Exploring the Diversity and Ecology of Bacteriophages in the Dairy Farm Environment

Ryan Cook

Student Number 4342363

Thesis submitted to the University of Nottingham

for the Degree of Doctor of Philosophy

School of Veterinary Medicine and Science,

University of Nottingham, UK

November 2022

List of Figures

List of Tables

List of Appendices

The supplementary files below are available on FigShare at https://figshare.com/s/b7ad8c844288a4325deb

Supplementary File 1: In-house Perl script that modifies fasta headers so that previously predicted genes can be used as input for MetaPop.

Supplementary File 2: Fasta file of nucleotide sequences for putative MBLs.

Supplementary File 3: Excel workbook containing supplementary tables that are referenced within this work.

List of Abbreviations

(p)ppGpp: Guanosine tetraphosphate or guanosine pentaphosphate

AMG: Auxiliary metabolic gene

AMR: Antimicrobial resistance

ANI: Average nucleotide identity

ARG: Antimicrobial resistance gene

ASR: Assimilatory sulfate reduction

BOD: Biological oxygen demand

CAZYme: Carbohydrate-active enzyme

CDS: Coding sequence

CF: Cystic fibrosis

COG: Clusters of orthologous groups

DEFRA: Department for Environment, Food and Rural Affairs

DGR: Diversity-generating retroelement

DTR: Direct terminal repeat

eggNOG: Evolutionary genealogy of genes: Non-supervised Orthologous Groups

ESBL: Extended-spectrum beta-lactamase

HGT: Horizontal gene transfer

HMW: High molecular weight

LASL: Linker-amplified shotgun library

LB: Luria-Bertani

MAG: Metagenome assembled genome

MazG: Nucleoside triphosphate pyrophosphohydrolase

MBL: Metallo-beta-lactamase

MDR: Multidrug resistant

MLS: Macrolides, lincosamides, and streptogramines

MNP: Multiple nucleotide polymorphism

NMDS: Non-metric dimensional scaling

ONT: Oxford Nanopore Technologies

ORF: Open reading frame

PAPS: 3′-Phosphoadenosine-5′-phosphosulfate

pVOG: Prokaryotic virus orthologous groups

SNP: Single nucleotide polymorphism

SOC: Super optimal catabolite

SRB: Sulfate-reducing bacteria

TerL: Terminase large subunit

VapE: Virulence-associated protein E

VC: Viral cluster

VHG: Viral hallmark gene

VLP: Virus-like particle

vMAG: Viral metagenome assembled genome

vOTU: Viral operational taxonomic unit

Zot: *Zona occludens* toxin

Acknowledgments

I would like to thank my incredible supervisory team (Mike Jones, Andy Millard, Dov Stekel, Chris Hudson, and Jon Hobman) for all of their support and mentoring during the project. Andy wasn't even based at the same institution but went over and above to make sure I actually at least somewhat knew what I was doing (at least, I hope I reached that point some time near the end). I would also like to thank the wider research groups at Nottingham, Leicester, and other collaborating institutions who have contributed to this work and taught me so much.

Thank you to Izzy for keeping me sane and stable, and to her mum Caroline, for keeping me alive and allowing me use of a lovely garden during the 2020 COVID-19 pandemic. And of course, thank you to my parents for their support (emotionally and financially!) during my seven years at the University of Nottingham.

Declaration

I declare that the work in this dissertation was carried out in accordance with the regulations of the University of Nottingham. The work is original and has not been submitted for any other degree at the University of Nottingham or elsewhere. For chapters which represent manuscripts, both published and ready for submission, the work of other authors has been described in chapter prefaces and the corresponding authors approve of how these contributions have been reflected.

Name: Ryan Cook

Signature: 

Date: 27th November 2022

Name: Michael Jones

Signature: 

Date: 1st December 2022

Name: Andrew Millard

Signature: 

Date: 1st December 2022

Summary

Bacteriophages, viruses that obligately infect bacteria, represent the most abundant and diverse biological entities on the planet, with key and complex ecological roles in all environments where they have been studied. Agricultural wastes and manures (i.e., cattle slurry) are economically important fertilisers that are applied to land. Despite the widespread use of slurries, there is a paucity of knowledge regarding the microbial composition within them.

The first part of this thesis sought to optimise viral metagenomics, for the study of viral communities in nature. As the study of uncultivated viral genomes is underpinned by known complete viral genomes, I assessed the current extent of viral sequencing to provide the most complete reference database possible in an updated and reproducible fashion. This led to the INPHARED database, now published in PHAGE and available on GitHub; a community resource that provides genomes and annotation files to aid in viral genomic and metagenomic analyses. Furthermore, I investigated biases in the current collection of phage genomes and demonstrated that clear biases towards phages of a small subset of clinically relevant bacteria. Subsequently, I sought to benchmark sequencing technologies and assembly approaches for the recovery of viral genomes from viral metagenomes. This work, in part published in Microbiome and under review in Microbial Genomics, demonstrated that choice of sequencing technology and assembly algorithm will have significant impacts on downstream analyses and estimates of viral diversity. Overall, these analyses demonstrated that a combination of long and short read sequencing approaches performed best at recovering viral genomes in a mixed community.

The second part of this thesis applied the understandings described above to the dairy farm environment. I utilised long- and short-read sequencing to characterise the viral community of agricultural slurry in a longitudinal study, as well as modelled slurry tanks that contained agricultural antimicrobials, and the dairy cattle gut across life stages. Analysis of the cattle slurry virome, now published in Microbiome, revealed a diverse and novel community that was stable over time, despite constant influx and efflux of material. Notably, there was widespread phage carriage of a virulence determinant—VapE—that is associated with bovine mastitis-causing pathogens such as *Streptococcus* spp. Subsequent experiments with mock slurry tanks revealed the slurry virome may be influenced by the presence of footwash, although the reasons for this remain unclear. Analysis of the dairy cow virome uncovered 1,338 predicted complete phage genomes, the most of a single virome study to date. The phages within the dairy cow gut were largely novel, and their community composition changed over key life stages.

The results within this thesis have advanced the methodology of viral metagenomics approaches in general, and show that viruses likely play important ecological roles within agricultural environments, including augmenting the virulence of relevant veterinary pathogens.

Chapter 1 Introduction

1.1 An Introduction to Bacterial Viruses

Bacteriophages, hereafter phages, are viruses that specifically infect bacteria. There's an often-used opening sentence in phage-related publications – "*Phages are the most abundant and diverse biological entities on the planet*" – it's a cliché, but you'd struggle to argue against it. There are thought to be 10^{31} phages within the biosphere; ubiquitous within all environments where their bacterial hosts can be found (Suttle, 2007; Comeau *et al.*, 2008; Clokie *et al.*, 2011; Cobián Güemes *et al.*, 2016).

First discovered independently by Frederick Twort in 1915 (Twort, 1915) and Félix d'Hérelle in 1917 (D'Hérelle, 2007), phages are obligate intracellular parasites of bacteria. Although their structure and genomes vary greatly, all phages consist of nucleic acids encapsulated within a protein coat and rely on host-cell machinery to produce progeny viral particles.

1.2 The Lifecycles of Phages

The lifecycles and infection strategies of phages are diverse and complex, although they generally fall into three main categories: phages may be (1) obligately lytic (hereafter lytic; sometimes described as virulent); (2) temperate, whereby they have access to both the lytic and lysogenic lifecycles; or (3) chronic, whereby a phage that may or may not be temperate continually produces and releases viral progeny without lysing the host cell (Rakonjac *et al.*, 2011; Salmond and Fineran, 2012). All three life cycles begin with the phage attaching to specific cell surface host receptors and injecting their DNA into the host cytoplasm (Orlova, 2012). After this, the three cycles differ.

1.2.1 The Lytic Cycle

Following injection of their genome into host cytoplasm, lytic phages will redirect (or "hijack") host metabolism to produce viral progeny. The viral genome will be replicated and viral proteins are synthesised, from which new viral particles are subsequently produced (Ofir and Sorek, 2018). Following this, the host cell will undergo lysis due to the expression of phage-derived holins and lysins, killing the host cell and releasing the viral progeny (Ofir and Sorek, 2018). This life cycle may be accessed by both lytic and temperate phages, and is exemplified by the widely studied obligately lytic bacteriophage T4 (Miller *et al.*, 2003).

1.2.2 The Lysogenic Cycle

Whereas lytic phages exclusively follow the lytic lifecycle, temperate phages can access both the lytic cycle and the lysogenic cycle. In the lysogenic cycle, following injection of genetic material into host cytoplasm, a latent infection is established. The viral genome is incorporated within the bacterial host genome and replicates alongside the host, with the phage genome being transmitted vertically to all bacterial progeny, as demonstrated by the *Escherichia* phage λ (Casjens and Hendrix, 2015). However, in some instances, such as the *Leptospira biflexa* phage LE1, the integrated phage genome exists freely within the host cytoplasm as a circular replicon (Girons *et al.*, 2000). The integrated phage genome is described as a prophage, and the prophage-containing host cell is known as a lysogen. Changes in host-cell conditions (for example, environmental stressors such as radiation or nutrient depletion) can release the prophage, leading to proliferation of new viral progeny via the lytic cycle (Howard-Varona *et al.*, 2017).

1.2.3 The Chronic Cycle

Whilst phages in the lytic cycle lyse their hosts to release viral progeny, those in the chronic cycle will continually produce and release progeny without lysing the host cell (Russel and Model, 2006). For filamentous phages of the family *Inoviridae*, such as the *Escherichia* phage M13, a productive infection results in viral particles being secreted from the host cell without the need for lysis (Rakonjac *et al.*, 2011). Due to some chronic phages also being able to access the lysogenic cycle, there have been calls for phages to be classified based upon whether virions are released (e.g. productive infection versus lysogeny) and the means of release (e.g. lytic versus chronic) (Hobbs and Abedon, 2016).

1.3 The Diversity and Classification of Phages

Phages are thought to be the most diverse biological entities in the biosphere, and currently known viral diversity may represent only the tip of the iceberg. Their diversity encompasses a range of properties including morphology (e.g., tailed vs non-tailed and shape of capsid), genome molecule and replication strategy (e.g., dsDNA, ssRNA, etc.), host specificity and range, lifecycles used (e.g., temperate vs lytic), and genomic sequence similarity. Due to the absence of a universal phylogenetic marker, the success of microbial 16s rRNA gene sequencing for taxonomic classification cannot be applied to phages (Yarza *et al.*, 2014; Dion, Oechslin and Moineau, 2020). Phage classification is curated by the International Committee on Taxonomy of Viruses (ICTV) (Walker *et al.*, 2021).

1.3.1 Morphological Diversity and Historical Classification

Whilst the classification of phages is now based upon genomic similarity, historically, the classification of phages centred around morphological characteristics (Aiewsakun *et al.*, 2018; Walker *et al.*, 2021). Researchers would observe the phage using transmission electron microscopy (TEM) and the phage would be classified in a framework that examined capsid structure, genome molecule and the presence/absence of an envelope (Ackermann, 2009; King *et al.*, 2012).

The morphological diversity of phages is known to be wide, although the majority of currently cultured phages are tailed and possess dsDNA genomes, historically belonging to the now outdated *Caudovirales,* which was previously divided into three families based upon their morphological characteristics; *Myoviridae* (with long contractile tails), *Siphoviridae* (with flexible non-contractile tails), and *Podoviridae* (with short tails) (Ackermann, 2009; Fokine and Rossmann, 2014; Dion, Oechslin and Moineau, 2020). Whilst tailed phages are arguably the most widely studied, there is a wide range of observed non-tailed morphologies including: polyhedral phages (e.g., *Microviridae*), filamentous phages (e.g., *Inoviridae*), and pleomorphic phages (e.g., *Plasmaviridae*) (Ackermann, 2009; Fokine and Rossmann, 2014; Dion, Oechslin and Moineau, 2020). Furthermore, despite tailed phages comprising the majority of phages studied within the lab, electron microscopy has revealed that non-tailed phages dominate the oceans and their diversity may be under-represented within current databases and collections (Borsheim, Bratbak and Heldal, 1990; Wommack *et al.*, 1992; Brum, Schenck and Sullivan, 2013).

1.3.2 Genomic Diversity and Current Classification

The genome structure and replication strategies of viruses are varied and include genomes comprised of dsDNA, ssDNA, dsRNA and ssRNA (Fokine and Rossmann, 2014; Dion, Oechslin and Moineau, 2020). Our current understanding of phage genomic diversity is primarily based upon those with dsDNA genomes, as these are the most widely cultivated (Cook, Brown, *et al.*, 2021). However, exploration of global transcriptome datasets has uncovered a previously unknown diversity of RNA viral genomes (Wolf *et al.*, 2020; Neri *et al.*, 2022). The current collection of available phage genomes within publicly available databases is therefore likely biased towards particular types of phages.

Currently available phage genomes obtained from cultured isolates range in size from 2.3 kb (*Pseudomonas* phage phi12, accession NC_004174) to 497.5 kb (*Bacillus* phage G, accession NC_023719). Additionally, putative phage genomes >500 kb of so-called “mega-phages” have been assembled from metagenomes, although these have not been brought into culture (Devoto *et al.*, 2019; Michniewski *et al.*, 2021).

The composition of phage genomes is equally diverse, with many phages sharing little or no sequence similarity with others. Furthermore, those with similar morphology may share little sequence similarity and vice versa. For this reason, viral classification has moved away from morphology and towards a genome-organisation based taxonomy (Aiewsakun *et al.*, 2018).

Genome-based taxonomic frameworks based upon nucleotide and/or protein sequence and proteome comparisons have been suggested and there are notable examples of their implementation. Proteomic approaches have successfully been

used to classify members of the now redundant order *Caudovirales*, including the notable families *Myoviridae* (Lavigne *et al.*, 2009), *Podoviridae* (Lavigne *et al.*, 2008), and *Siphoviridae* (Adriaenssens *et al.*, 2015), which resulted in the introduction of sub-families and genera that were ratified by the ICTV. Later, these frameworks were built upon and made available as online tools such as ViPTree (Nishimura *et al.*, 2017) and VICTOR (Meier-Kolthoff and Göker, 2017) which are able to rapidly classify a user's sequence(s) based upon shared proteins. Similarly, a hierarchical cluster based approach based upon the presence/absence of shared proteins, dubbed vConTACT2, was developed for the classification of uncultivated viruses and is scalable to large numbers of genomes (Bin Jang *et al.*, 2019). Other protein-based approaches have been built around the phylogeny of so-called "viral hallmark genes" (VHGs) that are highly conserved across diverse groups of viruses. For example, a framework that concatenates single-copy protein markers has been developed for the classification of dsDNA phages belonging to the historical order *Caudovirales* (Low *et al.*, 2019), and was recently implemented on large-scale datasets of uncultivated viruses (Nayfach *et al.*, 2021). Furthermore, the proposed viral "megataxonomy" from Koonin *et al.* is a hierarchical taxonomy based upon the phylogeny of VHGs (Koonin *et al.*, 2020). Alternatively to the protein-based method, VICTOR is able to classify phages using the nucleotide sequence of the whole genome (Meier-Kolthoff and Göker, 2017). Another approach, VIRIDIC, uses nucleotide-based intergenomic similarity and can help to classify phage to the levels of genus and species, but is less effective for more distantly related phages for which protein based metrics are suggested (Moraru, Varsani and Kropinski, 2020).

Therefore, genome-based frameworks for the classification of phage have been successful despite the absence of a universal phylogenetic marker (Dion, Oechslin and Moineau, 2020), and extensive horizontal gene transfer (or mosaicism) between phages (Lawrence, Hatfull and Hendrix, 2002; Iranzo, Krupovic and Koonin, 2016).

As of September 2022, the ICTV recognises 50 families, 100 sub-families, and 1,652 genera of viruses that infect prokaryotes (https://ictv.global/taxonomy).

1.4 Distribution and Abundance of Phages

The distribution and abundance of phages has most extensively been studied in the oceans, where the number of virus-like particles (VLPs) was found to range between 10^5 and 10^7 per millilitre of seawater in 95% of samples, and the number of putative viruses typically outnumber microbial cells by 10:1 (Wigington *et al.*, 2016). Other environments where phages have been found to be abundant include soils, with each gram of soil (dry weight) typically containing 10^9 VLPs (Swanson *et al.*, 2009), and the human gut, with each gram of human faeces containing up to 10^{10} VLPs (Sutton and Hill, 2019).

Whilst phages are ubiquitous within the marine environment, their distribution is not homogenous. Phages are known to form distinct communities within different marine environments, and the composition of this viral community has been used to distinguish between different aquatic samples (Hayes *et al.*, 2017; Parmar *et al.*, 2018). Furthermore, analysis of the Pacific Ocean Virome dataset has revealed significant variability of community composition based upon season, depth and proximity to land (Hayes *et al.*, 2017).

As phages are a natural predator of bacteria, fully reliant on their hosts for viral replication, it would therefore make sense that the composition of the viral community is shaped by the bacterial community and vice versa. For example, studies of the infant gut show a strong temporal correlation between phages and their predicted hosts (Beller *et al.*, 2022). Furthermore, phages of the human gut are potentially induced from early colonising bacteria (Liang *et al.*, 2020; Beller *et al.*, 2022).

1.5 Phages as Agents of Horizontal Gene Transfer

Alongside plasmids, transposons, and other integrative and conjugative elements (ICEs), phages are known to be widespread mediators of horizontal gene transfer (HGT) (Arnold, Huang and Hanage, 2021). The transfer of genetic material between cells facilitated by phages is broadly referred to as transduction (Canchaya *et al.*, 2003). However, there are many forms of transduction which rely upon entirely different biological processes. The two most well characterised are generalised and specialised transduction, although lateral transduction and auto-transduction are also described.

1.5.1 Generalised Transduction

In the later stages of phage replication, those with dsDNA genomes typically form concatemers that are cut by the terminase protein during packaging into the capsid (Black, 1989). There are four widely characterised mechanisms by which dsDNA phages recognise and cleave their own DNA prior to packaging into the capsid: (1) for phages such as λ (Feiss *et al.*, 1983) and HK97 (Juhala *et al.*, 2000), the terminase recognises a specific cohesive end site (*cos* site) where it introduces a staggered cut, consistently generating DNA with fixed termini at a fixed length. (2) For phages such

as P1 (Bächi and Arber, 1977) and P22 (Tye, Huberman and Botstein, 1974), the terminase recognises a specific packaging site (*pac* site) to initiate packaging and the DNA is cleaved once the capsid (or "head") is full. This is described as headful packaging and may lead to variable lengths of DNA being packaged into the capsid. (3) For phages such as T5 (Wang *et al.*, 2005) and T7 (Dunn, Studier and Gottesman, 1983), the DNA is cut at a fixed position to generate direct terminal repeats, leading to packaging of circularly permuted genomes that may be re-circularised upon injection into host cytoplasm. (4) For T4-like phages (Kalinski and Black, 1986), a variant of headful packaging is used during which no *pac* site is recognised and packaging of DNA is initiated randomly. Although these four mechanisms are the most widely characterised, other mechanisms have been described (e.g. those in phages P2, Mu, and phi29) (Pruss and Calendar, 1978; George and Bukhari, 1981; Bjornsti, Reilly and Anderson, 1983), and many more likely exist in nature.

Generalised transduction, first discovered in the *Salmonella* phage P22, was the first phage mediated HGT mechanism to be identified and is mediated by phages that utilise *pac* site initiated headful packaging (Zinder and Lederberg, 1952; Thierauf, Perez and Maloy, 2009). During generalised transduction, the *pac*-terminase will recognise pseudo-*pac* sites (*pac* site homologues) on the bacterial genome and subsequently package host DNA into the viral capsid, rather than a viral genome (Chelala and Margolin, 1976; Schmieger, 1982; Thierauf, Perez and Maloy, 2009). The host DNA containing particles may go on to infect other cells, upon which the DNA is injected into the cytoplasm of recipient cells (Figure 1.1A). Although generalised transduction can be performed by phages with a *cos*-terminase, the chances of two pseudo-*cos* sites occurring on the host DNA and being separated by the optimum

distance is highly unlikely, and so *cos*-terminase facilitated generalised transduction is thought to be rare (Chiang, Penadés and Chen, 2019).

1.5.2 Specialised Transduction

Specialised transduction results from the aberrant excision of a prophage from the bacterial chromosome, and is exemplified by the archetypal *Escherichia* phage λ (Morse, 1954; Morse, Lederberg and Lederberg, 1956b, 1956a). During specialised transduction, bacterial host DNA flanking the prophage attachment site (*att*B) is excised alongside the prophage. The phage-host hybrid DNA may then be replicated and packaged into viral progeny, which may horizontally transfer the host DNA to recipient cells (Figure 1.1B). Specialised transduction can be performed by both *cos*- and *pac*-terminase phages. However, specialised transduction is thought to be rare, as aberrant prophage excision is not a common event (Chiang, Penadés and Chen, 2019).

1.5.3 Lateral Transduction

In 2018, a third type of transduction was described and termed “lateral transduction” (Chen *et al.*, 2018). In lateral transduction, DNA packaging is initiated from *pac* sites of integrated prophages which have delayed excision and underwent bi-directional (theta-form) replication. As a result, part of the phage genome is packaged into a capsid, alongside up to seven headfuls of host DNA. Meanwhile, some prophages will undergo excision and follow typical maturation (Figure 1.1C). Lateral transduction therefore results in normal titres of phage progeny, alongside numbers of transducible particles that may transfer host DNA at far higher frequencies than those reported for generalised and specialised transduction (Chen *et al.*, 2018; Fillol-Salom *et al.*, 2021).

