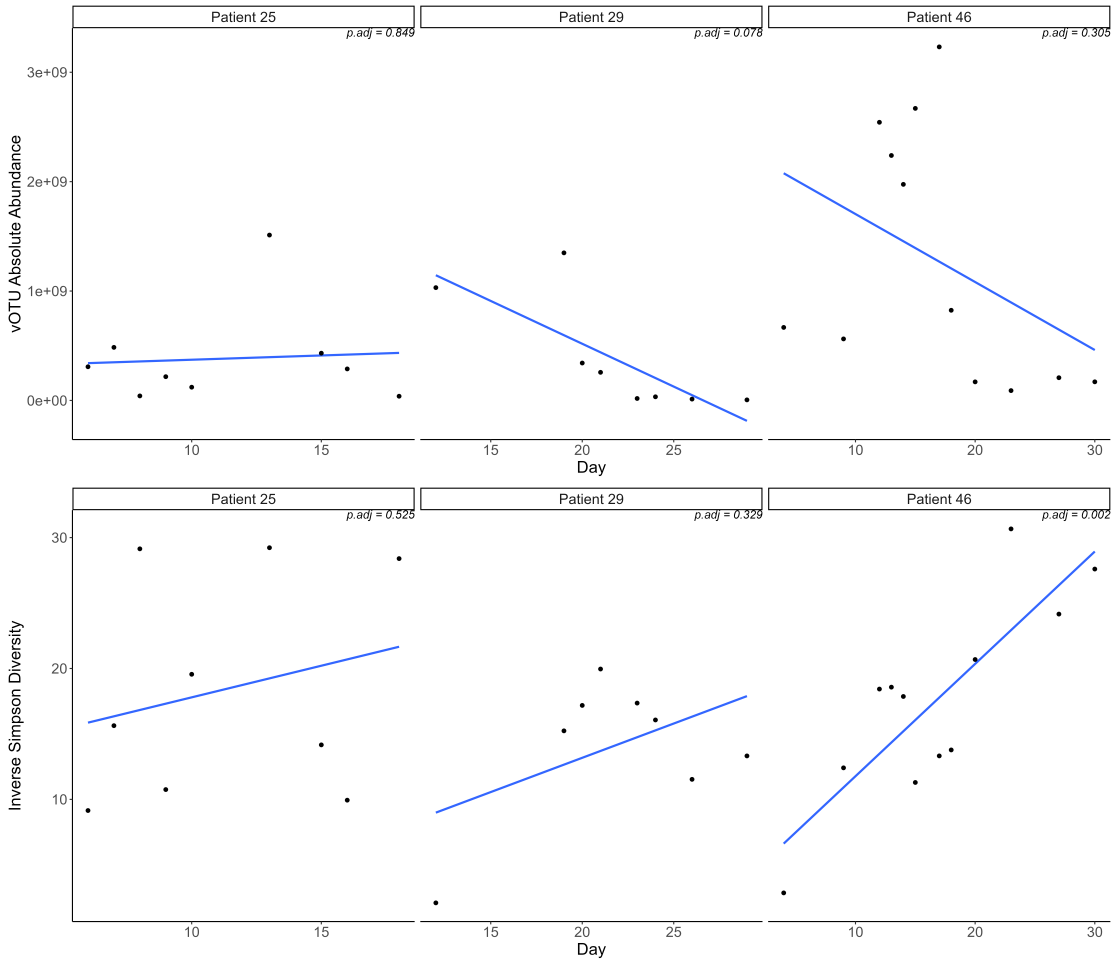


Figure 4.8: (Caption on next page.)

Figure 4.8: **Linear regression of vOTU abundance or vOTU alpha diversity and length of stay in the intensive care unit.** Patients 21 and 26 were omitted from this analysis as they had too few samples. Linear models were fit using `geom_smooth` from ggplot2. Multiple comparison corrections was applied to p-values using the Benjamini-Hochberg procedure.

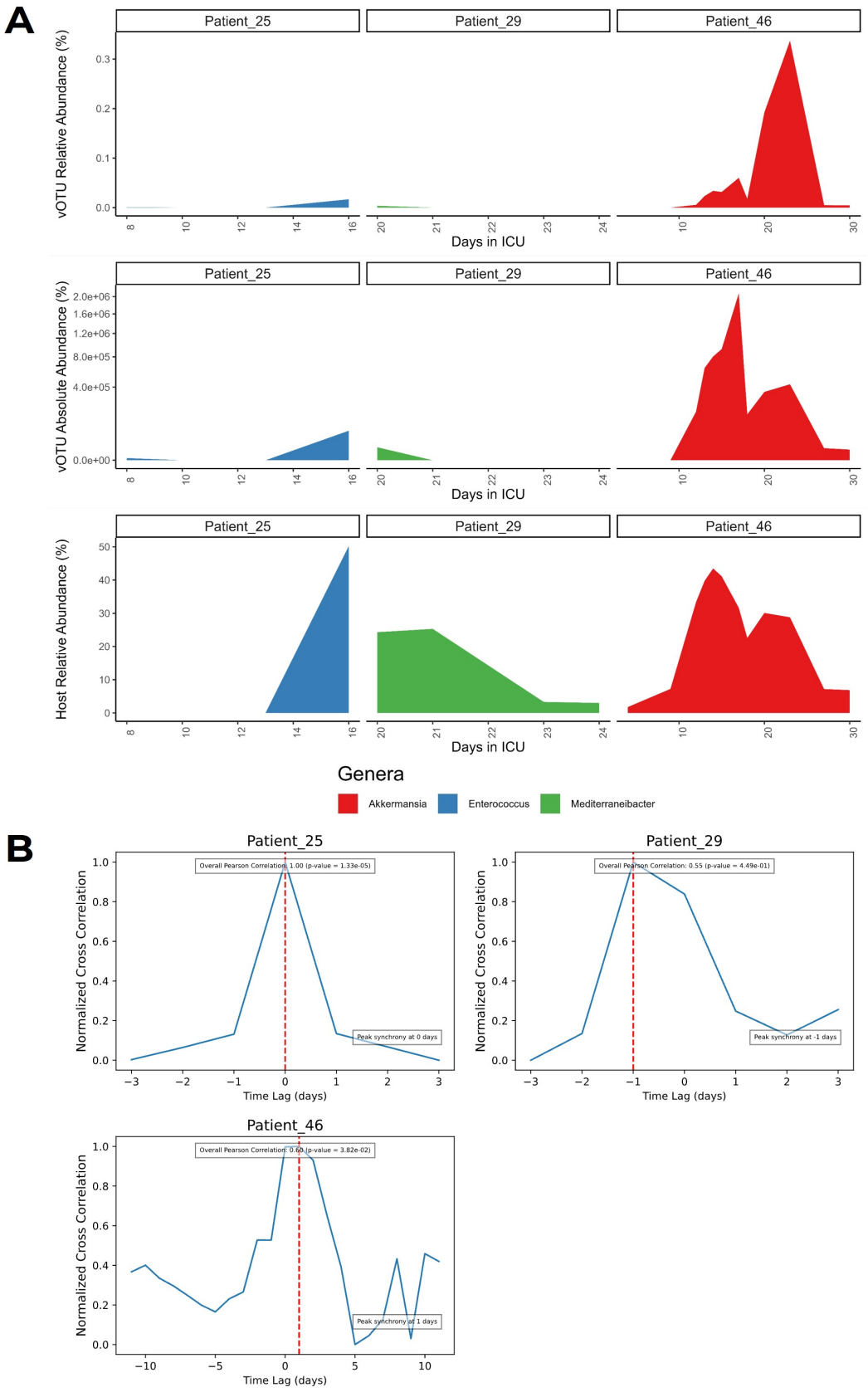


Figure 4.9: (Caption on next page.)

Figure 4.9: **Dynamics of dominant host genus and corresponding vOTUs. A.** Comparison of dominant host relative abundance, and the relative/absolute abundance of their corresponding vOTUs. Y axis of vOTU absolute abundance is square root transformed. *Mediterraneibacter* was used for patient 29 as the most dominant taxa Clostridia was not classified to genus level. **B.** Time lagged cross correlation between dominant pathogens and their phages. The relative abundance of dominant hosts and their predicted infecting vOTUs were calculated for patients 25, 29, and 46. Data for patients 21 and 26 were omitted as there were not enough samples. Pearson correlation and time-lagged cross-correlation (TLCC) coefficients were calculated between host and phage relative abundance, with the latter using days in the ICU as the lag unit. Coefficients were normalised to between 0 and 1 for visualisation. The TLCC was then plotted for each patient, with a red dashed vertical line indicating the lag at peak correlation.

vOTUs across our patients (mean Pearson's r = 0.717±0.247). To identify whether vOTU and host abundances were driving each other or synchronised, we conducted TLCC analysis for each patient (Figure 4.9B). For patient 25, the correlation was highest at a lag of zero days, indicating that the abundances were synchronised. In patient 29, the cross-correlation was highest when the time lag was set to -1 days, i.e. when the host abundance was shifted back by 1 day, the correlation was strongest. The opposite occurred for patient 46, with the strongest correlation at a time lag of 1 day, although the correlation coefficient at a lag of 0 days was virtually identical (1.00 versus 0.998 respectively).

4.5 Discussion

No previous studies on the gut virome of non-healthy individuals have utilised long-read sequencing. The work in this chapter aimed to optimise a VLP nucleic acid extraction protocol for use with the ONT sequencing platform, in order to recover more contiguous viral genomes, for high resolution analysis of host-phage dynamics and AMR carriage. This builds on the work in Chapter 3, as we are experimentally removing bacterial hosts and long-read sequencing the viral fraction, instead of having to remove bacterial sequences computationally. Importantly, the viral fraction sequenced in this chapter is made up of only actively replicating viruses, in contrast to our previous chapter, where we concluded that the majority of our vOTUs were derived from integrated prophages, some of which may be cryptic (see page 129).

To use the ONT sequencing platform on stool samples from ICU patients, we modified Shkoporov et al., 2018b's VLP nucleic acid extraction protocol. The use of ZymoBIOMICS DNA Microprep kit was the most substantial amendment that we made, as it allowed us to remove a persistent contaminant that was interfering with the library preparation process. We also decided to use a PCR-based ONT kit due to the low DNA yields from all of our extraction attempts. Whilst this kit does not produce the highly contiguous reads that are possible with ONT's non-PCR based methods (e.g. SQK-LSK114, SQK-RAD114), its expected read lengths of approximately 2 Kb are still significantly longer than those produced using the Illumina platform. These modifications coupled with some other small changes result in an efficient

protocol for long-read sequencing of VLPs when input DNA is low.

Despite the extensive modifications and optimisation attempts, we had mixed success when we applied it on stool samples from ICU patients, with over half the samples not having a sufficiently high DNA yield, despite our inclusion of a PCR amplification step. It is notable that the majority (70.3%) of these failed samples were from one patient, which suggests that either this patient had very low phage counts in their gut, or their stool contained something that interfered with our protocol. It is unclear what this could be, as metagenomic sequencing largely worked with the same samples. As we only had one virome sequence from patient 21, it was omitted from some downstream analyses. Nevertheless, we were able to sequence VLPs from samples that covered a substantial portion of 4 out of 5 of our patient's ICU stays.

We also carried out whole community metagenomic sequencing for the same stool samples, to provide bacterial host context for our virome sequences. From this, we were able to observe microbiome domination by a single species in 80% of our patients, a similar percentage to what we observed in Chapter 3. *E. faecium*, and *E. coli* which were seen dominating the gut of multiple patients in Chapter 3 were also seen here. We also identified dominating species that were not seen in Ravi et al., 2019, with *P. aeruginosa*, *E. faecalis*, and *A. muciniphilia* blooming to $\geq$30% relative abundance. Whilst the latter two species are gut commensals (Belzer & de Vos, 2012; Sava et al., 2010), *P. aeruginosa* is not thought to be so (Didelot et al., 2016), and intestinal colonisation by it is associated with subsequent lung infections (Markou & Apidianakis, 2014), gut-derived sepsis (Osmon et al., 2004) and increased ICU mortality (de Jonge

et al., 2003). *P. aeruginosa* is a problematic pathogen due to its high levels of intrinsic antibiotic resistance (López-Causapé et al., 2018), as well as its ability to evolve resistance in hospitalised patients (Wheatley et al., 2021). Although *E. faecalis* is commonly, and asymptomatically, found in the human gut microbiome, it is also a major cause of nosocomial infections, especially in critically ill patients who are immunocompromised or have a severe underlying disease(Kristich et al., 2014; Sava et al., 2010). *Enterococcus* species also have high levels of intrinsic antibiotic resistance, and acquired resistance to aminoglycosides and vancomycin have been reported for decades (Hegstad et al., 2010; Leclercq et al., 1988). *A. muciniphila* is a Gram-negative anaerobic bacterium that utilises mucin as a carbon and nitrogen source (Derrien et al., 2004). It comprises approximately 1 to 3% of the microbiome in healthy adults (Derrien et al., 2008). By degrading mucin, *A. muciniphila* produces propionate and acetate, which are used as substrates by other bacteria, and by the host (Derrien et al., 2004). It is commonly proposed as a probiotic for certain metabolic and autoimmune diseases due to its lower abundance in individuals with these diseases (Depommier et al., 2019; Jian et al., 2023; Luo et al., 2022; Rodrigues et al., 2022), and pasteurised *A. muciniphila* has been recognised as a safe food supplement by the European Union (EFSA Panel on Nutrition, Novel Foods and Food Allergens (NDA) et al., 2021). *A. muciniphila*, like most other anaerobes, has been shown to have intrinsic resistance to both ciprofloxacin and aminoglycosides (Filardi et al., 2022; Pence, 2019). This variety in dominating species suggests that both pathogens and commensal bacteria have the ability to fill the niche and bloom when the microbial diversity drops during an ICU stay. This phenomenon was also seen in Chapter 3, and has been reported by

Hildebrand et al., 2019.

As virome studies are often hampered by bacterial contamination, even when restrictive methods like CsCl density gradients are used, we employed ViromeQC, a tool that quantifies contamination in viromes by alignment to SSU rRNA, LSU rRNA, and single copy bacterial markers. It then produces an enrichment score by comparing the minimum alignment rate across the three categories to the median observed in human gut metagenomes. For our samples, the average enrichment score was extremely low, estimated to be two and a half times less enriched than a typical whole community metagenome. This score was calculated solely from the SSU rRNA alignment rate, as this was consistently lower than the LSU rRNA and single-copy bacterial marker alignment rates. The authors of ViromeQC showed that these alignment rates should correlate with each other, but this was not the case with our samples. Indeed, if the enrichment score was calculated using single-copy bacterial markers, on average our viromes were 41.6 times more enriched. This discordance between the rRNA and single copy marker abundances was noted in some datasets analysed by Zolfo et al., 2019 and by Shkoporov et al., 2018b, and was hypothesised by both papers to be a result of co-purification of ribosomes during the VLP extraction process, which were then reverse transcribed into cDNA before sequencing. Our protocol does not include a reverse transcription step, so it is unclear why atypically high levels of rRNA were found across our samples. We also assessed bacterial DNA contamination using a method described by (Shkoporov et al., 2018b), where viromes are aligned to the *cpn60* gene database. *cpn60* genes are present in virtually all bacteria as a single copy and was not one of the 31 genes used in

ViromeQC. This method estimated that the median bacterial contamination in our viromes was 16.8%. Assuming the typical gut microbiome contains 5.8% viral DNA, this constitutes a 14.4x enrichment.

Despite the considerable enrichment, it is clear that significant bacterial contamination remains in our viromes. Therefore, we used the same viral identification process that we used in Chapter 3. This resulted in 684 vOTUs which had a median length of 37.7 Kb, over double the median length of vOTUs identified in Chapter 3. The average quality of these vOTUs was also considerably higher than those in Chapter 3, with 67.7% medium quality or higher in the former, and only 18.3% in the latter. We obtained similar improvements with completeness, which was three times higher in the vOTUs identified in this chapter (30.8% vs 92.3%). Host contamination within these vOTUs was also lower (1.9% vs 10.6%). The improvement in these metrics demonstrates the benefits of ONT long-read sequencing versus Illumina next generation sequencing. When we clustered our vOTUs in a gene-sharing network, we found that a higher percentage of vOTUs clustered with reference genome at family or genus level than in Chapter 3 (70.0% vs 31% respectively). This lower proportion of putative novel taxa is likely due to either, (1) an overestimation of novel genera in Chapter 3 due to the shorter vOTU genomes being more difficult to cluster, (2) or the actively replicating phages that were captured in virome sequencing are on average better characterised in reference databases, or a combination of the two. The proportion of vOTUs predicted to have a virulent lifestyle was 3.4 times higher than in Chapter 3 at 78.9%. This is in line with current estimates of temperate phages making up 20-50% of free phages in the

gut (Reyes et al., 2010; Sausset et al., 2020).

Only two of our vOTUs encoded ARGs, which is in line with the current understanding that ARGs are rare in phage genomes (Enault et al., 2017; Pfeifer et al., 2022). These vOTUs encoded the beta-lactamases *cfxA2* and *cfxA5* which confer resistance to the cephamycin group of beta-lactam antibiotics. Our use of ONT long-read sequencing allowed us to analyse the sequence of these vOTUs at a high resolution. B1.5_contig_500 was the first vOTU we investigated, and it contained a *cfxa2* gene. The only other non-hypothetical proteins we found were an excisionase, DnaB-like replicative helicase, and an integrase. The sequence has very high homology to *P. distasonis*, which is known to carry *cfxa2* (Ezeji et al., 2021). As we could not identify phage termini in this vOTU, we aligned it with our metagenomic assembly set. This would also allow us to identify if the phage was present in multiple host contexts, which would suggest that HGT had occurred. A complete alignment to an 81.0 Kb metagenomic contig was identified. This contig was shown to be derived from the same *P. disastonis* chromosome, and again, no phage termini or reference was found by CheckV. We also identified that the *cfxa2* gene in B1.5_contig_500 was encoded within a Tn*4555*-like transposon. This Tn*4555*-like transposon is likely an ancestral version of Tn*4555*, as it does not contain the IS*Bf8* which is present in the reference sequence. There is also a further 226 bp section of IS*Bf8* that is missing, towards the gamma terminus. It is unclear why this extra section is missing, with no homology to any transposable elements. Using the alpha and gamma terminus features that were described in the original paper that identified Tn*4555*, we were able to identify the boundaries of our Tn*4555*-like transposon. Other than the 81.0 Kb metagenomic contig, one other

contig had significant homology to B1.5_contig_500, although the Tn*4555*-like transposon section was missing. We also aligned B1.5_contig_500 with the GPD reference that was used to identify it as a vOTU. This revealed that the Tn*4555*-like transposon region was the only part of the vOTU that had homology to the reference. From this set of analyses, we cannot be certain that B1.5_contig_500 is a phage derived sequence as we could not identify any phage termini, the four non-hypothetical annotated CDS are not exclusive to phage elements, and its basis for vOTU classification was a transposon region of a GPD reference. Nevertheless, we were able to identify a novel Tn*4555*-like transposon that encodes a beta-lactamase.

With the second ARG encoding vOTU D3_contig_193, a number of phage specific structural or packaging proteins were identified, including a terminase large subunit, which is commonly used as a phylogenetic marker for construction of *Caudoviricetes* phage trees due to its high conservation and prevalence among phage genomes. No phage termini were found in this vOTU, although CheckV analysis identified that the vOTU was 93.3% complete based on AAI to a reference. D3_contig_193 had high sequence homology to a number of different NCBI sequences, including a *Bacteroides* species, and a *Caudoviricetes* vMAG. However, these alignments all had a gap at 2,512 bp on the vOTU, with the alignments coming from entirely different sections of the genome. This was also seen when we aligned D3_contig_193 to our metagenomic sequences and its GPD reference. This 2,512 bp sequence was identified to be identical to a middle section of Tn*4555*, except for a 350 bp region, which was a fragment of IS*942*. This middle section contains the *cfxa5* gene as well as part of *mobA*. No repeat

sequences were identified nearby. As the junction sequence between this 2,512 bp sequence and the rest of the vOTU could not be found in our metagenomic contig set, or in NCBI's nr/nt database, we believe that this vOTU may be the result of a misassembly, although it is possible that it is a genuine recombination event that was not captured by the metagenomic sequencing or assembly process.

We did not find any VFs across our vOTUs, in contrast to the relatively high number we identified in Chapter 3. From the literature, it is not clear how widespread VFs are in gut-derived phage genomes, although many papers have described the importance of prophage-encoded VFs (Fortier & Sekulovic, 2013; Mathieu et al., 2020; Shuwen & Kefeng, 2022; X. Zhang et al., 2000).

During our VLP extraction protocol, we spiked-in a known quantity of Q33 phage into each of our stool samples, which allowed us to calculate the absolute abundance of our vOTUs. The importance of this was exemplified when we plotted the absolute abundance of our vOTUs against their relative abundance. The conclusions that would be drawn from the relative abundance of vOTUs across each patient's ICU stay were entirely different for all four patients. For example, in patient 29, the relative abundance of vOTUs is stable across the patient's stay, but is characterised by an increase in *Parabacteroides* infecting phages over time, with a peak around day 20 and day 30. Conversely, the absolute abundance of vOTUs sharply drops off around day 19 of patient 29's ICU stay, and only one peak in *Parabacteroides* phages occurs. Whilst relative abundances are useful for comparing the abundance of taxa in relation to each other, it is well known that absolute abundances allow researchers to properly investigate population dynamics (Vandeputte et al., 2017; Yao et al., 2022).

The most abundant host predicted for our vOTUs by absolute abundance - *Parabacteroides*, *Bacteroides*, *Alistipes*, *Prevotella*, and *Flavonifractor* - are all common gut commensals. Notably, these predicted hosts did not overlap with the most abundant genera in our metagenomic reads, such as *Enteroccous* and *Pseudomonas* which were dominating the microbiome of our ICU patients. This contrasts what we observed in Chapter 3, where the most abundant vOTU hosts, were the bacteria that were most abundant in the metagenomic data. We believe that this is in line with our hypothesis that the majority of vOTUs that we identified in the previous chapter are integrated prophages. This would result in a high abundance of vOTUs when their host is also sequenced. In this chapter, as we are only sequencing free phage particles, this bias is eliminated.

For the three patients that had more than two VLP sequencing samples, we conducted linear regression analysis for both total vOTU absolute abundance, and alpha diversity of vOTUs, over each patient's ICU stay. No significant linear model could be fitted for total vOTU absolute abundance for any of the patients, although the adjusted p-value for patient 29 was close to our α level. For alpha diversity, no linear trend was found for patient 25 or 29, but a strong positive trend was found for patient 46, indicating that alpha diversity of vOTUs increased over their ICU stay.

When we correlated the most dominant host genera with their corresponding vOTUs, a significant correlation was found for all patients, albeit not as strong as the correlations found in Chapter3. TLCC analysis resulted in inconclusive results, with one patient having the highest cross-correlation at a lag of zero days, another patient at negative one days, and the last patient having

equally high correlation at 0 and 1 days lag. Although there is a correlation between these vOTUs and their hosts, their absolute abundance is low, especially when compared to vOTUs infecting gut commensals such as *Parabacteroides*. Together, these results generally align with the "piggyback-the-winner" model for gut phage dynamics (Silveira & Rohwer, 2016), as the abundances of free vOTUs infecting dominant microbes are low, suggesting that phages are taking advantage of the proliferation of their hosts by replicating together, as opposed to entering the lytic cycle. However, with the exception of patient 26, the absolute abundance of vOTU population is far from stable, implying that factors other than host abundance are at play.

4.5.1 Limitations

The results in this chapter show the benefits of utilising long-read sequencing to profile the gut of critically ill patients. However, due to low DNA yield, we had to use a PCR based library preparation. PCR based methods in Illumina sequencing have been shown to result in small biases in the resulting virome compared to PCR-free methods, especially in regards to the recovery of low abundance viral genomes (Hsieh et al., 2021). This bias is likely to also occur in ONT PCR based kits. The output of the Rapid PCR Barcoding kit we used is also significantly shorter than ONT's PCR-free kits, so the number of complete and high-quality vOTUs was not as high as expected. Nevertheless, our reads were still significantly longer than those produced by Illumina next generation sequencing, and resulted in vOTUs that were twice as long as those in Chapter 3. Preferably, in the future,

we would collect much larger quantities of stool in order to scale up our protocol to produce a higher DNA yield, although this is limited by the amount of stool produced by patients at each time point. The modified VLP nucleic acid extraction protocol that we used may have also introduced biases into the final virome. For example, a chloroform extraction was used to remove excess PEG from our filtrates, and enveloped phages and non-enveloped phages belonging to families *Inoviridae*, *Corticoviridae*, and *Plasmaviridae*, have previously been shown to be susceptible to chloroform (Kleiner et al., 2015; Thurber et al., 2009) [see Chapter 1, section 1.5, 35]. However, when we attempted to remove this step from the protocol, almost no DNA could be recovered. Alternatively, if time allowed we could have attempted to remove the excess PEG using another method such as dialysis (Castro-Mejía et al., 2015), which would not eliminate chloroform sensitive phages. Other modifications that we tried to make such as the addition of a CTAB extraction step may have improved or even had desirable results if we had further optimised the protocols or identified the root cause of the issues that we faced - e.g. mass spectrometry to identify inhibitor molecules. However, time constraints in this study meant that further optimisation of an already complex protocol was untenable.

Although Q33 spike-ins were validated by Shkoporov et al., 2018b, we did modify the protocol substantially, so it is possible that the changes we made resulted in a non-quantitative recovery of the spike-in. If we had more time, we could have carried out the same analysis as Shkoporov et al., 2018b, and spiked in varying amounts of Q33 to different samples, and validated that the read alignment rate was a linear function of spike-in concentration.

As metagenomic DNA extraction failed on six samples that we virome

sequenced, we decided to not polish our long-reads with our metagenomic reads. This meant that our average read quality after filtering was 13.7, which is much lower than that of our metagenomic reads. This in turn may have impacted our ability to annotate and align our vOTUs. If we had more time, we would have ideally troubleshooted why these samples did not produce sufficient high quality DNA, and repeated the sequencing. We could have also attempted to polish the dereplicated vOTUs with all of our metagenomic reads and compared the results to see if the read quality improved significantly, and whether any micro diversity was eliminated.

In comparison to Ravi et al., 2019, our study had a much smaller cohort of patients. In addition to this, we were also unable to include every sample that was collected over each patient's stay in the ICU. This mainly resulted from financial and time constraints, as we were virome and metagenomic sequencing each sample. Due to our smaller sample size, we may have missed any phage-mediated HGT transfer events in the ICU as they are likely rare events. We also had inconclusive results from our linear regression and TLCC analysis, which may have been resolved by using a larger cohort.

One of the metrics we used to identify vOTUs was when the MASH distance to GPD was below 0.05, equivalent to an ANI of 95%. MASH distances are used to measure the overall similarity of sequences and are often used to find related sequences. When we locally aligned B1_contig_500, a vOTU identified using MASH distances, to its GPD reference we found the only significant homologous region was a transposon sequence. This is an issue in cases where the GPD reference contains elements that can appear in viral and non-viral contexts,

such as with MGEs. It therefore may have been more suitable to use a local alignment method such as BLAST, even though it would be significantly slower.

Inadequacies in viral databases are still a large concern in phage research and impacted our research when we predicted hosts for our vOTUs. A large proportion of our vOTUs could not be predicted as a host at the genus level, and so they were grouped as unclassified. It is possible that trends in vOTUs infecting certain hosts were not identified due to this.

Finally, in this study, we compared the difference between using ONT virome sequencing and extracting viral sequences from whole community metagenomes in Chapter 3. Ideally, we would have liked to complement our study by identifying vOTUs within the metagenomic sequencing data generated in this chapter. This would have allowed for a direct comparison between the methods, enabling us to determine if the vOTUs detected in the virome were also present in the metagenomic data. Nonetheless, our comparative analysis of the two methods across these datasets still yields valuable insights into their respective benefits and limitations.

4.6 Conclusion

Our results provide important insights into the host-phage dynamics in the critically ill gut. We produced a modified version of a preexisting VLP nucleic acid extraction protocol, specifically designed for long-read sequencing from low abundance stool samples. This protocol was utilised successfully for samples

from four patients and produced a virome profile of their gut during their stay in the ICU. The resulting vOTUs were significantly more contiguous and higher quality than those produced in the previous results chapter. We also carried out metagenomic sequencing for the same samples to provide a view of the wider gut microbiome community. From this, we were able to identify microbial domination in four out of five of our patients. These dominating microbial taxa did not overlap with the most abundant hosts we predicted for our vOTUs. Despite their low abundance, the vOTUs infecting the dominant taxa still correlated with their host's abundance. Finally, we identified two vOTUs which encoded *cfxA* beta-lactamases. Within the first, we discovered a novel Tn*4555*-like transposon which included the *cfxa* gene, but no contextual information supported that the sequence was of phage origin. The second vOTU likely is the result of a misassembly, resulting in a section of Tn*4555*, including the *cfxA* gene, being connected to our vOTU. Overall, it is clear that ARG encoding phages are rare, and no evidence of transduction of ARGs was observed.

CHAPTER 5

General Discussion

Phages are important members of the gut microbiome and play a crucial role in shaping their bacterial host community via predation, lysogeny, and transduction. Previous studies have identified that certain alterations in gut phage composition are associated with diverse health conditions, although very little is known about how specific phage-host dynamics play a part in this. Metagenomics and viromics have allowed for the direct identification and analysis of gut phages, regardless of their host's cultivability. However, this process remains complex with significant

challenges the scientific community has yet to definitively overcome. This thesis therefore aimed to advance our understanding of these areas by leveraging novel bioinformatic and sequencing technologies.

The work in Chapter 2 compared the performance of ten commonly used metagenomic virus/phage identification tools and provided valuable metrics and insights for other researchers to use and build on. By using a mixture of artificial datasets and mock communities we were able to estimate the recall, precision, and biases of these tools. We were also able to optimise the performance of each tool by adjusting their scoring thresholds to achieve the highest F1-score possible. Whilst these optimal thresholds may not necessarily be ideal for all other datasets, they consistently outperformed the default parameters, and we believe they can be utilised as a starting point for further application of these tools.

When we benchmarked the tools using positive and negative datasets derived from NCBI RefSeq, most tools performed well by detecting the majority of the phage sequences, whilst keeping false positive predictions low. On average, tools that utilise k-mer-based approaches demonstrated superior performance compared to tools relying on reference similarity and homology. This trend was primarily influenced by two homology-based tools, Seeker and ViralVerify, which exhibited relatively low F1-scores, contributing to an overall lower average. Notably, machine learning-based tools PPR Meta, VIBRANT, and VirSorter2 - where the latter two are homology-based - achieved the highest F1-scores in the benchmark. We also randomly shuffled the positive dataset as a negative control and retested each tool. All of the machine learning tools identified a significant proportion of the sequences as viral with the exception of VirSorter2. Three of

the other four tools returned zero false positives and Kraken2 only returned three. This highlights that whilst these machine/deep learning algorithms have the capability to detect novel phages, their heuristics may be prone to varying performance when exposed to novel datasets with features that differ from those in our training sets. The tools were also benchmarked on real metagenomic data, where they performed significantly worse than on the artificial contig set, with the exception of Kraken2. The drop in performance is likely due to the comparatively short phage sequences that were assembled from the metagenomic reads compared to those we created from RefSeq genomes. Overall, Kraken2 and DeepVirFinder had the best performance in this benchmark, with the former having higher precision at the expense of lower recall, and the latter *vice versa*. We also used alpha and beta diversity, as well as ordination to gauge any potential biases each tool had on the resulting composition of its predicted viral population. With the exception of Seeker, each tool's output was within 10% of the original population's alpha diversity. The same phenomenon was seen with beta diversity, with only Seeker's output having significant dissimilarity from the other tools. However, our benchmarking of these tools was not fully comprehensive, as we did not assess the performance of combining multiple tools, which has been attempted by the authors of What the Phage (Marquet et al., 2020), nor did we take into consideration each tools' ability to identify viral sequences belonging to different taxonomic groups. Whilst no taxonomic bias was found with any of the tools in our RefSeq benchmark, with the exception of Seeker classifying fewer Ackermannviridae fragments, RefSeq only contains exemplar phage species as defined by the ICTV and thus does not represent the enormous diversity of phage genomes. Furthermore, 95.8% of phages from RefSeq

belong to class *Caudoviricetes* (Shang et al., 2021), and Schackart et al., 2023 found that all tools they benchmarked had lower recall on non-*Caudoviricetes* sequences, with homology-based tools having especially poor performance.

Including Schackart et al., 2023, a number of other metagenomic phage identification tool benchmarking studies have been published since the completion of this chapter. L.-Y. Wu et al., 2023 benchmarked ten virus identification tools on paired real-world metagenomic datasets from three different biomes: soil, human gut, and Antarctic seawater. All tools included in L.-Y. Wu et al., 2023 were also benchmarked in our study with the exception of Sourmash (Pierce et al., 2019) and Phigaro. The results found by L.-Y. Wu et al., 2023 were consistent with findings: tool performance could be significantly improved by tweaking detection thresholds and parameters, true positive and false positive rates highly varied between tools, and PPR-Meta, DeepVirFinder, VirSorter2, and VIBRANT had the best performance. Pratama et al., 2021 used both simulated viral sequences from RefSeq genomes and mock communities to benchmark five viral identification tools: MARVEL, DeepVirFinder, MetaPhinder, VIBRANT, and VirSorter. The simulated data had varying sequence lengths ranging from 500 bp - 20 Kb, and the authors found that viral identification was improved in all tools as sequence length increased. They also found that the performance of homology based tools dropped off when sequences were $<$3 Kb. K-mer based tools and BLAST-based tool MetaPhinder were good at identifying viruses from shorter viral genome fragments, although they were more prone to false positives. Hegarty et al., 2023 investigated combining different tools using a set of rules for each method. The tools they included were

DeepVirFinder, VIBRANT, VirSorter, and VirSorter2, whilst Kaiju (Menzel et al., 2016) and CheckV were used to tune the outputs. In total, 63 distinct rulesets were created using combinations of these tools. The authors found that VirSorter2 and their tuning removal rules were included in the top five performing rulesets, whilst their tuning addition rules were included in the top ten performing rulesets. The tuning removal rules contained four subrules for removing a positive prediction based on CheckV results and the tuning addition contained four subrules for including predictions based on Kaiju, CheckV, and VirSorter2 outputs. Importantly, the authors recommended that the tools should not be used in isolation based on the poor performance of the single tool rulesets.

In addition to recent benchmarking studies, new viral identification tools continue to be developed and adopted by researchers. PhaMer, not to be confused with PhaMers (Deaton et al., 2017), uses an adapted Transformer model for context-aware embedding which allows the model to learn the importance of protein clusters (Shang et al., 2022). The authors benchmarked PhaMer on five different datasets, including a RefSeq based dataset, simulated metagenomes, and mock. PhaMer outperformed the other five viral identification tools they included in all five datasets. INHERIT uses a pre-trained BERT based Transformer model to identify phage sequences (Bai et al., 2022). The authors opted to use the RefSeq test set that we designed in Chapter 2, except instead of fragmenting the genomes randomly to lengths between 1-15 Kb, the authors decided to split them into five length intervals. The shortest interval for phage genomes was $\leq$42 Kb. Considering that Pratama et al., 2021 found that tools were generally able to identify all true positive viral sequences $\geq$10 Kb, we find

Bai et al., 2022's decision to split the sequences in this way to be puzzling. Nevertheless, the authors found that INHERIT outperformed the four other tools on this dataset. GeNomad is a classification and annotation tool for identifying plasmid and viral sequences using both information from over 200,000 marker proteins and a deep neural network (Camargo et al., 2023). The author's compared geNomad's performance with seven other viral identification tools, including VirSorter2, DeepVirfinder, and PPR-Meta, geNomad outperformed all other tools at all sequence length ranges. GeNomad can also be used to identify prophages and demarcate their boundaries. Along with the majority of tools we benchmarked in Chapter 2, the above tools all claim to have superior performance compared to other state-of-the-art tools in their respective publications. This highlights the importance of third-party benchmarking studies to provide researchers with an unbiased view of each tool's capability. As more tools are developed and databases improve in size and quality, further benchmarking studies will be vital.

Following the benchmarking of metagenomic phage identification tools, in Chapter 3, we used these insights to identify vOTUs from a metagenomic sequencing set obtained by a previous study that investigated the gut microbiome of critically ill patients. This study identified episodes of gut domination by a single species but did not define what their criteria were. Therefore, we defined it as when the relative abundance of at least one species is $\geq$30%. We also reanalysed the taxonomic classification of the metagenomic reads, as the method that was used by the original authors had become outdated. This was showcased by the identification of a bloom of *Candidatus*

Borkfalkia ceftriaxoniphila in patient 35, which had been completely missed in the original study due to the recency of its discovery.

Out of the tools we benchmarked in Chapter 2, we settled on using Kraken2 to identify vOTUs, due to its high precision. To supplement this prediction we also used a prophage detection tool Phigaro, and MASH distances to GPD. The majority of the vOTUs we identified were genome fragments according to MIUVIG guidelines. Over half of these did not cluster with any reference genomes in our gene sharing network analysis, suggesting that these were novel genera. We found double the number of VFs than ARGs, although these were concentrated in fewer vOTUs. ARGs included class A beta-lactamases such as *cfxA2* and *cfxA4*, which we believe to be the first identification within a phage genome. VFs included the *E. coli* exotoxin haemolysin gene *hlyE*. The abundance of vOTUs by their predicted hosts generally paralleled the abundance of the corresponding host genera in each sample, and when we focused on the dominant genus in each patient, we found a strong, near-perfect correlation between its abundance and the vOTUs that were predicted to infect it. TLCC analysis suggested that the abundance of the dominant genus and its vOTUs were synchronised instead of one driving the other. We hypothesise that this synchronicity is due to the sequencing of integrated prophages and their respective hosts, as the vast majority of our vOTUs were predicted to be temperate. From the prophages identified by Phigaro, we found that approximately a fifth of them are likely to be actively replicating based on their differential coverage. None of these active prophage vOTUs encoded any VFs or ARGs. There were a number of limitations to this study: The vOTUs we

identified were considerably shorter than the length of high-quality phage genomes in the literature, due to the limitations of metagenomic sequencing on the Illumina platform. Whilst we were able to identify VFs and ARGs encoded by a small number of vOTUs, we were not able to confirm whether these were involved in any HGT event, or even if the vOTUs themselves were capable of replicating.

In Chapter 4, we built on the previous chapter by experimentally removing bacterial hosts from our stool samples in order to long-read sequence the free viral fraction alone. This viral fraction only included particles that are actively replicating, in contrast to Chapter 3, and the long-read sequencing would allow us to get higher quality and more complete vOTUs. To use the ONT sequencing platform on viral fractions of stool samples, we modified Shkoporov et al., 2018b's VLP nucleic acid extraction protocol and coupled it with ONT's Rapid PCR Barcoding Kit. This protocol worked for the majority of our samples, except for samples from patient 21, where all but one sample failed. In parallel, we also carried out whole community metagenomic sequencing for the same stool samples, to provide bacterial host context for our virome sequences. From this, we were able to identify microbiome domination by a single species in four out of five of our patients, although some of the culpable species were distinct from those identified in Chapter 3.

We estimated the bacterial DNA contamination in our viromes using ViromeQC. This suggested that our viromes were less enriched than even a typical whole-community metagenomic gut microbiome. However, this was due to unusually high levels of rRNA in our viromes, and when only single-copy

marker genes were used, the average enrichment of our viromes was 41.6x. Further analysis using the *cpn60* gene database suggests that our viromes contained 16.8% bacterial DNA. We therefore used the same vOTU identification method as in Chapter 3 to remove this contamination. When compared to the vOTU set in Chapter 3, the vOTUs identified in this study were on average twice as long, completeness was three times higher, substantially more of them were classified as high or medium quality, and a higher proportion of these vOTUs clustered with reference genomes. The majority of vOTUs were classified as virulent, as expected due to the viral fraction we sequenced theoretically only containing free phages. Only two ARGs (and zero VFs) were encoded by our vOTUs, both of which were *cfxA* beta-lactamases. From the first vOTU, which had high homology to a strain of *P. distasonis*, we characterised a novel Tn*4555*-like transposon, which is likely an ancestral form of Tn*4555* before it acquired IS*Bf8*. This vOTU was originally identified due to its similarity to a genome in GPD, but this homology only covers the transposon region. This coupled with the lack of hallmark phage genes within the vOTU means that we can not be certain whether it is a phage-derived sequence. The second vOTU had high similarity to *Bacteriodes* species, as well as a *Caudoviricetes* vMAG. However, these similarities did not include a 2,512 bp sequence at the 3' end which included the CDS for *cfxA*. This 2,512 bp sequence aligns to a middle section of Tn*4555* which includes a part of the *mobA* gene as well as the beta-lactamase. The junction sequence between this 2,512 bp sequence and the rest of the vOTU could not be found in our metagenomic contigs, or in NCBI's nr/nt database. The most likely explanations for this are that this vOTU was a result of a misassembly, where a genuine phage sequence was connected to a

partial sequence of Tn*4555*, or that it was a result of a genuine recombination event, that was not captured by the metagenomic sequencing or assembly process. With both ARG encoding vOTUs found in this chapter having contentious origins, this brings into doubt whether the 54 ARG encoding vOTUs that we identified in Chapter 3 were genuine phage-encoded ARGs. Despite implementing extensive vOTU identification criteria in Chapter 3, it is clear from Chapter 4 that some of our vOTUs may not be derived from viral genomes. Previous studies have shown that ARGs are encoded in phage genomes and their over-representation in other studies is due to host contamination or relaxed thresholds (E-value $<10^3$) for ARG detection (Enault et al., 2017; Pfeifer et al., 2022). Whilst we used conservative thresholds for ARG detection (E-value $<10^{-5}$), we opted to not investigate the genomic context of each ARG encoding vOTU in the same manner as in Chapter 4 due to time constraints. However, in future analyses we recommend implementing a tuning removal step, as suggested by Hegarty et al., 2023 and Pratama et al., 2021, using an automated genomic context-aware tool such as CheckV to further reduce false positive vOTUs.

Absolute concentrations of our vOTUs were calculated using a Q33 phage spike-in. When we compared the results from vOTU absolute abundance and relative abundance, notable differences were identified. For example, in one patient, the relative abundance of their vOTUs stayed stable over time, but the absolute abundance of vOTUs dropped steeply around day 19 of their stay in the ICU. In contrast to Chapter 3, the most abundant vOTUs by predicted host, had no overlap with the most abundant bacteria in the metagenomic data. This supports our hypothesis that the majority of vOTUs that we identified in the

previous chapter were integrated prophages. As the abundance of free vOTUs infecting the dominant microbes is low, this indicates that phages are taking advantage of the proliferation of their hosts by replicating with them, rather than entering the lytic cycle, which aligns with the "piggyback-the-winner" model for gut phage dynamics.

5.1 Conclusion

The work in this thesis yielded valuable insights into the field of metagenomic phage identification tools, offering a robust set of metrics that will serve as a foundation for future studies. The benchmarking tests devised herein not only contribute to our understanding of current tools but also provide a framework for assessing the performance of future tools. Through the application of these tools, we identified phage sequences from the gut microbiome of ICU patients. This highlighted the limitations of using metagenomic approaches to study phage dynamics, as we cannot currently predict activity for the majority of vOTUs. We therefore, showcased the usage of long-read virome sequencing coupled with whole-community metagenomic sequencing and used this to investigate the host-phage dynamics within a critically ill gut. Overall, the significance of this work extends beyond the results of this thesis. We have progressed our understanding of metagenomic phage sequence identification and provided support for a model of gut phage dynamics in the context of critical illness.

References

Aagaard, K., Petrosino, J., Keitel, W., Watson, M., Katancik, J., Garcia, N., Patel, S., Cutting, M., Madden, T., Hamilton, H., Harris, E., Gevers, D., Simone, G., McInnes, P., & Versalovic, J. (2013). The Human Microbiome Project strategy for comprehensive sampling of the human microbiome and why it matters. *The FASEB Journal*, *27*(3), 1012–1022. https://doi.org/10.1096/fj.12-220806

Abdelkareem, A. O., Khalil, M. I., Elaraby, M., Abbas, H., & Elbehery, A. H. A. (2018). VirNet: Deep attention model for viral reads identification. *2018 13th International Conference on Computer Engineering and Systems (ICCES)*, 623–626.

Abedon, S. T., & LeJeune, J. T. (2007). Why Bacteriophage Encode Exotoxins and other Virulence Factors. *Evolutionary Bioinformatics Online*, *1*, 97–110. Retrieved May 2, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2658872/

Ackermann, H. W. (2003). Bacteriophage observations and evolution. *Research in Microbiology*, *154*(4), 245–251. https://doi.org/10.1016/S0923-2508(03)00067-6

Ackermann, H. W. (2007). 5500 Phages examined in the electron microscope [Number: 2]. *Archives of Virology*, *152*(2), 227–243. https://doi.org/10.1007/s00705-006-0849-1

Ackermann, H.-W., Audurier, A., Berthiaume, L., Jones, L. A., Mayo, J. A., & Vidaver, A. K. (1978). Guidelines for Bacteriophage Characterization. In M. A. Lauffer, F. B. Bang, K. Maramorosch, & K. M. Smith (Eds.), *Advances in Virus Research* (pp. 1–24). Academic Press. https://doi.org/10.1016/S0065-3527(08)60096-2

Adriaenssens, E. M., Farkas, K., Harrison, C., Jones, D. L., Allison, H. E., & McCarthy, A. J. (2018a). Viromic Analysis of Wastewater Input to a River Catchment Reveals a Diverse Assemblage of RNA Viruses. *mSystems*, *3*(3), e00025–18. https://doi.org/10.1128/mSystems.00025-18

Adriaenssens, E. M., Sullivan, M. B., Knezevic, P., van Zyl, L. J., Sarkar, B. L., Dutilh, B. E., Alfenas-Zerbini, P., Łobocka, M., Tong, Y., Brister, J. R., Moreno Switt, A. I., Klumpp, J., Aziz, R. K., Barylski, J., Uchiyama, J., Edwards, R. A., Kropinski, A. M., Petty, N. K., Clokie, M. R. J., … Krupovic, M. (2020). Taxonomy of prokaryotic viruses: 2018-2019 update

from the ICTV Bacterial and Archaeal Viruses Subcommittee. *Archives of Virology*, *165*(5), 1253–1260. https://doi.org/10.1007/s00705-020-04577-8

Adriaenssens, E. M., Wittmann, J., Kuhn, J. H., Turner, D., Sullivan, M. B., Dutilh, B. E., Jang, H. B., van Zyl, L. J., Klumpp, J., Lobocka, M., Moreno Switt, A. I., Rumnieks, J., Edwards, R. A., Uchiyama, J., Alfenas-Zerbini, P., Petty, N. K., Kropinski, A. M., Barylski, J., Gillis, A., ... Rodney Brister, J. (2018b). Taxonomy of prokaryotic viruses: 2017 update from the ICTV Bacterial and Archaeal Viruses Subcommittee. *Archives of Virology*, *163*(4), 1125–1129. https://doi.org/10.1007/s00705-018-3723-z

Aggarwala, V., Liang, G., & Bushman, F. D. (2017). Viral communities of the human gut: Metagenomic analysis of composition and dynamics [Number: 1]. *Mobile DNA*, *8*(1), 12. https://doi.org/10.1186/s13100-017-0095-y

Agudelo-Ochoa, G. M., Valdés-Duque, B. E., Giraldo-Giraldo, N. A., Jaillier-Ramírez, A. M., Giraldo-Villa, A., Acevedo-Castaño, I., Yepes-Molina, M. A., Barbosa-Barbosa, J., & Benítez-Paéz, A. (2020). Gut microbiota profiles in critically ill patients, potential biomarkers and risk variables for sepsis. *Gut Microbes*, *12*(1), 1707610. https://doi.org/10.1080/19490976.2019.1707610

Ahlgren, N., Ren, J., Lu, Y. Y., Fuhrman, J. A., & Sun, F. (2016). Alignment-free d_2 oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences [ISBN: 2076792171 _eprint: 1611.06654]. *Nucleic Acids Research*, *45*(1), 39–53. https://doi.org/10.1093/nar/gkw1002

Akagawa, S., Tsuji, S., Onuma, C., Akagawa, Y., Yamaguchi, T., Yamagishi, M., Yamanouchi, S., Kimata, T., Sekiya, S.-I., Ohashi, A., Hashiyada, M.,

Akane, A., & Kaneko, K. (2019). Effect of Delivery Mode and Nutrition on Gut Microbiota in Neonates. *Annals of Nutrition & Metabolism*, *74*(2), 132–139. https://doi.org/10.1159/000496427

Akhter, S., Aziz, R. K., & Edwards, R. A. (2012). PhiSpy: A novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies [Number: 16]. *Nucleic Acids Research*, *40*(16), e126. https://doi.org/10.1093/nar/gks406

Albenberg, L., Esipova, T., Judge, C., Bittinger, K., Chen, J., Laughlin, A., Grunberg, S., Baldassano, R., Lewis, J., Li, H., Thom, S., Bushman, F., Vinogradov, S., & Wu, G. (2014). Correlation Between Intraluminal Oxygen Gradient and Radial Partitioning of Intestinal Microbiota in Humans and Mice. *Gastroenterology*, *147*(5), 1055–1063.e8. https://doi.org/10.1053/j.gastro.2014.07.020

Alcock, B. P., Huynh, W., Chalil, R., Smith, K. W., Raphenya, A. R., Wlodarski, M. A., Edalatmand, A., Petkau, A., Syed, S. A., Tsang, K. K., Baker, S. J. C., Dave, M., McCarthy, M. C., Mukiri, K. M., Nasir, J. A., Golbon, B., Imtiaz, H., Jiang, X., Kaur, K., … McArthur, A. G. (2023). CARD 2023: Expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Research*, *51*(D1), D690–D699. https://doi.org/10.1093/nar/gkac920

Allard, J. D., Gibson, M. L., Vu, L. H., Nguyen, T. T., & Bertrand, K. P. (1993). Nucleotide sequence of class D tetracycline resistance genes from Salmonella ordonez. *Molecular & general genetics: MGG*, *237*(1-2), 301–305. https://doi.org/10.1007/BF00282811

Al-Shayeb, B., Sachdeva, R., Chen, L.-X., Ward, F., Munk, P., Devoto, A., Castelle, C. J., Olm, M. R., Bouma-Gregson, K., Amano, Y., He, C., Méheust, R., Brooks, B., Thomas, A., Lavy, A., Matheus-Carnevali, P., Sun, C., Goltsman, D. S. A., Borton, M. A., … Banfield, J. F. (2020). Clades of huge phages from across Earth's ecosystems [Number: 7795 Publisher: Nature Publishing Group]. *Nature*, *578*(7795), 425–431. https://doi.org/10.1038/s41586-020-2007-4

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Alverdy, J. C., & Krezalek, M. A. (2017). Collapse of the microbiome, emergence of the pathobiome and the immunopathology of sepsis. *Critical care medicine*, *45*(2), 337–347. https://doi.org/10.1097/CCM.0000000000002172

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, *21*(1), 30. https://doi.org/10.1186/s13059-020-1935-5

Amgarten, D., Braga, L. P. P., da Silva, A. M., & Setubal, J. C. (2018). MARVEL, a Tool for Prediction of Bacteriophage Sequences in Metagenomic Bins [Publisher: Frontiers]. *Frontiers in Genetics*, *9*. https://doi.org/10.3389/fgene.2018.00304

Andrews, S., Krueger, F., Segonds-Pichon, A., Biggins, L., Krueger, C., & Wingett, S. (2019). FastQC [Programmers: _:n46].