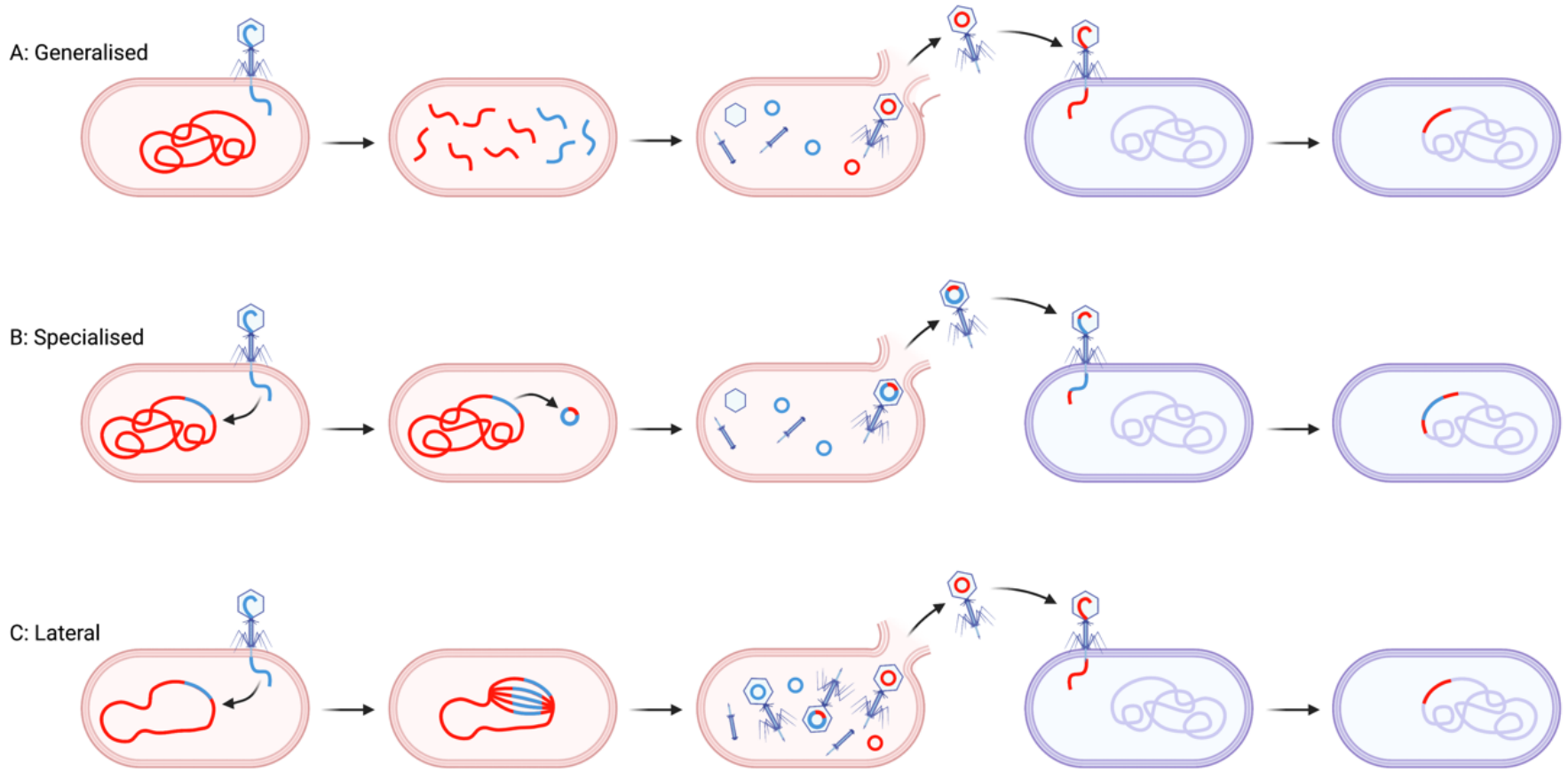


Figure 1.1 Phage transduction

The transfer of bacterial DNA mediated by phages through **(A)** generalised, **(B)** specialised, and **(C)** lateral transduction. Note that auto-transduction (see 1.5.4) is also described but not shown in this figure.

1.5.4 Auto-transduction

A fourth type of transduction, termed “auto-transduction”, has also been described (Haaber *et al.*, 2016). In auto-transduction, a sub-population of lysogenic bacteria release active phages which go on to infect a susceptible population of bacteria that do not contain the prophage. The phage undergoes lytic replication, lyses the susceptible cells, and encapsulates some of the susceptible host DNA into viral progeny. The progeny may then re-infect the lysogenic sub-population of bacteria, transferring some of the susceptible population host DNA to the lysogenic population in the process (Haaber *et al.*, 2016).

1.6 Lysogenic Conversion

In addition to facilitating the HGT of bacterial DNA, phages can alter the phenotype of their hosts through lysogenic conversion. Lysogenic conversion (or phage conversion) is the process by which a prophage alters the metabolism of its host through the expression of phage-encoded genes. These accessory genes, or “morons” (“more on” the phage genome), may provide a fitness advantage to the bacterial cell in which the prophage resides which is mutually advantageous for the prophage (Juhala *et al.*, 2000; Brüssow, Canchaya and Hardt, 2004).

1.6.1 Phage Morons

Most bacteria harbour multiple prophages and the genomes of some bacteria are composed of up to 20% prophage sequences (Casjens, 2003). As the prophage and lysogen exist in a stable mutualistic relationship, it is therefore advantageous for the prophage to encode genes that provide a fitness advantage to its host. These genes, or “morons”, may modulate host metabolism in a number of ways, including but not

limited to: mediation of resistance to further phage infection (Panis, Méjean and Ansaldi, 2007; Ali *et al.*, 2014; Cumby *et al.*, 2015); regulating the expression of genes relating to motility phenotypes (Su *et al.*, 2010; Addy *et al.*, 2012; Tsao *et al.*, 2018); effecting quorum sensing pathways (Hargreaves, Kropinski and Clokie, 2014a, 2014b); and augmenting bacterial virulence (see 1.6.2).

1.6.2 Phage-Encoded Virulence Determinants

Whilst the term moron was first described by Roger Hendrix in 2000, phages have been known to have significant impacts on the virulence of bacteria for decades (Freeman, 1951; Juhala *et al.*, 2000). The impact of lysogenic conversion by phage-encoded toxins and virulence factors is significant, with phages contributing to the pathogenesis of clinically relevant pathogens including *Clostridium botulinum, Vibrio Cholerae, Escherichia coli* and *Shigella* spp. (Freeman, 1951; Eklund *et al.*, 1974; Waldor and Mekalanos, 1996; Wagner *et al.*, 2002; Fortier and Sekulovic, 2013; Khalil *et al.*, 2016). Furthermore, there are many examples where the expression of phage-encoded toxins cause otherwise harmless commensal bacteria to convert into a pathogen, including multidrug-resistant ST11 strains of *Pseudomonas aeruginosa* (van Belkum *et al.*, 2015; Tsao *et al.*, 2018), and the Shiga-toxin encoding *Escherichia coli* (O'Brien *et al.*, 1984).

1.7 Phage-Encoded Auxiliary Metabolic Genes

The terms lysogenic conversion factor and moron were coined to describe genes encoded within prophages that effect host metabolism. However, obligately lytic phages that do not form a prophage are also known to possess non-phage genes that modulate host metabolism, and these have been dubbed auxiliary metabolic genes

(AMGs) (Breitbart *et al.*, 2007). The term AMGs broadly encompasses all phage-encoded “host genes” that may augment the metabolism of their bacterial hosts (therefore, whilst morons and lysogenic conversion factors are AMGs, not all AMGs are morons or lysogenic conversion factors) (Breitbart *et al.*, 2007). Unlike morons and lysogenic conversion factors, phage-encoded AMGs may simply drive the metabolism of their host towards their own purposes. For example, lytic phages of cyanobacteria are known to possess photosynthetic genes to ensure photosynthesis is continued during viral replication (Mann *et al.*, 2003; Lindell *et al.*, 2004).

Since their first description (Breitbart *et al.*, 2007), AMGs have been described in a plethora of environments, including oceans and soils, with potential impacts on bacterial metabolism including augmentation of photosynthesis, carbon metabolism, sulphur metabolism, nitrogen uptake and complex carbohydrate metabolism (Yooseph *et al.*, 2007; Dinsdale *et al.*, 2008; Sharon *et al.*, 2011; Hurwitz, Hallam and Sullivan, 2013; Anantharaman *et al.*, 2014; Zhang, Wei and Cai, 2014; Hurwitz, Brum and Sullivan, 2015; Hurwitz and U’Ren, 2016; Roux, Brum, *et al.*, 2016; York, 2017; Monier *et al.*, 2017; Jin *et al.*, 2019). The widespread presence of AMGs within the genomes of phages is thought to have significant impacts on the ecology and metabolic processes of bacteria, and their subsequent role within global biogeochemical cycling.

1.8 Phages as Vectors for Antimicrobial Resistance

Given the widespread roles of phages in shaping bacterial phenotypes through lysogenic conversion and HGT, it may be logical to assume a similar pattern is observed for the transfer of antimicrobial resistance genes (ARGs). However, this is not necessarily the case. The significance of phages in the spread of antimicrobial

resistance (AMR) in the wider environment is unclear, as prominent studies within this field have reached polarising conclusions (Enault *et al.*, 2017; Debroas and Siguret, 2019).

1.8.1 Phage-Encoded ARGs

Screening the genomes of all cultured phages (n=16,928; 01/March/2022; (Cook, Brown, *et al.*, 2021)) against the Virulence Factor Database and ResFinder database using Abricate with default parameters results in 534 phage genomes predicted to contain a virulence factor (3.15%) and only 53 to contain an ARG (0.31%) (Seemann, no date a; Zankari *et al.*, 2012; Chen *et al.*, 2016). Moreover, ARGs are 10-fold less abundant in free phages than in prophages (Kleinheinz, Joensen and Larsen, 2014), and a number of well characterised prophages that do encode ARGs have been shown to exhibit no lytic activity and are likely unable to facilitate HGT (Banks, Lei and Musser, 2003; Brenciani *et al.*, 2010; Wang *et al.*, 2010; Billard-Pomares *et al.*, 2014; Iannelli *et al.*, 2014; Wipf, Schwendener and Perreten, 2014). However, a recently characterised prophage of *Streptococcus pyogenes* was found to encode the *mef*(A)-*msr*(D) macrolide resistance gene pair (Santoro *et al.*, 2022). Whilst free phages frequently encode genes that act as fitness factors for their bacterial host within their ecological niche, including the augmentation of virulence, they seem to rarely encode ARGs.

1.8.2 Phage Transfer of ARGs

Despite phages rarely encoding ARGs, there are known cases of clinically relevant bacterial species acquiring antimicrobial resistance through mechanisms mediated by phages. With regard to generalised transduction (see 1.5.1), one study reported the

phage mediated transfer of ampicillin resistance in *E. coli* at frequencies between 10^{-4} and 10^{-3} transductants per plaque forming unit (PFU) (Kenzaka *et al.*, 2007), and more recently, a study investigating 243 coliphages isolated from chicken meat found that 24.7% were able to transduce one or more ARGs (encoding resistance to ampicillin, chloramphenicol, kanamycin, and/or tetracycline) to a laboratory strain of *E. coli* (ATCC 13706) (Shousha *et al.*, 2015).

Beyond generalised transduction, temperate phages of *Staphylococcus aureus* have been implicated in the spread of antimicrobial resistance through auto-transduction (see 1.5.4) (Haaber *et al.*, 2016), and lateral transduction has likely played a significant role in the acquisition of AMR for clinically relevant bacteria such as *Salmonella* spp. and *Staphylococcus aureus* (see 1.5.3) (Chen *et al.*, 2018; Fillol-Salom *et al.*, 2021). Furthermore, although not directly mediated by the phages, a number of “super-spreader” phages have been shown to promote HGT of plasmids by transformation through lysing bacteria and leaving large quantities of intact transformable plasmid DNA (Keen *et al.*, 2017). Additionally, a recent analysis of so-called "phage-plasmids” (elements that both phage and plasmid) found them to commonly carry ARGs (Pfeifer, Bonnin and Rocha, 2022). Phages therefore have the potential to transfer ARGs between bacteria belonging to well characterised genera, however the role of phages in the transfer of ARGs in the wider environment is a topic of debate.

1.8.3 Viral ARGs in the Wider Environment

Whilst individual phages have been demonstrated to facilitate HGT of AMR through a number of mechanisms, studies of individual phages do little to demonstrate the importance of phages in the transfer of AMR in the wider environment. To investigate

this, much research is focussed on investigating the total viral community within an environmental sample. Common methods involve isolating the total viral DNA from an environmental sample of interest and either sequencing the DNA (see 1.10) or using it as template in (q)PCR reactions for the detection of specific ARGs of interest.

Despite the widespread belief that HGT of ARGs is mediated by plasmids, ICEs and generalised transduction (Munita and Arias, 2016; Haudiquet *et al.*, 2022), a number of prominent virome analyses suggested that ARG carriage within phages was much higher than previously thought and a paradigm shift was needed. These studies include: analyses of viromes produced from human pulmonary samples of cystic fibrosis (CF) patients concluding phage-carriage of ARGs to be at high levels (Fancello *et al.*, 2011; Rolain *et al.*, 2011); an analysis of viromes from murine faecal samples concluding phages were likely key drivers of multidrug resistance, and that the extent of which was increased after treatment with antibiotics (Modi *et al.*, 2013); a metagenomic analysis of hospital wastewater concluding ARGs were more prevalent in the viral DNA fraction (0.26%) than the bacterial DNA fraction (0.18%) (Subirats *et al.*, 2016); and analyses of viromes from a plethora of environments concluding that non-human viromes were key reservoirs of ARGs that may be disseminating AMR in the wider environment (Lekunberri *et al.*, 2017). However, all of these analyses used read-based approaches for the quantification of ARGs. Typically, a read-based approach involves comparison of reads against a database of ARGs (e.g., CARD or Arg-annot) using an aligner algorithm (e.g., BLASTx or DIAMOND).

The use of read-based approaches for the quantification of ARGs within viromes was brought into question in a prominent re-analysis of the Fancello, Rolain and Modi

datasets (Enault *et al.*, 2017). The re-analysis concluded that the carriage of ARGs within phage genomes was likely over-estimated in the original analyses. The quantification of ARGs within the CF samples was likely misled by high levels of all bacterial DNA and not specifically ARGs (Fancello *et al.*, 2011; Rolain *et al.*, 2011; Enault *et al.*, 2017). Furthermore, the level of ARGs estimated in the murine samples was likely inflated due to exploratory cut-offs being used for the detection of ARGs which led to a number of false-positive ARGs that were later found to not confer an AMR phenotype (Modi *et al.*, 2013; Enault *et al.*, 2017).

The conclusions from the Enault *et al*., re-analysis provided a cautionary tale and offered suggestions to guide the identification of ARGs within viromes, including: (1) bacterial contamination should be quantified using methods outlined in their analyses or other automated methods (a dedicated programme for this, ViromeQC, is now available (Zolfo *et al.*, 2019)), (2) conservative thresholds should be used for the identification of putative ARGs to avoid false-positives, (3) assembly in contigs should be used where possible to confirm that the ARG is on a contig of demonstrably viral origin to avoid being misled by generalised transduction or contaminating bacterial DNA, and (4) only experimental testing will validate the predicted open reading frame (ORF) as a true ARG (Enault *et al.*, 2017).

Conversely, a more recent analysis of bacterial genomes and viromes from a range of environments that used conservative thresholds for the prediction of ARGs concluded that phages were key reservoirs of AMR in the wider environment (Debroas and Siguret, 2019), furthering the difficulty to determine the importance of phages in the transfer of AMR in the environment. Despite conflicting reports, it is still widely believed

that phage-carriage of ARGs is rare and that generalised transduction is the most widespread phage-mediated mechanism for the HGT of ARGs in the environment, although the contribution of generalised transduction is still minimal when compared to transformation and conjugation. However, if ARGs truly are rarely carried in phage genomes, it posits the question; why have they not been selected for?

1.9 Phage Sequencing

The advent of genome sequencing has expanded our understanding of phage-host interactions, and the level of this understanding has increased with the number of phages to be sequenced. Since the genome of ΦX174, the first DNA phage to be sequenced, was sequenced in 1977 using Sanger sequencing (Sanger *et al.*, 1977), the number of phage genomes to be sequenced has increased massively due to the ease of high-throughput sequencing and a resurgence of interest in the therapeutic potential of bacteriophages (Hatfull, 2008; Perez Sepulveda *et al.*, 2016; Luong, Salabarria and Roach, 2020). Furthermore, the relatively simple nature of phage genomes means that the vast majority of isolated phage genomes can be fully assembled using short-read next-generation sequencing approaches only (Rihtman *et al.*, 2016).

This expansion in the number of sequenced phage genomes has accelerated our understanding of the diversity, size, and composition of phage genomes. For example, between 2013 and 2016 a number of phages with surprisingly large genomes (>200 kb) were sequenced and named “jumbo-phages” (Yuan and Gao, 2017). However, isolation of so-called jumbo-phages is thought to be rare, and a recent analysis of

jumbo-phage genomes suggested that 180 kb was a more informative cut-off than the previous 200 kb cut-off (Iyer *et al.*, 2021).

In addition to deepening our understanding of phage genomics, the larger number of phage genomes to be sequenced allows for common analyses that advance the field of bacteriophage research in a number of ways, such as: (1) comparative genomics, where the sequencing of cyanophages has uncovered novel AMGs within their genomes (Mann *et al.*, 2003; Lindell *et al.*, 2004), provided insights to their phylogeny by identifying niche-differentiating genes (Gregory *et al.*, 2016), and combined with proteomic analysis to inform the identification of tail fibres likely responsible for host range (Michniewski *et al.*, 2019); (2) informed continual improvements and advances of viral taxonomy, including the revision of N4-like viruses into the family *Schitoviridae* (Wittmann *et al.*, 2020) and the reclassification of the *Spounavirinae* subfamily of the former family *Myoviridae* to form the new family *Herelleviridae* (Barylski *et al.*, 2019); (3) known phage genome sequences are typically used to inform and train software for prediction of novel phages from metagenomic sequence data (Akhter, Aziz and Edwards, 2012; Roux *et al.*, 2015; Arndt *et al.*, 2017; Bolduc *et al.*, 2017; Ren *et al.*, 2017, 2018) and to subsequently predict their bacterial hosts (Villarroel *et al.*, 2016; Ahlgren *et al.*, 2017; Galiez *et al.*, 2017; Leite *et al.*, 2018, 2019; Boeckaerts *et al.*, 2021; Roux *et al.*, 2022; Ruohan *et al.*, 2022); and (4) often the first step in the analysis of viral metagenomics (hereafter, viromics) is the comparison of sequences with a database of known phage genomes. Therefore, a greater number of publicly available viral genomes helps to inform the field of bacteriophage research as a whole.

1.10 Viral Metagenomics

Whilst the sequencing of individual phages has provided invaluable insight to how phages may alter the metabolism of their hosts, the power of sequencing in the exploration of phage diversity and ecology has been exemplified by the field of viromics.

The isolation, cultivation and sequencing of individual phages relies upon the cultivation of their bacterial hosts. However, due to technical challenges in culturing fastidious bacteria, it is thought that most bacteria remain uncultured (Steen *et al.*, 2019; Thrash, 2021); and therefore, so are their phages. Viromics offers an elegant solution to uncover the unseen diversity of prokaryotic viruses, as it allows for the high-throughput analysis of large numbers of uncultivated viruses (predominantly phages).

In short, viromics involves separating the viral particles from an environmental sample through methods such as centrifugation and filtration to remove environmental debris and cellular organisms. The resultant filtrate may be concentrated, using a centrifugal filter column for example, and the nucleic acids are extracted for downstream applications (e.g., sequencing). However, as viruses are far smaller than bacteria and their genomes are much shorter, the amount of viral DNA extracted from environmental samples is typically very low and therefore insufficient to be sequenced directly. Thus, early viromics work-throughs developed methods for the amplification of viral DNA prior to sequencing.

Notably, the linker amplified shotgun library (LASL) was developed to sequence the first virome (Breitbart *et al.*, 2002). The LASL approach involved randomly shearing

viral DNA, ligating dsDNA linkers to repaired ends, amplifying the fragments by DNA polymerase, and ligating into a vector prior to electroporation into recipient cells. Plasmids were isolated from the resultant clones and sequenced. As viral genomes often contain modified nucleotides that cannot be directly cloned into *E. coli* and many viral genes are toxic to bacteria and must be disrupted prior to cloning (e.g., holins and lysins), the LASL approach offered an elegant solution to these issues in addition to generating sufficient material for sequencing. This approach was implemented to study viromes in a plethora of environments, including the ocean (Breitbart *et al.*, 2002; Bench *et al.*, 2007), human gut/faeces (Breitbart *et al.*, 2003, 2008), blood (Breitbart and Rohwer, 2005), and soil (Fierer *et al.*, 2007); revealing a previously unknown diversity of phage-encoded genes. However, the LASL approach is time-consuming and still required relatively high input quantities of DNA that may be inhibitory from some environments.

Alternatively to LASL, multiple displacement amplification (MDA) has been used for the amplification of viral DNA to perform viromics (Angly *et al.*, 2006). MDA utilises the Φ29 DNA polymerase to amplify DNA isothermally and has been used to study viromes from diverse environments, including the ocean, (Angly *et al.*, 2006), human gut (Reyes *et al.*, 2010), and an Antarctic lake (López-Bueno *et al.*, 2009). Much like LASL, the use of MDA overcame issues with low yields of viral DNA from environmental samples and allowed for exploration of previously unseen viral diversity. Furthermore, the implementation of MDA is less technically difficult to perform than cloning based approaches and requires even lower starting DNA concentrations (Polson, Wilhelm and Wommack, 2011). However, MDA has been associated with the preferential amplification of ssDNA genomes (Kim *et al.*, 2008),

formation of chimeras (Lasken and Stockwell, 2007), and quantitative biases that make inter-sample abundance comparisons impossible (Yilmaz, Allgaier and Hugenholtz, 2010). Whilst LASL and MDA offered new insights into viral communities, a notable comparison of the two methods applied to the same surface seawater sample found that the resulting sequence data, and subsequent taxonomic and functional assignments, varied widely between the two (Kim and Bae, 2011).

Whilst LASL and MDA approaches differed in how the viral DNA was amplified, most viromes of this era were sequenced using the same platform: Roche 454 pyrosequencing, which was typically favoured over Illumina platforms available at the time due to its longer read lengths. The application of this era of viromics is exemplified by the study of aquatic samples which provided early estimates of richness and diversity across globally distributed viral communities (Rodriguez-Brito *et al.*, 2010; Roux *et al.*, 2012; Hurwitz, Hallam and Sullivan, 2013; Hurwitz, Brum and Sullivan, 2015), and human gut samples, shedding light on a previously unseen component of the human microbiome and uncovering the enigmatic crAssphage (Reyes *et al.*, 2010; Kim *et al.*, 2011; Minot *et al.*, 2011; Dutilh *et al.*, 2014).

In 2014, Dutilh *et al.* re-analysed the faecal viromes described in the Reyes *et al.* (2010) dataset using a cross-assembly approach. The reads from the viromes were pooled and assembled *de novo* using gsAssembler and crAss (Margulies *et al.*, 2005; Dutilh *et al.*, 2012). Upon examination of the cross-assembly, the researchers observed a contig which was comprised of reads from all 12 individuals in the original dataset, suggesting it may be derived from a universal viral entity. To find other contigs derived from the same potential genome, they used depth-profile binning and

homology binning, which unveiled a number of contigs that had significant similarity to unknown sequences from unrelated gut metagenomes; further suggesting the universal viral genome was present in other human datasets. Subsequently, they carefully re-assembled the reads from the individual virome in which most reads of the ubiquitous contig were derived from. This re-assembly yielded a complete ~97 kb genome that was designated crAssphage (Dutilh *et al.*, 2014). Since its first assembly, the crAssphage was found to be the most abundant phage in the human gut and has subsequently been brought into culture, four years after its initial discovery (Dutilh *et al.*, 2014; Guerin *et al.*, 2018; Shkoporov *et al.*, 2018). Thus, the discovery of crAssphage provides an elegant example as to how viromic studies allow for the exploration of ecologically important viruses, before they can be brought into culture.

However, like the sequencing technologies that came before it, pyrosequencing has since been superseded by high throughput sequencing (HTS) platforms (e.g., the Illumina MiSeq, HiSeq and NovaSeq). In the early 2010's, Illumina platforms became the sequencer of choice as they offered much greater sequence coverage at a lower cost than pyrosequencing, and generated far fewer sequencing errors (Loman *et al.*, 2012). The accessibility of HTS and improvements to genome assembly have driven a viromics revolution. The greater sequence coverage obtained from these platforms has facilitated deeper understanding of the micro-diversity of distinct groups of globally distributed viral communities (Gregory *et al.*, 2019), and allowed for the construction of so-called "mega-phages" with genomes >500 kb (Devoto *et al.*, 2019).

Although the use of short-read viromics has expanded our understanding of viruses within nature, short-read sequencing is not without its limitations. Phage genomes that

contain genomic islands and/or have high micro-diversity, such as those of the ubiquitous *Pelagibacterales* (Zhao *et al.*, 2013; Martinez-Hernandez *et al.*, 2019), may cause genome fragmentation during assembly (Temperton and Giovannoni, 2012; Mizuno, Ghai and Rodriguez-Valera, 2014; Roux *et al.*, 2017; Olson *et al.*, 2019). Furthermore, the choice of assembler will have a large impact on the quality of the final assembly (Sutton *et al.*, 2019). Other approaches, such as cloning large fragments into fosmids or techniques involving single-cell and/or single-virus MDA have been used, although these methods are technically challenging (Mizuno *et al.*, 2013; Roux *et al.*, 2014; Martinez-Hernandez *et al.*, 2019). Long-read sequencing may offer a more convenient solution.

1.10.1 Long-Read and Hybrid Viromics

The two predominant technologies for long-read sequencing are Oxford Nanopore Technology (ONT) and PacBio. ONT sequencing relies upon lengths of DNA being pulled through nanoscale protein pores (or nanopores) which are embedded into a membrane that separates a differential charge (Wang *et al.*, 2021). The DNA is pulled from the negatively charged side to the positively charged side, and differences in ionic current are measured and attributed to individual nucleotide bases (Wang *et al.*, 2021) (Figure 1.2). PacBio utilise an approach termed circular consensus sequencing (CCS). This approach involves attaching ssDNA hairpin adapters to target dsDNA, primers are annealed to the adapter and a DNA polymerase binds. The now circular sequence is amplified and sequenced in multiple passes to form subreads, which are then used to generate a consensus read (Kanwar *et al.*, 2021) (Figure 1.3). Whilst ONT and PacBio are both able to generate individual reads hundreds of kb long, the theoretical maximum for ONT is higher, with reads over 2 Mb having been obtained

(Payne *et al.*, 2019; Kanwar *et al.*, 2021; Wang *et al.*, 2021). Long read sequencing is therefore able to produce reads that span the entire length of phage genomes, potentially overcoming issues associated with fragmentation and assembly.

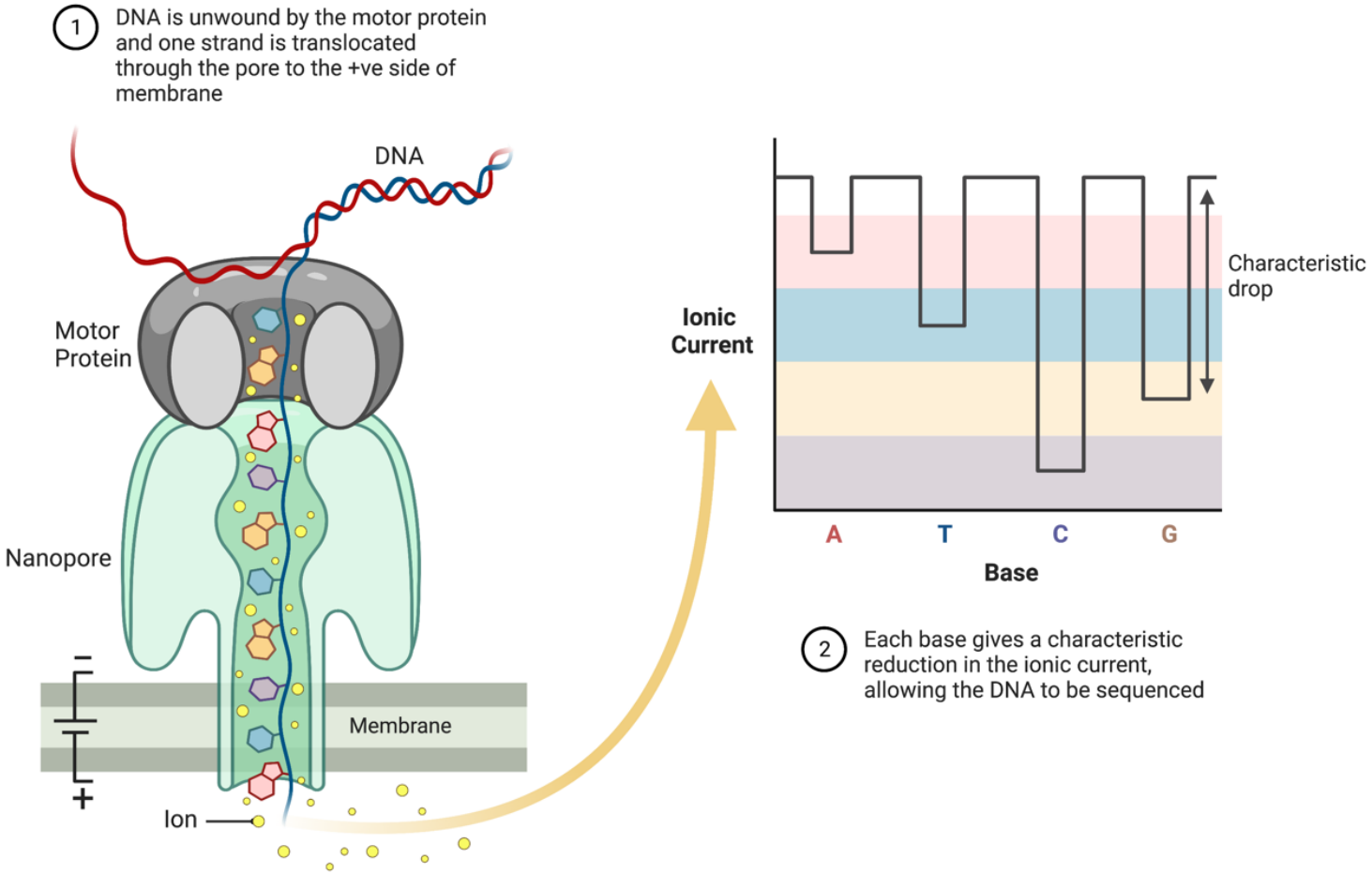


Figure 1.2 Nanopore sequencing principals

A schematic showing the main principals of Nanopore sequencing.

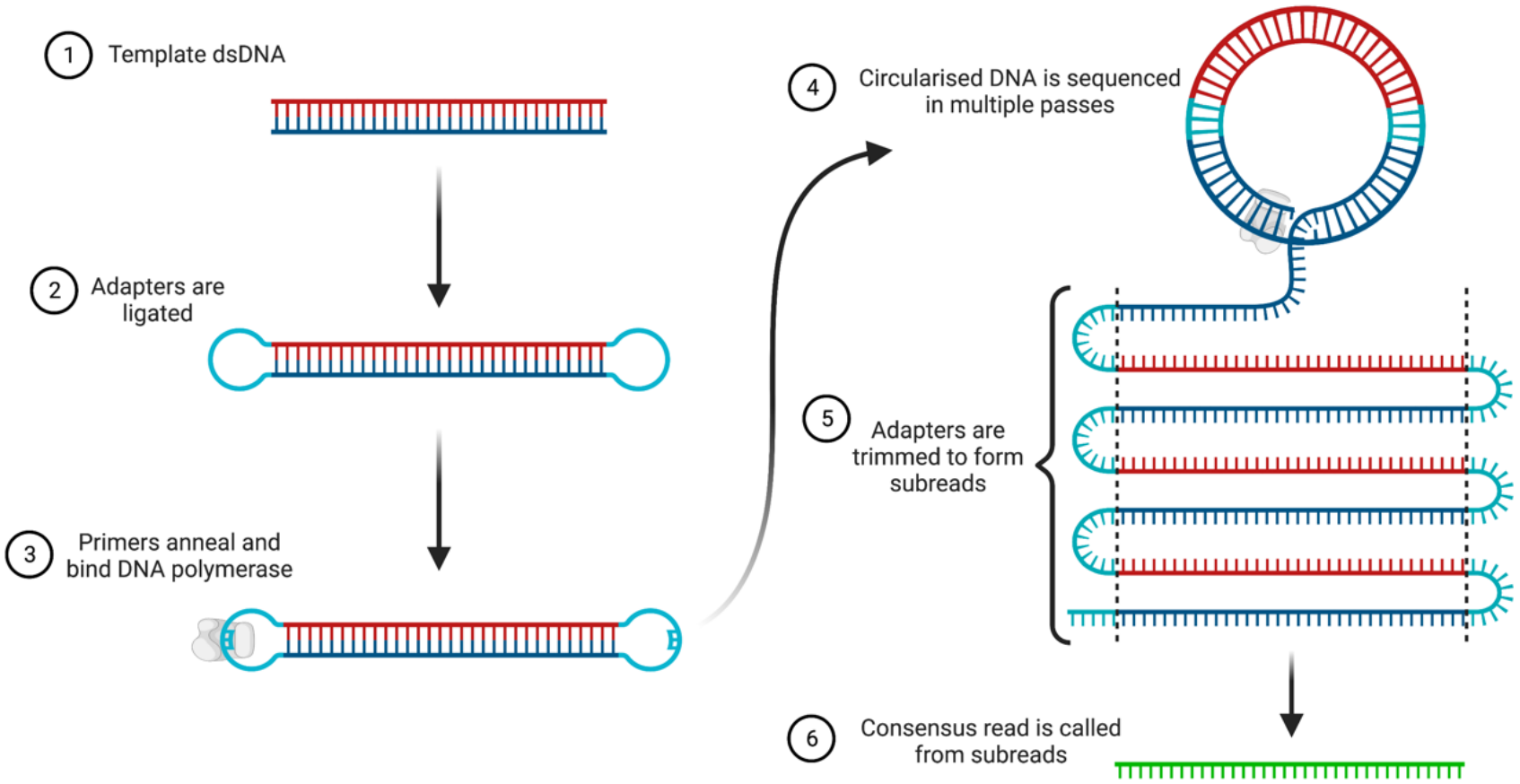


Figure 1.3 PacBio sequencing principals

A schematic showing the main principals of PacBio sequencing.

To date, there are limited but notable examples of ONT sequencing being used for the analysis of viromes. ONT sequencing of a human gut virome uncovered novel phage genomes and was able to detect epigenetic modifications that would be undetectable with short reads (Cao *et al.*, 2020), and novel oral bacteriophages have been uncovered from ONT sequenced metagenomes (Yahara *et al.*, 2021). However, the input requirements for ONT sequencing require large quantities of DNA. To overcome this, the human gut study isolated total nucleic acids and amplified them using a reverse transcription PCR reaction, and the oral metagenome sequenced DNA from the bacterial fraction, which yields higher quantities of DNA but offers less information about total viral diversity (Cao *et al.*, 2020; Yahara *et al.*, 2021). Alternatively, an assembly-free single-molecule approach was used to uncover previously unknown viral diversity within seawater samples (Beaulaurier *et al.*, 2020). This approach overcame the input requirements by filtering ~100 litres of seawater per sample and concentrating the resulting filtrate via tangential flow filtration (TFF) (Beaulaurier *et al.*, 2020). Whilst this could be achieved with pristine seawater, it can not from more viscous and heterogeneous environments.

It is possible to amplify DNA prior to sequencing with ONT, however, amplification techniques such as MDA may preferentially amplify some genomes (e.g., ssDNA) more than others and introduce biases in the abundance of genomes (Kim *et al.*, 2008; Yilmaz, Allgaier and Hugenholtz, 2010). Furthermore, there are other limitations beyond the input requirements. Long read sequencing platforms are known to have higher sequence error rates than Illumina (Buck *et al.*, 2017), which may in turn affect coding sequence (CDS) prediction and functional annotation (Watson and Warr,

2019). Due to the input requirements and higher error rate associated with long read sequencing, there is an emerging interest in the use of hybrid sequencing approaches.

Hybrid sequencing is an approach which utilises both short and long read sequencing platforms to sequence the same DNA sample. This approach aims to overcome the assembly fragmentation issues associated with short reads, and the higher error rates associated with long reads. A typical hybrid work-through may involve producing a long-read assembly, mapping short reads to the assembly and then using the short reads to correct, or "polish", errors in the long read assembly.

Hybrid sequencing approaches have been shown to increase the completeness and quality of bacterial metagenome assembled genomes (MAGs) from human and environmental samples (Bertrand *et al.*, 2019; Liu *et al.*, 2020, 2021; Brown *et al.*, 2021; Jin *et al.*, 2022). However, there are limited examples of hybrid sequencing being applied to viromes.

A hybrid approach that combined Illumina and ONT reads found inclusion of ONT reads improved the quality of MAGs for bacterial and viral genomes from groundwater samples (Overholt *et al.*, 2020). However, this study was of metagenomes and not VLP-enriched viromes. With regard to viromes, the most notable hybrid virome was sequenced and analysed using an approach dubbed virION (Warwick-Dugdale *et al.*, 2019). To overcome issues associated with input DNA requirements and biases of amplification, the virION method combined MinION sequencing with a long-read LASL, alongside Illumina sequencing (Warwick-Dugdale *et al.*, 2019). The virION approach improved recovery of high-quality genomes and was later improved to lower

the required quantity of input DNA (Zablocki *et al.*, 2021). Moreover, recent improvements were made to the metaFlye assembler (Kolmogorov *et al.*, 2020) to produce viralFlye with the specific aim of assembling viral genomes from long-read sequence data (Antipov *et al.*, 2022). This is the first long read assembler designed specifically for viral metagenomes.

Whilst there are a couple of examples of ONT sequencing being used to explore viral diversity, there are no notable equivalents conducted with PacBio sequencing. There are examples of single phage genomes being sequenced with PacBio (Akhwale *et al.*, 2019), and a study of prophages predicted from a PacBio bacterial metagenome (Zaragoza-Solas *et al.*, 2022), however there are no prominent viromes that have been sequenced with PacBio. The reasons for the absence of PacBio viromes are currently unclear. Furthermore, there is no robust comparison of different sequencing platforms for the recovery of viral genomes with a virome.

1.11 Cattle Manure and its Microbial Composition

Manure is an unavoidable by-product from the rearing of livestock. As manures are rich in nitrates and phosphates, they are a valuable source of organic fertiliser, which is typically applied to land in the form of semi-solid slurry. To produce slurry from manure, solids are separated using apparatus such as a screw press. The liquid fraction forms the basis of slurry, which is stored in a tank or lagoon, where it is mixed with water and other agricultural wastes before its application to land.

In the UK, dairy farms are estimated to be responsible for 80% of livestock manure production (Smith and Williams, 2016). There are ~ 2.7 million dairy cattle in the UK,

with ~ 1.8 million in milking herds (AHDB, 2018). An adult milking cow produces 7–8% of their own bodyweight as manure per day (Font-Palma, 2019), leading to an estimated 67 million tonnes produced annually (Smith and Williams, 2016). The economic value of cattle slurry is thought to be significant, estimated at an average value of £78 per cow per year (AHDB, no date b).

Despite their importance as a fertiliser, agricultural manures and slurries can be an environmental pollutant. Inadequate storage and agricultural run-off may lead to an increased biological oxygen demand (BOD) of freshwaters, leading to algal blooms and eutrophication (Sandars *et al.*, 2003; Thomassen *et al.*, 2008; Prapaspongsa *et al.*, 2010; De Vries, Groenestein and De Boer, 2012). Areas particularly at risk of nitrate pollution of ground or surface waters are classified as nitrate vulnerable zones (NVZs), and these constitute 55% of land in England (UK Government, 2013). To prevent pollution of freshwater, the application of organic fertilisers to fields in the UK is strictly controlled and can only be applied during certain times of the year (UK Government, no date). Thus, there is the requirement to store vast volumes of slurry for several months.

As cattle manure is the primary input of slurry, what is being excreted by the cow will likely be found within the slurry. Antibiotics given to livestock comprise 73% of global antibiotic sales (Van Boeckel *et al.*, 2017), the use of which has been implicated in the emergence of drug-resistant infections in humans (O'Neill, 2015) and animals (Aarestrup *et al.*, 2000). In the UK, dairy cattle are routinely treated with antibiotics for common illnesses including mastitis and respiratory illnesses (Oliver, Murinda and Jayarao, 2011). Furthermore, lameness—the costliest disease to UK dairy cattle

(CHAWG, 2020)—is typically prevented by treatment with footbaths that contain antimicrobial metals (e.g., copper and zinc) and/or other chemicals (e.g., formalin and glutaraldehyde) that are known to co-select for AMR (Pal *et al.*, 2015; Griffiths, White and Oikonomou, 2018; Davies and Wales, 2019). Therefore, dairy cattle slurries may contain selective and co-selective pressures for the transmission of AMR. Consequently, there is much interest in how antimicrobial resistant bacteria and ARGs associated with livestock may enter humans, whether directly through consuming animal products, or indirectly though animal wastes that are applied to the wider environment.

Culture-based techniques for the identification of multidrug resistant (MDR) bacteria within dairy wastes have found extended-spectrum beta-lactamase (ESBL) producing *E. coli* (Seiffert *et al.*, 2013; Ibrahim *et al.*, 2016), and metagenomic analyses have uncovered a diverse range of ARGs within the bacterial fraction of manure (Wichmann *et al.*, 2014; Zhou *et al.*, 2016). To ameliorate the risk of AMR transmission from cattle slurry into the wider environment, current recommendations are to store slurry for three months prior to application to ameliorate the risk of AMR (UK Government, 2016). However, there is no evidence given to justify this guidance.

To investigate the potential role of dairy cattle slurry in the transmission of AMR into the wider environment, a recent study profiled the bacterial fraction of a slurry tank over six months using a mixed methods approach (Baker *et al.*, 2022). Metagenomic analysis of the slurry tank samples revealed the tank to contain a diverse community of bacteria that was stable over time, with the two most dominant phyla being *Bacteroidetes* and *Firmicutes* (Baker *et al.*, 2022). Predicted ARGs in the

metagenomes were also stable over time, with the most common predicted ARGs conferring resistance to multiple drugs, tetracycline, MLS antibiotics (macrolides, lincosamides, and streptogramines), aminoglycosides, and beta-lactams. The diverse but stable profile of ARGs predicted in the metagenomes was mirrored in phenotypic resistance patterns observed in *E. coli* (Baker *et al.*, 2022). Furthermore, “mock” slurry tanks which did not receive regular influent after initial setup were profiled over time. These mock tanks showed that the abundance of many classes of resistance and clinically relevant bacterial genera decreased over time when the tank did not receive further influent (Baker *et al.*, 2022). Overall, this study concluded that dairy slurry tanks did not necessarily represent an AMR “hotspot”, but rather, that good management practises and storage of slurry could ameliorate the risk of transmission of AMR from livestock into the wider environment (Baker *et al.*, 2022).

Whilst there is emerging research into the bacterial fraction of cattle slurry, very little is known about the viral fraction. Viromic analyses of cattle has largely focused on rumen samples, which are now known to have a diverse and largely novel viral community that may augment host metabolism to aid the breakdown of complex carbohydrates (Berg Miller *et al.*, 2012; Ross *et al.*, 2013; Anderson, Sullivan and Fernando, 2017). Moreover, a recent analysis of pig faeces viromes that focused on the abundance of ARGs concluded that phage carriage of ARGs within pig samples was a rare event (Billaud *et al.*, 2021). However, the composition of the virome within cattle wastes and slurry is poorly studied. Individual phages infecting *Escherichia coli* have been isolated from slurry and characterised (Smith *et al.*, 2015; Sazinas *et al.*, 2018; Besler *et al.*, 2020), and there are limited studies into the rumen and cattle gut viromes (Ross *et al.*, 2013; Park and Kim, 2019), but total viral diversity within cattle

slurry remains largely unexplored. Given the widespread use of slurry in the environment, this paucity of knowledge is alarming.

1.12 Project Overview

It is increasingly clear that phages have significant ecological impacts, as this has been demonstrated in every environment in which they have been studied in detail. Phages are known to augment the metabolism of their hosts in a plethora of environments, however their impact in the transfer of ARGs remains a subject of debate. Our understanding of the specific mechanisms by which phages may augment the metabolism of their hosts is underpinned by genomics and viromics. The continual improvement of sequencing platforms and bioinformatic pipelines will continue to deepen our understanding of viruses in nature. Whilst we are implementing these work-flows in environments such as the human gut and ocean, there is a paucity of knowledge concerning phages within agricultural settings.

1.13 Research Objectives

The aim of this PhD project was to determine the diversity and ecological roles of bacteriophages within the dairy farm environment, with an emphasis on their potential roles in augmenting the metabolism of their bacterial hosts. Therefore, the objectives were to:

1. To develop methods to retrieve all currently sequenced bacteriophage genomes and investigate any biases within the current collection of complete phage genomes

2. To determine which sequencing platforms and bioinformatic approaches are best at recovering viral communities in nature
3. To characterise the viral community of agricultural slurry and determine the potential for bacteriophages to disseminate ARGs and virulence determinants in the wider environment through application of slurry to land
4. To determine if the presence of antimicrobials influences the viral communities within agricultural slurry
5. To characterise the viral community of the healthy dairy cow gut across key life stages and investigate the role of diet and age on the natural gut virome
6. To determine the functionality of putative phage-encoded ARGs

Chapter 2 INfrastructure for a PHAge REference Database

2.1 Chapter Preface

This chapter presents the work previously published in a paper format 'INfrastructure for a PHAge REference Database: Identification of Large-Scale Biases in the Current Collection of Cultured Phage Genomes. Cook, R. et al (2021) PHAGE.' https://doi.org/10.1089/phage.2021.0007. The text and figures from the published paper have been inserted into this chapter *verbatim*. As this work is not mine alone, the contribution of other authors is outlined below.

2.1.1 Author Contributions

Study design and the writing of an early version of the script that would go on to become INPHARED were performed prior to commencement of this PhD project. Andrew Millard, Martha Clokie, Branko Rihtman and Nathan Brown conceived the study. Ryan Cook and Andrew Millard carried out the bioinformatic analysis. Ryan Cook, Nathan Brown, Michael Jones and Andrew Millard drafted the manuscript. All authors approved and contributed to the final manuscript.

2.1.2 Chapter Objectives

The aim of this work was to determine how many phage genomes have been sequenced from cultured isolates to date, and to investigate any biases within the current collection of publicly available genomes. Therefore, the objectives were to:

1. To develop a method for the automatic retrieval of bacteriophage genomes and associated metadata
2. To create re-producible community resources to aid phage genomic analysis
3. To investigate biases in the collection of publicly available genomes (e.g., are bacteriophages from different hosts represented equally)

2.2 Abstract

Background: With advances in sequencing technology and decreasing costs, the number of phage genomes that have been sequenced has increased markedly in the past decade.

Materials and Methods: We developed an automated retrieval and analysis system for phage genomes (https://github.com/RyanCook94/inphared) to produce the INfrastructure for a PHAge REference Database (INPHARED) of phage genomes and associated metadata.

Results: As of January 2021, 14,244 complete phage genomes have been sequenced. The INPHARED data set is dominated by phages that infect a small number of bacterial genera, with 75% of phages isolated on only 30 bacterial genera. There is further bias, with significantly more lytic phage genomes (~70%) than temperate (~30%) within our database. Collectively, this results in ~54% of temperate phage genomes originating from just three host genera. With much debate on the carriage of antibiotic resistance genes and their potential safety in phage therapy, we searched for putative antibiotic resistance genes. Frequency of antibiotic resistance gene carriage was found to be higher in temperate phages than in lytic phages and again varied with host.