Antipov, D., Raiko, M., Lapidus, A., & Pevzner, P. A. (2020). metaviralSPAdes: Assembly of viruses from metagenomic data. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btaa490

Arisdakessian, C. G., Nigro, O. D., Steward, G. F., Poisson, G., & Belcaid, M. (2021). CoCoNet: An efficient deep learning tool for viral metagenome binning. *Bioinformatics*, *37*(18), 2803–2810. https://doi.org/10.1093/bioinformatics/btab213

Arndt, D., Marcu, A., Liang, Y., & Wishart, D. S. (2017). PHAST, PHASTER and PHASTEST: Tools for finding prophage in bacterial genomes [Number: 4]. *Briefings in Bioinformatics*, *20*(4), 1560–1567. https://doi.org/10.1093/bib/bbx121

Arumugam, M., Raes, J., Pelletier, E., Paslier, D. L., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J.-M., Bertalan, M., Borruel, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., Kleerebezem, M., ... Bork, P. (2011). Enterotypes of the human gut microbiome [Number: 7346 Publisher: Nature Publishing Group]. *Nature*, *473*(7346), 174–180. https://doi.org/10.1038/nature09944

Ashelford, K. E., Day, M. J., & Fry, J. C. (2003). Elevated Abundance of Bacteriophage Infecting Bacteria in Soil [Publisher: American Society for Microbiology]. *Applied and Environmental Microbiology*, *69*(1), 285–289. https://doi.org/10.1128/AEM.69.1.285-289.2003

Auslander, N., Gussow, A. B., Benler, S., Wolf, Y. I., & Koonin, E. V. (2020). Seeker: Alignment-free identification of bacteriophage genomes by deep learning [Number: 21]. *Nucleic Acids Research*, *48*(21), e121–e121. https://doi.org/10.1093/nar/gkaa856

Aziz, R. K., Dwivedi, B., Akhter, S., Breitbart, M., & Edwards, R. A. (2015). Multidimensional metrics for estimating phage abundance, distribution,

gene density, and sequence coverage in metagenomes. *Frontiers in Microbiology*, *6*, 381. https://doi.org/10.3389/fmicb.2015.00381

Bai, Z., Zhang, Y.-z., Miyano, S., Yamaguchi, R., Fujimoto, K., Uematsu, S., & Imoto, S. (2022). Identification of bacteriophage genome sequences with representation learning. *Bioinformatics*, *38*(18), 4264–4270. https://doi.org/10.1093/bioinformatics/btac509

Barcik, W., Boutin, R. C. T., Sokolowska, M., & Finlay, B. B. (2020). The Role of Lung and Gut Microbiota in the Pathology of Asthma. *Immunity*, *52*(2), 241–255. https://doi.org/10.1016/j.immuni.2020.01.007

Bardou, M., Quenot, J.-P., & Barkun, A. (2015). Stress-related mucosal disease in the critically ill patient. *Nature Reviews. Gastroenterology & Hepatology*, *12*(2), 98–107. https://doi.org/10.1038/nrgastro.2014.235

Bartlau, N., Wichels, A., Krohne, G., Adriaenssens, E. M., Heins, A., Fuchs, B. M., Amann, R., & Moraru, C. (2022). Highly diverse flavobacterial phages isolated from North Sea spring blooms [Number: 2 Publisher: Nature Publishing Group]. *The ISME Journal*, *16*(2), 555–568. https://doi.org/10.1038/s41396-021-01097-4

Bartosch, S., Fite, A., Macfarlane, G. T., & McMurdo, M. E. T. (2004). Characterization of Bacterial Communities in Feces from Healthy Elderly Volunteers and Hospitalized Elderly Patients by Using Real-Time PCR and Effects of Antibiotic Treatment on the Fecal Microbiota. *Applied and Environmental Microbiology*, *70*(6), 3575–3581. https://doi.org/10.1128/AEM.70.6.3575-3581.2004

Barylski, J., Kropinski, A. M., Alikhan, N.-F., & Adriaenssens, E. M. (2020). ICTV Virus Taxonomy Profile: Herelleviridae. *The Journal of General Virology*, *101*(4), 362–363. https://doi.org/10.1099/jgv.0.001392

Beaulaurier, J., Luo, E., Eppley, J. M., Uyl, P. D., Dai, X., Burger, A., Turner, D. J., Pendleton, M., Juul, S., Harrington, E., & DeLong, E. F. (2020). Assembly-free single-molecule sequencing recovers complete virus genomes from natural microbial communities [Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab]. *Genome Research*, gr.251686.119. https://doi.org/10.1101/gr.251686.119

Belzer, C., & de Vos, W. M. (2012). Microbes inside—from diversity to function: The case of Akkermansia. *The ISME Journal*, *6*(8), 1449–1458. https://doi.org/10.1038/ismej.2012.6

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1995.tb02031.x]. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Benler, S., Yutin, N., Antipov, D., Rayko, M., Shmakov, S., Gussow, A. B., Pevzner, P., & Koonin, E. V. (2021). Thousands of previously unknown phages discovered in whole-community human gut metagenomes. *Microbiome*, *9*(1), 78. https://doi.org/10.1186/s40168-021-01017-w

Bergh, Ø., BØrsheim, K. Y., Bratbak, G., & Heldal, M. (1989). High abundance of viruses found in aquatic environments. *Nature*, *340*(6233), 467–468. https://doi.org/10.1038/340467a0

Bertani, G., & Nice, S. J. (1954). Studies on lysogenesis. II. The effect of temperature on the lysogenization of Shigella dysenteriae with phage P1. *Journal of Bacteriology*, *67*(2), 202–209. https://doi.org/10.1128/jb.67.2.202-209.1954

Beutin, L., Bode, L., Ozel, M., & Stephan, R. (1990). Enterohemolysin production is associated with a temperate bacteriophage in Escherichia coli serogroup O26 strains. *Journal of bacteriology*, *172*(11), 6469–6475. https://doi.org/10.1128/jb.172.11.6469-6475.1990

Bi, L., Yu, D.-T., Du, S., Zhang, L.-M., Zhang, L.-Y., Wu, C.-F., Xiong, C., Han, L.-L., & He, J.-Z. (2021). Diversity and potential biogeochemical impacts of viruses in bulk and rhizosphere soils [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1462-2920.15010]. *Environmental Microbiology*, *23*(2), 588–599. https://doi.org/10.1111/1462-2920.15010

Blackman Northwood, J., Del Grosso, M., Cossins, L. R., Coley, M. D., Creti, R., Pantosti, A., & Farrell, D. J. (2009). Characterization of Macrolide Efflux Pump mef Subclasses Detected in Clinical Isolates of Streptococcus pyogenes Isolated between 1999 and 2005. *Antimicrobial Agents and Chemotherapy*, *53*(5), 1921–1925. https://doi.org/10.1128/AAC.01065-08

Blanco, L., Bernad, A., Lázaro, J. M., Martín, G., Garmendia, C., & Salas, M. (1989). Highly Efficient DNA Synthesis by the Phage 29 DNA Polymerase: Symmetrical Mode of DNA Replication [Publisher: Elsevier]. *Journal of*

Biological Chemistry, *264*(15), 8935–8940. https://doi.org/10.1016/S0021-9258(18)81883-X

Blanco-Miguez, A., Beghini, F., Cumbo, F., McIver, L. J., Thompson, K. N., Zolfo, M., Manghi, P., Dubois, L., Huang, K. D., Thomas, A. M., Piccinno, G., Piperni, E., Punčochář, M., Valles-Colomer, M., Tett, A., Giordano, F., Davies, R., Wolf, J., Berry, S. E., … Segata, N. (2022). Extending and improving metagenomic taxonomic profiling with uncharacterized species with MetaPhlAn 4 [Pages: 2022.08.22.504593 Section: New Results]. https://doi.org/10.1101/2022.08.22.504593

Blot, M. (1994). Transposable elements and adaptation of host bacteria. *Genetica*, *93*(1-3), 5–12. https://doi.org/10.1007/BF01435235

Boiten, K. E., Kuijper, E. J., Schuele, L., van Prehn, J., Bode, L. G. M., Maat, I., van Asten, S. A. V., Notermans, D. W., Rossen, J. W. A., & Veloo, A. C. M. (2023). Characterization of mobile genetic elements in multidrug-resistant Bacteroides fragilis isolates from different hospitals in the Netherlands. *Anaerobe*, *81*, 102722. https://doi.org/10.1016/j.anaerobe.2023.102722

Bokulich, N. A., Chung, J., Battaglia, T., Henderson, N., Jay, M., Li, H., D Lieber, A., Wu, F., Perez-Perez, G. I., Chen, Y., Schweizer, W., Zheng, X., Contreras, M., Dominguez-Bello, M. G., & Blaser, M. J. (2016). Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Science Translational Medicine*, *8*(343), 343ra82. https://doi.org/10.1126/scitranslmed.aad7121

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data [Number: 15]. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bonilla, N., Santiago, T., Marcos, P., Urdaneta, M., Domingo, J. S., & Toranzos, G. A. (2010). Enterophages, a group of phages infecting Enterococcus faecalis, and their potential as alternate indicators of human faecal contamination. *Water Science and Technology: A Journal of the International Association on Water Pollution Research*, *61*(2), 293–300. https://doi.org/10.2166/wst.2010.815

Borre, Y. E., O'Keeffe, G. W., Clarke, G., Stanton, C., Dinan, T. G., & Cryan, J. F. (2014). Microbiota and neurodevelopmental windows: Implications for brain disorders. *Trends in Molecular Medicine*, *20*(9), 509–518. https://doi.org/10.1016/j.molmed.2014.05.002

Boucher, Y., Labbate, M., Koenig, J. E., & Stokes, H. W. (2007). Integrons: Mobilizable platforms that promote genetic diversity in bacteria. *Trends in Microbiology*, *15*(7), 301–309. https://doi.org/10.1016/j.tim.2007.05.004

Bouras, G., Nepal, R., Houtak, G., Psaltis, A. J., Wormald, P.-J., & Vreugde, S. (2023). Pharokka: A fast scalable bacteriophage annotation tool. *Bioinformatics*, *39*(1), btac776. https://doi.org/10.1093/bioinformatics/btac776

Boyd, E. F. (2012). Bacteriophage-encoded bacterial virulence factors and phage-pathogenicity island interactions. *Advances in Virus Research*, *82*, 91–118. https://doi.org/10.1016/B978-0-12-394621-8.00014-5

Brooks, A. W., Priya, S., Blekhman, R., & Bordenstein, S. R. (2018). Gut microbiota diversity across ethnicities in the United States. *PLoS Biology*, *16*(12), e2006842. https://doi.org/10.1371/journal.pbio.2006842

Brown, T. L., Charity, O. J., & Adriaenssens, E. M. (2022). Ecological and functional roles of bacteriophages in contrasting environments: Marine, terrestrial and human gut. *Current Opinion in Microbiology*, *70*, 102229. https://doi.org/10.1016/j.mib.2022.102229

Brum, J. R., Ignacio-Espinoza, J. C., Roux, S., Doulcier, G., Acinas, S. G., Alberti, A., Chaffron, S., Cruaud, C., de Vargas, C., Gasol, J. M., Gorsky, G., Gregory, A. C., Guidi, L., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Poulos, B. T., ... Sullivan, M. B. (2015). Patterns and ecological drivers of ocean viral communities [Publisher: American Association for the Advancement of Science]. *Science*, *348*(6237), 1261498. https://doi.org/10.1126/science.1261498

Brum, J. R., Schenck, R. O., & Sullivan, M. B. (2013). Global morphological analysis of marine viruses shows minimal regional variation and dominance of non-tailed viruses. *The ISME journal*, *7*(9), 1738–1751. https://doi.org/10.1038/ismej.2013.67

Bushnell, B. (2014). BBMap: A Fast, Accurate, Splice-Aware Aligner. https://www.osti.gov/biblio/1241166

Bushnell, B., Rood, J., & Singer, E. (2017). BBMerge – Accurate paired shotgun read merging via overlap [Number: 10 Publisher: Public Library of Science]. *PLOS ONE*, *12*(10), e0185056. https://doi.org/10.1371/journal.pone.0185056

Callanan, J., Stockdale, S. R., Shkoporov, A., Draper, L. A., Ross, R. P., & Hill, C. (2020). Expansion of known ssRNA phage genomes: From tens to over a thousand [Publisher: American Association for the Advancement of Science]. *Science Advances*, *6*(6), eaay5981. https://doi.org/10.1126/sciadv.aay5981

Callanan, J., Stockdale, S. R., Shkoporov, A., Draper, L. A., Ross, R. P., & Hill, C. (2021). Biases in Viral Metagenomics-Based Detection, Cataloguing and Quantification of Bacteriophage Genomes in Human Faeces, a Review [Number: 3 Publisher: Multidisciplinary Digital Publishing Institute]. *Microorganisms*, *9*(3), 524. https://doi.org/10.3390/microorganisms9030524

Camargo, A. P., Roux, S., Schulz, F., Babinski, M., Xu, Y., Hu, B., Chain, P. S. G., Nayfach, S., & Kyrpides, N. C. (2023). Identification of mobile genetic elements with geNomad [Publisher: Nature Publishing Group]. *Nature Biotechnology*, 1–10. https://doi.org/10.1038/s41587-023-01953-y

Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D., & Lawley, T. D. (2021). Massive expansion of human gut bacteriophage diversity. *Cell*, *184*(4), 1098–1109.e9. https://doi.org/10.1016/j.cell.2021.01.029

Cao, J., Zhang, Y., Dai, M., Xu, J., Chen, L., Zhang, F., Zhao, N., & Wang, J. (2020). Profiling of Human Gut Virome with Oxford Nanopore Technology. *Medicine in Microecology*, *4*, 100012. https://doi.org/10.1016/j.medmic.2020.100012

Cao, Z., Sugimura, N., Burgermeister, E., Ebert, M. P., Zuo, T., & Lan, P. (2022). The gut virome: A new microbiome component in health and

disease [Publisher: Elsevier]. *eBioMedicine*, *81*. https://doi.org/10.1016/j.ebiom.2022.104113

Casjens, S. R., & Gilcrease, E. B. (2009). Determining DNA packaging strategy by analysis of the termini of the chromosomes in tailed-bacteriophage virions. *Methods in molecular biology (Clifton, N.J.)*, *502*, 91–111. https://doi.org/10.1007/978-1-60327-565-1_7

Castro-Mejía, J. L., Muhammed, M. K., Kot, W., Neve, H., Franz, C. M. A. P., Hansen, L. H., Vogensen, F. K., & Nielsen, D. S. (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

Cervantes-Echeverría, M., Gallardo-Becerra, L., Cornejo-Granados, F., & Ochoa-Leyva, A. (2023). The Two-Faced Role of crAssphage Subfamilies in Obesity and Metabolic Syndrome: Between Good and Evil [Number: 1 Publisher: Multidisciplinary Digital Publishing Institute]. *Genes*, *14*(1), 139. https://doi.org/10.3390/genes14010139

Che, Y., Xia, Y., Liu, L., Li, A.-D., Yang, Y., & Zhang, T. (2019). Mobile antibiotic resistome in wastewater treatment plants revealed by Nanopore metagenomic sequencing. *Microbiome*, *7*(1), 44. https://doi.org/10.1186/s40168-019-0663-0

Chen, J., Quiles-Puchalt, N., Chiang, Y. N., Bacigalupe, R., Fillol-Salom, A., Chee, M. S. J., Fitzgerald, J. R., & Penadés, J. R. (2018). Genome hypermobility by lateral transduction [Number: 6411 Publisher: American Association for the Advancement of Science Section: Research Article]. *Science*, *362*(6411), 207–212. https://doi.org/10.1126/science.aat5867

Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y., & Jin, Q. (2005). VFDB: A reference database for bacterial virulence factors. *Nucleic Acids Research*, *33*(Database issue), D325–328. https://doi.org/10.1093/nar/gki008

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Chen, X., Ma, R., Yang, Y., Jiao, N., & Zhang, R. (2019). Viral Regulation on Bacterial Community Impacted by Lysis-Lysogeny Switch: A Microcosm Experiment in Eutrophic Coastal Waters. *Frontiers in Microbiology*, *10*. Retrieved February 1, 2024, from https://www.frontiersin.org/articles/10.3389/fmicb.2019.01763

Chiang, Y. N., Penadés, J. R., & Chen, J. (2019). Genetic transduction by phages and chromosomal islands: The new and noncanonical [Number: 8]. *PLoS Pathogens*, *15*(8). https://doi.org/10.1371/journal.ppat.1007878

Clarridge, J. E. (2004). Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases. *Clinical Microbiology Reviews*, *17*(4), 840–862. https://doi.org/10.1128/CMR.17.4.840-862.2004

Colomer-Lluch, M., Jofre, J., & Muniesa, M. (2011). Antibiotic Resistance Genes in the Bacteriophage DNA Fraction of Environmental Samples. *PLoS ONE*, *6*(3), e17549. https://doi.org/10.1371/journal.pone.0017549

Conceição-Neto, N., Zeller, M., Lefrère, H., De Bruyn, P., Beller, L., Deboutte, W., Yinda, C. K., Lavigne, R., Maes, P., Ranst, M. V., Heylen, E., & Matthijnssens, J. (2015). Modular approach to customise sample preparation procedures for viral metagenomics: A reproducible

protocol for virome analysis. *Scientific Reports*, *5*, 16532. https://doi.org/10.1038/srep16532

Connor, T. R., Loman, N. J., Thompson, S., Smith, A., Southgate, J., Poplawski, R., Bull, M. J., Richardson, E., Ismail, M., Thompson, S. E., Kitchen, C., Guest, M., Bakke, M., Sheppard, S. K., & Pallen, M. J. (2016). CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): An online resource for the medical microbiology community [Number: 9 Publisher: Microbiology Society,]. *Microbial Genomics, 2*(9), e000086. https://doi.org/10.1099/mgen.0.000086

Cook, R., Brown, N., Redgwell, T., Rihtman, B., Barnes, M., Clokie, M., Stekel, D. J., Hobman, J., Jones, M. A., & Millard, A. (2021). INfrastructure for a PHAge REference Database: Identification of Large-Scale Biases in the Current Collection of Cultured Phage Genomes [Publisher: Mary Ann Liebert, Inc., publishers]. *PHAGE*, *2*(4), 214–223. https://doi.org/10.1089/phage.2021.0007

Costerton, J. W., Stewart, P. S., & Greenberg, E. P. (1999). Bacterial biofilms: A common cause of persistent infections. *Science (New York, N.Y.)*, *284*(5418), 1318–1322. https://doi.org/10.1126/science.284.5418.1318

Coutinho, F. H., Zaragoza-Solas, A., López-Pérez, M., Barylski, J., Zielezinski, A., Dutilh, B. E., Edwards, R., & Rodriguez-Valera, F. (2021). RaFAH: Host prediction for viruses of Bacteria and Archaea based on protein content. *Patterns*, *2*(7). https://doi.org/10.1016/j.patter.2021.100274

Coutinho, H. B., Robalinho, T. I., Coutinho, V. B., Amorim, A. M., Furtado, A. F., Ferraz, A., Ferraz, E., Walker, F., King, G., Sewell, H. F., & Wakelin, D. (1997). Intra-abdominal sepsis: An immunocytochemical study of the small

intestine mucosa. *Journal of Clinical Pathology*, *50*(4), 294–298. Retrieved April 24, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC499878/

Daly, R. A., Roux, S., Borton, M. A., Morgan, D. M., Johnston, M. D., Booker, A. E., Hoyt, D. W., Meulia, T., Wolfe, R. A., Hanson, A. J., Mouser, P. J., Moore, J. D., Wunch, K., Sullivan, M. B., Wrighton, K. C., & Wilkins, M. J. (2019). Viruses control dominant bacteria colonizing the terrestrial deep biosphere after hydraulic fracturing. *Nature Microbiology*, *4*(2), 352–361. https://doi.org/10.1038/s41564-018-0312-6

D'Amore, R., Ijaz, U. Z., Schirmer, M., Kenny, J. G., Gregory, R., Darby, A. C., Shakya, M., Podar, M., Quince, C., & Hall, N. (2016). A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC Genomics*, *17*, 55. https://doi.org/10.1186/s12864-015-2194-9

David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., Ling, A. V., Devlin, A. S., Varma, Y., Fischbach, M. A., Biddinger, S. B., Dutton, R. J., & Turnbaugh, P. J. (2014). Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, *505*(7484), 559–563. https://doi.org/10.1038/nature12820

Davison, J. M., & Wischmeyer, P. E. (2019). Probiotic and synbiotic therapy in the critically ill: State of the art. *Nutrition (Burbank, Los Angeles County, Calif.)*, *59*, 29–36. https://doi.org/10.1016/j.nut.2018.07.017

De Gottardi, A., & McCoy, K. D. (2011). Evaluation of the gut barrier to intestinal bacteria in non-alcoholic fatty liver disease. *Journal of Hepatology*, *55*(6), 1181–1183. https://doi.org/10.1016/j.jhep.2011.05.003

Deaton, J., Yu, F. B., & Quake, S. R. (2017). PhaMers identifies novel bacteriophage sequences from thermophilic hot springs [Publisher: Cold Spring Harbor Laboratory Section: New Results]. *bioRxiv*, 169672. https://doi.org/10.1101/169672

de Haan, J. J., Lubbers, T., Derikx, J. P., Relja, B., Henrich, D., Greve, J.-W., Marzi, I., & Buurman, W. A. (2009). Rapid development of intestinal cell damage following severe trauma: A prospective observational cohort study. *Critical Care*, *13*(3), R86. https://doi.org/10.1186/cc7910

de Jong, N. W. M., Vrieling, M., Garcia, B. L., Koop, G., Brettmann, M., Aerts, P. C., Ruyken, M., van Strijp, J. A. G., Holmes, M., Harrison, E. M., Geisbrecht, B. V., & Rooijakkers, S. H. M. (2018). Identification of a staphylococcal complement inhibitor with broad host specificity in equid Staphylococcus aureus strains. *The Journal of Biological Chemistry*, *293*(12), 4468–4477. https://doi.org/10.1074/jbc.RA117.000599

de Jonge, E., Schultz, M. J., Spanjaard, L., Bossuyt, P. M. M., Vroom, M. B., Dankert, J., & Kesecioglu, J. (2003). Effects of selective decontamination of digestive tract on mortality and acquisition of resistant bacteria in intensive care: A randomised controlled trial. *Lancet (London, England)*, *362*(9389), 1011–1016. https://doi.org/10.1016/S0140-6736(03)14409-1

de Muinck, E. J., Trosvik, P., Gilfillan, G. D., Hov, J. R., & Sundaram, A. Y. M. (2017). A novel ultra high-throughput 16S rRNA gene amplicon sequencing library preparation method for the Illumina HiSeq platform. *Microbiome*, *5*(1), 68. https://doi.org/10.1186/s40168-017-0279-1

Depommier, C., Everard, A., Druart, C., Plovier, H., Van Hul, M., Vieira-Silva, S., Falony, G., Raes, J., Maiter, D., Delzenne, N. M., de Barsy, M., Loumaye, A., Hermans, M. P., Thissen, J.-P., de Vos, W. M., & Cani, P. D. (2019). Supplementation with Akkermansia muciniphila in overweight and obese human volunteers: A proof-of-concept exploratory study. *Nature medicine*, *25*(7), 1096–1103. https://doi.org/10.1038/s41591-019-0495-2

Derrien, M., Collado, M. C., Ben-Amor, K., Salminen, S., & de Vos, W. M. (2008). The Mucin Degrader Akkermansia muciniphila Is an Abundant Resident of the Human Intestinal Tract. *Applied and Environmental Microbiology*, *74*(5), 1646–1648. https://doi.org/10.1128/AEM.01226-07

Derrien, M., Vaughan, E. E., Plugge, C. M., & de Vos, W. M. (2004). Akkermansia muciniphila gen. nov., sp. nov., a human intestinal mucin-degrading bacterium [Publisher: Microbiology Society,]. *International Journal of Systematic and Evolutionary Microbiology*, *54*(5), 1469–1476. https://doi.org/10.1099/ijs.0.02873-0

de Vree, J. M., Romijn, J. A., Mok, K. S., Mathus-Vliegen, L. M., Stoutenbeek, C. P., Ostrow, J. D., Tytgat, G. N., Sauerwein, H. P., Oude Elferink, R. P., & Groen, A. K. (1999). Lack of enteral nutrition during critical illness is associated with profound decrements in biliary lipid concentrations. *The American Journal of Clinical Nutrition*, *70*(1), 70–77. https://doi.org/10.1093/ajcn/70.1.70

d'Herelle, M. F. (1917). Sur un microbe invisible antagoniste des bacilles dysenteriques [Accepted: 2019-08-12T18:52:43Z Publisher: Ministerio de Asistencia Social y Salud Pública; Instituto Nacional de Microbiología.].

Comptes rendus de l'Académie des Sciences. Retrieved April 17, 2020, from http://sgc.anlis.gob.ar/handle/123456789/753

Dhillon, T. S., Dhillon, E. K., Chau, H. C., Li, W. K., & Tsang, A. H. (1976). Studies on bacteriophage distribution: Virulent and temperate bacteriophage content of mammalian feces [Number: 1]. *Applied and Environmental Microbiology*, *32*(1), 68–74.

d'Humières, C., Touchon, M., Dion, S., Cury, J., Ghozlane, A., Garcia-Garcera, M., Bouchier, C., Ma, L., Denamur, E., & P. C.Rocha, E. (2019). A simple, reproducible and cost-effective procedure to analyse gut phageome: From phage isolation to bioinformatic approach [Number: 1 Publisher: Nature Publishing Group]. *Scientific Reports*, *9*(1), 11331. https://doi.org/10.1038/s41598-019-47656-w

Dickson, R. P. (2016). The microbiome and critical illness. *The Lancet. Respiratory medicine*, *4*(1), 59–72. https://doi.org/10.1016/S2213-2600(15)00427-0

Didelot, X., Walker, A. S., Peto, T. E., Crook, D. W., & Wilson, D. J. (2016). Within-host evolution of bacterial pathogens. *Nature reviews. Microbiology*, *14*(3), 150–162. https://doi.org/10.1038/nrmicro.2015.13

Dion, M. B., Oechslin, F., & Moineau, S. (2020). Phage diversity, genomics and phylogeny [Number: 3 Publisher: Nature Publishing Group]. *Nature Reviews Microbiology*, *18*(3), 125–138. https://doi.org/10.1038/s41579-019-0311-5

Dominguez-Bello, M. G., Costello, E. K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., & Knight, R. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proceedings of the National Academy of Sciences of the United*

States of America, *107*(26), 11971–11975. https://doi.org/10.1073/pnas.1002601107

Dominguez-Huerta, G., Zayed, A. A., Wainaina, J. M., Guo, J., Tian, F., Pratama, A. A., Bolduc, B., Mohssen, M., Zablocki, O., Pelletier, E., Delage, E., Alberti, A., Aury, J.-M., Carradec, Q., da Silva, C., Labadie, K., Poulain, J., Tara Oceans Coordinators, Bowler, C., ... Sullivan, M. B. (2022). Diversity and ecological footprint of Global Ocean RNA viruses [Publisher: American Association for the Advancement of Science]. *Science*, *376*(6598), 1202–1208. https://doi.org/10.1126/science.abn6358

Dorward, D. W., Garon, C. F., & Judd, R. C. (1989). Export and intercellular transfer of DNA via membrane blebs of Neisseria gonorrhoeae. *Journal of Bacteriology*, *171*(5), 2499–2505. Retrieved November 8, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC209926/

Doyle, J., & Doyle, J. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue [Place: 1987]. *PHYTOCHEMICAL BULLETIN*, (RESEARCH). http://worldveg.tind.io/record/33886

Dubey, G. P., & Ben-Yehuda, S. (2011). Intercellular Nanotubes Mediate Bacterial Communication. *Cell*, *144*(4), 590–600. https://doi.org/10.1016/j.cell.2011.01.015

Dutilh, B. E., Cassman, N., McNair, K., Sanchez, S. E., Silva, G. G. Z., Boling, L., Barr, J. J., Speth, D. R., Seguritan, V., Aziz, R. K., Felts, B., Dinsdale, E. A., Mokili, J. L., & Edwards, R. A. (2014). A highly abundant bacteriophage discovered in the unknown sequences of human

faecal metagenomes. *Nature Communications*, *5*, 4498. https://doi.org/10.1038/ncomms5498

Edgar, R., Rokney, A., Feeney, M., Semsey, S., Kessel, M., Goldberg, M. B., Adhya, S., & Oppenheim, A. B. (2008). Bacteriophage infection is targeted to cellular poles [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2958.2008.06205.x]. *Molecular Microbiology*, *68*(5), 1107–1116. https://doi.org/10.1111/j.1365-2958.2008.06205.x

Edwards, R. A., Vega, A. A., Norman, H. M., Ohaeri, M., Levi, K., Dinsdale, E. A., Cinek, O., Aziz, R. K., McNair, K., Barr, J. J., Bibby, K., Brouns, S. J. J., Cazares, A., de Jonge, P. A., Desnues, C., Díaz Muñoz, S. L., Fineran, P. C., Kurilshikov, A., Lavigne, R., … Dutilh, B. E. (2019). Global phylogeography and ancient evolution of the widespread human gut virus crAssphage [Number: 10 Publisher: Nature Publishing Group]. *Nature Microbiology*, *4*(10), 1727–1736. https://doi.org/10.1038/s41564-019-0494-6

EFSA Panel on Nutrition, Novel Foods and Food Allergens (NDA), Turck, D., Bohn, T., Castenmiller, J., De Henauw, S., Hirsch-Ernst, K. I., Maciuk, A., Mangelsdorf, I., McArdle, H. J., Naska, A., Pelaez, C., Pentieva, K., Siani, A., Thies, F., Tsabouri, S., Vinceti, M., Cubadda, F., Frenzel, T., Heinonen, M., … Knutsen, H. K. (2021). Safety of pasteurised Akkermansia muciniphila as a novel food pursuant to Regulation (EU) 2015/2283 [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.2903/j.efsa.2021.6780]. *EFSA Journal*, *19*(9), e06780. https://doi.org/10.2903/j.efsa.2021.6780

Ellis, E. L., & Delbrück, M. (1939). THE GROWTH OF BACTERIOPHAGE. *The Journal of General Physiology*, *22*(3), 365–384. Retrieved March 8, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2141994/

Emerson, J. B., Roux, S., Brum, J. R., Bolduc, B., Woodcroft, B. J., Jang, H. B., Singleton, C. M., Solden, L. M., Naas, A. E., Boyd, J. A., Hodgkins, S. B., Wilson, R. M., Trubl, G., Li, C., Frolking, S., Pope, P. B., Wrighton, K. C., Crill, P. M., Chanton, J. P., . . . Sullivan, M. B. (2018). Host-linked soil viral ecology along a permafrost thaw gradient [Number: 8]. *Nature Microbiology*, *3*(8), 870–880. https://doi.org/10.1038/s41564-018-0190-y

Enault, F., Briet, A., Bouteille, L., Roux, S., Sullivan, M. B., & Petit, M.-A. (2017). Phages rarely encode antibiotic resistance genes: A cautionary tale for virome analyses. *The ISME Journal*, *11*(1), 237–247. https://doi.org/10.1038/ismej.2016.90

Entcheva, P., Phillips, D. A., & Streit, W. R. (2002). Functional analysis of Sinorhizobium meliloti genes involved in biotin synthesis and transport. *Applied and Environmental Microbiology*, *68*(6), 2843–2848. https://doi.org/10.1128/AEM.68.6.2843-2848.2002

Erez, Z., Steinberger-Levy, I., Shamir, M., Doron, S., Stokar-Avihail, A., Peleg, Y., Melamed, S., Leavitt, A., Savidor, A., Albeck, S., Amitai, G., & Sorek, R. (2017). Communication between viruses guides lysis–lysogeny decisions [Number: 7638 Publisher: Nature Publishing Group]. *Nature*, *541*(7638), 488–493. https://doi.org/10.1038/nature21049

Evans, T., Ali, U., Anderton, R., Raby, E., Manning, L., & Litton, E. (2023). Lower gut dysbiosis and mortality in acute critical illness: A systematic

review and meta-analysis. *Intensive Care Medicine Experimental*, *11*(1), 6. https://doi.org/10.1186/s40635-022-00486-z

Ezeji, J. C., Sarikonda, D. K., Hopperton, A., Erkkila, H. L., Cohen, D. E., Martinez, S. P., Cominelli, F., Kuwahara, T., Dichosa, A. E. K., Good, C. E., Jacobs, M. R., Khoretonenko, M., Veloo, A., & Rodriguez-Palacios, A. (2021). Parabacteroides distasonis: Intriguing aerotolerant gut anaerobe with emerging antimicrobial resistance and pathogenic and probiotic roles in human health. *Gut Microbes*, *13*(1), 1922241. https://doi.org/10.1080/19490976.2021.1922241

Fang, Z., Tan, J., Wu, S., Li, M., Xu, C., Xie, Z., & Zhu, H. (2019). PPR-Meta: A tool for identifying phages and plasmids from metagenomic fragments using deep learning [Number: 6 Publisher: Oxford Academic]. *GigaScience*, *8*(6). https://doi.org/10.1093/gigascience/giz066

Filardi, R., Gargari, G., Mora, D., & Arioli, S. (2022). Characterization of antibiotic-resistance traits in Akkermansia muciniphila strains of human origin. *Scientific Reports*, *12*, 19426. https://doi.org/10.1038/s41598-022-23980-6

Fortier, L.-C., & Sekulovic, O. (2013). Importance of prophages to evolution and virulence of bacterial pathogens [Publisher: Taylor & Francis _eprint: https://doi.org/10.4161/viru.24498]. *Virulence*, *4*(5), 354–365. https://doi.org/10.4161/viru.24498

Fouts, D. E. (2006). Phage_finder: Automated identification and classification of prophage regions in complete bacterial genome sequences [Number: 20]. *Nucleic Acids Research*, *34*(20), 5839–5851. https://doi.org/10.1093/nar/gkl732

Frosini, S. M., Bond, R., McCarthy, A. J., Feudi, C., Schwarz, S., Lindsay, J. A., & Loeffler, A. (2020). Genes on the Move: In Vitro Transduction of Antimicrobial Resistance Genes between Human and Canine Staphylococcal Pathogens. *Microorganisms*, *8*(12), 2031. https://doi.org/10.3390/microorganisms8122031

Fyfe, C., Grossman, T. H., Kerstein, K., & Sutcliffe, J. (2016). Resistance to Macrolide Antibiotics in Public Health Pathogens. *Cold Spring Harbor Perspectives in Medicine*, *6*(10), a025395. https://doi.org/10.1101/cshperspect.a025395

Galas, D., Chandler, M., Berg, D., & Howe, M. (1989). Bacterial insertion sequences. *American Society for Microbiology, Washington, DC*.