Conclusions: Given the bias of currently sequenced phage genomes, we suggest to fully understand phage diversity, efforts should be made to isolate and sequence a larger number of phages, in particular temperate phages, from a greater diversity of hosts.

2.3 Introduction

Bacteriophages (hereafter phages) are viruses that specifically infect bacteria and are thought to be the most abundant biological entities in the biosphere (Suttle, 2007). Phages may be obligately lytic (hereafter lytic) or temperate, whereby they have access to both the lytic and lysogenic cycle. Phages have many roles; in the oceans they are important in diverting the flow of carbon into dissolved and particulate organic matter through the lysis of their hosts (Suttle, 2007), or directly halting the fixation of CO_2 carried out by their cyanobacterial hosts (Puxty *et al.*, 2016). In the human microbiome, it is becoming increasingly clear that phages play roles in the severity and symptoms of several diseases. Many recent studies have shown disease-specific alterations can be seen in the gut virome community in both gastrointestinal and systemic conditions, including irritable bowel disease (Norman *et al.*, 2015), AIDS (Monaco *et al.*, 2016), malnutrition (Reyes *et al.*, 2015), and diabetes (Ma *et al.*, 2018).

Phages alter the physiology of their bacterial hosts such as by causing increased virulence, a notable example being phage CTX that actually encodes the toxins within the genome of *Vibrio cholerae*, which cause cholera (Waldor and Mekalanos, 1996). Furthermore, there are many cases where the expression of phage-encoded toxins cause otherwise harmless commensal bacteria to convert into a pathogen, including multidrug-resistant ST11 strains of *Pseudomonas aeruginosa* (van Belkum *et al.*, 2015; Tsao *et al.*, 2018), and the Shiga-toxin encoding *Escherichia coli* (O'Brien *et al.*, 1984). As well as increasing the virulence of host bacteria, phages can also utilize parts of their genomes known as auxiliary metabolic genes, homologues of host metabolic genes, to modulate their host's metabolism that can again have profound impacts on bacterial physiology and disease (Breitbart *et al.*, 2007).

Our understanding of how phages alter host metabolism has increased as the number of phage genomes has been sequenced. The first phage genome in 1977 (Sanger *et al.*, 1977), and since then, the relative ease of high-throughput sequencing combined with the resurgence of interest in this topic, has led to a rapid increase in the number of sequenced phage genomes (Hatfull, 2008; Perez Sepulveda *et al.*, 2016). The relatively simple nature of phage genomes means that the vast majority of isolated phage genomes can be fully assembled using short-read next-generation sequencing approaches (Rihtman *et al.*, 2016). As temperate phages can integrate into the genomes of their bacterial hosts as prophages, it is possible to predict prophage genomes within their bacterial hosts. However, not all predicted prophages can produce virions. Therefore, for the purposes of this study, phage genomes are those that have been experimentally verified to produce virions.

As sequencing capacity has increased, our understanding of the size of phage genomes has also increased. Between 2013 and 2016, a significant number of phages with genomes >200 kb were sequenced and dubbed “jumbo phages” (Yuan and Gao, 2017), although the isolation of “jumbo phages” is still thought to be rare. More recently, phages with genomes >500 kb have been reconstructed from metagenomes and referred to as megaphages, further expanding the known size of phage genomes (Devoto *et al.*, 2019).

The greater number of phage genomes available results in common analyses, including (1) comparative genomic analyses (Michniewski *et al.*, 2019; Rezaei Javan *et al.*, 2019), (2) taxonomic classification (Rohwer and Edwards, 2002; Adriaenssens *et al.*, 2018; Barylski *et al.*, 2019; Chibani *et al.*, 2019), (3) software for prediction of

novel phages (Akhter, Aziz and Edwards, 2012; Roux *et al.*, 2015; Arndt *et al.*, 2017; Bolduc *et al.*, 2017; Ren *et al.*, 2017, 2020), and (4) often the first step in analysis of viromes is the comparison of sequences with a known database. The huge amount of potential resource within phage genomes requires a comprehensive set of complete and consistently curated genomes from cultured isolates that can be used to build databases for further analyses.

When analyzing new phage genomes, it is important to know exactly how many phage genomes you are comparing the search with, and any biases (or not) inherent in that data set. Although this should be a relatively trivial question to answer, it is not because there are currently no such databases that contain only complete phage genomes that allow extraction in an automated reproducible manner. Although RefSeq provides well annotated complete phage genomes, it is not representative of the diversity of complete phage genomes. RefSeqs are only created for exemplar phage species, as defined by the International Committee on Taxonomy of Viruses (ICTV). Despite the tremendous work from the ICTV, the process of taxonomy approval is done annually and many phages remain without taxonomy. Thus, RefSeqs will always be catching up with the submission of new phage genomes and lag behind latest submissions. We have created an automated method for researchers to extract complete phage genomes from GenBank in a reproducible manner for use in genomic and metagenomic analyses, and provide general properties of the data set, thus allowing for better understanding of its features and limitations.

2.4 Materials and Methods

Phage genomes were download using the “PHG” identifier along with minimum and maximum length cutoffs. We also assume the genomes are from phages that have been shown to produce virions and are not predictions of prophages, a requirement of submitting phage genomes. Genomes were filtered based on several parameters to identify complete and near complete phage genomes. This includes initial searching for the term “Complete” and “Genome” in the phage description, followed by “Complete” and (“Genome” or “Sequence”) or a genome length of >10 kb. The list of genomes was then manually curated to identify obviously incomplete phage genomes, the accession numbers of genomes that are obviously incomplete were added to an exclusion list. As new genomes are added to GenBank continually, the INfrastructure for a PHAge REference Database (INPHARED) is designed to be updated continually. The use of an exclusion list allows the same incomplete genomes to be identified each time it is updated. An exclusion list is maintained on GitHub that can be added by the community. Although this process is not perfect, it provides a mechanism for the community to manually curate complete phage genomes that is better than one individual checking thousands of genomes repeatedly. Efforts to identify “false hits” were reported by many researchers, we would like to thank all members of the phage community who helped in initial curation.

After filtering, genes are called using Prokka with the --noanno flag, with a small number of phages using --gcode 15 (Seemann, 2014; Devoto *et al.*, 2019). Gene calling was repeated to provide consistency across all genomes, which is essential for comparative genomics. A prebuilt database (https://doi.org/10.25392/leicester.data.14242085) is provided so gene calling only

occurs on newly deposited genomes. The original GenBank files are used to gather metadata including taxa and bacterial host, and the Prokka output files are used to gather data relating to genomic features. The gathered data are summarized in a tab-delimited file that includes the following: accession number, description of the phage genome, GenBank classification, genome length (bp), molecular GC (%), modification date, number of coding sequences (CDS), proportion of CDS on positive sense strand (%), proportion of CDS on negative sense strand (%), coding capacity (%), number of transferRNAs (tRNAs), bacterial host, viral genus, viral subfamily, viral family, viral realm, Baltimore group (derived from phylum), and the lowest viral taxa available (from genus, subfamily, and family). Coding capacity was calculated by comparing the genome length with the sum length of all coding features within the Prokka output, and tRNAs were identified by the use of tRNA identifier. Other outputs include a fasta file of all phage genomes, a MASH index for rapid comparison of new sequences, vConTACT2 input files, and various annotation files for IToL and vConTACT2. The vConTACT2 input files produced from the script were processed using vConTACT2 v0.9.13 with --rel-mode Diamond --db “None” --pcs-mode MCL --vcs-mode ClusterONE --min-size 1 and the resultant network was visualized using Cytoscape v3.8.0 (Shannon *et al.*, 2003; Bin Jang *et al.*, 2019).

To identify genes indicative of a temperate lifestyle within genomes, we used a set of protein families Hidden Markov Models (HMM) as described previously (Clooney *et al.*, 2019; Cook, Hooton, *et al.*, 2021). These HMMs cover the integrase and transposase genes that are associated with the known integration methods of phages into bacterial genomes (PF07508, PF00589, PF01609, PF03184, PF02914, PF01797, PF04986, PF00665, PF07825, PF00239, PF13009, PF16795, PF01526, PF03400,

PF01610, PF03050, PF04693, PF07592, PF12762, PF13359, PF13586, PF13610, PF13612, PF13701, PF13737, PF13751, PF13808, PF13843, and PF13358) (Clooney *et al.*, 2019; Cook, Hooton, *et al.*, 2021). If a genome encoded one of these genes, it was assumed to be temperate. Antimicrobial resistance genes (ARGs) and virulence factors were identified using Abricate with the resfinder and VFDB databases using 95% identity and 75% coverage cutoffs (Seemann, no date a; Zankari *et al.*, 2012; Chen *et al.*, 2016).

The phylogeny of “jumbo phages” was constructed from the amino acid sequence of the TerL protein, extracted from 313/314 of the “jumbo phage” genomes. Sequences were queried against a database of proteins from non “jumbo phages” using Blastp and the top 5 hits were extracted with redundant sequences being removed (Altschul *et al.*, 1990). Sequences were aligned with MAFFT, with a phylogenetic tree being produced using IQ-Tree with “-m WAG -bb 1000” that was visualized using IToL (Nguyen *et al.*, 2015; Nakamura *et al.*, 2018; Letunic and Bork, 2019). Additional information was overlaid using IToL templates that are generated through INPHARED.

Rarefaction analysis was carried out for phage genomes from the top 10 most common hosts. Phage genomes were clustered at the level of genus if they belonged to the same vConTACT2 subcluster, and species using ClusterGenomes v5.1 (95% ID over 95% length) on the final set of nondeduplicated genomes, although RefSeq duplicates had been removed at this point (*GitHub - simroux/ClusterGenomes: Archive for ClusterGenomes scripts*, no date). An additional set of these genomes pooled together was included. Rarefaction curves and species richness estimates were produced using Vegan in R (Team, 2018; Oksanen *et al.*, 2020).

All data from January 2021 are available at Figshare https://doi.org/10.25392/leicester.data.14242085 and the script used for downloading and analyzing genomes is available on GitHub (https://github.com/RyanCook94/).

2.5 Results

The output of the INPHARED script provides a set of complete phage genomes, where all genes have been called in a consistent manner that allows comparative genomics and phylogenetic analysis. Unlike RefSeq, it will include all complete phage genomes, including those that have not been classified by the ICTV, and strains of the same phage species (or genome neighbors as they are referred to in the National Center for Biotechnology Information [NCBI] Viral Genomes Resource). In addition, it provides a MASH database to allow rapid comparison of new phage genomes against to identify close relatives, along with formatted databases for input into vConTACT2 to allow identification of more distant relatives. The host data (genus) for each phage are extracted along with summary information for each genome, which is reformatted to allow overlay onto trees in IToL so that the most common analyses for classification of new phages can be easily produced (Figure 2.5).

For this study, we used a lenient definition of “complete” to identify complete phage genomes. Strictly speaking, a complete phage genome would include the terminal ends of the phage genome, but because many phages are sequenced using a transposon-based library preparation (Rihtman *et al.*, 2016; Michniewski *et al.*, 2019), these terminal bases are never obtained (as transposons have to insert between bases). Another limitation for completeness is that for phage genomes with long terminal repeats; if the length of the repeat is larger than the library insert size, the repeats cannot be resolved. Details of library preparation, and if terminal ends have been confirmed, are not included in GenBank files, thus preventing automated retrieval of this information.

We then identify how many phage genomes have been sequenced to date and 18,134 genomes were extracted from GenBank. Of these, 3890 phage genomes are RefSeq entries that are derived from primary submissions, resulting in 14,244 complete phage genomes.

Current recommendations by the ICTV are that phages are uniquely named (Adriaenssens and Rodney Brister, 2017). Assuming a unique name represents a unique phage there are 12,127 phages. However, there are multiple examples of phages with the same name that are not genetically identical. Thus, phage names are not a suitable method for determining the number of unique phage genomes. As an alternative, deduplication of genomes at 100%, 97%, and 95% identity results in 13,830, 12,845, and 12,770 genomes, respectively.

Having established a data set of “complete” phage genomes, we then analyzed these data to look at how the number of phage genomes being sequenced over time is changing, the host they are isolated on, and overall genomic properties. First, we looked at the increase in the number of phage genomes that are sequenced over time. Although the number of phage genomes has rapidly increased over the past 20 years, the rate of increase has slowed in the past decade (Figure 2.1), with the number of phage genomes doubling every 2–3 years.

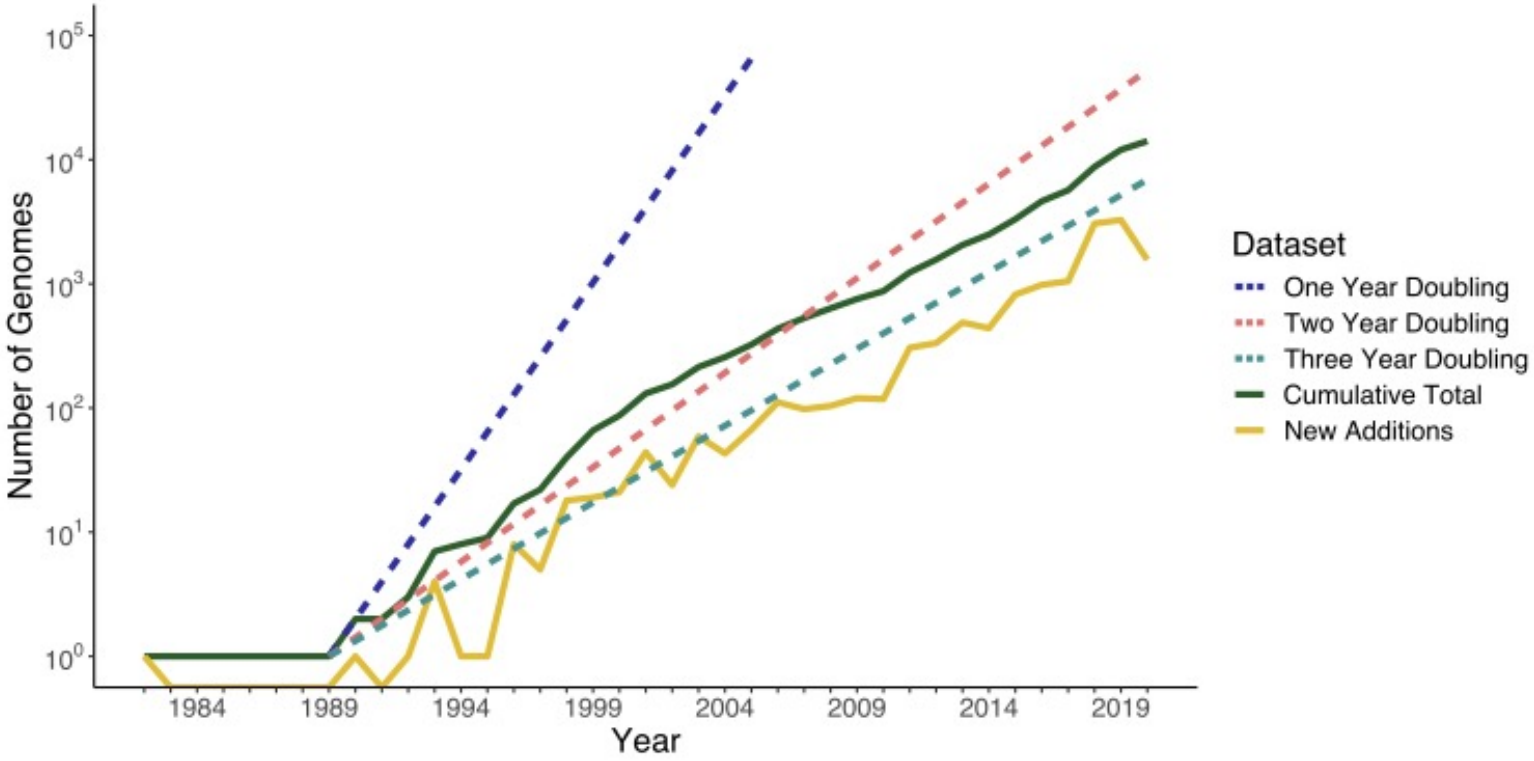


Figure 2.1 Number of complete phage genomes in GenBank over time

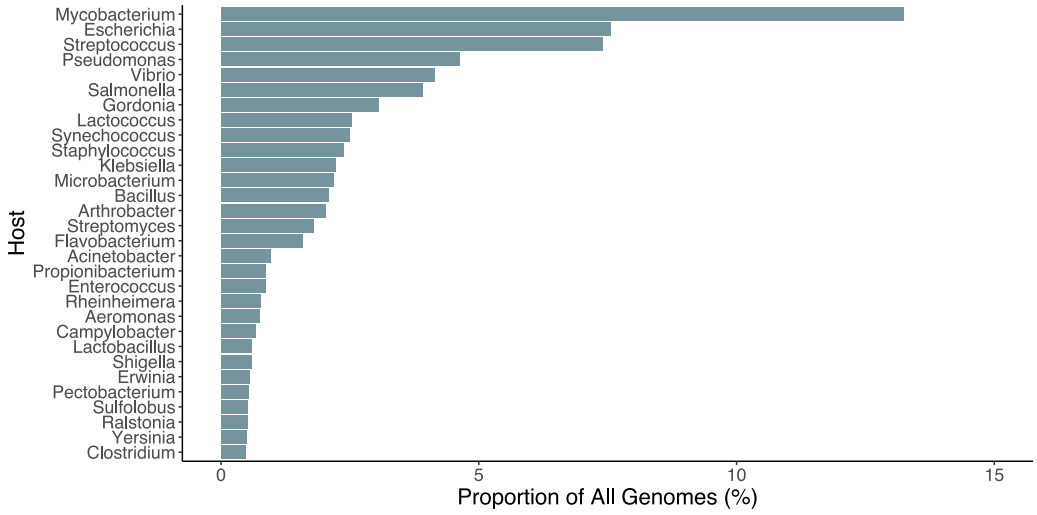
Dates were estimated based on date of submission (for 235 genomes, the date of update was used as no submission date was available). The reference lines showing doubling rates (dashed) begin in 1989, as this is when the number of phage genomes increased beyond the first submission in 1982.

2.5.1 Phage hosts and predicted gene function

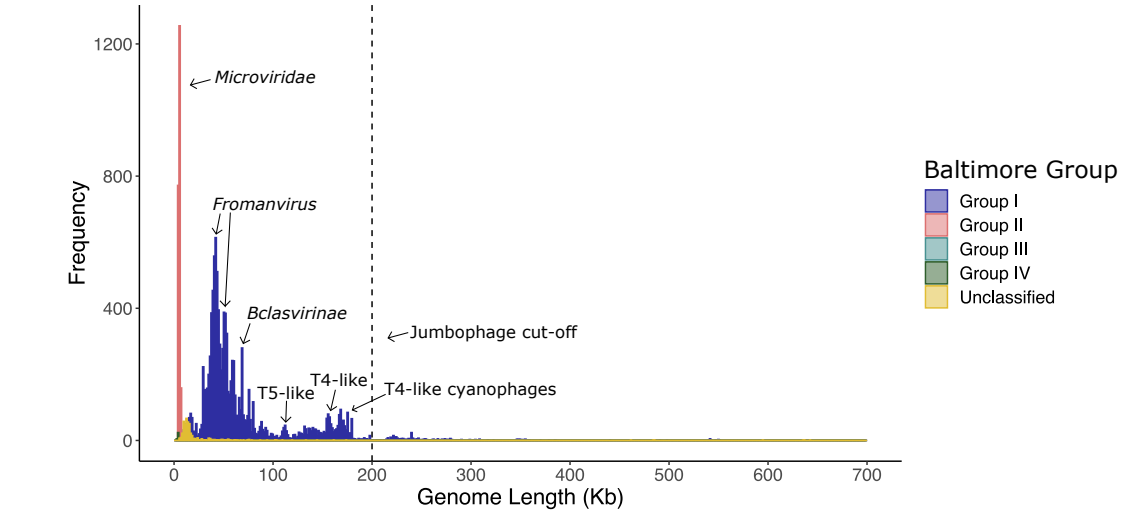
Utilizing our INPHARED database, we extracted the bacterial hosts and information on the predicted number of “hypothetical” proteins for each phage (i.e., so those with no predicted function). Across all phages, 56% of genes encoded hypothetical proteins, supporting the often quoted idea that the majority of genes encode proteins within unknown function (Edwards and Rohwer, 2005).

The host of 87% (12,402/14,244) of phages could be identified, with 13% of phages not having a known host or identifiable host information in the GenBank file, resulting in the genomes of phages infecting 234 different hosts (bacterial genera) having been sequenced. However, there is a clear bias in the isolation of phages against the same host (Figure 2.2A). Phages that infect *Mycobacterium* spp. are the most commonly deposited genomes (~13%), largely due to the pioneering work of the Science Education Alliance-Phage Hunters Advancing Genomics and Evolutionary Science (SEA-PHAGES) program (Hatfull *et al.*, 2006). This is followed by *Escherichia* spp., *Streptococcus* spp., and *Pseudomonas* spp. (Figure 2.2A).

A


Mycobacterium
Escherichia
Streptococcus
Pseudomonas
Vibrio
Salmonella
Gordonia
Lactococcus
Synechococcus
Staphylococcus
Klebsiella
Microbacterium
Bacillus
Arthrobacter
Streptomyces
Flavobacterium
Acinetobacter
Propionibacterium
Enterococcus
Rheinheimera
Aeromonas
Campylobacter
Lactobacillus
Shigella
Erwinia
Pectobacterium
Sulfolobus
Ralstonia
Yersinia
Clostridium
Host
0
5
10
15
Proportion of All Genomes (%)

B


1200
800
400
0
Frequency
Microviridae
Fromanvirus
Bclasvirinae
Jumbophage cut-off
T4-like
T5-like
T4-like cyanophages
0
100
200
300
400
500
600
700
Genome Length (Kb)
Baltimore Group
Group I
Group II
Group III
Group IV
Unclassified

C

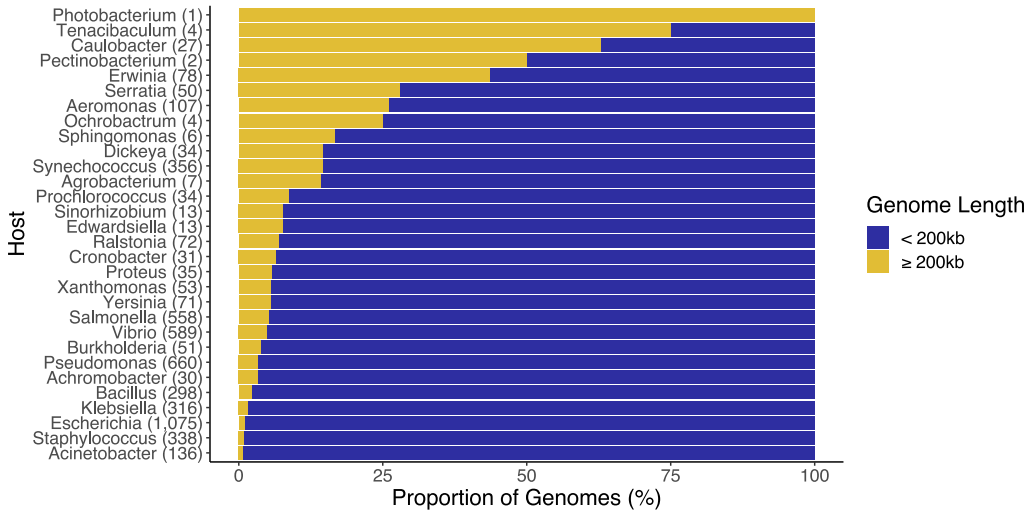

Photobacterium (1)
Tenacibaculum (4)
Caulobacter (27)
Pectinobacterium (2)
Erwinia (78)
Serratia (50)
Aeromonas (107)
Ochrobactrum (4)
Sphingomonas (6)
Dickeya (34)
Synechococcus (356)
Agrobacterium (7)
Prochlorococcus (34)
Sinorhizobium (13)
Edwardsiella (13)
Ralstonia (72)
Cronobacter (31)
Proteus (35)
Xanthomonas (53)
Yersinia (71)
Salmonella (558)
Vibrio (589)
Burkholderia (51)
Pseudomonas (660)
Achromobacter (30)
Bacillus (298)
Klebsiella (316)
Escherichia (1,075)
Staphylococcus (338)
Acinetobacter (136)
Host
0
25
50
75
100
Proportion of Genomes (%)
Genome Length
< 200kb
≥ 200kb

Figure 2.2 Overall properties of phages

(A) Proportion of phages isolated on the top 30 most abundant hosts. **(B)** Distribution of phage genome sizes with colors indicating Baltimore group and labels indicating typical phages for prominent peaks. **(C)** Proportion of “jumbo phages” on top 30 hosts for which at least one “jumbo phage” has been isolated with the total number of phages isolated against that host shown in brackets after its name.