Galiez, C., Siebert, M., Enault, F., Vincent, J., Söding, J., Enault, F., Vincent, J., & Söding, J. (2017). WIsH: Who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics*, *33*(19), 3113–14. https://doi.org/10.1093/bioinformatics/btx383

Ganime, A. C., Carvalho-Costa, F. A., Mendonça, M. C. L., Vieira, C. B., Santos, M., Costa Filho, R., Miagostovich, M. P., & Leite, J. P. G. (2012). Group A rotavirus detection on environmental surfaces in a hospital intensive care unit. *American Journal of Infection Control*, *40*(6), 544–547. https://doi.org/10.1016/j.ajic.2011.07.017

Gao, R., Zhu, Y., Kong, C., Xia, K., Li, H., Zhu, Y., Zhang, X., Liu, Y., Zhong, H., Yang, R., Chen, C., Qin, N., & Qin, H. (2021). Alterations, Interactions, and Diagnostic Potential of Gut Bacteria and Viruses in Colorectal Cancer. *Frontiers in Cellular and Infection Microbiology*, *11*. Retrieved February 14,

2023, from https://www.frontiersin.org/articles/10.3389/fcimb.2021.657867

García-López, R., Vázquez-Castellanos, J. F., & Moya, A. (2015). Fragmentation and Coverage Variation in Viral Metagenome Assemblies, and Their Effect in Diversity Calculations. *Frontiers in Bioengineering and Biotechnology*, *3*, 141. https://doi.org/10.3389/fbioe.2015.00141

Garretto, A., Hatzopoulos, T., & Putonti, C. (2019). virMine: Automated detection of viral sequences from complex metagenomic samples. *PeerJ*, *7*, e6695. https://doi.org/10.7717/peerj.6695

Gay, K., & Stephens, D. S. (2001). Structure and Dissemination of a Chromosomal Insertion Element Encoding Macrolide Efflux in Streptococcus pneumoniae. *The Journal of Infectious Diseases*, *184*(1), 56–65. https://doi.org/10.1086/321001

Ghosh, D., Roy, K., Williamson, K. E., Srinivasiah, S., Wommack, K. E., & Radosevich, M. (2009). Acyl-homoserine lactones can induce virus production in lysogenic bacteria: An alternative paradigm for prophage induction. *Applied and Environmental Microbiology*, *75*(22), 7142–7152. https://doi.org/10.1128/AEM.00950-09

Ghoshal, U. C., Shukla, R., Ghoshal, U., Gwee, K.-A., Ng, S. C., & Quigley, E. M. M. (2012). The gut microbiota and irritable bowel syndrome: Friend or foe? *International Journal of Inflammation*, *2012*, 151085. https://doi.org/10.1155/2012/151085

Ghurye, J. S., Cepeda-Espinoza, V., & Pop, M. (2016). Metagenomic Assembly: Overview, Challenges and Applications. *The Yale Journal of Biology and*

Medicine, *89*(3), 353–362. Retrieved May 3, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5045144/

Gilchrist, C. L. M., & Chooi, Y.-H. (2021). Clinker & clustermap.js: Automatic generation of gene cluster comparison figures. *Bioinformatics*, *37*(16), 2473–2475. https://doi.org/10.1093/bioinformatics/btab007

Gill, D. M., Uchida, T., & Singer, R. A. (1972). Expression of diphtheria toxin genes carried by integrated and nonintegrated phage beta. *Virology*, *50*(3), 664–668. https://doi.org/10.1016/0042-6822(72)90420-5

Girbovan, A., Sur, G., Samasca, G., & Lupan, I. (2017). Dysbiosis a risk factor for celiac disease. *Medical Microbiology and Immunology*, *206*(2), 83–91. https://doi.org/10.1007/s00430-017-0496-z

Göller, P. C., Haro-Moreno, J. M., Rodriguez-Valera, F., Loessner, M. J., & Gómez-Sanz, E. (2020). Uncovering a hidden diversity: Optimized protocols for the extraction of dsDNA bacteriophages from soil. *Microbiome*, *8*(1), 17. https://doi.org/10.1186/s40168-020-0795-2

Gómez, P., Bennie, J., Gaston, K. J., & Buckling, A. (2015). The Impact of Resource Availability on Bacterial Resistance to Phages in Soil [Publisher: Public Library of Science]. *PLOS ONE*, *10*(4), e0123752. https://doi.org/10.1371/journal.pone.0123752

Gorbalenya, A. E., Krupovic, M., Mushegian, A., Kropinski, A. M., Siddell, S. G., Varsani, A., Adams, M. J., Davison, A. J., Dutilh, B. E., Harrach, B., Harrison, R. L., Junglen, S., King, A. M. Q., Knowles, N. J., Lefkowitz, E. J., Nibert, M. L., Rubino, L., Sabanadzovic, S., Sanfaçon, H., … International Committee on Taxonomy of Viruses Executive Committee. (2020). The new scope of virus taxonomy:

Partitioning the virosphere into 15 hierarchical ranks [Number: 5 Publisher: Nature Publishing Group]. *Nature Microbiology*, *5*(5), 668–674. https://doi.org/10.1038/s41564-020-0709-x

Gordon, H. A., & Bruckner-Kardoss, E. (1961). Effect of normal microbial flora on intestinal surface area. *The American Journal of Physiology*, *201*, 175–178. https://doi.org/10.1152/ajplegacy.1961.201.1.175

Grazziotin, A. L., Koonin, E. V., & Kristensen, D. M. (2017). Prokaryotic Virus Orthologous Groups (pVOGs): A resource for comparative genomics and protein family annotation [Number: Database issue]. *Nucleic Acids Research*, *45*(Database issue), D491–D498. https://doi.org/10.1093/nar/gkw975

Gregory, A. C., Zablocki, O., Zayed, A. A., Howell, A., Bolduc, B., & Sullivan, M. B. (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

Gregory, A. C., Zayed, A. A., Conceição-Neto, N., Temperton, B., Bolduc, B., Alberti, A., Ardyna, M., Arkhipova, K., Carmichael, M., Cruaud, C., Dimier, C., Domínguez-Huerta, G., Ferland, J., Kandels, S., Liu, Y., Marec, C., Pesant, S., Picheral, M., Pisarev, S., ... Wincker, P. (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

Grindley, N. D. (2002). The Movement of Tn3-Like Elements: Transposition and Cointegrate Resolution [Publisher: Washington, DC: American Society for Microbiology, 1916-]. *Journal of Bacteriology*, *184*(12).

Grootjans, J., Lenaerts, K., Derikx, J. P., Matthijsen, R. A., de Bruïne, A. P., van Bijnen, A. A., van Dam, R. M., Dejong, C. H., & Buurman, W. A. (2010). Human Intestinal Ischemia-Reperfusion–Induced Inflammation Characterized. *The American Journal of Pathology*, *176*(5), 2283–2291. https://doi.org/10.2353/ajpath.2010.091069

Guerrero-Ferreira, R. C., Viollier, P. H., Ely, B., Poindexter, J. S., Georgieva, M., Jensen, G. J., & Wright, E. R. (2011). Alternative mechanism for bacteriophage adsorption to the motile bacterium Caulobacter crescentus [Publisher: Proceedings of the National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *108*(24), 9963–9968. https://doi.org/10.1073/pnas.1012388108

Guo, J., Bolduc, B., Zayed, A. A., Varsani, A., Dominguez-Huerta, G., Delmont, T. O., Pratama, A. A., Gazitúa, M. C., Vik, D., Sullivan, M. B., & Roux, S. (2021). VirSorter2: A multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses [Number: 1]. *Microbiome*, *9*(1), 37. https://doi.org/10.1186/s40168-020-00990-y

Gupta, V. K., Paul, S., & Dutta, C. (2017). Geography, Ethnicity or Subsistence-Specific Variations in Human Microbiome Composition and Diversity. *Frontiers in Microbiology*, *8*. Retrieved February 14, 2023, from https://www.frontiersin.org/articles/10.3389/fmicb.2017.01162

Gyles, C. L. (2007). Shiga toxin-producing Escherichia coli: An overview. *Journal of Animal Science*, *85*(13 Suppl), E45–62. https://doi.org/10.2527/jas.2006-508

Hammami, R., Fernandez, B., Lacroix, C., & Fliss, I. (2013). Anti-infective properties of bacteriocins: An update. *Cellular and molecular life sciences: CMLS*, *70*(16), 2947–2967. https://doi.org/10.1007/s00018-012-1202-3

Han, B., Sivaramakrishnan, P., Lin, C.-C. J., Neve, I. A., He, J., Tay, L. W. R., Sowa, J. N., Sizovs, A., Du, G., Wang, J., Herman, C., & Wang, M. C. (2017). Microbial Genetic Composition Tunes Host Longevity. *Cell*, *169*(7), 1249–1262.e13. https://doi.org/10.1016/j.cell.2017.05.036

Handal, T., Giraud-Morin, C., Caugant, D. A., Madinier, I., Olsen, I., & Fosse, T. (2005). Chromosome- and Plasmid-Encoded -Lactamases in Capnocytophaga spp. *Antimicrobial Agents and Chemotherapy*, *49*(9), 3940–3943. https://doi.org/10.1128/AAC.49.9.3940-3943.2005

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chemistry & Biology*, *5*(10), R245–R249. https://doi.org/10.1016/S1074-5521(98)90108-9

Harmer, C. J., Moran, R. A., & Hall, R. M. (2014). Movement of IS26-Associated Antibiotic Resistance Genes Occurs via a Translocatable Unit That Includes a Single IS26 and Preferentially Inserts Adjacent to Another IS26 [Publisher: American Society for Microbiology]. *mBio*, *5*(5), 10.1128/mbio.01801–14. https://doi.org/10.1128/mbio.01801-14

Harris, C. R., Millman, K. J., Walt, S. J. v. d., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Kerkwijk, M. H. v., Brett, M., Haldane, A., Río, J. F. d., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy [Publisher: Springer Science and Business

Media LLC]. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hasegawa, M., Osaka, T., Tawaratsumida, K., Yamazaki, T., Tada, H., Chen, G. Y., Tsuneda, S., Núñez, G., & Inohara, N. (2010). Transitions in oral and intestinal microflora composition and innate immune receptor-dependent stimulation during mouse development. *Infection and Immunity*, *78*(2), 639–650. https://doi.org/10.1128/IAI.01043-09

Hay, I. D., & Lithgow, T. (2019). Filamentous phages: Masters of a microbial sharing economy. *EMBO reports*, *20*(6), e47427. https://doi.org/10.15252/embr.201847427

Hegarty, B., V, J. R., Bastien, E., Langenfeld, K., Lindback, M., Saini, J. S., Wing, A., Zhang, J., & Duhaime, M. (2023). Benchmarking informatics approaches for virus discovery: Caution is needed when combining in silico identification methods [Pages: 2023.08.07.552334 Section: New Results]. https://doi.org/10.1101/2023.08.07.552334

Hegstad, K., Mikalsen, T., Coque, T. M., Werner, G., & Sundsfjord, A. (2010). Mobile genetic elements and their contribution to the emergence of antimicrobial resistant Enterococcus faecalis and Enterococcus faecium. *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases*, *16*(6), 541–554. https://doi.org/10.1111/j.1469-0691.2010.03226.x

Hildebrand, F., Moitinho-Silva, L., Blasche, S., Jahn, M. T., Gossmann, T. I., Huerta-Cepas, J., Hercog, R., Luetge, M., Bahram, M., Pryszlak, A., Alves, R. J., Waszak, S. M., Zhu, A., Ye, L., Costea, P. I., Aalvink, S., Belzer, C., Forslund, S. K., Sunagawa, S., … Bork, P. (2019). Antibiotics-induced

monodominance of a novel gut bacterial order [Publisher: BMJ Publishing Group Section: Gut microbiota]. *Gut*, *68*(10), 1781–1790. https://doi.org/10.1136/gutjnl-2018-317715

Hill, C. J., Lynch, D. B., Murphy, K., Ulaszewska, M., Jeffery, I. B., O'Shea, C. A., Watkins, C., Dempsey, E., Mattivi, F., Tuohy, K., Ross, R. P., Ryan, C. A., O' Toole, P. W., & Stanton, C. (2017). Evolution of gut microbiota composition from birth to 24 weeks in the INFANTMET Cohort. *Microbiome*, *5*(1), 4. https://doi.org/10.1186/s40168-016-0213-y

Hill, J. E., Penny, S. L., Crowell, K. G., Goh, S. H., & Hemmingsen, S. M. (2004). cpnDB: A Chaperonin Sequence Database. *Genome Research*, *14*(8), 1669–1675. https://doi.org/10.1101/gr.2649204

Hillary, L. S., Adriaenssens, E. M., Jones, D. L., & McDonald, J. E. (2022). RNA-viromics reveals diverse communities of soil RNA viruses with the potential to affect grassland ecosystems across multiple trophic levels [Number: 1 Publisher: Nature Publishing Group]. *ISME Communications*, *2*(1), 1–10. https://doi.org/10.1038/s43705-022-00110-x

Holmfeldt, K., Middelboe, M., Nybroe, O., & Riemann, L. (2007). Large Variabilities in Host Strain Susceptibility and Phage Host Range Govern Interactions between Lytic Marine Phages and Their Flavobacterium Hosts. *Applied and Environmental Microbiology*, *73*(21), 6730–6739. https://doi.org/10.1128/AEM.01399-07

Hordijk, J., Bosman, A. B., van Essen-Zandbergen, A., Veldman, K., Dierikx, C., Wagenaar, J. A., & Mevius, D. (2011). qnrB19 gene bracketed by IS26 on a 40-kilobase IncR plasmid from an Escherichia coli isolate from a veal calf.

Antimicrobial Agents and Chemotherapy, *55*(1), 453–454. https://doi.org/10.1128/AAC.00866-10

Howard-Varona, C., Hargreaves, K. R., Solonenko, N. E., Markillie, L. M., White, R. A., Brewer, H. M., Ansong, C., Orr, G., Adkins, J. N., & Sullivan, M. B. (2018). Multiple mechanisms drive phage infection efficiency in nearly identical hosts [Number: 6 Publisher: Nature Publishing Group]. *The ISME Journal*, *12*(6), 1605–1618. https://doi.org/10.1038/s41396-018-0099-8

Hoyles, L., McCartney, A. L., Neve, H., Gibson, G. R., Sanderson, J. D., Heller, K. J., & van Sinderen, D. (2014). Characterization of virus-like particles associated with the human faecal and caecal microbiota [Number: 10]. *Research in Microbiology*, *165*(10), 803–812. https://doi.org/10.1016/j.resmic.2014.10.006

Hsieh, S.-Y., Tariq, M. A., Telatin, A., Ansorge, R., Adriaenssens, E. M., Savva, G. M., Booth, C., Wileman, T., Hoyles, L., & Carding, S. R. (2021). Comparison of PCR versus PCR-Free DNA Library Preparation for Characterising the Human Faecal Virome. *Viruses*, *13*(10), 2093. https://doi.org/10.3390/v13102093

Huang, D., Yu, P., Ye, M., Schwarz, C., Jiang, X., & Alvarez, P. J. J. (2021). Enhanced mutualistic symbiosis between soil phages and bacteria with elevated chromium-induced environmental stress. *Microbiome*, *9*(1), 150. https://doi.org/10.1186/s40168-021-01074-1

Hunt, S., Green, J., & Artymiuk, P. J. (2010). Hemolysin E (HlyE, ClyA, SheA) and related toxins. *Advances in Experimental Medicine and Biology*, *677*, 116–126. https://doi.org/10.1007/978-1-4419-6327-7_10

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment [Publisher: IEEE COMPUTER SOC]. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Huseyin, C. E., O'Toole, P. W., Cotter, P. D., & Scanlan, P. D. (2017). Forgotten fungi—the gut mycobiome in human health and disease [Number: 4 Publisher: Oxford Academic]. *FEMS Microbiology Reviews*, *41*(4), 479–511. https://doi.org/10.1093/femsre/fuw047

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., Creasy, H. H., Earl, A. M., FitzGerald, M. G., Fulton, R. S., Giglio, M. G., Hallsworth-Pepin, K., Lobos, E. A., Madupu, R., Magrini, V., Martin, J. C., Mitreva, M., Muzny, D. M., Sodergren, E. J., ... The Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human microbiome [Number: 7402 Publisher: Nature Publishing Group]. *Nature*, *486*(7402), 207–214. https://doi.org/10.1038/nature11234

Huxley, R. R., Ansary-Moghaddam, A., Clifton, P., Czernichow, S., Parr, C. L., & Woodward, M. (2009). The impact of dietary and lifestyle risk factors on risk of colorectal cancer: A quantitative overview of the epidemiological evidence. *International Journal of Cancer*, *125*(1), 171–180. https://doi.org/10.1002/ijc.24343

Iapichino, G., Callegari, M. L., Marzorati, S., Cigada, M., Corbella, D., Ferrari, S., & Morelli, L. (2008). Impact of antibiotics on the gut microbiota of critically ill patients. *Journal of Medical Microbiology*, *57*(Pt 8), 1007–1014. https://doi.org/10.1099/jmm.0.47387-0

Illumina. (2023). Nextera XT DNA Library Prep Kit | Sequence small genomes, plasmids, cDNA. Retrieved July 19, 2023, from https://emea.illumina.com/products/by-type/sequencing-kits/library-prep-kits/nextera-xt-dna.html

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., Richardson, H., Quinlan, A. R., ... Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads [Number: 4 Publisher: Nature Publishing Group]. *Nature Biotechnology*, *36*(4), 338–345. https://doi.org/10.1038/nbt.4060

Jang, H. B., Bolduc, B., Zablocki, O., Kuhn, J. H., Roux, S., Adriaenssens, E. M., Brister, J. R., Kropinski, A. M., Krupovic, M., Turner, D., & Sullivan, M. B. (2019). Gene sharing networks to automate genome-based prokaryotic viral taxonomy [Publisher: Cold Spring Harbor Laboratory Section: New Results]. *bioRxiv*, 533240. https://doi.org/10.1101/533240

Jansson, J. K., & Wu, R. (2023). Soil viral diversity, ecology and climate change [Number: 5 Publisher: Nature Publishing Group]. *Nature Reviews Microbiology*, *21*(5), 296–311. https://doi.org/10.1038/s41579-022-00811-z

Jazani, N. H., Savoj, J., Lustgarten, M., Lau, W. L., & Vaziri, N. D. (2019). Impact of Gut Dysbiosis on Neurohormonal Pathways in Chronic Kidney Disease [Number: 1 Publisher: Multidisciplinary Digital Publishing Institute]. *Diseases*, *7*(1), 21. https://doi.org/10.3390/diseases7010021

Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., Lago, B. A., Dave, B. M., Pereira, S., Sharma, A. N., Doshi, S., Courtot, M., Lo, R., Williams, L. E., Frye, J. G., Elsayegh, T., Sardar, D., Westman, E. L.,

Pawlowski, A. C., ... McArthur, A. G. (2017). CARD 2017: Expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Research*, *45*(D1), D566–D573. https://doi.org/10.1093/nar/gkw1004

Jian, H., Liu, Y., Wang, X., Dong, X., & Zou, X. (2023). Akkermansia muciniphila as a Next-Generation Probiotic in Modulating Human Metabolic Homeostasis and Disease Progression: A Role Mediated by Gut–Liver–Brain Axes? *International Journal of Molecular Sciences*, *24*(4), 3900. https://doi.org/10.3390/ijms24043900

Johanson, W. G., Pierce, A. K., & Sanford, J. P. (1969). Changing Pharyngeal Bacterial Flora of Hospitalized Patients [Publisher: Massachusetts Medical Society]. *New England Journal of Medicine*, *281*(21), 1137–1140. https://doi.org/10.1056/NEJM196911202812101

Johnson, J. R., Owens, K., Gajewski, A., & Kuskowski, M. A. (2005). Bacterial Characteristics in Relation to Clinical Source of Escherichia coli Isolates from Women with Acute Cystitis or Pyelonephritis and Uninfected Women. *Journal of Clinical Microbiology*, *43*(12), 6064–6072. https://doi.org/10.1128/JCM.43.12.6064-6072.2005

Jung, C. Y., & Bae, J. M. (2020). Pathophysiology and protective approaches of gut injury in critical illness. *Yeungnam University Journal of Medicine*, *38*(1), 27–33. https://doi.org/10.12701/yujm.2020.00703

Jurtz, V. I., Villarroel, J., Lund, O., Voldby Larsen, M., & Nielsen, M. (2016). MetaPhinder—Identifying Bacteriophage Sequences in Metagenomic Data Sets [Number: 9]. *PLoS ONE*, *11*(9). https://doi.org/10.1371/journal.pone.0163111

Kang, D., Li, F., Kirton, E. S., Thomas, A., Egan, R. S., An, H., & Wang, Z. (2019). MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. https://doi.org/10.7287/peerj.preprints.27522v1

Kang, K., Imamovic, L., Misiakou, M.-A., Bornakke Sørensen, M., Heshiki, Y., Ni, Y., Zheng, T., Li, J., Ellabaan, M. M. H., Colomer-Lluch, M., Rode, A. A., Bytzer, P., Panagiotou, G., & Sommer, M. O. (2021). Expansion and persistence of antibiotic-specific resistance genes following antibiotic treatment [Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19490976.2021.1900995]. *Gut Microbes*, *13*(1), 1900995. https://doi.org/10.1080/19490976.2021.1900995

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

Kembel, S. W., Wu, M., Eisen, J. A., & Green, J. L. (2012). Incorporating 16S Gene Copy Number Information Improves Estimates of Microbial Diversity and Abundance [Publisher: Public Library of Science]. *PLOS Computational Biology*, *8*(10), e1002743. https://doi.org/10.1371/journal.pcbi.1002743

Kieft, K., Adams, A., Salamzade, R., Kalan, L., & Anantharaman, K. (2022). vRhyme enables binning of viral genomes from metagenomes. *Nucleic Acids Research*, *50*(14), e83. https://doi.org/10.1093/nar/gkac341

Kieft, K., & Anantharaman, K. (2022). Deciphering Active Prophages from Metagenomes [Publisher: American Society for Microbiology]. *mSystems*, *7*(2), e00084–22. https://doi.org/10.1128/msystems.00084-22

Kieft, K., Zhou, Z., & Anantharaman, K. (2020). VIBRANT: Automated recovery, annotation and curation of microbial viruses, and evaluation of

viral community function from genomic sequences [Number: 1]. *Microbiome*, *8*(1), 90. https://doi.org/10.1186/s40168-020-00867-0

King, A. M., Adams, M. J., Carstens, E. B., & Lefkowitz, E. J. (2012). Virus taxonomy [Publisher: Elsevier Academic Press San Diego]. *Ninth report of the International Committee on Taxonomy of Viruses*, *9*.

Kingsford, C., Schatz, M. C., & Pop, M. (2010). Assembly complexity of prokaryotic genomes using short reads [Number: 1]. *BMC Bioinformatics*, *11*(1), 21. https://doi.org/10.1186/1471-2105-11-21

Kitsios, G. D., Morowitz, M. J., Dickson, R. P., Huffnagle, G. B., McVerry, B. J., & Morris, A. (2017). Dysbiosis in the ICU: Microbiome science coming to the bedside. *Journal of critical care*, *38*, 84–91. https://doi.org/10.1016/j.jcrc.2016.09.029

Kleiner, M., Bushnell, B., Sanderson, K. E., Hooper, L. V., & Duerkop, B. A. (2020). Transductomics: Sequencing-based detection and analysis of transduced DNA in pure cultures and microbial communities. *Microbiome*, *8*, 158. https://doi.org/10.1186/s40168-020-00935-5

Kleiner, M., Hooper, L. V., & Duerkop, B. A. (2015). Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes. *BMC Genomics*, *16*(1), 7. https://doi.org/10.1186/s12864-014-1207-4

Kleiner, M., Thorson, E., Sharp, C. E., Dong, X., Liu, D., Li, C., & Strous, M. (2017). Assessing species biomass contributions in microbial communities via metaproteomics. *Nature Communications*, *8*. https://doi.org/10.1038/s41467-017-01544-x

Knowles, B., Silveira, C. B., Bailey, B. A., Barott, K., Cantu, V. A., Cobián-Güemes, A. G., Coutinho, F. H., Dinsdale, E. A., Felts, B.,

Furby, K. A., George, E. E., Green, K. T., Gregoracci, G. B., Haas, A. F., Haggerty, J. M., Hester, E. R., Hisakawa, N., Kelly, L. W., Lim, Y. W., ... Rohwer, F. (2016). Lytic to temperate switching of viral communities. *Nature*, *531*(7595), 466–470. https://doi.org/10.1038/nature17193

Kolmogorov, M., Rayko, M., Yuan, J., Polevikov, E., & Pevzner, P. (2019). metaFlye: Scalable long-read metagenome assembly using repeat graphs. *bioRxiv*, 637637. https://doi.org/10.1101/637637

Kostic, A. D., Gevers, D., Siljander, H., Vatanen, T., Hyötyläinen, T., Hämäläinen, A.-M., Peet, A., Tillmann, V., Pöhö, P., Mattila, I., Lähdesmäki, H., Franzosa, E. A., Vaarala, O., de Goffau, M., Harmsen, H., Ilonen, J., Virtanen, S. M., Clish, C. B., Orešič, M., ... Xavier, R. J. (2015). The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell Host & Microbe*, *17*(2), 260–273. https://doi.org/10.1016/j.chom.2015.01.001

Kouakou, J.-L., Gonedelé-Bi, S., Assamoi, J.-B., & Assanvo N'Guetta, S.-P. (2022). Optimization of the Cetyltrimethylammonium bromide (CTAB) DNA extraction protocol using forest elephant dung samples. *MethodsX*, *9*, 101867. https://doi.org/10.1016/j.mex.2022.101867

Kourilsky, P. (1973). Lysogenization by bacteriophage lambda. I. Multiple infection and the lysogenic response. *Molecular & general genetics: MGG*, *122*(2), 183–195. https://doi.org/10.1007/BF00435190

Krezalek, M. A., Skowron, K., Guyton, K. L., Shakhsheer, B., Hyoju, S., & Alverdy, J. C. (2016). The Intestinal Microbiome and Surgical Disease. *Current problems in surgery*, *53*(6), 257–293. https://doi.org/10.1067/j.cpsurg.2016.06.001

Krezalek, M. A., Yeh, A., Alverdy, J. C., & Morowitz, M. (2017). Influence of nutrition therapy on the intestinal microbiome. *Current Opinion in Clinical Nutrition and Metabolic Care*, *20*(2), 131–137. https://doi.org/10.1097/MCO.0000000000000348

Kristich, C. J., Rice, L. B., & Arias, C. A. (2014). Enterococcal Infection—Treatment and Antibiotic Resistance. In M. S. Gilmore, D. B. Clewell, Y. Ike, & N. Shankar (Eds.), *Enterococci: From Commensals to Leading Causes of Drug Resistant Infection*. Massachusetts Eye; Ear Infirmary. Retrieved October 11, 2018, from http://www.ncbi.nlm.nih.gov/books/NBK190420/

Krüger, A., & Lucchesi, P. M. A. (2015). Shiga toxins and stx phages: Highly diverse entities. *Microbiology (Reading, England)*, *161*(Pt 3), 451–462. https://doi.org/10.1099/mic.0.000003

Kutter, E., & Sulakvelidze, A. (2006). Bacteriophages: Biology and Applications [Publisher: The University of Chicago Press]. *The Quarterly Review of Biology*, *81*(3), 280–281. https://doi.org/10.1086/509419

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

Labonté, J. M., Swan, B. K., Poulos, B., Luo, H., Koren, S., Hallam, S. J., Sullivan, M. B., Woyke, T., Wommack, K. E., & Stepanauskas, R. (2015). Single-cell genomics-based analysis of virus–host interactions in marine surface bacterioplankton [Number: 11 Publisher: Nature Publishing Group]. *The ISME Journal*, *9*(11), 2386–2399. https://doi.org/10.1038/ismej.2015.48

Laforest-Lapointe, I., & Arrieta, M.-C. (2018). Microbial Eukaryotes: A Missing Link in Gut Microbiome Studies [Number: 2 Publisher: American Society for Microbiology Journals Section: Perspective]. *mSystems*, *3*(2). https://doi.org/10.1128/mSystems.00201-17

Lampejo, T., Curtis, C., Ijaz, S., Haywood, B., Flores, A., Sudhanva, M., El Bouzidi, K., Patel, S., Dowling, M., & Zuckerman, M. (2022). Nosocomial transmission of hepatitis E virus and development of chronic infection: The wider impact of COVID-19. *Journal of Clinical Virology*, *148*, 105083. https://doi.org/10.1016/j.jcv.2022.105083

Lang, A. S., Westbye, A. B., & Beatty, J. T. (2017). The Distribution, Evolution, and Roles of Gene Transfer Agents in Prokaryotic Genetic Exchange [_eprint: https://doi.org/10.1146/annurev-virology-101416-041624]. *Annual Review of Virology*, *4*(1), 87–104. https://doi.org/10.1146/annurev-virology-101416-041624

Lang, A. S., Zhaxybayeva, O., & Beatty, J. T. (2012). Gene transfer agents: Phage-like elements of genetic exchange. *Nature Reviews. Microbiology*, *10*(7), 472–482. https://doi.org/10.1038/nrmicro2802

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2 [Number: 4]. *Nature methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Lankelma, J. M., Cranendonk, D. R., Belzer, C., de Vos, A. F., de Vos, W. M., van der Poll, T., & Wiersinga, W. J. (2017). Antibiotic-induced gut microbiota disruption during human endotoxemia: A randomised controlled study. *Gut*, *66*(9), 1623–1630. https://doi.org/10.1136/gutjnl-2016-312132

Lara, E., Vaqué, D., Sà, E. L., Boras, J. A., Gomes, A., Borrull, E., Díez-Vives, C., Teira, E., Pernice, M. C., Garcia, F. C., Forn, I., Castillo, Y. M., Peiró, A., Salazar, G., Morán, X. A. G., Massana, R., Catalá, T. S., Luna, G. M., Agustí, S., … Duarte, C. M. (2017). Unveiling the role and life strategies of viruses from the surface to the dark ocean. *Science Advances*, *3*(9), e1602565. https://doi.org/10.1126/sciadv.1602565

Larralde, M., & Zeller, G. (2023). PyHMMER: A Python library binding to HMMER for efficient sequence analysis. *Bioinformatics*, *39*(5), btad214. https://doi.org/10.1093/bioinformatics/btad214

Larsen, N., Vogensen, F. K., van den Berg, F. W. J., Nielsen, D. S., Andreasen, A. S., Pedersen, B. K., Al-Soud, W. A., Sørensen, S. J., Hansen, L. H., & Jakobsen, M. (2010). Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PloS One*, *5*(2), e9085. https://doi.org/10.1371/journal.pone.0009085

Lavigne, R., Darius, P., Summer, E. J., Seto, D., Mahadevan, P., Nilsson, A. S., Ackermann, H. W., & Kropinski, A. M. (2009). Classification of Myoviridae bacteriophages using protein sequence similarity. *BMC Microbiology*, *9*, 224. https://doi.org/10.1186/1471-2180-9-224

Lavigne, R., Seto, D., Mahadevan, P., Ackermann, H.-W., & Kropinski, A. M. (2008). Unifying classical and molecular taxonomic classification: Analysis of the Podoviridae using BLASTP-based tools. *Research in Microbiology*, *159*(5), 406–414. https://doi.org/10.1016/j.resmic.2008.03.005

Leavitt, J. C., Gilcrease, E. B., Wilson, K., & Casjens, S. R. (2013). Function and horizontal transfer of the small terminase subunit of the tailed

bacteriophage Sf6 DNA packaging nanomotor. *Virology*, *440*(2), 117–133. https://doi.org/10.1016/j.virol.2013.02.023

Leclercq, R., Derlot, E., Duval, J., & Courvalin, P. (1988). Plasmid-mediated resistance to vancomycin and teicoplanin in Enterococcus faecium. *The New England Journal of Medicine*, *319*(3), 157–161. https://doi.org/10.1056/NEJM198807213190307

Lee, C. Z., Zoqratt, M. Z. H. M., Phipps, M. E., Barr, J. J., Lal, S. K., Ayub, Q., & Rahman, S. (2022). The gut virome in two indigenous populations from Malaysia [Number: 1 Publisher: Nature Publishing Group]. *Scientific Reports*, *12*(1), 1824. https://doi.org/10.1038/s41598-022-05656-3

Lefkowitz, E. J., Dempsey, D. M., Hendrickson, R. C., Orton, R. J., Siddell, S. G., & Smith, D. B. (2018). Virus taxonomy: The database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Research*, *46*(D1), D708–D717. https://doi.org/10.1093/nar/gkx932

Lepage, P., Häsler, R., Spehlmann, M. E., Rehman, A., Zvirbliene, A., Begun, A., Ott, S., Kupcinskas, L., Doré, J., Raedler, A., & Schreiber, S. (2011). Twin study indicates loss of interaction between microbiota and mucosa of patients with ulcerative colitis. *Gastroenterology*, *141*(1), 227–236. https://doi.org/10.1053/j.gastro.2011.04.011

Lévesque, C., Brassard, S., Lapointe, J., & Roy, P. H. (1994). Diversity and relative strength of tandem promoters for the antibiotic-resistance genes of several integrons. *Gene*, *142*(1), 49–54. https://doi.org/10.1016/0378-1119(94)90353-0

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools [Number: 16]. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Liang, W., Feng, Z., Rao, S., Xiao, C., Xue, X., Lin, Z., Zhang, Q., & Qi, W. (2020). Diarrhoea may be underestimated: A missing link in 2019 novel coronavirus. *Gut*, *69*(6), 1141–1143. https://doi.org/10.1136/gutjnl-2020-320832

Lillehaug, D. (1997). An improved plaque assay for poor plaque-producing temperate lactococcal bacteriophages [Number: 1 _eprint: https://sfamjournals.onlinelibrary.wiley.com/doi/pdf/10.1046/j.1365-2672.1997.00193.x]. *Journal of Applied Microbiology*, *83*(1), 85–90. https://doi.org/10.1046/j.1365-2672.1997.00193.x

Lim, E. S., Zhou, Y., Zhao, G., Bauer, I. K., Droit, L., Ndao, I. M., Warner, B. B., Tarr, P. I., Wang, D., & Holtz, L. R. (2015). Early life dynamics of the human gut virome and bacterial microbiome in infants [Number: 10 Publisher: Nature Publishing Group]. *Nature Medicine*, *21*(10), 1228–1234. https://doi.org/10.1038/nm.3950

Lima-Mendez, G., Van Helden, J., Toussaint, A., & Leplae, R. (2008). Prophinder: A computational tool for prophage prediction in prokaryotic

genomes [Number: 6]. *Bioinformatics (Oxford, England)*, *24*(6), 863–865. https://doi.org/10.1093/bioinformatics/btn043

Lin, D. M., & Lin, H. C. (2019). A theoretical model of temperate phages as mediators of gut microbiome dysbiosis. *F1000Research*, *8*. https://doi.org/10.12688/f1000research.18480.1

Loman, N. J., Constantinidou, C., Christner, M., Rohde, H., Chan, J. Z.-M., Quick, J., Weir, J. C., Quince, C., Smith, G. P., Betley, J. R., Aepfelbacher, M., & Pallen, M. J. (2013). A Culture-Independent Sequence-Based Metagenomics Approach to the Investigation of an Outbreak of Shiga-Toxigenic Escherichia coli O104:H4 [Number: 14 Publisher: American Medical Association]. *JAMA*, *309*(14), 1502–1510. https://doi.org/10.1001/jama.2013.3231

Lombardo, L., Foti, M., Ruggia, O., & Chiecchio, A. (2010). Increased incidence of small intestinal bacterial overgrowth during proton pump inhibitor therapy. *Clinical Gastroenterology and Hepatology: The Official Clinical Practice Journal of the American Gastroenterological Association*, *8*(6), 504–508. https://doi.org/10.1016/j.cgh.2009.12.022

López-Causapé, C., Cabot, G., Del Barrio-Tofiño, E., & Oliver, A. (2018). The Versatile Mutational Resistome of Pseudomonas aeruginosa. *Frontiers in Microbiology*, *9*, 685. https://doi.org/10.3389/fmicb.2018.00685

Lu, C., Zhang, Z., Cai, Z., Zhu, Z., Qiu, Y., Wu, A., Jiang, T., Zheng, H., & Peng, Y. (2021). Prokaryotic virus Host Predictor: A Gaussian model for host prediction of prokaryotic viruses in metagenomics [Publisher: BMC Biology]. *BMC biology*, *19*(5), 1–11. https://doi.org/10.1101/2020.12.02.408310

Ludwig, A.-K., De Miroschedji, K., Doeppner, T. R., Börger, V., Ruesing, J., Rebmann, V., Durst, S., Jansen, S., Bremer, M., Behrmann, E., Singer, B. B., Jastrow, H., Kuhlmann, J. D., El Magraoui, F., Meyer, H. E., Hermann, D. M., Opalka, B., Raunser, S., Epple, M., … Giebel, B. (2018). Precipitation with polyethylene glycol followed by washing and pelleting by ultracentrifugation enriches extracellular vesicles from tissue culture supernatants in small and large scales. *Journal of Extracellular Vesicles*, *7*(1), 1528109. https://doi.org/10.1080/20013078.2018.1528109

Luo, Y., Lan, C., Li, H., Ouyang, Q., Kong, F., Wu, A., Ren, Z., Tian, G., Cai, J., Yu, B., He, J., & Wright, A.-D. G. (2022). Rational consideration of Akkermansia muciniphila targeting intestinal health: Advantages and challenges [Number: 1 Publisher: Nature Publishing Group]. *npj Biofilms and Microbiomes*, *8*(1), 1–11. https://doi.org/10.1038/s41522-022-00338-4

Lutgendorff, F., Akkermans, L. M. A., & Söderholm, J. D. (2008). The role of microbiota and probiotics in stress-induced gastro-intestinal damage. *Current Molecular Medicine*, *8*(4), 282–298. https://doi.org/10.2174/156652408784533779

Lwoff, A. (1953). Lysogeny [Number: 4]. *Bacteriological Reviews*, *17*(4), 269–337. Retrieved September 23, 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC180777/

Macfarlane, S., & Macfarlane, G. T. (2003). Regulation of short-chain fatty acid production. *The Proceedings of the Nutrition Society*, *62*(1), 67–72. https://doi.org/10.1079/PNS2002207

Maghini, D. G., Moss, E. L., Vance, S. E., & Bhatt, A. S. (2021). Improved high-molecular-weight DNA extraction, nanopore sequencing and metagenomic assembly from the human gut microbiome [Publisher: NIH Public Access]. *Nature protocols*, *16*(1), 458. https://doi.org/10.1038/s41596-020-00424-x

Mahillon, J., & Chandler, M. (1998). Insertion Sequences. *Microbiology and Molecular Biology Reviews*, *62*(3), 725–774. Retrieved August 11, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC98933/

Mahmoudabadi, G., & Phillips, R. (2018). A comprehensive and quantitative exploration of thousands of viral genomes (A. K. Chakraborty, Ed.) [Publisher: eLife Sciences Publications, Ltd]. *eLife*, *7*, e31955. https://doi.org/10.7554/eLife.31955

Makino, H., Kushiro, A., Ishikawa, E., Kubota, H., Gawad, A., Sakai, T., Oishi, K., Martin, R., Ben-Amor, K., Knol, J., & Tanaka, R. (2013). Mother-to-infant transmission of intestinal bifidobacterial strains has an impact on the early development of vaginally delivered infant's microbiota. *PloS One*, *8*(11), e78331. https://doi.org/10.1371/journal.pone.0078331

Mallawaarachchi, V., Roach, M. J., Papudeshi, B., Giles, S. K., Grigson, S. R., Decewicz, P., Bouras, G., Hesse, R. D., Inglis, L. K., Hutton, A. L. K., Dinsdale, E. A., & Edwards, R. A. (2023). Phables: From fragmented assemblies to high-quality bacteriophage genomes [Pages: 2023.04.04.535632 Section: New Results]. https://doi.org/10.1101/2023.04.04.535632

Mann, N. H., Cook, A., Millard, A., Bailey, S., & Clokie, M. (2003). Marine ecosystems: Bacterial photosynthesis genes in a virus. *Nature*, *424*(6950), 741. https://doi.org/10.1038/424741a

Manrique, P., Bolduc, B., Walk, S. T., Oost, J. v. d., Vos, W. M. d., & Young, M. J. (2016). Healthy human gut phageome [Number: 37 Publisher: National Academy of Sciences Section: Biological Sciences]. *Proceedings of the National Academy of Sciences*, *113*(37), 10400–10405. https://doi.org/10.1073/pnas.1601060113

Manzanilla, V. (2020). Guppy parameters and effect on final PHRED scores and number of reads? Retrieved May 23, 2023, from https://community.nanoporetech.com/posts/guppy-parameters-and-effec

Maqsood, R., Rodgers, R., Rodriguez, C., Handley, S. A., Ndao, I. M., Tarr, P. I., Warner, B. B., Lim, E. S., & Holtz, L. R. (2019). Discordant transmission of bacteria and viruses from mothers to babies at birth. *Microbiome*, *7*(1), 156. https://doi.org/10.1186/s40168-019-0766-7

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system [Number: 1 Publisher: Public Library of Science]. *PLOS Computational Biology*, *14*(1), e1005944. https://doi.org/10.1371/journal.pcbi.1005944

Markou, P., & Apidianakis, Y. (2014). Pathogenesis of intestinal Pseudomonas aeruginosa infection in patients with cancer. *Frontiers in Cellular and Infection Microbiology*, *3*, 115. https://doi.org/10.3389/fcimb.2013.00115

Marquet, M., Hölzer, M., Pletz, M. W., Viehweger, A., Makarewicz, O., Ehricht, R., & Brandt, C. (2020). What the Phage: A scalable workflow for the identification and analysis of phage sequences [Publisher: Cold Spring Harbor Laboratory Section: New Results]. *bioRxiv*, 2020.07.24.219899. https://doi.org/10.1101/2020.07.24.219899

Marquet, M., Zöllkau, J., Pastuschek, J., Viehweger, A., Schleußner, E., Makarewicz, O., Pletz, M. W., Ehricht, R., & Brandt, C. (2022). Evaluation of microbiome enrichment and host DNA depletion in human vaginal samples using Oxford Nanopore's adaptive sequencing. *Scientific Reports*, *12*, 4000. https://doi.org/10.1038/s41598-022-08003-8

Martel, A., Mans, C., Doss, G. A., & Williams, J. M. (2018). Effects of Midazolam and Midazolam-Butorphanol on Gastrointestinal Transit Time and Motility in Cockatiels (Nymphicus hollandicus). *Journal of Avian Medicine and Surgery*, *32*(4), 286–293. https://doi.org/10.1647/2017-266

Martin, J., Neurohr, C., Bauer, M., Weiss, M., & Schleppers, A. (2008). Cost of intensive care in a German hospital: Cost-unit accounting based on the InEK matrix. *Der Anaesthesist*, *57*(5), 505–512. https://doi.org/10.1007/s00101-008-1353-7

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads [Number: 1]. *EMBnet.journal*, *17*(1), 10–12. https://doi.org/10.14806/ej.17.1.200

Mathieu, A., Dion, M., Deng, L., Tremblay, D., Moncaut, E., Shah, S. A., Stokholm, J., Krogfelt, K. A., Schjørring, S., Bisgaard, H., Nielsen, D. S., Moineau, S., & Petit, M.-A. (2020). Virulent coliphages in 1-year-old children fecal samples are fewer, but more infectious than temperate coliphages [Number: 1 Publisher: Nature Publishing Group]. *Nature Communications*, *11*(1), 378. https://doi.org/10.1038/s41467-019-14042-z

Mazaheri Nezhad Fard, R., Barton, M. D., & Heuzenroeder, M. W. (2011). Bacteriophage-mediated transduction of antibiotic resistance in

enterococci [Number: 6]. *Letters in Applied Microbiology*, *52*(6), 559–564. https://doi.org/10.1111/j.1472-765X.2011.03043.x

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. v. d. Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a

McNair, K., Zhou, C., Dinsdale, E. A., Souza, B., & Edwards, R. A. (2019). PHANOTATE: A novel approach to gene identification in phage genomes. *Bioinformatics (Oxford, England)*, *35*(22), 4537–4542. https://doi.org/10.1093/bioinformatics/btz265

Menzel, P., Ng, K. L., & Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications*, *7*, 11257. https://doi.org/10.1038/ncomms11257

Messier, N., & Roy, P. H. (2001). Integron integrases possess a unique additional domain necessary for activity. *Journal of Bacteriology*, *183*(22), 6699–6706. https://doi.org/10.1128/JB.183.22.6699-6706.2001

Middelboe, M., & Lyck, P. G. (2002). Regeneration of dissolved organic matter by viral lysis in marine microbial communities. *Aquatic Microbial Ecology*, *27*(2), 187–194. https://doi.org/10.3354/ame027187

Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., Hingamp, P., Goto, S., & Ogata, H. (2016). Linking Virus Genomes with Host Taxonomy. *Viruses*, *8*(3), 66. https://doi.org/10.3390/v8030066

Mikheenko, A., Saveliev, V., & Gurevich, A. (2016). MetaQUAST: Evaluation of metagenome assemblies. *Bioinformatics*, *32*(7), 1088–1090. https://doi.org/10.1093/bioinformatics/btv697

Miki, T., Nakazawa, T., Yokokawa, T., & Nagata, T. (2008). Functional consequences of viral impacts on bacterial communities: A food-web model analysis [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2427.2007.01934.x]. *Freshwater Biology*, *53*(6), 1142–1153. https://doi.org/10.1111/j.1365-2427.2007.01934.x

Miller, R. V., & Day, M. J. (2008). Contribution of lysogeny, pseudolysogeny, and starvation to phage ecology. In S. T. Abedon (Ed.), *Bacteriophage Ecology: Population Growth, Evolution, and Impact of Bacterial Viruses* (pp. 114–144). Cambridge University Press. https://doi.org/10.1017/CBO9780511541483.008

Minot, S., Bryson, A., Chehoud, C., Wu, G. D., Lewis, J. D., & Bushman, F. D. (2013). Rapid evolution of the human gut virome [Number: 30 Publisher: National Academy of Sciences Section: Biological Sciences]. *Proceedings of the National Academy of Sciences*, *110*(30), 12450–12455. https://doi.org/10.1073/pnas.1300833110

Minot, S., Grunberg, S., Wu, G. D., Lewis, J. D., & Bushman, F. D. (2012). Hypervariable loci in the human gut virome. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(10), 3962–3966. https://doi.org/10.1073/pnas.1119061109

Minot, S., Sinha, R., Chen, J., Li, H., Keilbaugh, S. A., Wu, G. D., Lewis, J. D., & Bushman, F. D. (2011). The human gut virome: Inter-individual variation and dynamic response to diet [Number: 10 Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press

Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab]. *Genome Research*, *21*(10), 1616–1625. https://doi.org/10.1101/gr.122705.111

Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., & Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Research*, *41*(12), e121. https://doi.org/10.1093/nar/gkt263

Mittal, R., Debs, L. H., Patel, A. P., Nguyen, D., Patel, K., O'Connor, G., Grati, M., Mittal, J., Yan, D., Eshraghi, A. A., Deo, S. K., Daunert, S., & Liu, X. Z. (2017). Neurotransmitters: The critical modulators regulating gut-brain axis. *Journal of cellular physiology*, *232*(9), 2359–2372. https://doi.org/10.1002/jcp.25518

Modi, S. R., Lee, H. H., Spina, C. S., & Collins, J. J. (2013). Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature*, *499*(7457), 219–222. https://doi.org/10.1038/nature12212

Mohandas, S., & Pannaraj, P. S. (2020). Beyond the Bacterial Microbiome: Virome of Human Milk and Effects on the Developing Infant. *Nestle Nutrition Institute Workshop Series*, *94*, 86–93. https://doi.org/10.1159/000504997

Momose, Y., Hirayama, K., & Itoh, K. (2008). Competition for proline between indigenous Escherichia coli and E. coli O157:H7 in gnotobiotic mice associated with infant intestinal microbiota and its contribution to the colonization resistance against E. coli O157:H7. *Antonie van Leeuwenhoek*, *94*(2), 165–171. https://doi.org/10.1007/s10482-008-9222-6

Moron, R., Galvez, J., Colmenero, M., Anderson, P., Cabeza, J., & Rodriguez-Cabezas, M. E. (2019). The Importance of the Microbiome in Critically Ill Patients: Role of Nutrition. *Nutrients*, *11*(12), 3002. https://doi.org/10.3390/nu11123002

Morse, M. L., Lederberg, E. M., & Lederberg, J. (1956). Transduction in Escherichia Coli K-12 [Number: 1]. *Genetics*, *41*(1), 142–156. Retrieved May 6, 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1209761/

Mottawea, W., Chiang, C.-K., Mühlbauer, M., Starr, A. E., Butcher, J., Abujamel, T., Deeke, S. A., Brandel, A., Zhou, H., Shokralla, S., Hajibabaei, M., Singleton, R., Benchimol, E. I., Jobin, C., Mack, D. R., Figeys, D., & Stintzi, A. (2016). Altered intestinal microbiota-host mitochondria crosstalk in new onset Crohn's disease. *Nature Communications*, *7*, 13419. https://doi.org/10.1038/ncomms13419

Müller, M., Canfora, E. E., & Blaak, E. E. (2018). Gastrointestinal Transit Time, Glucose Homeostasis and Metabolic Health: Modulation by Dietary Fibers. *Nutrients*, *10*(3), 275. https://doi.org/10.3390/nu10030275

Na, O., Mw, W., Jr, B., S, C., D, H., R, M., B, R., B, R., B, S.-W., D, A.-A., A, A., A, B., Y, B., O, B., V, B., V, C., J, C., E, C., O, E., ... Kd, P. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation [ISSN: 1362-4962 Issue: D1 Publisher: Nucleic Acids Res Volume: 44]. https://doi.org/10.1093/nar/gkv1189

Naor, A., & Gophna, U. (2013). Cell fusion and hybrids in Archaea: Prospects for genome shuffling and accelerated strain development for biotechnology. *Bioengineered*, *4*(3), 126–129. https://doi.org/10.4161/bioe.22649

Nayfach, S., Camargo, A. P., Eloe-Fadrosh, E., Roux, S., & Kyrpides, N. (2020). CheckV: Assessing the quality of metagenome-assembled viral genomes [Publisher: Cold Spring Harbor Laboratory Section: New Results]. *bioRxiv*, 2020.05.06.081778. https://doi.org/10.1101/2020.05.06.081778

Nayfach, S., Páez-Espino, D., Call, L., Low, S. J., Sberro, H., Ivanova, N. N., Proal, A. D., Fischbach, M. A., Bhatt, A. S., Hugenholtz, P., & Kyrpides, N. C. (2021). Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome [Number: 7 Publisher: Nature Publishing Group]. *Nature Microbiology*, *6*(7), 960–970. https://doi.org/10.1038/s41564-021-00928-6

Nayfach, S., & Pollard, K. S. (2015). Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biology*, *16*(1), 51. https://doi.org/10.1186/s13059-015-0611-7

Nguyen, L. T., & Vogel, H. J. (2016). Staphylokinase has distinct modes of interaction with antimicrobial peptides, modulating its plasminogen-activation properties. *Scientific Reports*, *6*, 31817. https://doi.org/10.1038/srep31817

Nicholls, S. M., Quick, J. C., Tang, S., & Loman, N. J. (2018). Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *bioRxiv*. https://doi.org/10.1101/487033

Nishimura, Y., Watai, H., Honda, T., Mihara, T., Omae, K., Roux, S., Blanc-Mathieu, R., Yamamoto, K., Hingamp, P., Sako, Y., Sullivan, M. B., Goto, S., Ogata, H., & Yoshida, T. (2017). Environmental Viral Genomes Shed New Light on Virus-Host Interactions in the Ocean. *mSphere*, *2*(2), e00359–16. https://doi.org/10.1128/mSphere.00359-16