Phages isolated on just 30 different bacterial genera account for ~75% of all phage genomes in the database (Supplementary Table S2.1). For non-deduplicated genomes isolated against the top 10 hosts specified in the GenBank file, we used rarefaction analysis to determine the diversity of these genomes and establish redundancy with respect to host. Using a cutoff of 95% identity over 95% length to define a species and vConTACT2 subclusters to define a genus, the number of phages continues to increase with the number of genomes sequenced (Figure 2.3). Suggesting that there is little redundancy within the database and we are not reaching the point where identifying new phage species is a rare event. Utilizing the rarefaction data for the top 10 hosts, we estimated how many different species of phages might infect each of these different bacterial genera (Supplementary Table S2.2). For *Mycobacterium*, there are 695 current phage species that lead to an estimation of 2132–2282 species that might infect *Mycobacterium*. Thus, even for hosts wherein thousands of phages have been isolated, we are only just scratching the surface of total phage diversity. We are also likely underestimating the total number of different phage species. In the case of *Mycobacterium*, a large proportion of phages have been isolated on only a single strain as part of the SEA-PHAGES program (Hatfull *et al.*, 2006). Thus, these phages are unlikely to be representative of phages that infect all bacterial species within the genus *Mycobacterium*. Increasing the diversity of the host *Mycobacterium*, that is, using more species of *Mycobacterium* for phage isolation, is likely to lead to more species of phage being isolated, increasing our estimates.

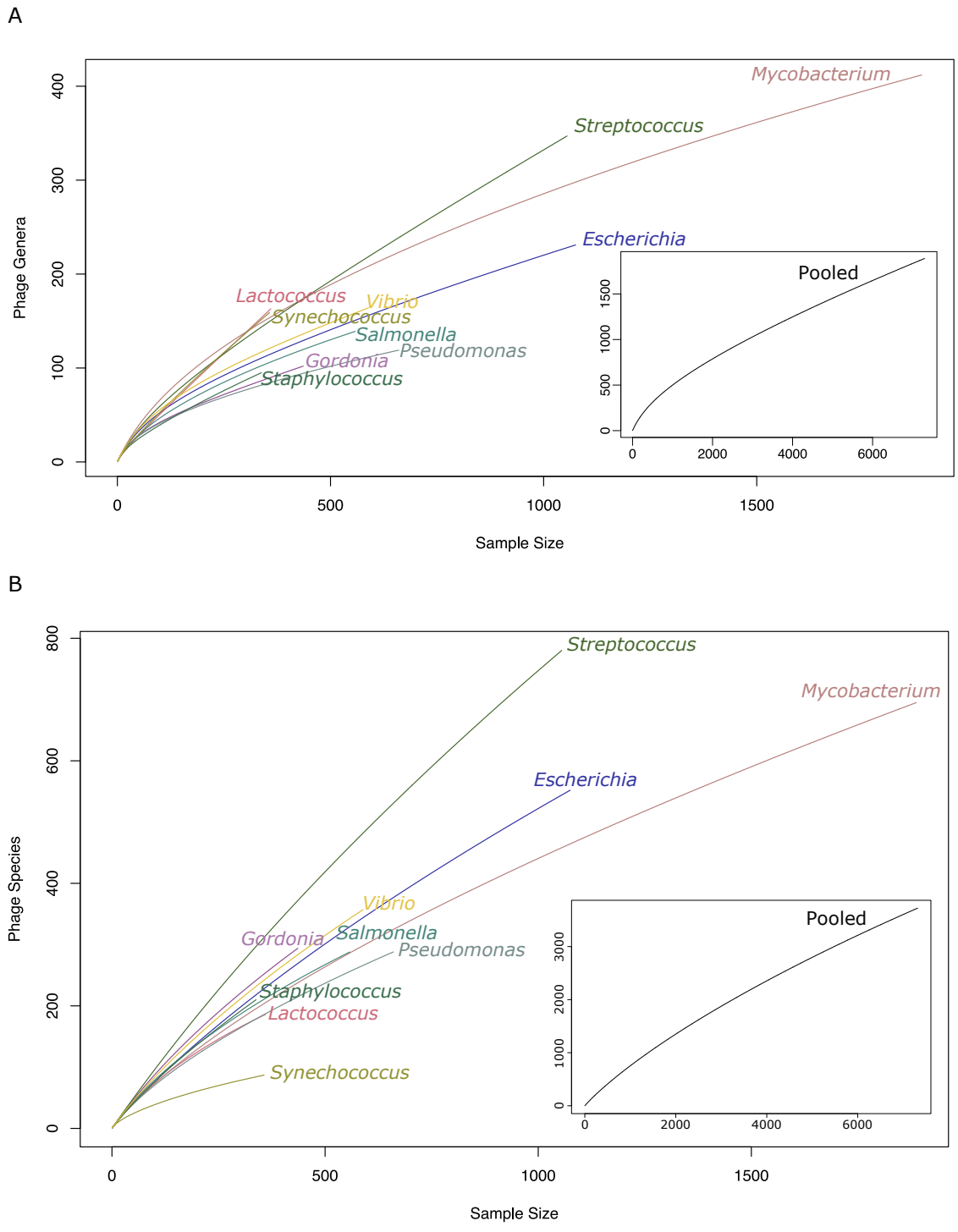


Figure 2.3 Genome diversity of phages on the top 10 most abundant hosts

Rarefaction curve of phage genera (**A**). Genera were defined by vConTACT2 clustering. Rarefaction curve of phage species (**B**). Species were defined as 95% identity over 95% of genome length.

2.5.2 Lytic and temperate phages

To identify whether phages are lytic or temperate, we searched for genes that facilitate a temperate lifestyle (e.g., integrase and recombinase) that have been used in previous studies to predict lytic/temperate phages (Clooney *et al.*, 2019; Cook, Hooton, *et al.*, 2021). This process is only a prediction and having such genes does not always mean the phage will enter a lysogenic cycle. However, it is a useful starting point that facilitates large scale comparative analyses when experimental data for all phages are either not available or readily accessible on such a scale.

Within the INPHARED data set, 4258 (~30%) phages have the potential to access a lysogenic lifecycle. The frequency of putative temperate phages was highly variable depending on the host (Figure 2.6A). The number of putative temperate phages is also biased toward a small number of hosts with 1217, 846, and 214 isolated on *Mycobacterium*, *Streptococcus*, and *Gordonia*, respectively. Collectively, these three hosts account for ~54% of all putative temperate phage genomes sequenced to date (Figure 2.6A).

2.5.3 Genome Sizes

Phage genomes ranged from 3.1 to 642.4 kb in size, with a wide distribution in the size of genomes with several observable peaks in genome size. The most prominent peaks are at 5-10, 40, 50, and ~165 kb (Figure 2.2B).

2.5.4 Coding capacity

The mean and median coding capacity was 90.45% and 91.52%, respectively (Figure 2.6B). Of the 14,244 genomes, 5731 (~40%) have ≥90% of coding features on one

strand and 3293 (~23%) of these are entirely on one strand (Figure 2.6C). The number of phages with genes encoding tRNAs was 4590 (~32%) and the number of tRNAs ranged from 1 to 62 with a median of 3 (mean of 7.23, and mode of 1). Although there is much literature on phage-encoded tRNAs, the roles they play remain unclear (Bailly-Bechet, Vergassola and Rocha, 2007).

2.5.5 Jumbo phages

Phages with genomes >200 kb are often referred to as "jumbo phages" and are reported to be "rarely isolated" and indeed only 314 genomes (~2.2%) fitting this definition were identified, suggesting that they are indeed rare (Yuan and Gao, 2017). To further investigate whether "jumbo phages" are as rare as is thought, we looked at the distribution in the context of the previously identified host bias. "Jumbo phages" have only been isolated on 31 of 234 identifiable bacterial hosts (Supplementary Table S2.1) and are far more commonly isolated on some hosts than others. Noticeably absent are any "jumbo phages" that infect *Mycobacterium*, *Gordonia*, *Lactococcus*, *Arthrobacter*, and *Streptococcus*, with >4000 phages having been sequenced from these bacterial hosts (Figure 2.2C).

For host bacteria that have had far fewer phages isolated on them such as *Caluobacter*, *Sphingomonas*, *Erwinia*, *Areomonas*, *Dickeya*, and *Ralstonia*, the frequency of "jumbo phage" isolation is far higher (Figure 2.2C). Owing to the small sampling depth of some of these hosts (e.g., *Photobacterium* and *Tenacibaclum*), it is not possible to determine whether the high proportion of genomes is merely a result of the low number of genomes sequenced. However, for other hosts such as *Aeromonas*, *Erwinia*, and *Caulobacter* from which >20 phages have been isolated,

~26%, ~44%, and ~63% are categorized as “jumbo,” respectively. Therefore suggesting “jumbo phages” are not always rare on particular hosts.

We further investigated the phylogeny of “jumbo phages” using the translated sequence of the *terL* gene. The “jumbo phages” are well distributed across the tree and do not form a single monophyletic clade, suggesting that they have arisen on multiple occasions with 14 clades containing at least one “jumbo phage.” Of these 14 clades, 12 also contain a non-“jumbo phage”. Furthermore, not all “jumbo phages” are equal, with “jumbo” cyanophages infecting the cyanobacteria *Synechococcus* and *Prochlorococcus* only marginally larger than their non-“jumbo” cyanophage relatives. These “jumbo phages” are also more closely related to their non-“jumbo” cyanophages relatives than other “jumbo phages” (Figure 2.4). A closer relationship of “jumbo phages” with non-“jumbo” phages than other “jumbo phages” is not limited to cyanophages (Figure 2.4). A similar pattern of grouping non-“jumbo” with “jumbo phages” is observed when a reticulate approach is used to look at the relatedness of phage genomes using vConTACT2 (Figure 2.7).

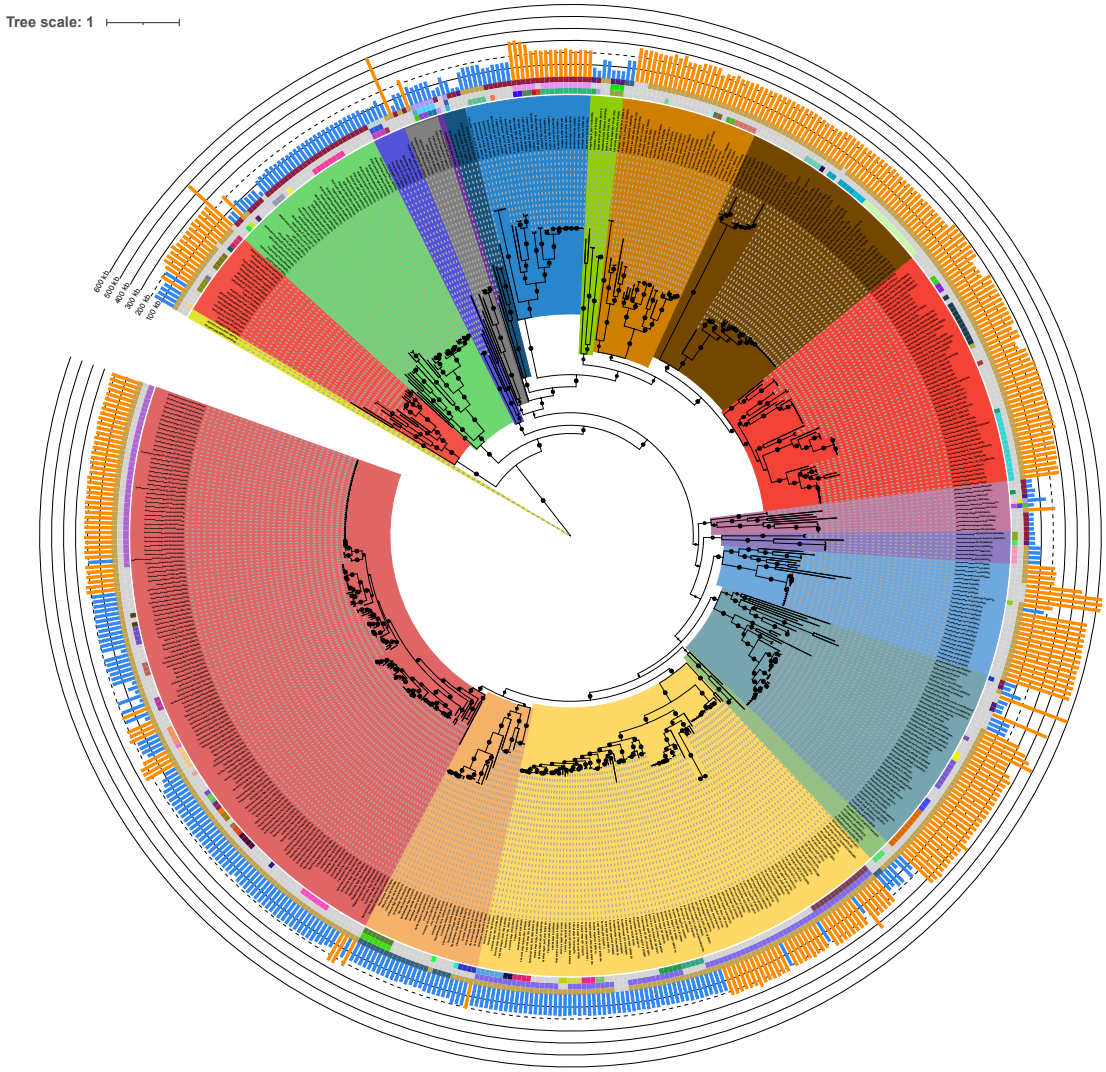


Figure 2.4 Phylogenetic tree of translated terL gene for 313 “jumbo phages” and their closest relatives

The alignment was produced using MAFFT and tree produced using IQTree using WAG model with 1000 bootstrap repeats (Nguyen *et al.*, 2015; Nakamura *et al.*, 2018). Colored regions indicate viral clades, colored rings indicate viral genus, subfamily, and family (innermost to outermost), and bars indicate genome length with blue and orange bars belonging to non-“jumbo” and “jumbo” phages, respectively. Bootstrap values indicated by black circles are scaled to the bootstrap value, with a minimum value of 70% displayed. Tree is rooted at the mid-point.

2.5.6 Virulence factors and ARGs

The presence of ARGs and virulence factors is a major concern for phage therapy, as the use of phages carrying such genes may make the populations of bacteria they are intended to kill more virulent or resistant to antibiotics. We therefore, used this database to investigate the frequency and diversity of phage-encoded virulence factors and ARGs. In total, 235 genomes (~1.6%) were found to encode a putative virulence factor and 43 genomes (~0.3%) to encode a putative ARG. The most common virulence genes were the *stx*$_{2A}$ (72 genomes) and *stx*$_{2B}$ (71 genomes) genes that encode subtypes of the Shiga toxin (Supplementary Table S2.3). The most common ARGs were the *mef*(A) (14 genomes) and *msr*(D) genes that confer resistance to macrolide antibiotics (Supplementary Table S2.4) (Daly *et al.*, 2004). Most genomes encoding a virulence factor were predicted to be from temperate phages (222/235), and were found to infect six bacterial genera, with the three most abundant hosts being *Streptococcus*, *Staphylococcus*, and *Escherichia*, respectively. The hosts for some genomes could not be determined (55/235). The genomes encoding virulence factors were widely distributed over 26 putative genera (Figure 2.7). All genomes encoding an ARGs were predicted to be temperate and were found to be isolated from eight bacterial genera, with the majority of phages linked to *Streptococcus* spp. (27/43).

2.6 Discussion

Defining how many different complete phage genomes have been sequenced is not as simple a question as it might appear. Based on accession numbers, there are 14,244 phage genomes, once RefSeq duplicates have been removed. Using unique names results in 12,127 phages, however, using names alone does not give an accurate estimate of the number of different phages, as genetically different phages have the same name. The use of deduplication at 100% identity suggests 13,830 unique phage genomes (January 2021) from cultured isolates. This also highlights that although RefSeq is a valuable resource, it is not at all representative of phage diversity. INPHARED provides a more comprehensive set of complete phage genomes from cultured phage isolates than RefSeq, in an easily accessible format. There are other resources that provide more comprehensive sets of phage genomes than RefSeq, including the NCBI Viral Genomes Resource (O'Leary *et al.*, 2016; Sayers *et al.*, 2020). The NCBI Viral Genome Resource allows manual filtering of phages through a graphical user interface and access to the same genomes in INPHARED. The automated filtering provided by INPHARED is a key difference, which prevents a user having to exclude the same genomes every time the database is updated. The integrated microbial genomes viral resource (IMG/VR) provides access to >2 million viral genomes, including phages, through a graphical interface (Roux *et al.*, 2021). The overwhelming majority of genomes in IMG/VR are constructed from metagenomes and have never been cultured. INPHARED is not designed to replace these valuable resources. The INPHARED provides rapid access to complete phage genomes from cultured phage isolates, without the need for continued manual filtering and provides metadata in an accessible format to allow initial analysis commonly used with phages to be carried out.

The INPHARED reveals clear patterns in phage genomes and biases in the selection of phage genomes that are currently available, but not always discussed in the analysis of genomes. The first is that the number of phage genomes is relatively small. Even for hosts wherein the highest number of phages have been isolated on, our estimates suggest thousands of new phage species remain to be isolated and sequenced. If we consider there are now >300,000 assembled representative bacterial genomes in GenBank, with many hundreds of thousands more for particular genera (e.g., >300,000 *Salmonella* and *Escherichia* genomes alone) compared with only 558 and 1075 of their respective phages, the representation of phage genomes to date is tiny compared with their bacterial hosts (Zhou *et al.*, 2020). Furthermore, the rate at which phage genomes are being sequenced is slowing down rather than increasing. Given the renewed interest in phages, and increased accessibility of sequencing, the decrease in the rate over time was surprising.

The second point of note is the bias in phage genomes. There is a clear bias in the isolation of phages from a small number hosts, with far more lytic than temperate phages. Thus, these phages are representative of these particular hosts, rather than phages in their entirety. Owing to the enormous success of the SEA-PHAGES program, many phages have been isolated on *Mycobacterium* and *Gordonia* (Hanauer *et al.*, 2017). This in turn results in approximately one-third of all temperate phage genomes being isolated on these two bacterial genera, whereas the remaining two-thirds are distributed across 142 different hosts.

The overrepresentation of phages infecting particular hosts can lead to truisms that may not be correct. For instance, “jumbo phages,” those that have genomes >200 kb, are rarely isolated (Yuan and Gao, 2017). Analysis of the INPHARED data set suggests ~2.2% of genomes fall into this category. However, this needs to be viewed in the context of the large bias in the hosts used for isolation, with ~75% of phages isolated on only ~16% of bacterial hosts that could be identified. When the number of “jumbo phages” is expressed as a percentage of all phage genomes, their isolation is clearly rare. For some hosts, such as *Mycobacterium*, many hundreds of phages isolated on the same host strain have been sequenced without the isolation of a “jumbo phage,” suggesting they are truly rare for this host (Hatfull *et al.*, 2006). However, for other hosts such as *Prochlorococcus*, *Synechococcus*, *Caulobacter*, and *Erwinia*, the isolation of “jumbo phages” is not a rare event. Although methodological adjustments of decreasing agar viscosity and large pore size filters may increase the number of phages isolated that have larger genome sizes, we suggest that using a wider variety of hosts may increase the number of “jumbo phages” isolated (Yuan and Gao, 2017). Phylogenetic analysis demonstrated that many “jumbo phages” are more closely related to non-“jumbo” phages than other “jumbo phages.” Thus, as the number of phage genomes has increased, an arbitrary descriptor of “jumbo” for phages with genomes >200 kb in length has less meaning. Recent comparative analysis of 224 “jumbo phages” used proteome size and analysis of protein length to determine a cutoff of 180 kb to separate “jumbo phages” from other phages. Using a clustering-based approach, three major clades of “jumbo phages” were identified (Iyer *et al.*, 2021). In this study, using *terL* as a phylogenetic marker to determine the phylogeny of 313 “jumbo phages” and their closely related phages suggests they have arisen on multiple occasions, as has been demonstrated

previously (Iyer *et al.*, 2021). “Jumbo phages” are clearly not monophyletic and what applies to one “jumbo phage” does not hold true for many others (Iyer *et al.*, 2021). As the number and diversity of “jumbo phages” increase, the use of the term seems to have less meaning.

With the increasing interest and use of phages for therapy, the isolation of phages that do not contain known virulence factors or ARGs is imperative. How frequently phages encode antibiotic resistance genes is a topic of much debate (Enault *et al.*, 2017; Debroas and Siguret, 2019). A previous study of 1181 phage genomes found that they are rarely encoded by phages, with only 13 candidate genes, of which 4 were experimentally tested and found to have no functional antibiotic activity (Enault *et al.*, 2017). We estimate that ~0.3% of phage genomes encode a putative ARG (none have been experimentally tested), a finding that is consistent with previous reports of low-level carriage in phage genomes in a data set that is ~10 × larger using similarly stringent cutoffs (Enault *et al.*, 2017). Critically, all of these ARGs were found in phages that are predicted to be temperate or have been engineered to carry ARGs as a marker for selection. With the frequency of carriage in temperate phages being ~1% overall. However, these data are still biased by the majority of temperate phages being isolated on only three bacterial genera. Notably no ARGs were detected on phages of *Mycobacterium*, which accounts for ~28% of temperate phages. In comparison, ~2.6% (27/1055) of temperate phages of *Streptococcus* carry putative ARGs and 50% of phages from *Erysipelothrix* (1/2) carry putative ARGs. Clearly a much deeper sampling of temperate phages from a broader range of hosts is required to get an accurate understanding of the role of phage in the carriage of ARGs. Based on the skewed data available to date, it seems unlikely there will be issues in the isolation of

lytic phages for therapeutic use that carry known ARGs within their genomes. However, we cannot determine whether these lytic phages can spread ARGs through transduction, or through carriage of as-yet uncharacterized ARGs.