Nonaka, L., Maruyama, F., Miyamoto, M., Miyakoshi, M., Kurokawa, K., & Masuda, M. (2012). Novel Conjugative Transferable Multiple Drug Resistance Plasmid pAQU1 from *Photobacterium damselae* subsp. *damselae* Isolated from Marine Aquaculture Environment. *Microbes and Environments*, *27*(3), 263–272. https://doi.org/10.1264/jsme2.ME11338

Norman, A., Hansen, L. H., & Sørensen, S. J. (2009). Conjugative plasmids: Vessels of the communal gene pool [Number: 1527]. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1527), 2275–2289. https://doi.org/10.1098/rstb.2009.0037

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: A new versatile metagenomic assembler. *Genome Research*, *27*(5), 824–834. https://doi.org/10.1101/gr.213959.116

Oh, J. S., Davies, D. R., Lawson, J. D., Arnold, G. E., & Dunker, A. K. (1999). Isolation of chloroform-resistant mutants of filamentous phage: Localization in models of phage structure11Edited by M. F. Moody. *Journal of Molecular Biology*, *287*(3), 449–457. https://doi.org/10.1006/jmbi.1999.2599

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., & Wagner, H. (2020). Vegan: Community Ecology Package [Programmers: _:n2469]. Retrieved February 1, 2021, from https://CRAN.R-project.org/package=vegan

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biology*, *17*(1), 132. https://doi.org/10.1186/s13059-016-0997-x

Oppenheim, A. B., Kobiler, O., Stavans, J., Court, D. L., & Adhya, S. (2005). Switches in bacteriophage lambda development. *Annual Review of Genetics*, *39*, 409–429. https://doi.org/10.1146/annurev.genet.39.073003.113656

Oscarsson, J., Westermark, M., Beutin, L., & Uhlin, B. E. (2002). The bacteriophage-associated Ehly1 and Ehly2 determinants from Escherichia coli O26:H strains do not encode enterohemolysins per se but cause release of the ClyA cytolysin. *International Journal of Medical Microbiology*, *291*(8), 625–631. https://doi.org/10.1078/1438-4221-00176

Osmon, S., Ward, S., Fraser, V. J., & Kollef, M. H. (2004). Hospital mortality for patients with bacteremia due to Staphylococcus aureus or Pseudomonas aeruginosa. *Chest*, *125*(2), 607–616. https://doi.org/10.1378/chest.125.2.607

Oxford Nanopore Technologies. (2020). R10.3: The newest nanopore for high accuracy nanopore sequencing – now available in store [Library Catalog: nanoporetech.com Section: News]. Retrieved May 28, 2020, from http://nanoporetech.com/about-us/news/r103-newest-nanopore-high-accuracy-nanopore-sequencing-now-available-store

Oxford Nanopore Technologies. (2023a). Accuracy [Section: Flexible]. Retrieved March 21, 2023, from https://nanoporetech.com/accuracy

Oxford Nanopore Technologies. (2023b). Ligation Sequencing Kit. Retrieved July 19, 2023, from https://store.nanoporetech.com/ligation-sequencing-kit110.html

Oxford Nanopore Technologies. (2023c). Rapid Sequencing Kit. Retrieved July 19, 2023, from https://store.nanoporetech.com/rapid-sequencing-kit.html

Parks, D. H., Rigato, F., Vera-Wolf, P., Krause, L., Hugenholtz, P., Tyson, G. W., & Wood, D. L. A. (2021). Evaluation of the Microba Community Profiler for Taxonomic Profiling of Metagenomic Datasets From the Human Gut Microbiome. *Frontiers in Microbiology*, *12*. Retrieved May 5, 2023, from https://www.frontiersin.org/articles/10.3389/fmicb.2021.643682

Partridge, S. R., Recchia, G. D., Scaramuzzi, C., Collis, C. M., Stokes, H. W., & Hall, R. M. (2000). Definition of the attI1 site of class 1 integrons. *Microbiology (Reading, England)*, *146 (Pt 11)*, 2855–2864. https://doi.org/10.1099/00221287-146-11-2855

Pastores, S. M., Dakwar, J., & Halpern, N. A. (2012). Costs of critical care medicine. *Critical Care Clinics*, *28*(1), 1–10, v. https://doi.org/10.1016/j.ccc.2011.10.003

Pence, M. A. (2019). Antimicrobial Resistance in Clinically Important Anaerobes. *Clinical Microbiology Newsletter*, *41*(1), 1–7. https://doi.org/10.1016/j.clinmicnews.2018.12.003

Pfeifer, E., Bonnin, R. A., & Rocha, E. P. C. (2022). Phage-Plasmids Spread Antibiotic Resistance Genes through Infection and Lysogenic Conversion. *mBio*, *13*(5), e01851–22. https://doi.org/10.1128/mbio.01851-22

Pfeifer, E., Moura de Sousa, J. A., Touchon, M., & Rocha, E. P. C. (2021). Bacteria have numerous distinctive groups of phage–plasmids with conserved phage and variable plasmid gene repertoires. *Nucleic Acids Research*, *49*(5), 2655–2673. https://doi.org/10.1093/nar/gkab064

Pierce, N. T., Irber, L., Reiter, T., Brooks, P., & Brown, C. T. (2019). *Large-scale sequence comparisons with sourmash* (tech. rep. 8:1006) [Type: article]. F1000Research. https://doi.org/10.12688/f1000research.19675.1

Pons, J. C., Paez-Espino, D., Riera, G., Ivanova, N., Kyrpides, N. C., & Llabrés, M. (2021). VPF-Class: Taxonomic assignment and host prediction of uncultivated viruses based on viral protein families. *Bioinformatics*, *37*(13), 1805–1813. https://doi.org/10.1093/bioinformatics/btab026

Ponsero, A. J., & Hurwitz, B. L. (2019). The Promises and Pitfalls of Machine Learning for Detecting Viruses in Aquatic Metagenomes [Publisher: Frontiers]. *Frontiers in Microbiology*, *10*. https://doi.org/10.3389/fmicb.2019.00806

Poulsen, J. L., Nilsson, M., Brock, C., Sandberg, T. H., Krogh, K., & Drewes, A. M. (2016). The Impact of Opioid Treatment on Regional Gastrointestinal Transit. *Journal of Neurogastroenterology and Motility*, *22*(2), 282–291. https://doi.org/10.5056/jnm15175

Pratama, A. A., Bolduc, B., Zayed, A. A., Zhong, Z.-P., Guo, J., Vik, D. R., Gazitúa, M. C., Wainaina, J. M., Roux, S., & Sullivan, M. B. (2021). Expanding standards in viromics: In silico evaluation of dsDNA viral genome identification, classification, and auxiliary metabolic gene curation [Publisher: PeerJ Inc.]. *PeerJ*, *9*, e11447. https://doi.org/10.7717/peerj.11447

Pratama, A. A., & van Elsas, J. D. (2018). The 'Neglected' Soil Virome - Potential Role and Impact. *Trends in Microbiology*, *26*(8), 649–662. https://doi.org/10.1016/j.tim.2017.12.004

Prehn-Kristensen, A., Zimmermann, A., Tittmann, L., Lieb, W., Schreiber, S., Baving, L., & Fischer, A. (2018). Reduced microbiome alpha diversity in

young patients with ADHD. *PloS One*, *13*(7), e0200728. https://doi.org/10.1371/journal.pone.0200728

Prestel, E., Salamitou, S., & DuBow, M. S. (2008). An examination of the bacteriophages and bacteria of the Namib desert. *The Journal of Microbiology*, *46*(4), 364–372. https://doi.org/10.1007/s12275-008-0007-4

Ptashne, M., Jeffrey, A., Johnson, A. D., Maurer, R., Meyer, B. J., Pabo, C. O., Roberts, T. M., & Sauer, R. T. (1980). How the lambda repressor and cro work. *Cell*, *19*(1), 1–11. https://doi.org/10.1016/0092-8674(80)90383-9

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., ... Wang, J. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, *464*(7285), 59–65. https://doi.org/10.1038/nature08821

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Ramesh, V., Fralick, J. A., & Rolfe, R. D. (1999). Prevention of Clostridium difficile -induced ileocecitis with Bacteriophage. *Anaerobe*, *5*(2), 69–78. https://doi.org/10.1006/anae.1999.0192

Rao, V. B., & Black, L. W. (2013). DNA Packaging in Bacteriophage T4. *Madame Curie Bioscience Database [Internet]*. Landes Bioscience. Retrieved December 4, 2023, from https://www.ncbi.nlm.nih.gov/books/NBK6418/

Ravi, A., Halstead, F. D., Bamford, A., Casey, A., Thomson, N. M., van Schaik, W., Snelson, C., Goulden, R., Foster-Nyarko, E., Savva, G. M., Whitehouse, T., Pallen, M. J., & Oppenheim, B. A. (2019). Loss of microbial diversity and pathogen domination of the gut microbiota in critically ill patients [Number: 9 Publisher: Microbiology Society,]. *Microbial Genomics, 5*(9), e000293. https://doi.org/10.1099/mgen.0.000293

Reche, I., D'Orta, G., Mladenov, N., Winget, D. M., & Suttle, C. A. (2018). Deposition rates of viruses and bacteria above the atmospheric boundary layer. *The ISME Journal*, *12*(4), 1154–1162. https://doi.org/10.1038/s41396-017-0042-4

Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A., & Sun, F. (2017). VirFinder: A novel k-mer based tool for identifying viral sequences from assembled metagenomic data [Number: 1]. *Microbiome*, *5*(1), 69. https://doi.org/10.1186/s40168-017-0283-5

Ren, J., Song, K., Deng, C., Ahlgren, N. A., Fuhrman, J. A., Li, Y., Xie, X., Poplin, R., & Sun, F. (2020). Identifying viruses from metagenomic data using deep learning [Number: 1]. *Quantitative Biology*, *8*(1), 64–77. https://doi.org/10.1007/s40484-019-0187-4

Reyes, A., Haynes, M., Hanson, N., Angly, F. E., Heath, A. C., Rohwer, F., & Gordon, J. I. (2010). Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature*, *466*(7304), 334–338. https://doi.org/10.1038/nature09199

Reyes, A., Wu, M., McNulty, N. P., Rohwer, F. L., & Gordon, J. I. (2013). Gnotobiotic mouse model of phage–bacterial host dynamics in the human gut [Number: 50 Publisher: National Academy of Sciences Section:

Biological Sciences]. *Proceedings of the National Academy of Sciences*, *110*(50), 20236–20241. https://doi.org/10.1073/pnas.1319470110

Ripp, S., & Miller, R. V. (1997). The role of pseudolysogeny in bacteriophage-host interactions in a natural freshwater environment [Number: 6 Publisher: Microbiology Society,]. *Microbiology, 143*(6), 2065–2070. https://doi.org/10.1099/00221287-143-6-2065

Roach, M. J., McNair, K., Michalczyk, M., Giles, S. K., Inglis, L. K., Pargin, E., Barylski, J., Roux, S., Decewicz, P., & Edwards, R. A. (2022). Philympics 2021: Prophage Predictions Perplex Programs [version 2; peer review: 1 approved, 1 approved with reservations].

Roberts, L. M., & Dunker, A. K. (1993). Structural changes accompanying chloroform-induced contraction of the filamentous phage fd. *Biochemistry*, *32*(39), 10479–10488. https://doi.org/10.1021/bi00090a026

Roberts, M. C., Sutcliffe, J., Courvalin, P., Jensen, L. B., Rood, J., & Seppala, H. (1999). Nomenclature for Macrolide and Macrolide-Lincosamide-Streptogramin B Resistance Determinants. *Antimicrobial Agents and Chemotherapy*, *43*(12), 2823–2830. Retrieved December 20, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC89572/

Rodrigues, V. F., Elias-Oliveira, J., Pereira, Í. S., Pereira, J. A., Barbosa, S. C., Machado, M. S. G., & Carlos, D. (2022). Akkermansia muciniphila and Gut Immune System: A Good Friendship That Attenuates Inflammatory Bowel Disease, Obesity, and Diabetes. *Frontiers in Immunology*, *13*, 934695. https://doi.org/10.3389/fimmu.2022.934695

Rodríguez-Rubio, L., Serna, C., Ares-Arroyo, M., Matamoros, B. R., Delgado-Blas, J. F., Montero, N., Bernabe-Balas, C., Wedel, E. F., Mendez, I. S., Muniesa, M., & Gonzalez-Zorn, B. (2020). Extensive antimicrobial resistance mobilization via multicopy plasmid encapsidation mediated by temperate phages. *Journal of Antimicrobial Chemotherapy*, *75*(11), 3173–3180. https://doi.org/10.1093/jac/dkaa311

Rodriguez-Valera, F., Martin-Cuadrado, A.-B., Rodriguez-Brito, B., Pasić, L., Thingstad, T. F., Rohwer, F., & Mira, A. (2009). Explaining microbial population genomics through phage predation. *Nature Reviews. Microbiology*, *7*(11), 828–836. https://doi.org/10.1038/nrmicro2235

Rohwer, F., & Edwards, R. (2002). The Phage Proteomic Tree: A Genome-Based Taxonomy for Phage. *Journal of Bacteriology*, *184*(16), 4529–4535. https://doi.org/10.1128/JB.184.16.4529-4535.2002

Rohwer, F., & Thurber, R. V. (2009). Viruses manipulate the marine environment [Number: 7244 Publisher: Nature Publishing Group]. *Nature*, *459*(7244), 207–212. https://doi.org/10.1038/nature08060

Rokney, A., Kobiler, O., Amir, A., Court, D. L., Stavans, J., Adhya, S., & Oppenheim, A. B. (2008). Host responses influence on the induction of lambda prophage [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2958.2008.06119.x]. *Molecular Microbiology*, *68*(1), 29–36. https://doi.org/10.1111/j.1365-2958.2008.06119.x

Romanowski, K., Zaborin, A., Valuckaite, V., Rolfes, R. J., Babrowski, T., Bethel, C., Olivas, A., Zaborina, O., & Alverdy, J. C. (2012). Candida albicans isolates from the gut of critically ill patients respond to phosphate limitation

by expressing filaments and a lethal phenotype. *PloS One*, *7*(1), e30119. https://doi.org/10.1371/journal.pone.0030119

Rosner, J. L. (1972). Formation, induction, and curing of bacteriophage P1 lysogens. *Virology*, *48*(3), 679–689. https://doi.org/10.1016/0042-6822(72)90152-3

Roux, S., Adriaenssens, E. M., Dutilh, B. E., Koonin, E. V., Kropinski, A. M., Krupovic, M., Kuhn, J. H., Lavigne, R., Brister, J. R., Varsani, A., Amid, C., Aziz, R. K., Bordenstein, S. R., Bork, P., Breitbart, M., Cochrane, G. R., Daly, R. A., Desnues, C., Duhaime, M. B., … Eloe-Fadrosh, E. A. (2019a). Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nature Biotechnology*, *37*(1), 29–37. https://doi.org/10.1038/nbt.4306

Roux, S., Camargo, A. P., Coutinho, F. H., Dabdoub, S. M., Dutilh, B. E., Nayfach, S., & Tritt, A. (2023). iPHoP: An integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria [Publisher: Public Library of Science]. *PLOS Biology*, *21*(4), e3002083. https://doi.org/10.1371/journal.pbio.3002083

Roux, S., & Emerson, J. B. (2022). Diversity in the soil virosphere: To infinity and beyond? *Trends in Microbiology*, *30*(11), 1025–1035. https://doi.org/10.1016/j.tim.2022.05.003

Roux, S., Emerson, J. B., Eloe-Fadrosh, E. A., & Sullivan, M. B. (2017). Benchmarking viromics: An in silico evaluation of metagenome-enabled estimates of viral community composition and diversity [Publisher: PeerJ Inc.]. *PeerJ*, *5*, e3817. https://doi.org/10.7717/peerj.3817

Roux, S., Enault, F., Hurwitz, B. L., & Sullivan, M. B. (2015). VirSorter: Mining viral signal from microbial genomic data. *PeerJ*, *3*, e985. https://doi.org/10.7717/peerj.985

Roux, S., Krupovic, M., Daly, R. A., Borges, A. L., Nayfach, S., Schulz, F., Sharrar, A., Matheus Carnevali, P. B., Cheng, J.-F., Ivanova, N. N., Bondy-Denomy, J., Wrighton, K. C., Woyke, T., Visel, A., Kyrpides, N. C., & Eloe-Fadrosh, E. A. (2019b). Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. *Nature Microbiology*, *4*(11), 1895–1906. https://doi.org/10.1038/s41564-019-0510-x

Roux, S., Tournayre, J., Mahul, A., Debroas, D., & Enault, F. (2014). Metavir 2: New tools for viral metagenome comparison and assembled virome analysis [Number: 1]. *BMC Bioinformatics*, *15*(1), 76. https://doi.org/10.1186/1471-2105-15-76

RStudio Team. (2020). *RStudio: Integrated Development Environment for R*. RStudio, PBC. http://www.rstudio.com/

Rutsch, A., Kantsjö, J. B., & Ronchi, F. (2020). The Gut-Brain Axis: How Microbiota and Host Inflammasome Influence Brain Physiology and Pathology. *Frontiers in Immunology*, *11*. Retrieved February 14, 2023, from https://www.frontiersin.org/articles/10.3389/fimmu.2020.604179

Ruzin, A., Lindsay, J., & Novick, R. P. (2001). Molecular genetics of SaPI1–a mobile pathogenicity island in Staphylococcus aureus. *Molecular Microbiology*, *41*(2), 365–377. https://doi.org/10.1046/j.1365-2958.2001.02488.x

Santos, F., Wegkamp, A., de Vos, W. M., Smid, E. J., & Hugenholtz, J. (2008). High-Level folate production in fermented foods by the B12 producer

Lactobacillus reuteri JCM1112. *Applied and Environmental Microbiology*, *74*(10), 3291–3294. https://doi.org/10.1128/AEM.02719-07

Santos-Medellin, C., Zinke, L. A., ter Horst, A. M., Gelardi, D. L., Parikh, S. J., & Emerson, J. B. (2021). Viromes outperform total metagenomes in revealing the spatiotemporal patterns of agricultural soil viral communities [Number: 7 Publisher: Nature Publishing Group]. *The ISME Journal*, *15*(7), 1956–1970. https://doi.org/10.1038/s41396-021-00897-y

Satoh-Takayama, N., Vosshenrich, C. A. J., Lesjean-Pottier, S., Sawa, S., Lochner, M., Rattis, F., Mention, J.-J., Thiam, K., Cerf-Bensussan, N., Mandelboim, O., Eberl, G., & Di Santo, J. P. (2008). Microbial flora drives interleukin 22 production in intestinal NKp46+ cells that provide innate mucosal immune defense. *Immunity*, *29*(6), 958–970. https://doi.org/10.1016/j.immuni.2008.11.001

Sausset, R., Petit, M. A., Gaboriau-Routhiau, V., & De Paepe, M. (2020). New insights into intestinal phages. *Mucosal Immunology*, *13*(2), 205–215. https://doi.org/10.1038/s41385-019-0250-5

Sava, I. G., Heikens, E., & Huebner, J. (2010). Pathogenesis and immunity in enterococcal infections. *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases*, *16*(6), 533–540. https://doi.org/10.1111/j.1469-0691.2010.03213.x

Schackart, K. E., Graham, J. B., Ponsero, A. J., & Hurwitz, B. L. (2023). Evaluation of computational phage detection tools for metagenomic datasets. *Frontiers in Microbiology*, *14*. Retrieved January 25, 2024, from https://www.frontiersin.org/articles/10.3389/fmicb.2023.1078760

Schimel, J. P. (2018). Life in Dry Soils: Effects of Drought on Soil Microbial Communities and Processes [_eprint: https://doi.org/10.1146/annurev-ecolsys-110617-062614]. *Annual Review of Ecology, Evolution, and Systematics*, *49*(1), 409–432. https://doi.org/10.1146/annurev-ecolsys-110617-062614

Schleper, C., Holz, I., Janekovic, D., Murphy, J., & Zillig, W. (1995). A multicopy plasmid of the extremely thermophilic archaeon Sulfolobus effects its transfer to recipients by mating. *Journal of Bacteriology*, *177*(15), 4417–4426. https://doi.org/10.1128/jb.177.15.4417-4426.1995

Schmieder, R., & Edwards, R. (2012). Insights into antibiotic resistance through metagenomic approaches [Publisher: Future Medicine]. *Future Microbiology*, *7*(1), 73–89. https://doi.org/10.2217/fmb.11.135

Schmieger, H. (1982). Packaging signals for phage P22 on the chromosome of Salmonella typhimurium [Number: 3]. *Molecular & general genetics: MGG*, *187*(3), 516–518. https://doi.org/10.1007/BF00332637

Schneider, C. L. (2021). Bacteriophage-Mediated Horizontal Gene Transfer: Transduction. In D. R. Harper, S. T. Abedon, B. H. Burrowes, & M. L. McConville (Eds.), *Bacteriophages: Biology, Technology, Therapy* (pp. 151–192). Springer International Publishing. https://doi.org/10.1007/978-3-319-41986-2_4

Secor, P. R., Sweere, J. M., Michaels, L. A., Malkovskiy, A. V., Lazzareschi, D., Katznelson, E., Rajadas, J., Birnbaum, M. E., Arrigoni, A., Braun, K. R., Evanko, S. P., Stevens, D. A., Kaminsky, W., Singh, P. K., Parks, W. C., & Bollyky, P. L. (2015). Filamentous Bacteriophage Promote Biofilm

Assembly And Function. *Cell host & microbe*, *18*(5), 549–559. https://doi.org/10.1016/j.chom.2015.10.013

Sender, R., Fuchs, S., & Milo, R. (2016). Revised Estimates for the Number of Human and Bacteria Cells in the Body [Number: 8 Publisher: Public Library of Science]. *PLOS Biology*, *14*(8), e1002533. https://doi.org/10.1371/journal.pbio.1002533

Shah, E., Rezaie, A., Riddle, M., & Pimentel, M. (2014). Psychological disorders in gastrointestinal disease: Epiphenomenon, cause or consequence? *Annals of Gastroenterology : Quarterly Publication of the Hellenic Society of Gastroenterology*, *27*(3), 224–230. Retrieved February 14, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4073018/

Shang, J., Jiang, J., & Sun, Y. (2021). Bacteriophage classification for assembled contigs using graph convolutional network. *Bioinformatics*, *37*(Supplement_1), i25–i33. https://doi.org/10.1093/bioinformatics/btab293

Shang, J., Tang, X., Guo, R., & Sun, Y. (2022). Accurate identification of bacteriophages from metagenomic data using Transformer. *Briefings in Bioinformatics*, *23*(4), bbac258. https://doi.org/10.1093/bib/bbac258

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, *13*(11), 2498–2504. https://doi.org/10.1101/gr.1239303

Shen, A., & Millard, A. (2021). Phage Genome Annotation: Where to Begin and End. *Phage*, *2*(4), 183–193. https://doi.org/10.1089/phage.2021.0015

Sherrill-Mix, S. (2023). Taxonomizr: Functions to Work with NCBI Accessions and Taxonomy.

Shkoporov, A. N., Clooney, A. G., Sutton, T. D. S., Ryan, F. J., Daly, K. M., Nolan, J. A., McDonnell, S. A., Khokhlova, E. V., Draper, L. A., Forde, A., Guerin, E., Velayudhan, V., Ross, R. P., & Hill, C. (2019). The human gut virome is highly diverse, stable and individual-specific. *bioRxiv*, 657528. https://doi.org/10.1101/657528

Shkoporov, A. N., & Hill, C. (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, A. N., Khokhlova, E. V., Fitzgerald, C. B., Stockdale, S. R., Draper, L. A., Ross, R. P., & Hill, C. (2018a). CrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis [Number: 1 Publisher: Nature Publishing Group]. *Nature Communications*, *9*(1), 1–8. https://doi.org/10.1038/s41467-018-07225-7

Shkoporov, A. N., Ryan, F. J., Draper, L. A., Forde, A., Stockdale, S. R., Daly, K. M., McDonnell, S. A., Nolan, J. A., Sutton, T. D., Dalmasso, M., McCann, A., Ross, R. P., & Hill, C. (2018b). Reproducible protocols for metagenomic analysis of human faecal phageomes. *Microbiome*, *6*(1), 68. https://doi.org/10.1186/s40168-018-0446-z

Shuwen, H., & Kefeng, D. (2022). Intestinal phages interact with bacteria and are involved in human diseases. *Gut Microbes*, *14*(1), 2113717. https://doi.org/10.1080/19490976.2022.2113717

Silveira, C. B., Coutinho, F. H., Cavalcanti, G. S., Benler, S., Doane, M. P., Dinsdale, E. A., Edwards, R. A., Francini-Filho, R. B., Thompson, C. C., Luque, A., Rohwer, F. L., & Thompson, F. (2020). Genomic and ecological attributes of marine bacteriophages encoding bacterial virulence genes. *BMC Genomics*, *21*(1), 126. https://doi.org/10.1186/s12864-020-6523-2

Silveira, C. B., & Rohwer, F. L. (2016). Piggyback-the-Winner in host-associated microbial communities [Number: 1 Publisher: Nature Publishing Group]. *npj Biofilms and Microbiomes*, *2*(1), 1–5. https://doi.org/10.1038/npjbiofilms.2016.10

Simon, G. L., & Gorbach, S. L. (1984). Intestinal flora in health and disease. *Gastroenterology*, *86*(1), 174–193.

Sirén, K., Millard, A., Petersen, B., Gilbert, M. T. P., Clokie, M. R. J., & Sicheritz-Pontén, T. (2021). Rapid discovery of novel prophages using biological feature engineering and machine learning [Number: lqaa109]. *NAR Genomics and Bioinformatics*, *3*(lqaa109). https://doi.org/10.1093/nargab/lqaa109

Smits, S. L., Bodewes, R., Ruiz-Gonzalez, A., Baumgärtner, W., Koopmans, M. P., Osterhaus, A. D. M. E., & Schürch, A. C. (2014). Assembly of viral genomes from metagenomes. *Frontiers in Microbiology*, *5*, 714. https://doi.org/10.3389/fmicb.2014.00714

Soucy, S. M., Huang, J., & Gogarten, J. P. (2015). Horizontal gene transfer: Building the web of life [Number: 8 Publisher: Nature Publishing Group]. *Nature Reviews Genetics*, *16*(8), 472–482. https://doi.org/10.1038/nrg3962

Starikova, E. V., Tikhonova, P. O., Prianichnikov, N. A., Rands, C. M., Zdobnov, E. M., Ilina, E. N., & Govorun, V. M. (2020). Phigaro: High throughput prophage sequence annotation. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btaa250

Starr, E. P., Shi, S., Blazewicz, S. J., Koch, B. J., Probst, A. J., Hungate, B. A., Pett-Ridge, J., Firestone, M. K., & Banfield, J. F. (2021). Stable-Isotope-Informed, Genome-Resolved Metagenomics Uncovers Potential Cross-Kingdom Interactions in Rhizosphere Soil. *mSphere*, *6*(5), e00085–21. https://doi.org/10.1128/mSphere.00085-21

Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets [Number: 11 Publisher: Nature Publishing Group]. *Nature Biotechnology*, *35*(11), 1026–1028. https://doi.org/10.1038/nbt.3988

Stentz, R., Horn, N., Cross, K., Salt, L., Brearley, C., Livermore, D. M., & Carding, S. R. (2015). Cephalosporinases associated with outer membrane vesicles released by Bacteroides spp. protect gut pathogens and commensals against -lactam antibiotics. *Journal of Antimicrobial Chemotherapy*, *70*(3), 701–709. https://doi.org/10.1093/jac/dku466

Sternberg, N., & Weisberg, R. (1975). Packaging of prophage and host DNA by coliphage [Number: 5513 Publisher: Nature Publishing Group]. *Nature*, *256*(5513), 97–103. https://doi.org/10.1038/256097a0

St-Pierre, F., & Endy, D. (2008). Determination of cell fate selection during phage lambda infection. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(52), 20705–20710. https://doi.org/10.1073/pnas.0808831105

Sulakvelidze, A., Alavidze, Z., & Morris, J. G. (2001). Bacteriophage therapy. *Antimicrobial Agents and Chemotherapy*, *45*(3), 649–659. https://doi.org/10.1128/AAC.45.3.649-659.2001

Suttle, C. A. (2007). Marine viruses — major players in the global ecosystem [Number: 10 Publisher: Nature Publishing Group]. *Nature Reviews Microbiology*, *5*(10), 801–812. https://doi.org/10.1038/nrmicro1750

Suttle, S. W. (1999). Viruses and Nutrient Cycles in the Sea. *49*(10).