Although there is much debate on the presence and importance of ARGs in phage genomes, the role of genes encoding virulence factors is well studied and the process of lysogenic conversion is well known (O'Brien *et al.*, 1984; Waldor and Mekalanos, 1996; van Belkum *et al.*, 2015; Tsao *et al.*, 2018). However, how widespread known virulence genes are in phages is not widely reported. We estimate ~1.6% of phages encode at least one putative virulence factor, with the frequency of carriage far higher in temperate phages (5.5%) than in lytic phages (0.13%). Again, these overall percentages are skewed by host bias with no known virulence factors detected in *Mycobacterium* temperate phages (0/1217), in comparison, 72% of temperate phages of *Shigella* (5/7) and 7% (61/846) of *Streptococcus* contain virulence factors. It is currently impossible to determine whether the higher proportion of ARGs and virulence factors in phages of known pathogens is a feature of their biology, or a skew in the database toward phages of clinically relevant isolates.

Given the biases in the data set, it is not clear whether the general phage patterns (e.g., jumbo phages are rarely isolated, more temperate phages on particular hosts, and the carriage of ARGs and virulence genes) are linked to biology or chronic undersampling of phage genomes that results in some bias. We speculate the latter, which distorts some generalizations about phages. Therefore, far deeper sampling of phage genomes across different hosts is required at an increasing rate.

2.7 Conclusions

We have provided a method to automate the download of a curated set of complete genomes from cultured phage isolates, providing metadata in a format that can be used as a starting point for many common analyses. Analysis of the current data highlights what we know about phage genomes is skewed by the majority of phages having been isolated from a small number of bacterial genera. Furthermore, the rate at which phage genomes are being deposited is decreasing. Although understanding of genomic diversity is always influenced by the data available, this is particularly acute for phage genomes with so many phages isolated on a small number of hosts. To obtain a greater understanding of phage genomic diversity, larger number of phages, in particular temperate phages, isolated from a broader range of bacteria need to be sequenced.

2.8 Supplementary Figures

Below are supplementary figures from the publication 'INfrastructure for a PHAge REference Database: Identification of Large-Scale Biases in the Current Collection of Cultured Phage Genomes. Cook, R. et al (2021) PHAGE.' https://doi.org/10.1089/phage.2021.0007.

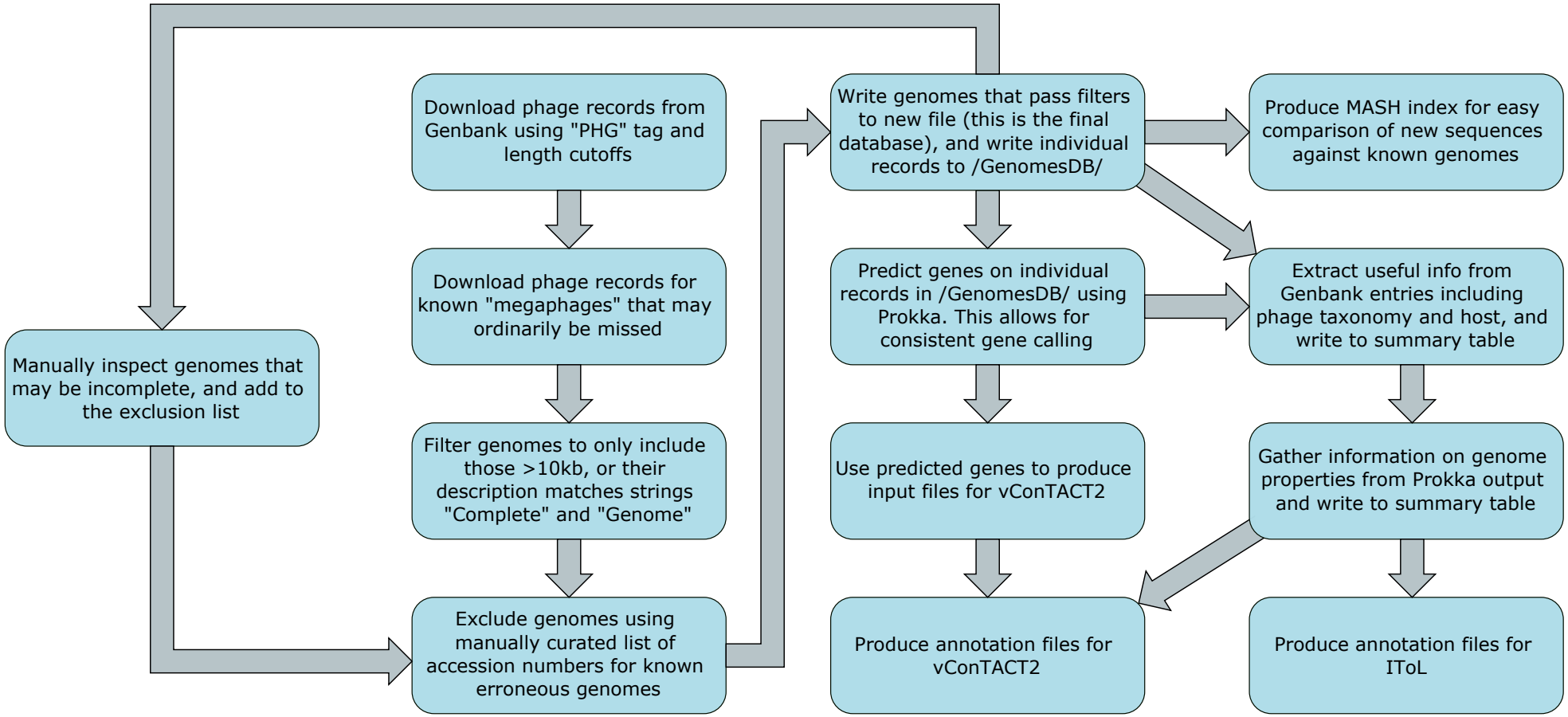
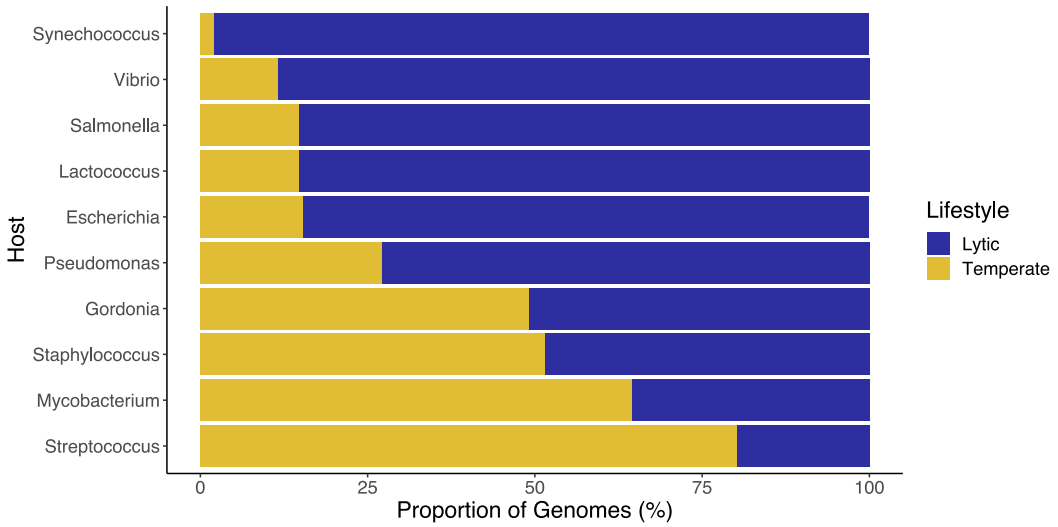


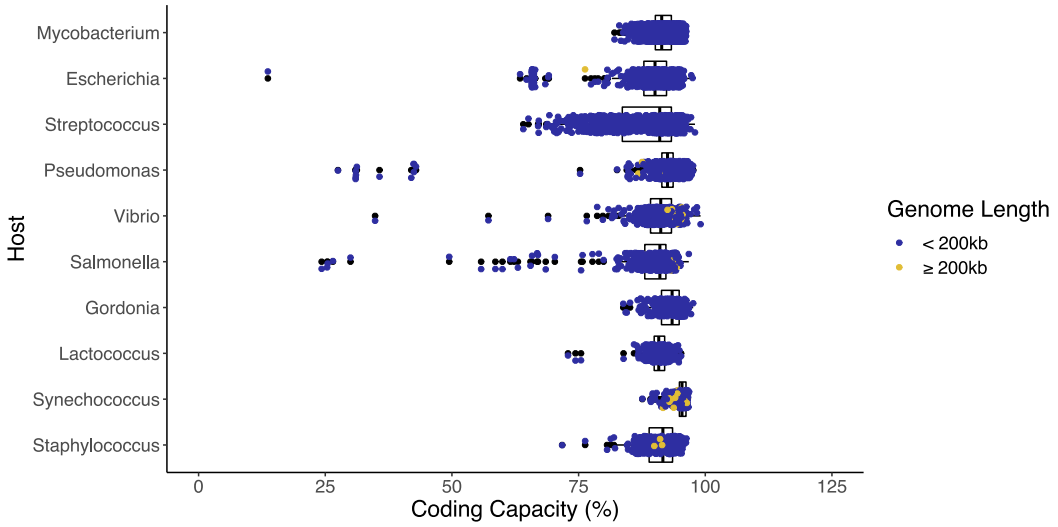
Figure 2.5 Outline of the INPHARED script

Simplified schematic showing the overall stages of the INPHARED script.

A


Synechococcus
Vibrio
Salmonella
Lactococcus
Escherichia
Pseudomonas
Gordonia
Staphylococcus
Mycobacterium
Streptococcus
Host
0
25
50
75
100
Proportion of Genomes (%)
Lifestyle
Lytic
Temperate

B


Mycobacterium
Escherichia
Streptococcus
Pseudomonas
Vibrio
Salmonella
Gordonia
Lactococcus
Synechococcus
Staphylococcus
Host
0
25
50
75
100
125
Coding Capacity (%)
Genome Length
< 200kb
≥ 200kb

C

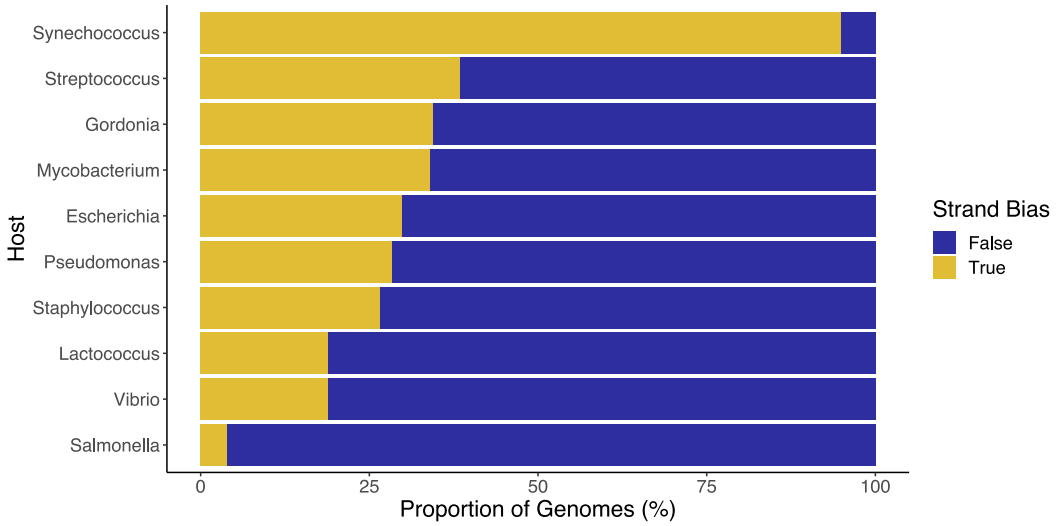

Synechococcus
Streptococcus
Gordonia
Mycobacterium
Escherichia
Pseudomonas
Staphylococcus
Lactococcus
Vibrio
Salmonella
Host
0
25
50
75
100
Proportion of Genomes (%)
Strand Bias
False
True

Figure 2.6 Genomic features for common hosts

Genomic features for phages of the ten most common bacterial hosts, showing (**A**) predicted lifestyle, (**B**) genome length, and (**C**) strand bias (defined as ≥90% of coding features on one strand).

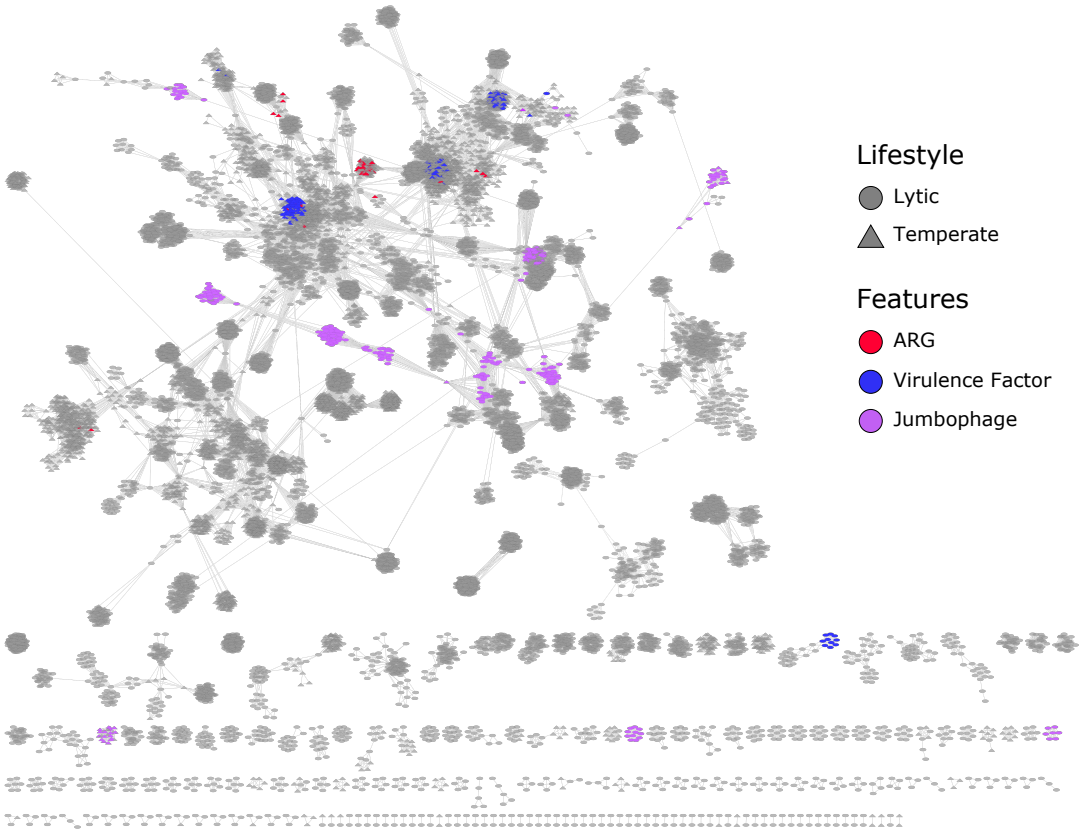


Figure 2.7 Distribution of ARGs, virulence factors, and jumbo-phages

vConTACT2 network showing large scale taxonomy of publicly available phage genomes, with shape indicating lifestyle, and colour indicating jumbo-phages, as well as the distribution of ARGs and virulence factors.

Chapter 3 Comparison of Illumina, Nanopore and PacBio sequencing for virome analysis

3.1 Chapter Preface

This chapter presents work prepared for submission to a journal in manuscript format 'Comparison of Illumina, Nanopore and PacBio sequencing for virome analysis'. The text and figures from the manuscript have been inserted into this chapter *verbatim*. As this work is not mine alone, the contribution of other authors is outlined below.

3.1.1 My Contributions

The study design, phage propagation and DNA extraction, sequencing, and some initial assemblies were performed prior to commencement of this PhD project. I completed the assemblies and performed the bioinformatic analyses. I drafted the manuscript with Andrew Millard. The version of the manuscript presented here has been contributed to and edited by Nathan Brown, Branko Rihtman, Slawomir Michniewski, Tamsin Redgwell, Martha Clokie, Dov J Stekel, Jon L Hobman, Michael A Jones, Darren Smith, and Andrew Millard.

3.1.2 Chapter Objectives

The aim of this work was to benchmark widely available sequencing technologies and assembly algorithms for the recovery of viral genomes from a mixed viral community. Therefore, the objectives were to:

1. To compare assemblies produced using different sequencing technology-assembler combinations against known genomes in a mock community
2. To determine the effect of polishing long-read assemblies (PacBio and ONT) with short-reads (Illumina)
3. To investigate whether the choice of sequencing platform and assembler affects common viromics analyses

3.2 Abstract

Viral metagenomics has fuelled a rapid change in our understanding of global viral diversity and ecology. Long-read sequencing and hybrid approaches that combine long and short read technologies are now being widely implemented in bacterial genomics and metagenomics. However, the use of long-read sequencing to investigate viral communities is still in its infancy. While Nanopore and PacBio technologies have been applied to viral metagenomics, it is not known to what extent different technologies will impact the reconstruction of the viral community.

Thus, we constructed a mock phage community of previously sequenced phage genomes and sequenced using Illumina, Nanopore, and PacBio sequencing technologies and tested a number of different assembly approaches. When using a single sequencing technology, Illumina assemblies were the best at recovering phage genomes. Nanopore- and PacBio-only assemblies performed poorly in comparison to Illumina in both genome recovery and error rates, which both varied with the assembler used. The best Nanopore assembly had errors that manifested as SNPs and INDELs at frequencies ~4x and 120x higher than found in Illumina only assemblies respectively. While the best PacBio assemblies had SNPs at frequencies ~3.5 x and 12x higher than found in Illumina only assemblies respectively. Despite high read coverage, long-read only assemblies failed to recover a complete genome for any of the 15 phage, down sampling of reads did increase the proportion of a genome that could be assembled into a single contig.

Overall the best approach was assembly by a combination of Illumina and Nanopore reads, which reduced error rates to levels comparable with short read only assemblies.

When using a single technology, Illumina only was the best approach. The differences in genome recovery and error rates between technology and assembler had downstream impacts on gene prediction, viral prediction, and subsequent estimates of diversity within a sample. These findings will provide a starting point for others in the choice of reads and assembly algorithms for the analysis of viromes.

3.3 Introduction

Due to the distribution and abundance of viruses, it is becoming increasingly apparent they play critical roles in all environments they are found. In particular viruses that infect bacteria, bacteriophages (from hereon in phages) are known to play important roles in regulating the abundance of their bacterial hosts, facilitating horizontal gene transfer and playing crucial roles in global biogeochemical cycles by augmenting host metabolism (Cobián Güemes *et al.*, 2016; Perez Sepulveda *et al.*, 2016; Breitbart *et al.*, 2018).

It is now over 40 years since the sequencing of the first DNA phage genome, by Sanger sequencing (Sanger *et al.*, 1977). The number of complete phage genomes from phage isolates is now >22,000 (Cook, Brown, *et al.*, 2021). However, millions more phage genomes have been sequenced through metagenome sequencing and are available through a variety of databases (Paez-Espino *et al.*, 2017; Gregory *et al.*, 2019; Roux *et al.*, 2021). Viral metagenomics (viromics) has revolutionised our understanding of the diversity of phages and their potential ability to augment host metabolism. Initial virome studies required DNA to be cloned into a vector and the clone sequenced by Sanger sequencing. As new sequencing technologies developed that did not require the cloning of DNA, such as Solexa (becoming Illumina), 454 and SOLiD, the field of viromics expanded. With Illumina sequencing becoming the dominant technology, more and more viromes have been sequenced from pristine ocean environments (Gregory *et al.*, 2019), the abyssal depths and from the faeces of a wide variety of animal species (Shan *et al.*, 2011; Babenko *et al.*, 2020; Camarillo-Guerrero *et al.*, 2021).

Whilst viromes produced using Illumina short-read sequencing have provided great insight into viral diversity, short reads are not able to resolve all viral genomes within a virome. Phages that contain hypervariable regions and or high microdiversity are known to cause virome assemblies to fragment, resulting in reduced contig size and exclusion from further analyses (Warwick-Dugdale *et al.*, 2019). To overcome these associated problems, alternative approaches to viromics can be taken, including the production of single cell viromics or the cloning of viral genomes into fosmids (Roux *et al.*, 2014). Whilst both of these approaches are beneficial, they are technologically challenging compared to more standard viromics workflows.

Recent technological developments have led to the production of long reads by both Oxford Nanopore Technology (ONT) (Wang *et al.*, 2021) and PacBio (Kanwar *et al.*, 2021). While the technologies differ in their approach, both platforms sequence single molecules and are capable of producing sequences of tens of kilobases in length (Kanwar *et al.*, 2021; Wang *et al.*, 2021). The ability to sequence long DNA molecules offers the ability to overcome the issues of microdiversity and or hypervariable regions found within phage genomes (Warwick-Dugdale *et al.*, 2019). To date there have been limited studies using ONT sequencing for viromics. One of the first studies to do so was able to acquire complete phage genomes from single ONT reads, utilising tangential flow filtration (TFF) of marine samples to obtain the required significant amounts of DNA for library input (Beaulaurier *et al.*, 2020). Extraction of such quantities of phage DNA is likely prohibitive from more viscous and heterogeneous environments where multiple displacement amplification (MDA) is already used to obtain enough DNA for library preparation for short read sequencing. While MDA provides a solution to the amount of input material, it does not come without problems.

It has been well documented that MDA can introduce biases in metagenomic libraries, in particular the over representation of ssDNA phages within samples (Yilmaz, Allgaier and Hugenholtz, 2010; Kim and Bae, 2011; Marine *et al.*, 2014). To overcome the problem of library input requirements, MDA for ONT library preparation, combined with unamplified short read libraries for quantification has been utilised (Cook, Hooton, *et al.*, 2021). Alternatively, ONT sequencing (minION) in combination with long-read linker amplified shotgun library (LASL) to sequence PCR products on a minION, combined with Illumina short reads were used in an approach dubbed virION (Warwick-Dugdale *et al.*, 2019; Zablocki *et al.*, 2021). Both approaches were successful in increasing the number and completeness of viral genomes.