Sutton, T. D. S., Clooney, A. G., Ryan, F. J., Ross, R. P., & Hill, C. (2019). Choice of assembly software has a critical impact on virome characterisation. *Microbiome*, *7*(1), 12. https://doi.org/10.1186/s40168-019-0626-5

Svenningsen, S. L., Costantino, N., Court, D. L., & Adhya, S. (2005). On the role of Cro in lambda prophage induction. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(12), 4465–4469. https://doi.org/10.1073/pnas.0409839102

Swanson, M., Fraser, G., Daniell, T., Torrance, L., Gregory, P., & Taliansky, M. (2009). Viruses in soils: Morphological diversity and abundance in the rhizosphere [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1744-7348.2009.00319.x]. *Annals of Applied Biology*, *155*(1), 51–60. https://doi.org/10.1111/j.1744-7348.2009.00319.x

Sydenham, T. V., Overballe-Petersen, S., Hasman, H., Wexler, H., Kemp, M., & Justesen, U. S. (2019). Complete hybrid genome assembly of clinical multidrug-resistant Bacteroides fragilis isolates enables comprehensive

identification of antimicrobial-resistance genes and plasmids. *Microbial Genomics*, *5*(11), e000312. https://doi.org/10.1099/mgen.0.000312

Synnott, A. J., Kuang, Y., Kurimoto, M., Yamamichi, K., Iwano, H., & Tanji, Y. (2009). Isolation from Sewage Influent and Characterization of Novel Staphylococcus aureus Bacteriophages with Wide Host Ranges and Potent Lytic Capabilities. *Applied and Environmental Microbiology*, *75*(13), 4483–4490. https://doi.org/10.1128/AEM.02641-08

Ta, L. D. H., Chan, J. C. Y., Yap, G. C., Purbojati, R. W., Drautz-Moses, D. I., Koh, Y. M., Tay, C. J. X., Huang, C.-H., Kioh, D. Y. Q., Woon, J. Y., Tham, E. H., Loo, E. X. L., Shek, L. P. C., Karnani, N., Goh, A., Van Bever, H. P. S., Teoh, O. H., Chan, Y. H., Lay, C., … Lee, B. W. (2020). A compromised developmental trajectory of the infant gut microbiome and metabolome in atopic eczema. *Gut Microbes*, *12*(1), 1–22. https://doi.org/10.1080/19490976.2020.1801964

Tamari, F., Hinkley, C. S., & Ramprashad, N. (2013). A Comparison of DNA Extraction Methods using Petunia hybrida Tissues. *Journal of Biomolecular Techniques : JBT*, *24*(3), 113–118. https://doi.org/10.7171/jbt.13-2403-001

Tampuu, A., Bzhalava, Z., Dillner, J., & Vicente, R. (2019). ViraMiner: Deep learning on raw DNA sequences for identifying viral genomes in human samples [Number: 9]. *PLoS ONE*, *14*(9). https://doi.org/10.1371/journal.pone.0222271

Tang, Q., Jin, G., Wang, G., Liu, T., Liu, X., Wang, B., & Cao, H. (2020). Current Sampling Methods for Gut Microbiota: A Call for More Precise Devices.

Frontiers in Cellular and Infection Microbiology, *10*. Retrieved February 14, 2023, from https://www.frontiersin.org/articles/10.3389/fcimb.2020.00151

Taur, Y., Xavier, J. B., Lipuma, L., Ubeda, C., Goldberg, J., Gobourne, A., Lee, Y. J., Dubin, K. A., Socci, N. D., Viale, A., Perales, M.-A., Jenq, R. R., van den Brink, M. R. M., & Pamer, E. G. (2012). Intestinal Domination and the Risk of Bacteremia in Patients Undergoing Allogeneic Hematopoietic Stem Cell Transplantation. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, *55*(7), 905–914. https://doi.org/10.1093/cid/cis580

ter Horst, A. M., Santos-Medellín, C., Sorensen, J. W., Zinke, L. A., Wilson, R. M., Johnston, E. R., Trubl, G., Pett-Ridge, J., Blazewicz, S. J., Hanson, P. J., Chanton, J. P., Schadt, C. W., Kostka, J. E., & Emerson, J. B. (2021). Minnesota peat viromes reveal terrestrial and aquatic niche partitioning for local and global viral populations. *Microbiome*, *9*(1), 233. https://doi.org/10.1186/s40168-021-01156-0

Terzian, P., Olo Ndela, E., Galiez, C., Lossouarn, J., Pérez Bucio, R. E., Mom, R., Toussaint, A., Petit, M.-A., & Enault, F. (2021). PHROG: Families of prokaryotic virus proteins clustered using remote homology. *NAR genomics and bioinformatics*, *3*(3), lqab067. https://doi.org/10.1093/nargab/lqab067

Thingstad, T. F. (2000). Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.4319/lo.2000.45.6.1320].

Limnology and Oceanography, *45*(6), 1320–1328. https://doi.org/10.4319/lo.2000.45.6.1320

Thingstad, T. F., Våge, S., Storesund, J. E., Sandaa, R.-A., & Giske, J. (2014). A theoretical analysis of how strain-specific viruses can control microbial species diversity [Publisher: Proceedings of the National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *111*(21), 7813–7818. https://doi.org/10.1073/pnas.1400909111

Thompson, L. R., Zeng, Q., Kelly, L., Huang, K. H., Singer, A. U., Stubbe, J., & Chisholm, S. W. (2011). Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(39), E757–764. https://doi.org/10.1073/pnas.1102164108

Thurber, R. V., Haynes, M., Breitbart, M., Wegley, L., & Rohwer, F. (2009). Laboratory procedures to generate viral metagenomes [Number: 4 Publisher: Nature Publishing Group]. *Nature Protocols*, *4*(4), 470–483. https://doi.org/10.1038/nprot.2009.10

Touchon, M., Moura de Sousa, J. A., & Rocha, E. P. (2017). Embracing the enemy: The diversification of microbial gene repertoires by phage-mediated horizontal gene transfer. *Current Opinion in Microbiology*, *38*, 66–73. https://doi.org/10.1016/j.mib.2017.04.010

Tribble, G. D., Parker, A. C., & Smith, C. J. (1997). The Bacteroides mobilizable transposon Tn4555 integrates by a site-specific recombination mechanism similar to that of the gram-positive bacterial element Tn916. *Journal of Bacteriology*, *179*(8), 2731–2739. https://doi.org/10.1128/jb.179.8.2731-2739.1997

Trubl, G., Jang, H. B., Roux, S., Emerson, J. B., Solonenko, N., Vik, D. R., Solden, L., Ellenbogen, J., Runyon, A., Bolduc, B., et al. (2018). Soil viruses are underexplored players in ecosystem carbon processing. mSystems 3: E00076-18.

Trucksis, M., Galen, J. E., Michalski, J., Fasano, A., & Kaper, J. B. (1993). Accessory cholera enterotoxin (Ace), the third toxin of a Vibrio cholerae virulence cassette. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(11), 5267–5271. Retrieved February 14, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC46697/

Turner, D., Kropinski, A. M., & Adriaenssens, E. M. (2021). A Roadmap for Genome-Based Phage Taxonomy. *Viruses*, *13*(3), 506. https://doi.org/10.3390/v13030506

Turner, D., Shkoporov, A. N., Lood, C., Millard, A. D., Dutilh, B. E., Alfenas-Zerbini, P., van Zyl, L. J., Aziz, R. K., Oksanen, H. M., Poranen, M. M., Kropinski, A. M., Barylski, J., Brister, J. R., Chanisvili, N., Edwards, R. A., Enault, F., Gillis, A., Knezevic, P., Krupovic, M., ... Adriaenssens, E. M. (2023). Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ICTV bacterial viruses subcommittee. *Archives of Virology*, *168*(2), 74. https://doi.org/10.1007/s00705-022-05694-2

Twort, F. W. (1915). AN INVESTIGATION ON THE NATURE OF ULTRA-MICROSCOPIC VIRUSES. *The Lancet*, *186*(4814), 1241–1243. https://doi.org/10.1016/S0140-6736(01)20383-3

Van de Peer, Y., Chapelle, S., & De Wachter, R. (1996). A quantitative map of nucleotide substitution rates in bacterial rRNA. *Nucleic Acids Research*,

24(17), 3381–3391. Retrieved January 18, 2024, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC146102/

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.

Vandeputte, D., Kathagen, G., D'hoe, K., Vieira-Silva, S., Valles-Colomer, M., Sabino, J., Wang, J., Tito, R. Y., De Commer, L., Darzi, Y., Vermeire, S., Falony, G., & Raes, J. (2017). Quantitative microbiome profiling links gut community variation to microbial load [Number: 7681 Publisher: Nature Publishing Group]. *Nature*, *551*(7681), 507–511. https://doi.org/10.1038/nature24460

Varga, M., Kuntová, L., Pantůček, R., Mašlaňová, I., Růžičková, V., & Doškař, J. (2012). Efficient transfer of antibiotic resistance plasmids by transduction within methicillin-resistant Staphylococcus aureus USA300 clone [Number: 2]. *FEMS microbiology letters*, *332*(2), 146–152. https://doi.org/10.1111/j.1574-6968.2012.02589.x

Vaser, R., Sovic, I., Nagarajan, N., & Sikic, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, gr.214270.116. https://doi.org/10.1101/gr.214270.116

Verbanic, S., Kim, C. Y., Deacon, J. M., & Chen, I. A. (2019). Improved single-swab sample preparation for recovering bacterial and phage DNA from human skin and wound microbiomes. *BMC Microbiology*, *19*, 214. https://doi.org/10.1186/s12866-019-1586-4

Vestergaard, G., Aramayo, R., Basta, T., Häring, M., Peng, X., Brügger, K., Chen, L., Rachel, R., Boisset, N., Garrett, R. A., & Prangishvili, D. (2008). Structure of the Acidianus Filamentous Virus 3 and Comparative

Genomics of Related Archaeal Lipothrixviruses. *Journal of Virology*, *82*(1), 371–381. https://doi.org/10.1128/JVI.01410-07

Viazis, S., Akhtar, M., Feirtag, J., Brabban, A., & Diez-Gonzalez, F. (2011). Isolation and characterization of lytic bacteriophages against enterohaemorrhagic Escherichia coli [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2672.2011.04989.x]. *Journal of Applied Microbiology*, *110*(5), 1323–1331. https://doi.org/10.1111/j.1365-2672.2011.04989.x

Vijay, A., & Valdes, A. M. (2022). Role of the gut microbiome in chronic diseases: A narrative review [Number: 4 Publisher: Nature Publishing Group]. *European Journal of Clinical Nutrition*, *76*(4), 489–501. https://doi.org/10.1038/s41430-021-00991-6

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Walker, A. W., Duncan, S. H., Louis, P., & Flint, H. J. (2014). Phylogeny, culturing, and metagenomics of the human gut microbiota [Number: 5 Publisher: Elsevier]. *Trends in Microbiology*, *22*(5), 267–274. https://doi.org/10.1016/j.tim.2014.03.001

Walker, P. J., Siddell, S. G., Lefkowitz, E. J., Mushegian, A. R., Adriaenssens, E. M., Alfenas-Zerbini, P., Dempsey, D. M., Dutilh, B. E., García, M. L.,

Curtis Hendrickson, R., Junglen, S., Krupovic, M., Kuhn, J. H., Lambert, A. J., Łobocka, M., Oksanen, H. M., Orton, R. J., Robertson, D. L., Rubino, L., ... Zerbini, F. M. (2022). Recent changes to virus taxonomy ratified by the International Committee on Taxonomy of Viruses (2022). *Archives of Virology*, *167*(11), 2429–2440. https://doi.org/10.1007/s00705-022-05516-5

Wallace, A. J., Stillman, T. J., Atkins, A., Jamieson, S. J., Bullough, P. A., Green, J., & Artymiuk, P. J. (2000). E. coli Hemolysin E (HlyE, ClyA, SheA): X-Ray Crystal Structure of the Toxin and Observation of Membrane Pores by Electron Microscopy [Publisher: Elsevier]. *Cell*, *100*(2), 265–276. https://doi.org/10.1016/S0092-8674(00)81564-0

Walters, W. A., Granados, A. C., Ley, C., Federman, S., Stryke, D., Santos, Y., Haggerty, T., Sotomayor-Gonzalez, A., Servellita, V., Ley, R. E., Parsonnet, J., & Chiu, C. Y. (2023). Longitudinal comparison of the developing gut virome in infants and their mothers. *Cell Host & Microbe*, *31*(2), 187–198.e3. https://doi.org/10.1016/j.chom.2023.01.003

Wang, L., Christophersen, C. T., Sorich, M. J., Gerber, J. P., Angley, M. T., & Conlon, M. A. (2012). Elevated fecal short chain fatty acid and ammonia concentrations in children with autism spectrum disorder. *Digestive Diseases and Sciences*, *57*(8), 2096–2102. https://doi.org/10.1007/s10620-012-2167-7

Warwick-Dugdale, J., Solonenko, N., Moore, K., Chittick, L., Gregory, A. C., Allen, M. J., Sullivan, M. B., & Temperton, B. (2019). Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands [Publisher: PeerJ Inc.]. *PeerJ*, *7*, e6800. https://doi.org/10.7717/peerj.6800

Weirather, J. L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.-J., Buck, D., & Au, K. F. (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research*, *6*, 100. https://doi.org/10.12688/f1000research.10571.2

Weitz, J. S., Mileyko, Y., Joh, R. I., & Voit, E. O. (2008). Collective Decision Making in Bacterial Viruses. *Biophysical Journal*, *95*(6), 2673–2680. https://doi.org/10.1529/biophysj.108.133694

Wells, P. M., Adebayo, A. S., Bowyer, R. C. E., Freidin, M. B., Finckh, A., Strowig, T., Lesker, T. R., Alpizar-Rodriguez, D., Gilbert, B., Kirkham, B., Cope, A. P., Steves, C. J., & Williams, F. M. K. (2020). Associations between gut microbiota and genetic risk for rheumatoid arthritis in the absence of disease: A cross-sectional study. *The Lancet. Rheumatology*, *2*(7), e418–e427. https://doi.org/10.1016/S2665-9913(20)30064-3

Wheatley, R., Diaz Caballero, J., Kapel, N., de Winter, F. H. R., Jangir, P., Quinn, A., del Barrio-Tofiño, E., López-Causapé, C., Hedge, J., Torrens, G., Van der Schalk, T., Xavier, B. B., Fernández-Cuenca, F., Arenzana, A., Recanatini, C., Timbermont, L., Sifakis, F., Ruzin, A., Ali, O., … MacLean, C. (2021). Rapid evolution and host immunity drive the rise and fall of carbapenem resistance during an acute Pseudomonas aeruginosa infection [Number: 1 Publisher: Nature Publishing Group]. *Nature Communications*, *12*(1), 2460. https://doi.org/10.1038/s41467-021-22814-9

Wickham, H. (2016). *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Dplyr: A Grammar of Data Manipulation*.

Wigington, C. H., Sonderegger, D., Brussaard, C. P. D., Buchan, A., Finke, J. F., Fuhrman, J. A., Lennon, J. T., Middelboe, M., Suttle, C. A., Stock, C., Wilson, W. H., Wommack, K. E., Wilhelm, S. W., & Weitz, J. S. (2016). Re-examination of the relationship between marine virus and microbial cell abundances [Number: 3 Publisher: Nature Publishing Group]. *Nature Microbiology*, *1*(3), 1–9. https://doi.org/10.1038/nmicrobiol.2015.24

Williamson, K. E., Fuhrmann, J. J., Wommack, K. E., & Radosevich, M. (2017). Viruses in Soil Ecosystems: An Unknown Quantity Within an Unexplored Territory [_eprint: https://doi.org/10.1146/annurev-virology-101416-041639]. *Annual Review of Virology*, *4*(1), 201–219. https://doi.org/10.1146/annurev-virology-101416-041639

Williamson, K. E., Radosevich, M., Smith, D. W., & Wommack, K. E. (2007). Incidence of lysogeny within temperate and extreme soil environments. *Environmental Microbiology*, *9*(10), 2563–2574. https://doi.org/10.1111/j.1462-2920.2007.01374.x

Williamson, K. E., Radosevich, M., & Wommack, K. E. (2005). Abundance and Diversity of Viruses in Six Delaware Soils [Publisher: American Society for Microbiology]. *Applied and Environmental Microbiology*, *71*(6), 3119–3125. https://doi.org/10.1128/AEM.71.6.3119-3125.2005

Willig, M., Kaufman, D., & Stevens, R. (2003). Latitudinal Gradients of Biodiversity: Pattern, Process, Scale, and Synthesis [_eprint: https://doi.org/10.1146/annurev.ecolsys.34.012103.144032]. *Annual Review of Ecology, Evolution, and Systematics*, *34*(1), 273–309. https://doi.org/10.1146/annurev.ecolsys.34.012103.144032

Winter, S. E., Winter, M. G., Xavier, M. N., Thiennimitr, P., Poon, V., Keestra, A. M., Laughlin, R. C., Gomez, G., Wu, J., Lawhon, S. D., Popova, I., Parikh, S. J., Adams, L. G., Tsolis, R. M., Stewart, V. J., & Bäumler, A. J. (2013). Host-derived nitrate boosts growth of E. coli in the inflamed gut. *Science (New York, N.Y.)*, *339*(6120), 708–711. https://doi.org/10.1126/science.1232467

Wommack, K. E., Bhavsar, J., Polson, S. W., Chen, J., Dumas, M., Srinivasiah, S., Furman, M., Jamindar, S., & Nasko, D. J. (2012). VIROME: A standard operating procedure for analysis of viral metagenome sequences [Number: 3]. *Standards in Genomic Sciences*, *6*(3), 427–439. https://doi.org/10.4056/sigs.2945050

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, *20*(1), 257. https://doi.org/10.1186/s13059-019-1891-0

Wood, M. W., Rosqvist, R., Mullan, P. B., Edwards, M. H., & Galyov, E. E. (1996). SopE, a secreted protein of Salmonella dublin, is translocated into the target eukaryotic cell via a sip-dependent mechanism and promotes bacterial entry. *Molecular Microbiology*, *22*(2), 327–338. https://doi.org/10.1046/j.1365-2958.1996.00116.x

Wrighton, C. J., & Strike, P. (1987). A pathway for the evolution of the plasmid NTP16 involving the novel kanamycin resistance transposon Tn4352. *Plasmid*, *17*(1), 37–45. https://doi.org/10.1016/0147-619x(87)90006-0

Wu, L.-Y., Pappas, N., Wijesekara, Y., Piedade, G. J., Brussaard, C. P. D., & Dutilh, B. E. (2023). Benchmarking Bioinformatic Virus Identification Tools Using Real-World Metagenomic Data across Biomes [Pages: 2023.04.26.538077 Section: New Results]. https://doi.org/10.1101/2023.04.26.538077

Wu, R., Davison, M. R., Gao, Y., Nicora, C. D., Mcdermott, J. E., Burnum-Johnson, K. E., Hofmockel, K. S., & Jansson, J. K. (2021). Moisture modulates soil reservoirs of active DNA and RNA viruses [Number: 1 Publisher: Nature Publishing Group]. *Communications Biology*, *4*(1), 1–11. https://doi.org/10.1038/s42003-021-02514-2

Wu, R., Davison, M. R., Nelson, W. C., Smith, M. L., Lipton, M. S., Jansson, J. K., McClure, R. S., McDermott, J. E., & Hofmockel, K. S. (2023). Hi-C metagenome sequencing reveals soil phage–host interactions [Number: 1 Publisher: Nature Publishing Group]. *Nature Communications*, *14*(1), 7666. https://doi.org/10.1038/s41467-023-42967-z

Wu, S., Fang, Z., Tan, J., Li, M., Wang, C., Guo, Q., Xu, C., Jiang, X., & Zhu, H. (2021). DeePhage: Distinguishing virulent and temperate phage-derived sequences in metavirome data with a deep learning approach. *GigaScience*, *10*(9), giab056. https://doi.org/10.1093/gigascience/giab056

Yang, K., Niu, J., Zuo, T., Sun, Y., Xu, Z., Tang, W., Liu, Q., Zhang, J., Ng, E. K. W., Wong, S. K. H., Yeoh, Y. K., Chan, P. K. S., Chan, F. K. L., Miao, Y., & Ng, S. C. (2021). Alterations in the Gut Virome in Obesity and

Type 2 Diabetes Mellitus [Publisher: Elsevier]. *Gastroenterology*, *161*(4), 1257–1269.e13. https://doi.org/10.1053/j.gastro.2021.06.056

Yang, Y., Du, L., Shi, D., Kong, C., Liu, J., Liu, G., Li, X., & Ma, Y. (2021). Dysbiosis of human gut microbiome in young-onset colorectal cancer [Number: 1 Publisher: Nature Publishing Group]. *Nature Communications*, *12*(1), 6757. https://doi.org/10.1038/s41467-021-27112-y

Yao, H., Lu, S., Williams, B. A., Flanagan, B. M., Gidley, M. J., & Mikkelsen, D. (2022). Absolute abundance values reveal microbial shifts and co-occurrence patterns during gut microbiota fermentation of dietary fibres in vitro. *Food Hydrocolloids*, *127*, 107422. https://doi.org/10.1016/j.foodhyd.2021.107422

Ye, S. H., Siddle, K. J., Park, D. J., & Sabeti, P. C. (2019). Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell*, *178*(4), 779–794. https://doi.org/10.1016/j.cell.2019.07.010

Yilmaz, S., Allgaier, M., & Hugenholtz, P. (2010). Multiple displacement amplification compromises quantitative analysis of metagenomes [Number: 12 Publisher: Nature Publishing Group]. *Nature Methods*, *7*(12), 943–944. https://doi.org/10.1038/nmeth1210-943

Yutin, N., Benler, S., Shmakov, S. A., Wolf, Y. I., Tolstoy, I., Rayko, M., Antipov, D., Pevzner, P. A., & Koonin, E. V. (2021). Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features [Number: 1 Publisher: Nature Publishing Group]. *Nature Communications*, *12*(1), 1044. https://doi.org/10.1038/s41467-021-21350-w

Zablocki, O., Michelsen, M., Burris, M., Solonenko, N., Warwick-Dugdale, J., Ghosh, R., Pett-Ridge, J., Sullivan, M. B., & Temperton, B. (2021). VirION2: A short- and long-read sequencing and informatics workflow to study the genomic diversity of viruses in nature. *PeerJ*, *9*, e11088. https://doi.org/10.7717/peerj.11088

Zaborin, A., Smith, D., Garfield, K., Quensen, J., Shakhsheer, B., Kade, M., Tirrell, M., Tiedje, J., Gilbert, J. A., Zaborina, O., & Alverdy, J. C. (2014). Membership and behavior of ultra-low-diversity pathogen communities present in the gut of humans during prolonged critical illness. *mBio*, *5*(5), e01361–01314. https://doi.org/10.1128/mBio.01361-14

Zeng, L., Skinner, S. O., Zong, C., Sippy, J., Feiss, M., & Golding, I. (2010). Decision Making at a Subcellular Level Determines the Outcome of Bacteriophage Infection. *Cell*, *141*(4), 682–691. https://doi.org/10.1016/j.cell.2010.03.034

Zhang, B.-W., Li, M., Ma, L.-C., & Wei, F.-W. (2006). A Widely Applicable Protocol for DNA Isolation from Fecal Samples. *Biochemical Genetics*, *44*(11), 494–503. https://doi.org/10.1007/s10528-006-9050-1

Zhang, K., Pankratz, K., Duong, H., Theodore, M., Guan, J., Jiang, A. (, Lin, Y., & Zeng, L. (2021). Interactions between Viral Regulatory Proteins Ensure an MOI-Independent Probability of Lysogeny during Infection by Bacteriophage P1 [Publisher: American Society for Microbiology]. *mBio*, *12*(5), e01013–21. https://doi.org/10.1128/mBio.01013-21

Zhang, X., McDaniel, A. D., Wolf, L. E., Keusch, G. T., Waldor, M. K., & Acheson, D. W. (2000). Quinolone antibiotics induce Shiga toxin-encoding

bacteriophages, toxin production, and death in mice. *The Journal of Infectious Diseases*, *181*(2), 664–670. https://doi.org/10.1086/315239

Zhang, Zhang, D., Jia, H., Feng, Q., Wang, D., Liang, D., Wu, X., Li, J., Tang, L., Li, Y., Lan, Z., Chen, B., Li, Y., Zhong, H., Xie, H., Jie, Z., Chen, W., Tang, S., Xu, X., … Wang, J. (2015). The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. *Nature Medicine*, *21*(8), 895–905. https://doi.org/10.1038/nm.3914

Zhao, G., Wu, G., Lim, E. S., Droit, L., Krishnamurthy, S., Barouch, D. H., Virgin, H. W., & Wang, D. (2017). VirusSeeker, a computational pipeline for virus discovery and virome composition analysis. *Virology*, *503*, 21–30. https://doi.org/10.1016/j.virol.2017.01.005

Zheng, T., Li, J., Ni, Y., Kang, K., Misiakou, M.-A., Imamovic, L., Chow, B. K. C., Rode, A. A., Bytzer, P., Sommer, M., & Panagiotou, G. (2019). Mining, analyzing, and integrating viral signals from metagenomic data. *Microbiome*, *7*. https://doi.org/10.1186/s40168-019-0657-y

Zinder, N. D., & Lederberg, J. (1952). GENETIC EXCHANGE IN SALMONELLA1. *Journal of Bacteriology*, *64*(5), 679–699. Retrieved November 8, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC169409/

Zolfo, M., Pinto, F., Asnicar, F., Manghi, P., Tett, A., Bushman, F. D., & Segata, N. (2019). Detecting contamination in viromes using ViromeQC [Number: 12 Publisher: Nature Publishing Group]. *Nature Biotechnology*, *37*(12), 1408–1412. https://doi.org/10.1038/s41587-019-0334-5

Zrelovs, N., Dislers, A., & Kazaks, A. (2021). Motley Crew: Overview of the Currently Available Phage Diversity. *Frontiers in Microbiology*, *11*.