While the number of viromes that utilise ONT alone or in combination with Illumina sequencing is slowly increasing (Warwick-Dugdale *et al.*, 2019; Cook, Hooton, *et al.*, 2021; Michniewski *et al.*, 2021; Yahara *et al.*, 2021; Zablocki *et al.*, 2021; Zaragoza-Solas *et al.*, 2022), reports of utilising PacBio sequencing for viromes are scarce (Zaragoza-Solas *et al.*, 2022). A recent study predicted phages from a bacterial metagenome assembled from PacBio reads, identifying phages not identified when the same sample was sequenced with short reads (Zaragoza-Solas *et al.*, 2022). Why there are not more viromes that are sequenced with long read technology, as has become commonplace for sequencing of bacterial metagenomes, is not clear. Even for the sequencing of individual phage isolates, there are relatively few studies that have utilised long reads (Akhwale *et al.*, 2019; Eckstein *et al.*, 2021; Kupritz *et al.*, 2021; Song *et al.*, 2021). In part, this is likely because the vast majority of phage genomes can be assembled from short read Illumina sequences alone (Rihtman *et al.*, 2016). Thus, unlike sequencing their bacterial hosts, long reads do not provide the

immediate benefit of a better genome assembly for an isolate and thus the need to use them is reduced. The lack of long-read data generally for phage isolates, combined with the lack of a comparative benchmarked dataset comparing different methods is likely contributing to long read sequencing not being widely adopted for viromes, despite clear benefits from the limited studies to date.

We have therefore sequenced a mock community of phages with three different sequencing technologies (PacBio, minION and Illumina) to benchmark the different approaches, in order to identify the benefits and limitations of each approach.

3.4 Materials and Methods

3.4.1 Mock Virome Preparation and Sequencing

Phages (vB_Eco_SLUR29, vB_EcoS_swan01, vB_Eco_mar001J1, vB_Eco_mar002J2, KUW1, PARMAL1, HP1, DSS3_PM1, vB_Eco_mar005P1, S-RSM4, vB_Eco_mar003J3, vB_Vpa_sm033, vB_VpaS_sm032, CDMH1) were propagated as previously described (Rihtman et al., 2016), and DNA was extracted using a standard phenol:chloroform method. DNA was quantified with Qubit dsDNA high sensitivity kit. ΦX174 DNA was obtained from the spike in control provided with Illumina library preparation kits. Genomic DNA was combined to produce a mock community of fifteen phages that covered a range of lengths (44,509 - 320,253 bp) and molGC content (38% - 61%). Genomes were combined across a range of abundances (169,000 - 684,329,545 genome copies) within the mock community (Supplementary Table S3.1). Genome copies were estimated by using the formula: (ng of DNA * 6.022 x 1023) / (Genome Length * 660 * 1 x 109). The genomes were chosen to include both highly divergent and highly similar phages (Supplementary Table S3.2; Figure 3.12).

Illumina library preparation was carried out using the NexteraXT library preparation kit, with a minor modification to the number of PCR cycles as described previously (Michniewski *et al.*, 2019). In addition, no ΦX174 spike was added to the library as is part of the normal Illumina library preparation protocol. Sequencing was carried out with a MiSeq 2 x 250 bp kit. For minION and PacBio sequencing, the DNA was amplified prior to sequencing with the GenomiPhi V3 DNA Amplification Kit, following the manufacturer's instructions. Eight individual amplification reactions were performed with 10 ng of DNA input for each amplification. Following amplification, DNA

was treated with S1 nuclease with 10 U per µg of input DNA and the enzyme deactivated, prior to cleanup and concentration with a DNA Clean & Concentrator-25 column (Zymo Research). Three independent amplification reactions were sequenced via PacBio or ONT sequencing.

Libraries were prepared for minION sequencing using SQK-LSK109 (Version: NBE_9065_v109_revB_23May2018) with the native barcoding kit, following the manufacturer's instructions (Oxford Nanopore Technologies, Oxford, UK) with omission of the initial g-tube fragmentation step. Base calling was carried out with Guppy v2.3.5, with reads demultiplexed using Porechop (https://github.com/rrwick/Porechop). PacBio sequencing was carried out at NUomics using the Sequel platform.

3.4.2 Bioinformatics Analyses

To determine coverage and depth, reads from each library were mapped to the 15 reference genomes using Minimap2 v2.14-r892-dirty with “-ax sr”, “-ax map-ont”, or “-ax map-pb” for Illumina, ONT and PacBio reads respectively (Li, 2018). Minimap2 output was piped and sorted using the Samtools sort command to produce sorted bam files (Li *et al.*, 2009). Coverage and depth were taken from the bam files using the Samtools coverage command (Li *et al.*, 2009).

Assemblies were separately produced for the three libraries, and additional assemblies were produced by pooling the three libraries together, resulting in four assemblies per read/assembler combination. Illumina reads were trimmed with Trim Galore v0.4.3 prior to assembly (https://github.com/FelixKrueger/TrimGalore). Illumina

reads were assembled using SPAdes v3.12.0 with parameters “--meta -t 16” (Nurk *et al.*, 2017). Flye assemblies were produced with parameters “--nano-raw/or --pacbio-raw --threads 90 --meta” (Kolmogorov *et al.*, 2019). Unicycler assembly of long reads was used with default parameters (Wick *et al.*, 2017), that utilise miniasm (Li, 2016) for an overlay consensus assembly followed by racon for polishing (Vaser *et al.*, 2017). wtdbg2 was used with the parameters “-p 21 -k 0 -AS 4 -K 0.05 -s 0.05 -L 1000 --edge-min 2 --rescue-low-cov-edges -t 90” (Ruan and Li, 2020).

To determine whether using long and short reads together improved the assemblies, three methods that utilised a hybrid approach were used. (1) Long read-only assemblies were polished with multiple rounds of polishing using Pilon (Walker *et al.*, 2014) (hereafter referred to as “polished”). (2) For a hybrid assembly with Unicycler, long and short reads were provided with default parameters (hereafter referred to as “hybrid”). (3) The hybrid Unicycler assemblies were combined with the Illumina-only assemblies and de-replicated at 95% average nucleotide identity (ANI) over 80% genome length using the ClusterGenomes script (*GitHub - simroux/ClusterGenomes: Archive for ClusterGenomes scripts*, no date) (hereafter referred to as “deduped”).

To assess completeness and quality assemblies were compared to the 15 reference genomes using metaQUAST v5.0.2 with default parameters (Mikheenko, Saveliev and Gurevich, 2016). All resultant plots were produced using ggplot2 in R v3.5.1. When investigating the fidelity of assemblies to the reference genomes, we included assemblies for which 50% of the genome was covered by contigs, no matter how fragmented the assembly was (i.e., if 100 individual contigs mapped to 50% of genome length, despite the longest contig only being 10% of genome length, this was still

included. This was to exclude misassembly data for which only small portions of genomes were assembled, potentially leading to under-estimation of error frequencies). To investigate the effect of sequence depth on long read assembly, reads mapping to the genome of interest were extracted and downsampled using seqtk sample with -s100 to the desired depth (https://github.com/lh3/seqtk).

To determine the effect of polishing long-read assemblies with short-reads on viral prediction software, we processed the long-read assemblies and their polished counterparts using VIBRANT v1.2.1 (Kieft, Zhou and Anantharaman, 2020) with the following parameters "-t 8 -l 10000 -virome" and compared against DeepVirFinder v1.0 (Ren *et al.*, 2020) with contigs >10 kb and a P-value <0.05. Prodigal v2.6.2 with default settings was used for predicting open reading frames on the vOTUs and the 15 reference genomes (Hyatt *et al.*, 2010).

To investigate the effect of different sequencing platforms and assemblers on estimates of viral diversity, we applied a typical virome analysis workflow to the assemblies. Each assembly was separately processed using DeepVirFinder v1.0 (Ren *et al.*, 2020). Contigs ≥10 kb or circular were included as viral operational taxonomic units (vOTUs) if they obtained a P-value of <0.05. Reads from the corresponding Illumina library were mapped to the assembly using Bbmap v38.69 at 90% minimum ID and the ambiguous=all flag (Bushnell, 2013). vOTUs were deemed as present in a sample if they obtained ≥ 1x coverage across ≥ 75% of contig length (Roux *et al.*, 2017). The number of reads mapped to present vOTUs were normalised to reads mapped per million. Relative abundance values were analysed using Phyloseq v1.26.1 (McMurdie and Holmes, 2013) in R v3.5.1 to calculate diversity statistics

(Team, 2018). The number of predicted vOTUs and alpha diversity statistics were compared to the genome copy numbers used in the original mock community.

3.5 Results

3.5.1 Mock Virome Composition

To assess the performance of short, long, and hybrid sequencing approaches for viromic analyses, we sequenced a mock community of 15 bacteriophage genomes with an Illumina MiSeq, PacBio Sequel, and ONT minION. For Illumina sequencing, no MDA was used to provide a library as free from bias as possible. For PacBio and ONT sequencing, the mock community was first amplified with MDA to obtain sufficient material for library preparation and sequencing. The Illumina and ONT libraries yielded similar amounts of data with 0.5 - 1.1 Gb and 0.6 - 1.1 Gb respectively, and 0.3 - 0.5 Gb from PacBio libraries. Pooling the libraries resulted in 2.4, 2.7 and 1.1 Gb for Illumina, ONT and PacBio libraries respectively (Supplementary Table S3.3).

3.5.2 Limits of Detection by Read Mapping

First, we assessed the limits of detection of each sequencing platform using a mapping-based approach, with detection of a genome set at 1x coverage across ≥97% of a genome. Four phage genomes were not detected at all (CDMH1, HP1, vB_Eco_mar005P1 and ΦX174) by any sequencing technology (Figure 3.1A). The Illumina and ONT libraries detected a similar number of genomes (8-10 genomes), with PacBio detecting between 7-8 genomes across the separate libraries (Figure 3.1B). Although Illumina and ONT both recovered between 8-10 genomes across all libraries, the average number of genomes detected in a single Illumina library was higher than that of a single ONT library (Figure 3.1B). The least abundant phage to be detected was S-RSM4 (465,530 copies) and was only detected by Illumina sequencing, although a small percentage of the genome was covered in the PacBio

and ONT libraries. The least abundant phages detected in ONT and PacBio libraries were vB_VpaS_sm032 (52,465,265 copies) and J1 (53,672,906 copies), respectively.

The use of unamplified DNA for Illumina libraries allowed any effects of MDA to be identified in the long read assemblies. Encouragingly, the abundance of a genome within a sample generally correlated across different sequencing platforms, even after MDA for PacBio and ONT sequencing (ONT vs Illumina r=0.9903948, PacBio vs Illumina r=0.9883086, ONT vs PacBio r=0.9996938) (Figure 3.2A, B, and C; Supplementary Table S3.4). Although, it should be noted that phage ΦX174 was not detected in any sample, suggesting we may have been overly cautious in the amount we added to the mock community.

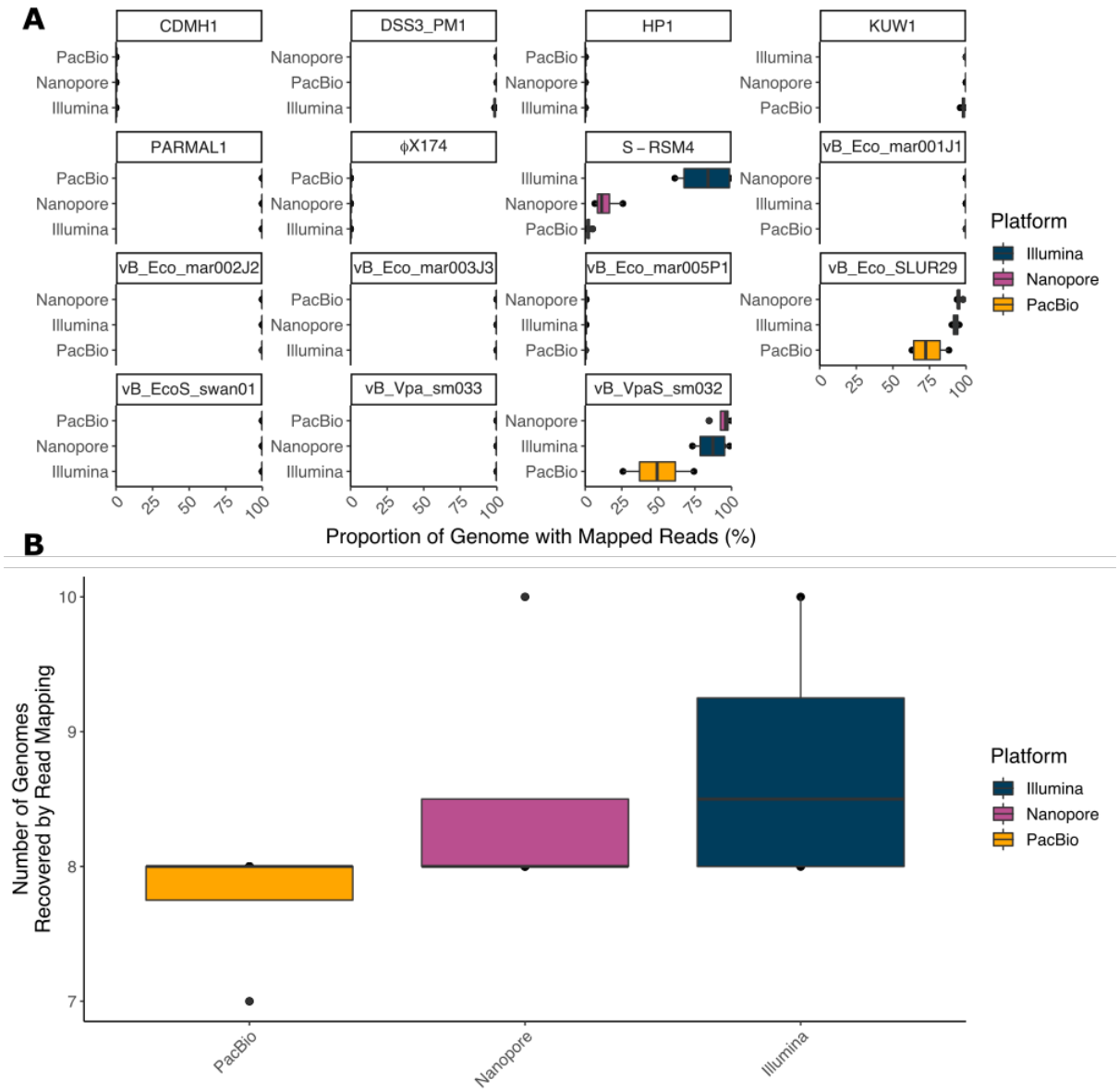


Figure 3.1 Detection of genomes by read mapping

(**A**) Boxplots showing the proportion of each genome to which reads were mapped from each of the three sequencing platforms for library repeats, and (**B**) the number of genomes detected by read mapping by each sequencing platform at 1x coverage over ≥97% of genome length.

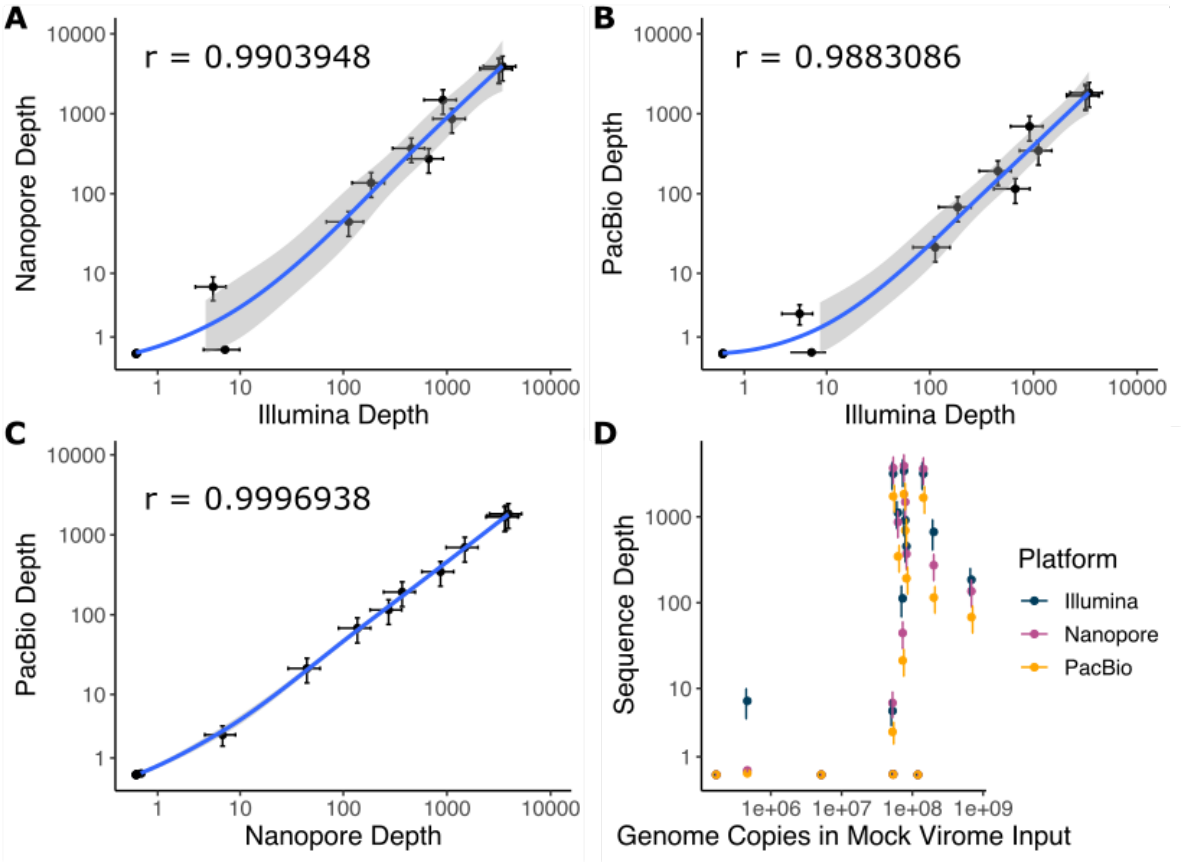


Figure 3.2 Comparison of sequencing depth between platforms

Correlation plots showing average sequence depth of a genome between (**A**) Illumina and ONT, (**B**) Illumina and PacBio, and (**C**) ONT and PacBio. An additional plot (**D**) shows sequence depth for the three sequencing platforms versus the estimated number of genome copies in the original mock community from which DNA libraries were prepared and sequenced. Values shown are the mean across three libraries and a pooled library, with bars showing standard error.

3.5.3 Assembly Results - Genome Recovery

As assembly options for each read type were tested to optimise assembly methods, assemblies were obtained for all samples and assemblers tested, with the exception of PacBio reads using Unicycler (miniasm + racon) so were excluded from further analysis. To investigate whether combining read technologies led to more complete assemblies, PacBio and ONT reads were separately assembled alongside Illumina reads using Unicycler to produce “hybrid” assemblies. The hybrid assemblies were separately combined with Illumina only assemblies and de-replicated at 95% average nucleotide identity (ANI) over 80% to produce “deduped” assemblies (Nayfach *et al.*, 2020).

For individual sequencing platforms, only short reads (Illumina) resulted in any completely assembled genomes (3-4) (Figure 3.3 and Figure 3.13). Despite having >1,000x coverage of some genomes in long-read-only libraries, the reads did not assemble into complete genomes, suggesting the coverage is not a limitation and may well be a hindrance to assembly. The Illumina + ONT hybrid assembly (Unicycler) recovered the most genomes (2-6 genomes) (Figure 3.3 and Figure 3.13). The addition of long reads to short reads increases the number of genomes recovered (particularly ONT).

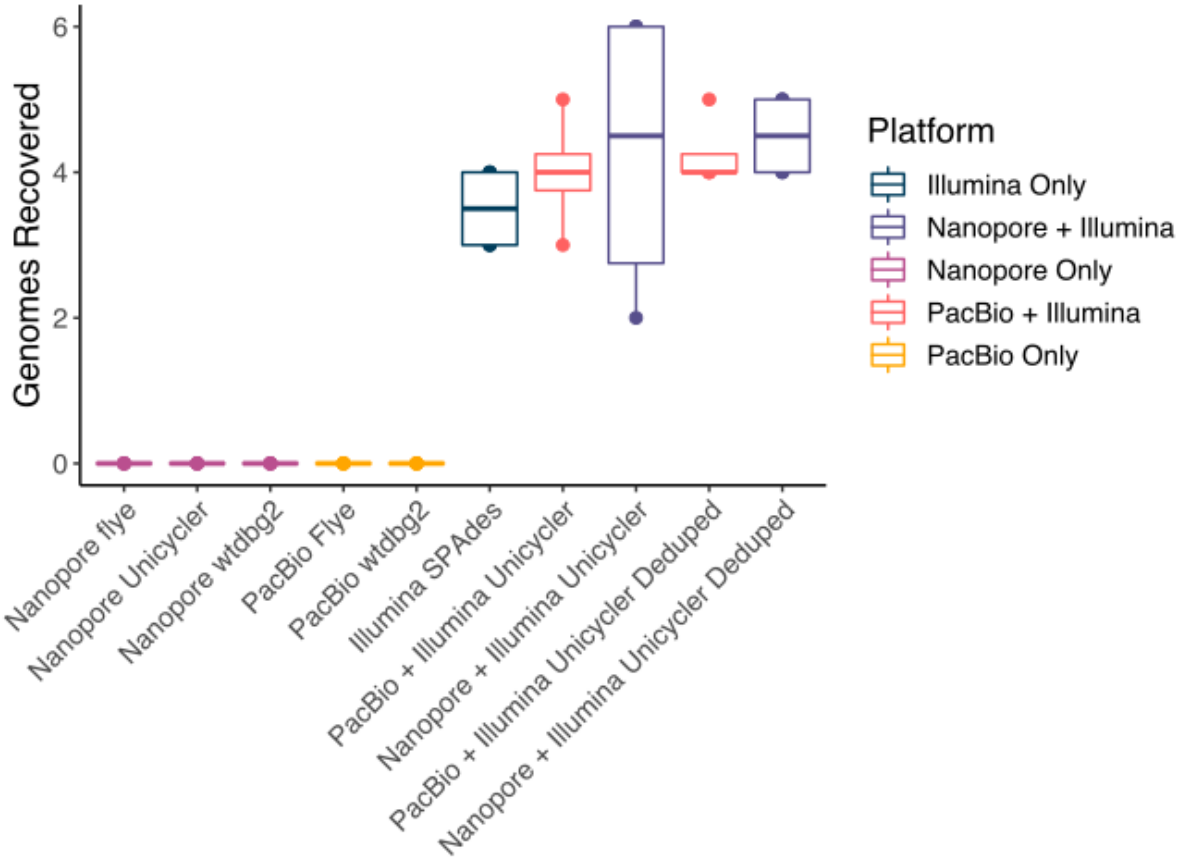


Figure 3.3 Comparison of genome recovery across sequencing technologies and assemblies

Boxplot showing the number of genomes fully assembled within each assembly (successful assembly defined as a single contig covering 97% of genome length), with the reads used for assembly shown in different colours. Boxes contain values for the 3 libraries and a pooled library.

3.5.4 Assembly Results - Limits of Detection for Assembled Genomes

The phage with the lowest input abundance to be recovered in a single contig within any assembly was vB_Eco_mar001J1 (53,672,905 genome copies), which was recovered in the ONT + Illumina hybrid assembly. The least abundant genome to be recovered from an Illumina only assembly, and a PacBio + Illumina hybrid assembly was KUW1 (72,995,151 genome copies), suggesting the addition of ONT reads to Illumina reads improves the recovery of lowly abundant genomes, but the addition of PacBio reads to the same Illumina reads did not.