Retrieved November 7, 2022, from https://www.frontiersin.org/articles/10.3389/fmicb.2020.579452

Zuo, T., Liu, Q., Zhang, F., Lui, G. C.-Y., Tso, E. Y., Yeoh, Y. K., Chen, Z., Boon, S. S., Chan, F. K., Chan, P. K., & Ng, S. C. (2021). Depicting SARS-CoV-2 faecal viral activity in association with gut microbiota composition in patients with COVID-19 [Publisher: BMJ Publishing Group Section: COVID-19]. *Gut*, *70*(2), 276–284. https://doi.org/10.1136/gutjnl-2020-322294

Zuo, T., Lu, X.-J., Zhang, Y., Cheung, C. P., Lam, S., Zhang, F., Tang, W., Ching, J. Y. L., Zhao, R., Chan, P. K. S., Sung, J. J. Y., Yu, J., Chan, F. K. L., Cao, Q., Sheng, J.-Q., & Ng, S. C. (2019). Gut mucosal virome alterations in ulcerative colitis [Publisher: BMJ Publishing Group Section: Gut microbiota]. *Gut*, *68*(7), 1169–1179. https://doi.org/10.1136/gutjnl-2018-318131

Appendix

5.2 Custom scripts

5.2.1 removesmalls.pl

```
## removesmalls.pl
#!/usr/bin/perl
use strict;
use warnings;

my $minlen = shift or die "Error: ‘minlen‘ parameter not
provided\n";
```

```
{
    local $/=">";
    while(<>) {
        chomp;
        next unless /\w/;
        s/>$//gs;
        my @chunk = split /\n/;
        my $header = shift @chunk;
        my $seqlen = length join "", @chunk;
        print ">$_" if($seqlen >= $minlen);
    }
    local $/="\n";
}
```

5.2.2 Time lagged cross-correlation python script

```
import pandas as pd
import numpy as np
from scipy.signal import correlate
from scipy.stats import pearsonr
import matplotlib.pyplot as plt
from statsmodels.stats import multitest as smm
```

```
# Define a list of csv files
csv_list = ["patient_24_sychrony.csv",
"patient_25_sychrony.csv","patient_29_sychrony.csv",
"patient_31_sychrony.csv","patient_37_sychrony.csv",
"patient_4_sychrony.csv","patient_41_sychrony.csv",
"patient_53_sychrony.csv", "patient_54_sychrony.csv",
"patient_55_sychrony.csv","patient_59_sychrony.csv"]

# Calculate the number of rows and columns
num_plots = len(csv_list)
num_rows = int(np.ceil(np.sqrt(num_plots)))
num_cols = int(np.ceil(num_plots / num_rows))



# Create a grid of subplots with the desired number of rows and
columns
fig, axs = plt.subplots(nrows=num_rows, ncols=num_cols,
figsize=(10, 4*num_rows))

# Flatten the array of subplots to make it easier to iterate over
them
axs = axs.flatten()

# Iterate over the csv files and plot the time-lagged
cross-correlation
```

```
for i, csv_file in enumerate(csv_list):
    # Load the data from the CSV file
    data = pd.read_csv(csv_file)

    # Extract the phage and host abundance data
    phage = data['Phage_Abundance']
    host = data['Host_Abundance']
    days = data['Day_in_ICU']

    # Calculate the Pearson correlation coefficient and its
significance
    pearson_corr, pearson_pvalue = pearsonr(phage, host)

    # Normalize the Pearson correlation coefficient to be between 0
and 1
    #pearson_corr_norm = (pearson_corr + 1) / 2

    # Calculate the time-lagged cross correlation
    corr = correlate(phage - np.mean(phage), host - np.mean(host),
mode='full', method='auto') / (np.std(phage) * np.std(host))

    # Normalize the cross correlation coefficient to be between 0
and 1
    corr_norm = (corr - np.min(corr)) / (np.max(corr) -
np.min(corr))
```

```
    # Determine the index of the peak correlation
    peak_idx = np.argmax(corr)

    # Determine the lag corresponding to the peak correlation
    lags = np.arange(-len(phage) + 1, len(phage))
    peak_lag = lags[peak_idx]



    # Plot the cross correlation
    axs[i].plot(lags, corr_norm)
    axs[i].axvline(x=peak_lag, linestyle='--', color='r')
    axs[i].set_ylabel('Normalized Cross Correlation', fontsize=8)
    axs[i].set_title(csv_file.replace('_sychrony.csv', ''),
fontsize=10)
    pearson_label = f"Overall Pearson Correlation:
{pearson_corr:.2f} (p-value = {pearson_pvalue:.2e})"
    axs[i].text(.5, .95, pearson_label, ha='center', va='top',
transform=axs[i].transAxes, fontsize=6,
bbox=dict(facecolor='white', alpha=0.5))
    axs[i].text(.95, .15, f"Peak synchrony at {peak_lag} days",
ha='right', va='center', transform=axs[i].transAxes, fontsize=6,
bbox=dict(facecolor='white', alpha=0.5))

# Add x-axis label and legend to the bottom plot
axs[-1].set_xlabel('Time Lag (days)')
```

```
axs[-1].legend()


# Adjust the plot layout
fig.tight_layout()
fig.subplots_adjust(hspace=0.4, wspace=0.2,top=0.95)
# Show the plot
plt.show()
```

5.3 Supplementary Tables

Table 5.1: **List of vOTUs from Chapter 3 that encode antibiotic resistance genes.** Gene predictions were carried out using pharokka and the CARD database.

vOTU	Gene	Sequence Identity
vOTU_1009	*lsaA*	0.922
vOTU_1023	*H-NS*	0.903
vOTU_1094	*ermB*	0.946
vOTU_1151	*cfxA2*	0.948
vOTU_125	*cfxA2*	0.948
vOTU_1268	*tet(Q)*	0.873
vOTU_1268	*tet(X)*	0.967
vOTU_1292	*vanU_in_vanG_cl*	0.816
vOTU_1388	*vanR_in_vanG_cl*	0.81
vOTU_1444	*tet(Q)*	0.932
vOTU_1512	*mel*	0.878
vOTU_1558	*cfxA4*	0.946
vOTU_156	*mel*	0.879
vOTU_156	*ermG*	0.948
vOTU_1563	*cfxA4*	0.946
vOTU_1566	*efrB*	0.812
vOTU_1567	*cfxA3*	0.947

vOTU	Gene	Sequence Identity
vOTU_1629	*cfxA4*	0.946
vOTU_1640	*marA*	0.912
vOTU_1673	*cfxA2*	0.948
vOTU_1814	*kpne_OmpK37*	0.824
vOTU_1851	*tet(Q)*	0.932
vOTU_1943	*kpne_KpnH*	0.937
vOTU_1943	*emrA*	0.919
vOTU_1943	*emrR*	0.948
vOTU_195	*tet(Q)*	0.932
vOTU_2006	*cfxA2*	0.946
vOTU_202	*cpxA*	0.941
vOTU_2089	*marA*	0.912
vOTU_2202	*vanU_in_vanG_cl*	0.816
vOTU_221	*cpxA*	0.945
vOTU_2211	*tet(Q)*	0.932
vOTU_2272	*emrK*	0.925
vOTU_2272	*emrY*	0.939
vOTU_2335	*MdtK*	0.816
vOTU_2343	*tolC*	0.905
vOTU_2345	*tet(Q)*	0.932
vOTU_2392	*cfxA2*	0.946
vOTU_246	*dfrF*	0.964
vOTU_2466	*cfxA2*	0.946

vOTU	Gene	Sequence Identity
vOTU_2508	*tet(O/W/32/O)*	0.821
vOTU_2558	*ecol_emrE*	0.97
vOTU_2568	*ecol_emrE*	0.96
vOTU_280	*tet(Q)*	0.935
vOTU_334	*tolC*	0.906
vOTU_338	*cfxA2*	0.948
vOTU_353	*cfiA2*	0.987
vOTU_369	*ecol_emrE*	0.96
vOTU_373	*MdtK*	0.816
vOTU_376	*tolC*	0.906
vOTU_433	*ecol_emrE*	0.96
vOTU_515	*tolC*	0.905
vOTU_568	*cfxA2*	0.948
vOTU_713	*tet(Q)*	0.936
vOTU_794	*tet(Q)*	0.932
vOTU_887	*emrY*	0.938
vOTU_887	*emrK*	0.921
vOTU_887	*evgA*	0.941
vOTU_887	*evgS*	0.939
vOTU_898	*tolC*	0.905
vOTU_923	*ant(6)-Ib*	0.8
vOTU_956	*tet(O)*	0.969

Table 5.2: **List of vOTUs from Chapter 3 that encode virulence factor genes.** Gene predictions were carried out using pharokka and the CARD database.

vOTU	Gene	Description	Sequence Identity
vOTU_101	*Z_RS10335*	fimbria/pilus periplasmic chaperone	0.834
vOTU_101	*flhe*	flagellar protein FlhE	0.911
vOTU_1026	*pla*	plasminogen activator protease precursor	0.913
vOTU_1026	*fimD*	Outer membrane usher protein fimD precursor	0.834
vOTU_1029	*rpoS*	RNA polymerase sigma factor RpoS	0.895
vOTU_1029	*sitA*	iron/manganese ABC transporter substrate-binding protein SitA	0.943
vOTU_1051	*ybtS*	yersiniabactin biosynthesis salicylate synthase YbtS	0.953
vOTU_1222	*sitB*	iron/manganese ABC transporter ATP-binding protein SitB	0.968
vOTU_1222	*ydeQ*	fimbrial protein	0.954

vOTU	Gene	Description	Sequence Identity
vOTU_1222	*fimF*	FimF protein precursor	1
vOTU_1222	*fimG*	FimG protein precursor	1
vOTU_1222	*ybtX*	yersiniabactin-associated zinc MFS transporter YbtX	0.964
vOTU_1222	*sitC*	iron/manganese ABC transporter permease subunit SitC	0.968
vOTU_1222	*ROD_RS04625*	phage tail assembly protein	0.981
vOTU_1222	*fimG*	FimG protein precursor	0.924
vOTU_1222	*fimF*	FimF protein precursor	0.923
vOTU_1222	*ydeQ*	fimbrial protein	0.901
vOTU_1222	*sitD*	iron/manganese ABC transporter permease subunit SitD	0.966
vOTU_1223	*ydeQ*	fimbrial protein	0.838
vOTU_1223	*phoQ*	two-component system sensor histidine kinase PhoQ	0.831
vOTU_1223	*rpoS*	RNA polymerase sigma factor RpoS	0.921
vOTU_1223	*fimG*	FimG protein precursor	0.849

vOTU	Gene	Description	Sequence Identity
vOTU_1223	*rcsB*	transcriptional regulator RcsB	0.801
vOTU_1230	*ybtQ*	yersiniabactin ABC transporter ATP-binding/permease protein YbtQ	0.832
vOTU_1230	*phoP*	two-component system response regulator PhoP	0.913
vOTU_1236	*fimF*	FimF protein precursor	0.947
vOTU_1236	*phoP*	two-component system response regulator PhoP	0.939
vOTU_1236	*phoQ*	two-component system sensor histidine kinase PhoQ	0.941
vOTU_1236	*fimD*	Outer membrane usher protein fimD precursor	0.951
vOTU_1236	*phoP*	two-component system response regulator PhoP	0.941
vOTU_1236	*nmpC*	phosphoporin PhoE	0.897

vOTU	Gene	Description	Sequence Identity
vOTU_1566	*ybtP*	yersiniabactin ABC transporter ATP-binding/permease protein YbtP	0.802
vOTU_1814	*ybtA*	yersiniabactin transcriptional regulator YbtA	0.824
vOTU_1864	*Z_RS10335*	fimbria/pilus periplasmic chaperone	0.803
vOTU_1864	*phoQ*	two-component system sensor histidine kinase PhoQ	0.845
vOTU_1864	*phoP*	two-component system response regulator PhoP	0.917
vOTU_1864	*fimD*	Outer membrane usher protein fimD precursor	0.832
vOTU_1864	*gtrB*	bactoprenol glucosyl transferase	0.842
vOTU_1943	*irp2*	yersiniabactin biosynthetic protein Irp2	0.885
vOTU_1944	*fimA*	type 1 fimbrial major subunit FimA	0.965

vOTU	Gene	Description	Sequence Identity
vOTU_1944	*phoQ*	two-component system sensor histidine kinase PhoQ	0.899
vOTU_1944	*Z_RS10335*	fimbria/pilus periplasmic chaperone	0.923
vOTU_1944	*rpoS*	RNA polymerase sigma factor RpoS	0.924
vOTU_1944	*fyuA*	siderophore yersiniabactin receptor FyuA	0.978
vOTU_1944	*espL1*	Type III secretion system effector espL1	0.967
vOTU_1944	*fimA*	type 1 fimbrial major subunit FimA	0.964
vOTU_1944	*phoP*	two-component system response regulator PhoP	1
vOTU_1944	*phoQ*	two-component system sensor histidine kinase PhoQ	1
vOTU_1944	*ybtE*	yersiniabactin biosynthesis salycil-AMP ligase YbtE	0.954
vOTU_1944	*ybtT*	yersiniabactin biosynthesis thioesterase YbtT	0.968

vOTU	Gene	Description	Sequence Identity
vOTU_1946	*allD*	ureidoglycolate dehydrogenase	0.909
vOTU_1946	*ybtU*	yersiniabactin biosynthesis oxidoreductase YbtU	0.833
vOTU_2233	*allA*	ureidoglycolate hydrolase	0.952
vOTU_2259	*allC*	allantoate amidohydrolase	0.914
vOTU_2259	*fimA*	type 1 fimbrial major subunit FimA	0.972
vOTU_2259	*allR*	DNA-binding transcriptional repressor AllR	0.904
vOTU_2259	*irp1*	yersiniabactin polyketide synthase HMWP1	0.933
vOTU_2259	*Z_RS10335*	fimbria/pilus periplasmic chaperone	0.905
vOTU_2259	*hlyE/clyA*	hemolysin E	0.893
vOTU_2259	*ecbA/fss3*	Collagen binding MSCRAMM, EcbA/Fss3 "	0.869
vOTU_2259	*allD*	ureidoglycolate dehydrogenase	0.967
vOTU_2259	*flhD*	flagellar transcriptional regulator FlhD	0.925

vOTU	Gene	Description	Sequence Identity
vOTU_2259	*flhC*	flagellar transcriptional regulator FlhC	0.933
vOTU_2259	*allC*	allantoate amidohydrolase	0.911
vOTU_2259	*fyuA*	siderophore yersiniabactin receptor FyuA	0.922
vOTU_2259	*motA*	flagellar motor stator protein MotA	0.802
vOTU_2259	*motB*	flagellar motor protein MotB	0.925
vOTU_2264	*ybtS*	salicylate synthase Irp9	0.951
vOTU_2264	*allB*	allantoinase	0.944
vOTU_2264	*ybtE*	yersiniabactin biosynthesis salycil-AMP ligase YbtE	0.949
vOTU_2264	*allB*	allantoinase	0.957
vOTU_2264	*cheA*	chemotaxis protein CheA	0.948
vOTU_2264	*ybtT*	yersiniabactin biosynthesis thioesterase YbtT	0.895
vOTU_2335	*fimA*	type 1 fimbrial major subunit FimA	0.952
vOTU_2335	*ybtX*	yersiniabactin-associated zinc MFS transporter YbtX	0.894

vOTU	Gene	Description	Sequence Identity
vOTU_2335	*Z_RS10335*	fimbria/pilus periplasmic chaperone	0.918
vOTU_2335	*ybtU*	yersiniabactin biosynthesis oxidoreductase YbtU	0.924
vOTU_2335	*cheW*	chemotaxis protein CheW	0.98
vOTU_2335	*irp1*	yersiniabactin polyketide synthase HMWP1	0.981
vOTU_2335	*fimD*	Outer membrane usher protein fimD precursor	0.963
vOTU_2335	*ybtQ*	yersiniabactin ABC transporter ATP-binding/permease protein YbtQ	1
vOTU_2335	*tar*	methyl-accepting chemotaxis protein II	1
vOTU_2335	*ybtU*	yersiniabactin biosynthesis oxidoreductase YbtU	0.951
vOTU_2335	*ybtP*	yersiniabactin ABC transporter ATP-binding/permease protein YbtP	0.968

vOTU	Gene	Description	Sequence Identity
vOTU_2376	*ybtA*	transcriptional regulator YbtA	0.809
vOTU_2386	*ybtT*	yersiniabactin biosynthesis thioesterase YbtT	0.91
vOTU_2410	*fimF*	FimF protein precursor	0.963
vOTU_2410	*fimG*	FimG protein precursor	0.933
vOTU_2410	*allR*	DNA-binding transcriptional repressor AllR	0.929
vOTU_2410	*irp2*	yersiniabactin biosynthetic protein Irp2	0.899
vOTU_268	*ybtE*	yersiniabactin biosynthesis salycil-AMP ligase YbtE	0.913
vOTU_268	*cheR*	protein-glutamate O-methyltransferase CheR	0.833
vOTU_269	*allC*	allantoate amidohydrolase	0.836
vOTU_269	*ydeQ*	fimbrial protein	0.834
vOTU_271	*allA*	ureidoglycolate hydrolase	0.91
vOTU_2860	*cheB*	protein-glutamate methylesterase/protein glutamine deamidase	0.878

vOTU	Gene	Description	Sequence Identity
vOTU_298	*allD*	ureidoglycolate dehydrogenase	0.936
vOTU_2996	*fyuA*	siderophore yersiniabactin receptor FyuA	0.804
vOTU_317	*cheY*	chemotaxis response regulator CheY	0.803
vOTU_317	*clpP*	ATP-dependent Clp protease proteolytic subunit	0.849
vOTU_317	*cheZ*	protein phosphatase CheZ	0.921
vOTU_317	*irp2*	yersiniabactin biosynthetic protein Irp2	0.834
vOTU_317	*flhB*	flagellar type III secretion system protein FlhB	0.836
vOTU_333	*ybtA*	yersiniabactin transcriptional regulator YbtA	0.944
vOTU_333	*ybtP*	yersiniabactin ABC transporter ATP-binding/permease protein YbtP	0.952

vOTU	Gene	Description	Sequence Identity
vOTU_333	*flhA*	flagellar biosynthesis protein FlhA	0.943
vOTU_333	*irp1*	yersiniabactin polyketide synthase HMWP1	0.939
vOTU_333	*ybtQ*	yersiniabactin ABC transporter ATP-binding/permease protein YbtQ	0.95
vOTU_369	*irp1*	yersiniabactin polyketide synthase HMWP1	0.936
vOTU_373	*flhe*	flagellar protein FlhE	0.966
vOTU_373	*ybtX*	yersiniabactin-associated zinc MFS transporter YbtX	0.955
vOTU_373	*ybtS*	yersiniabactin biosynthesis salicylate synthase YbtS	1
vOTU_373	*allA*	ureidoglycolate hydrolase	1
vOTU_373	*phoQ*	two-component system sensor histidine kinase PhoQ	0.933
vOTU_373	*allR*	DNA-binding transcriptional repressor AllR	0.98

vOTU	Gene	Description	Sequence Identity
vOTU_373	*phoP*	two-component system response regulator PhoP	0.981
vOTU_373	*irp2*	yersiniabactin biosynthetic protein Irp2	0.923
vOTU_373	*ybtA*	yersiniabactin transcriptional regulator YbtA	0.924
vOTU_373	*ybtP*	yersiniabactin ABC transporter ATP-binding/permease protein YbtP	0.901
vOTU_373	*ybtQ*	yersiniabactin ABC transporter ATP-binding/permease protein YbtQ	0.966
vOTU_387	*efaA*	endocarditis specific antigen	0.913
vOTU_387	*ybtU*	yersiniabactin biosynthesis oxidoreductase YbtU	0.834
vOTU_423	*ybtX*	yersiniabactin-associated zinc MFS transporter YbtX	0.912
vOTU_465	*allB*	allantoinase	0.974

vOTU	Gene	Description	Sequence Identity
vOTU_465	*ybtT*	yersiniabactin biosynthesis thioesterase YbtT	0.9
vOTU_465	*ybtE*	yersiniabactin biosynthesis salycil-AMP ligase YbtE	0.962
vOTU_465	*ybtS*	yersiniabactin biosynthesis salicylate synthase YbtS	0.978
vOTU_465	*efaA*	endocarditis specific antigen	0.945
vOTU_512	*fyuA*	siderophore yersiniabactin receptor FyuA	0.902
vOTU_514	*clpV*	ATP-dependent chaperone ClpB	0.891
vOTU_514	*allC*	allantoate amidohydrolase	0.936
vOTU_514	*allD*	ureidoglycolate dehydrogenase	0.951
vOTU_514	*PMI_RS09270*	fimbria/pilus outer membrane usher protein	0.943
vOTU_514	*PMI_RS09265*	fimbrial protein	0.939
vOTU_514	*PMI_RS09275*	molecular chaperone	0.949
vOTU_881	*PMI_RS09280*	fimbrial protein	0.91
vOTU_993	*PMI_RS09285*	type 1 fimbrial protein	0.946

5.4 Supplementary figures

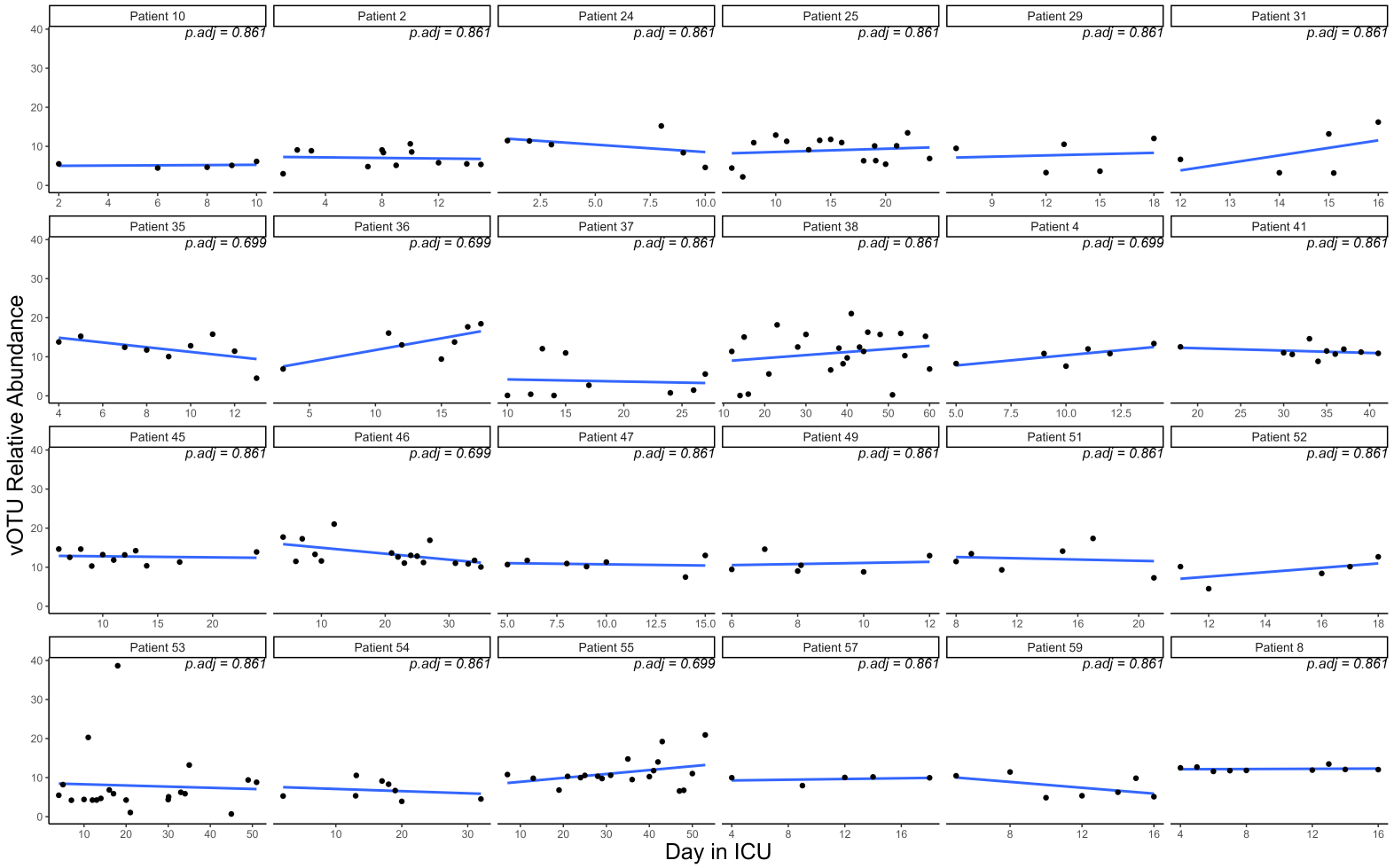


Figure 5.1: **Linear regression between vOTU abundance and length of stay of intensive care unit patients.** Linear models were fit using `geom_smooth` from ggplot2. Multiple comparisons correction was applied to p-values using the Benjamini-Hochberg procedure.

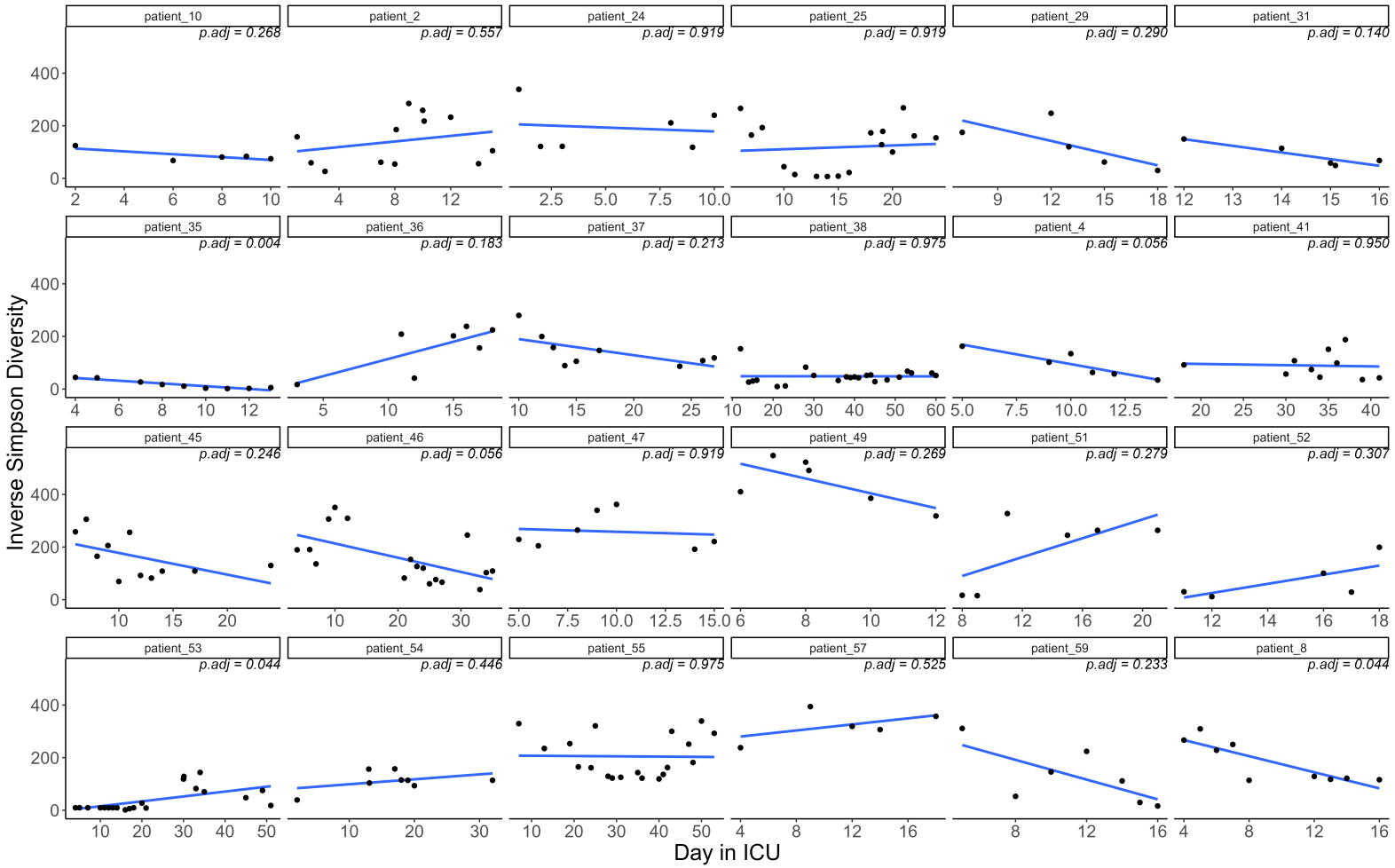


Figure 5.2: **Linear regression between Inverse Simpson alpha diversity of vOTUs and length of stay of intensive care unit patients.** Linear models were fit using `geom_smooth` from ggplot2. Multiple comparisons correction was applied to p-values using the Benjamini-Hochberg procedure.