KUW1 was recovered from Illumina libraries at a lower average sequence depth than any other genome (139 x coverage in the largest Illumina library, 225 x coverage in the pooled library), although it was not the genome with the lowest input abundance. Furthermore, KUW1 was not assembled in the two smaller Illumina libraries (37 and 49 x coverage obtained), suggesting that the depth of Illumina sequencing impacts the limits of detection.

As previously discussed (Section 3.5.2), the least abundant genomes to be detected by read mapping were vB_VpaS_sm032 and S-RSM4. Summed Illumina contigs from the pooled library mapped to 87% and 97% of vB_VpaS_sm032 and S-RSM4 respectively. However, the longest individual contigs only covered a small fraction of the genomes (22% and 9% respectively). The average read depth for vB_VpaS_sm032 and S-RSM4 contigs was 10 x over 98.7% and 14 x over 99.6% of genome lengths respectively in the pooled Illumina library. Manual inspection of alignments revealed that breaks in the assemblies typically coincided with a drop in read coverage which was often associated with a sudden and sharp change in molGC

(either upwards or downwards) (Cheung *et al.*, 2011; Sato *et al.*, 2019; Browne *et al.*, 2020) (Figure 3.4; Supplementary Table S3.4).

The longest genome to be recovered in a single contig, vB_Vpa_sm033 (320,253 bp), was assembled in Illumina-only, Illumina + PacBio, and Illumina + ONT assemblies. The shortest genome to be recovered in a single contig, KUW1 (44,509 bp), was assembled in Illumina-only, Illumina + PacBio, and Illumina + ONT assemblies. Whilst KUW1 was assembled from only one individual Illumina library, it was assembled in two each of the ONT + Illumina, and PacBio + Illumina hybrid assemblies. Furthermore, dereplicating these hybrid assemblies with the Illumina-only assemblies led to KUW1 being assembled in all individual libraries.

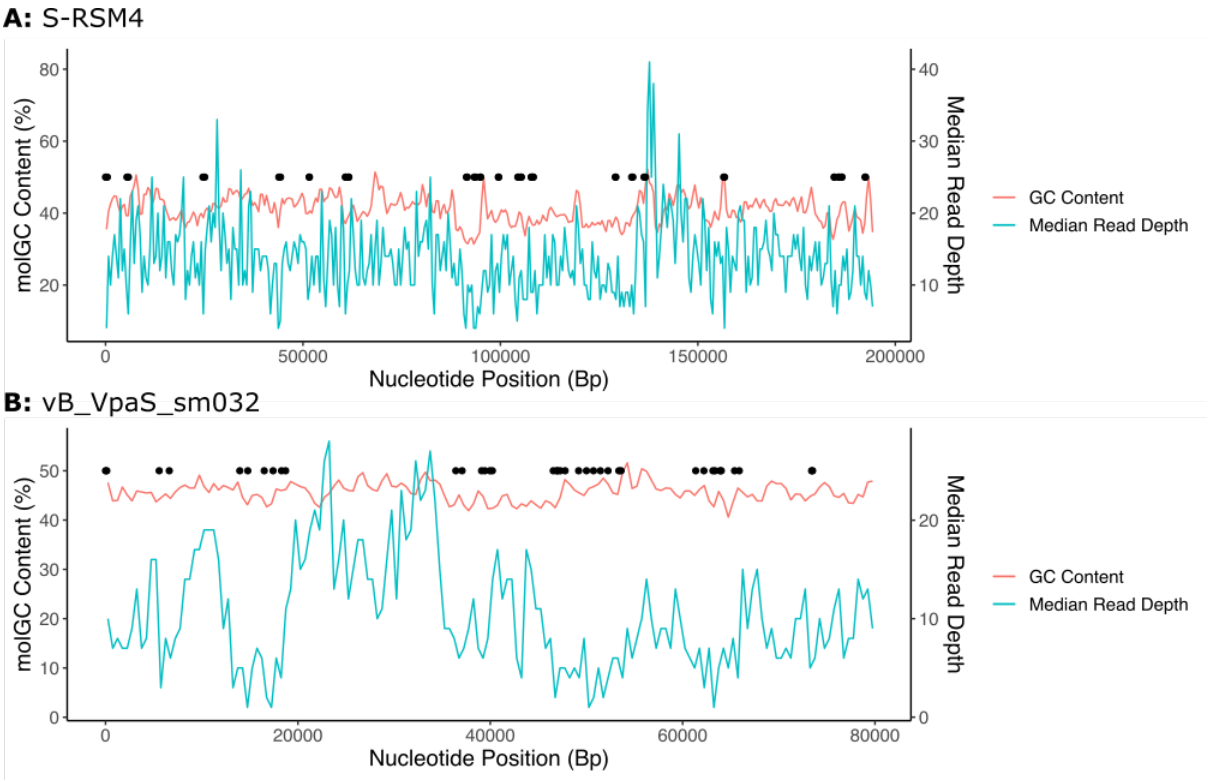


Figure 3.4 Fragmentation of Illumina assemblies

Plots showing molGC (%) content against median coverage (500bp sliding window) for (**A**) SRSM-4 and (**B**) vB_VpaS_sm032, which both fragmented in the pooled Illumina SPAdes assembly despite reads mapping to ~99% of both genomes. Breaks in the assembly are shown with black circles.

3.5.5 Assembly Results - Resolution of Highly Similar Genomes

It was possible to assemble a single genome that was representative of vB_Eco_mar001J1 or/and vB_Eco_mar002J2 from Illumina + ONT hybrid assemblies. As these genomes have >99% ANI between them, it was not surprising the assemblies contained a single genome that was a chimaera of both, rather than two individual genomes. It was also possible to obtain the genome of vB_EcoS_swan01 using an Illumina + ONT hybrid assembly (Figure 3.13), which has ~80% ANI with vB_Eco_SLUR29. However, the genome of vB_Eco_SLUR29 could not be resolved.

3.5.6 Assembly Results - Comparison of Long Read Assemblers

Despite high read coverage (Supplementary Table S3.4), long-read only assemblies failed to recover a complete genome (Figure 3.3). To identify the optimal long-read only assemblies, we used the NGA50 statistic (Figure 3.5). While nine genomes were detected by mapping long reads in at least one library, only eight are included in this analysis, due to the very low coverage of vB_Eco_SLUR29 recovered from any assembly. For this comparison, we also included long read assemblies that were polished with Illumina reads, as this was found to affect the results.

The NGA50 values averaged across the eight genomes and four libraries obtained from ONT assemblies were higher than those from PacBio, again this varied depending on the assembler used. ONT reads assembled with Flye, wtdbg2 and Unicycler obtained average NGA50 values of 28%, 10% and 8% respectively, whereas PacBio reads obtained values of 5% and 4% for wtdbg2 and Flye assemblies respectively. While ONT reads assembled with Flye typically produced the longest alignments in relation to reference genomes, its performance in the individual libraries

was higher than that in the pooled library; with the average NGA50 values as proportion of genome length being 27%, 39% and 30% for individual libraries, and only 16% for the pooled library (Figure 3.14). Conversely, the highest NGA50 values for ONT reads assembled with Unicycler were obtained from the pooled library (26%), and 3%, 1% and 0.1% from individual libraries (Figure 3.14). Therefore, whether pooling reads increases assembly length depends on the assembler being used.

For all five long-read only assemblies, its polished counterpart typically obtained more complete assemblies than before polishing (Figure 3.5 and Figure 3.14). This is particularly apparent with the individual ONT libraries assembled with Unicycler which went from obtaining some of the lowest average NGA50 values to having some of the highest (3%, 1% and 0.1% increasing to 43%, 35% and 29% respectively) (Figure 3.14). This suggests the ONT-Unicycler assemblies contained contigs below the 90% ANI threshold required for mapping and were only aligned to the reference genomes post-polishing (Figure 3.14). This post-polish increase was more modest in PacBio assemblies, which increased from 3.7% to 3.8% and from 4.9% to 5.1% for the Flye and wtdbg2 assemblies respectively (Figure 3.14). Manual inspection of contig alignments from long-read only assemblies to the reference genomes revealed large numbers of overlapping misassembled contigs that were not resolving into a single assembly. This is potentially due to the higher error frequency associated with ONT and PacBio reads.

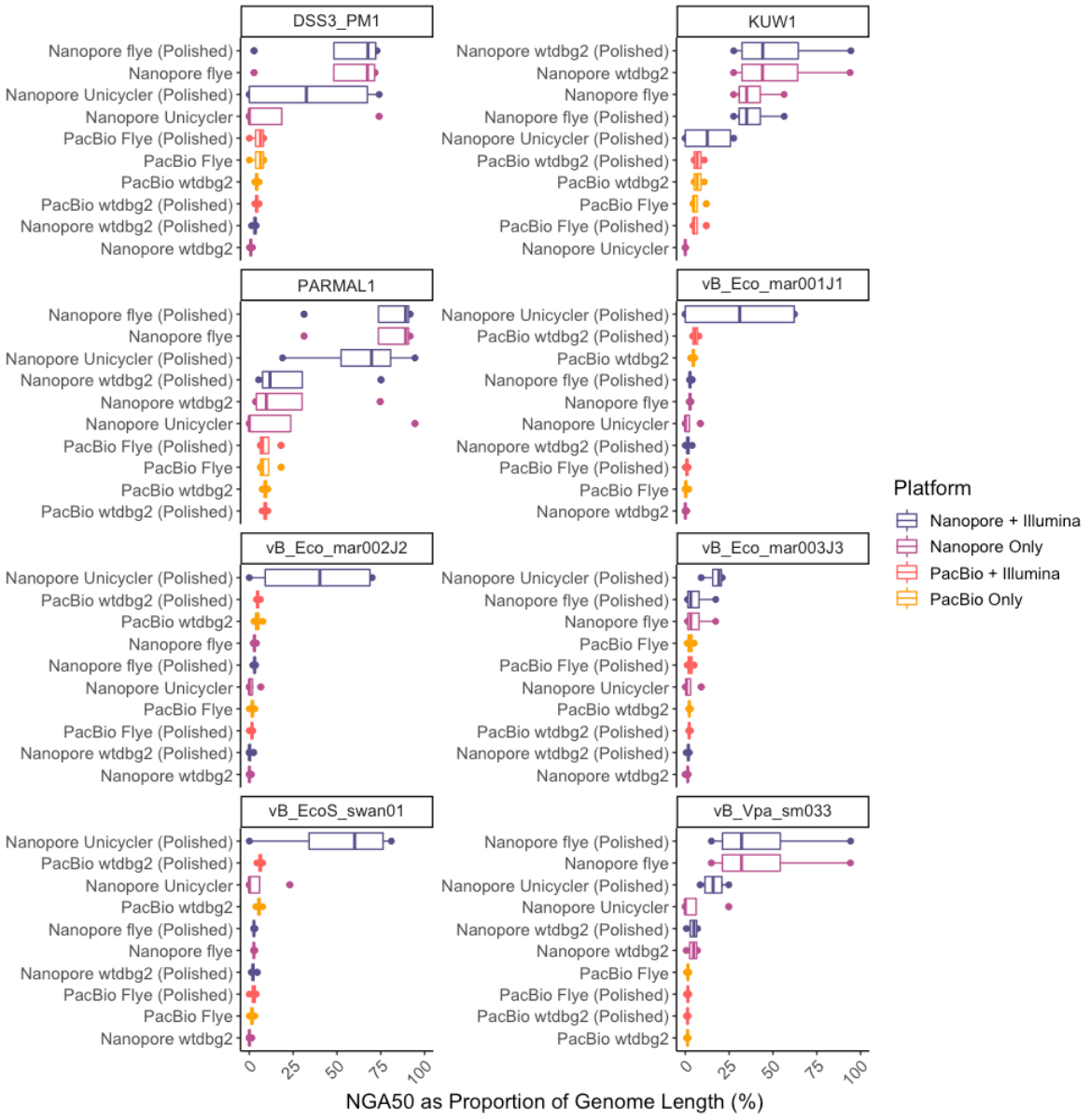


Figure 3.5 Comparison of genome assembly completeness for long-read assemblies

Boxplots showing the NGA50 statistic per genome as a percentage of genome length for each assembly, with the reads used for assembly shown in different colours. Boxes contain values for the three libraries and a pooled library.

To determine if the long-read assemblies were failing due to high sequencing depth, we individually extracted the reads mapping to genomes with ≥100 x coverage and re-assembled the mapped reads only, as well as randomly downsampled subsets. For both ONT and PacBio, and all assemblers used, downsampling the reads prior to re-assembly led to more complete assemblies (Figure 3.6; Supplementary Table S3.5). Furthermore, successful assemblies using PacBio reads with Unicycler were only obtained after downsampling. However, Nanopore reads assembled with Unicycler obtained the most complete assemblies using the original mixed community reads (i.e., rather than reads mapping to that genome only).

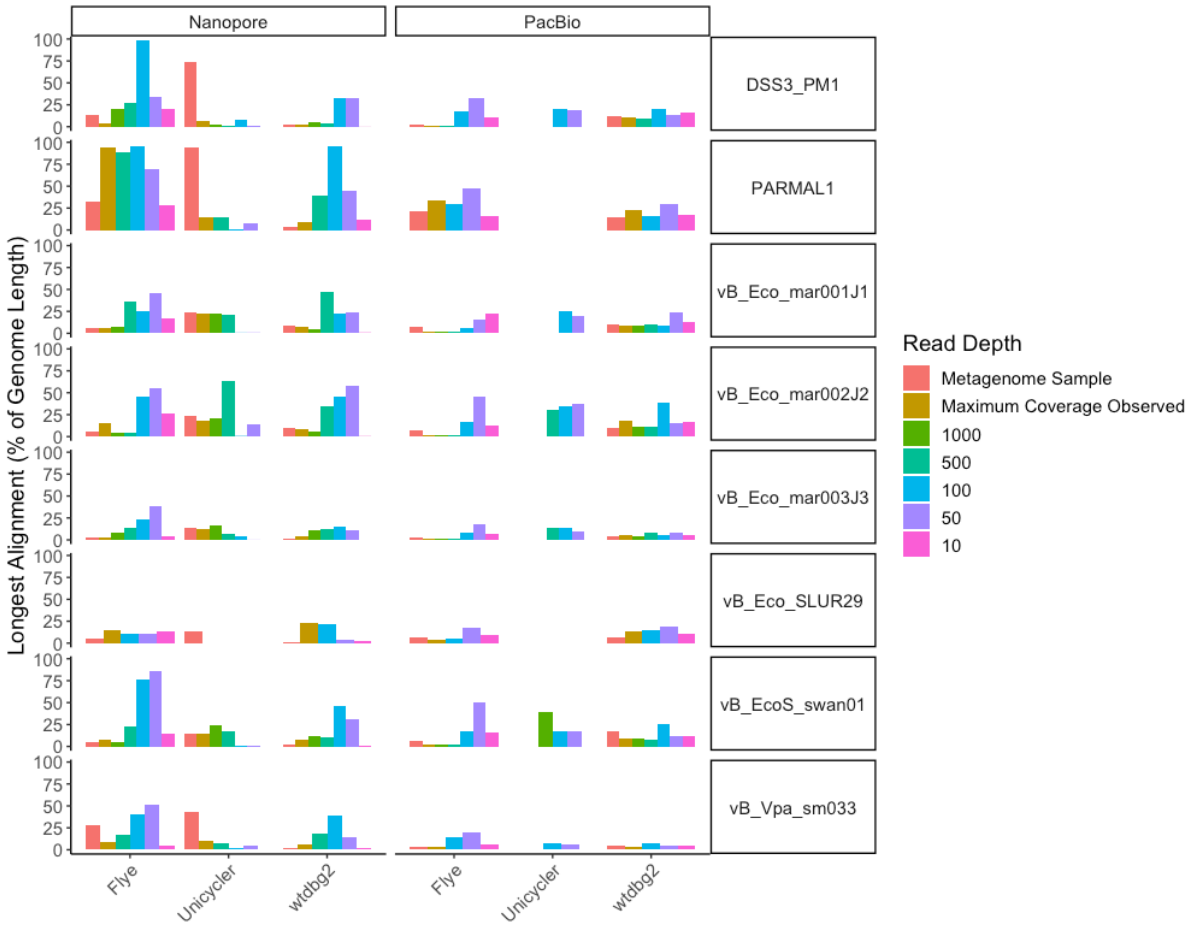


Figure 3.6 Effect of read depth on long-read assembly

The longest alignments per genome are shown for all long-read only assemblies using their pooled libraries (Metagenome Sample), reads mapping to the genome only (Maximum Coverage Observed), and randomly downsampled reads to approximate read depths.

3.5.7 Assembly Results - SNPs, INDELs and Misassemblies

To investigate the fidelity of assemblies, we compared assembled contigs to the mock community reference genomes to identify the frequency of SNPs and INDELs per 100 kb. Both SNPs and INDELs were calculated for genomes where ≥ 50% of the genome was covered by contigs. Using Illumina only reads resulted in the lowest number of SNPs per 100 kb (503) with ONT long-read only assemblies having the highest number of SNPs (2038-4159). The number of SNPs in long read assemblies was also dependent on the assembler used. Using ONT reads with Flye (2038) resulted in fewer SNPs than when wtdbg2 (3545) or Unicycler (4159) (Figure 3.7A and Figure 3.15). Conversely, PacBio reads assembled with Flye had a higher SNP frequency (2180) than those produced using wtdbg2 (1806) (Figure 3.7A and Figure 3.15).

A similar pattern of results was observed for the number of INDELs per 100 kb, although a much larger difference between the different read technologies was observed. Again, the assembler used had an impact on the frequently of INDELs. ONT assemblies produced by far the largest number of INDELS when using Unicycler (miniasm + racon; 4521) compared with Flye (1702) and wtdbg2 (1982) assemblies (Figure 3.7B and Figure 3.16). PacBio assemblies had far fewer INDELs than ONT with far fewer INDELs observed in Flye assemblies (176) than wtdbg2 assemblies (946). Illumina only assemblies had by far smallest number of INDELs (14) (Figure 3.7B and Figure 3.16).

Whilst long-read-only assemblies had a high frequency of SNPs and INDELs, hybrid assemblies produced with Unicycler that combined Illumina reads with ONT or PacBio

reads obtained SNP and INDEL levels comparable to Illumina only assemblies (Figure 3.7).

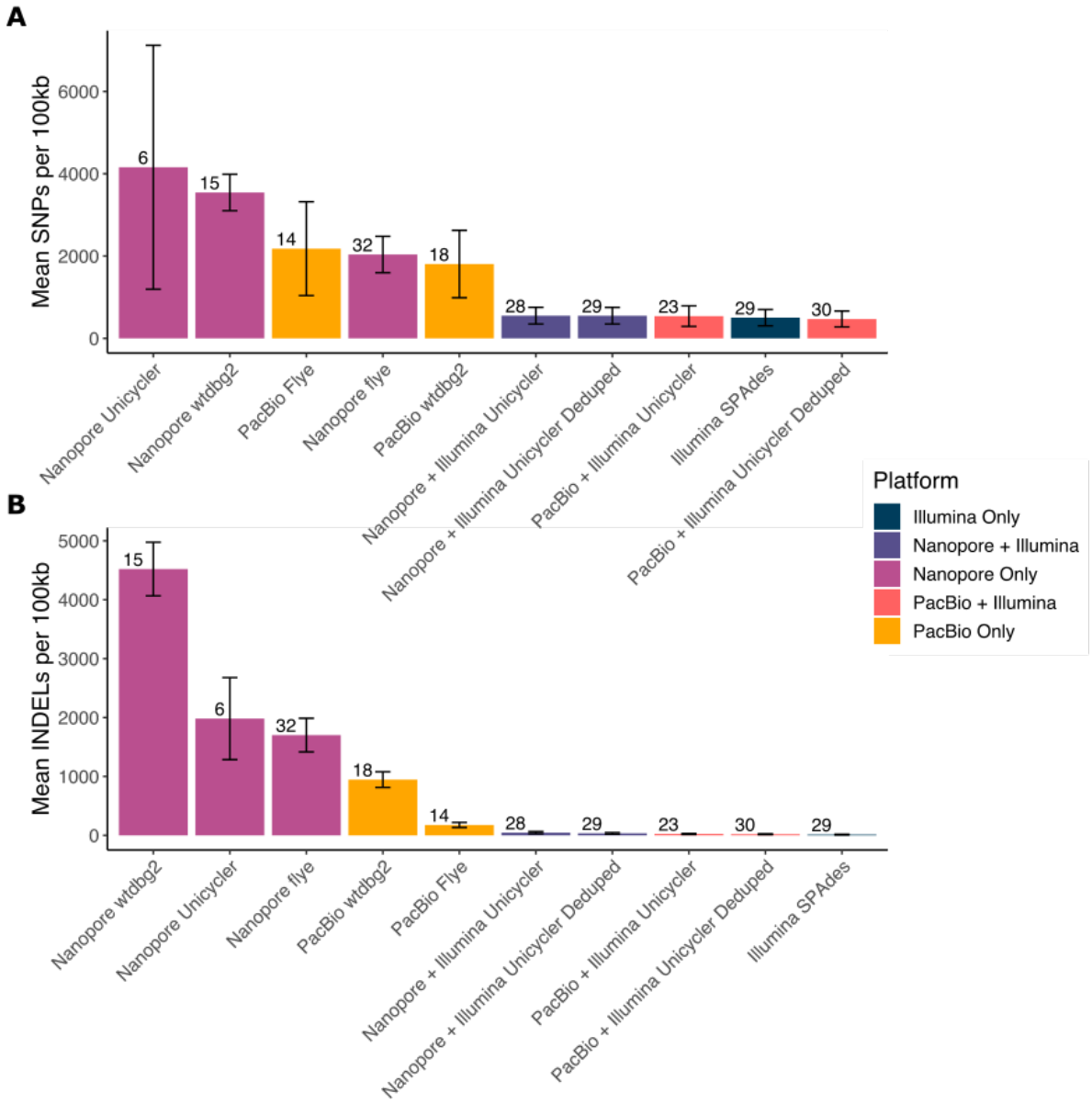


Figure 3.7 Effect of sequencing technology and assembler on error rate

Plots showing the frequency of SNPs (**A**) and INDELs (**B**) per 100kb of reference genome, where at least 50% of the reference genome was recovered by contigs. Error bars show standard error of mean and the number above the bar indicates the number of genomes included in mean calculation (from a total possible maximum of 60 (15 genomes, 4 assemblies)).

3.5.8 Effect of Polishing Long-Read Assemblies on SNPs, INDELs and ORF Prediction

Using short reads to polish contigs produced from long read assemblies generally reduced the number of SNPs per 100 kb, although this was dependent on the specific assembly. Polishing ONT assemblies produced with Unicycler and wtdbg2 decreased the frequency of SNPs by 42% and 26%, respectively (Figure 3.8A). The Flye assembly resulted in a small increase in the number of SNPs (Figure 3.8A). Rather than introducing errors, this is likely as a result of contigs prior to polishing having SNP frequencies that prevented recruitment to a reference genome at 90% identity by mapping. Post polishing, these contigs are now recruited to genomes, but still contain a number of SNPs (Figure 3.8A). With PacBio reads assembled with Flye, polishing had no effect on the number of SNPs (Figure 3.8A). For the PacBio wtdbg2 assembly, the number of SNPs increased, as observed with ONT reads assembled with Flye. Again, this increase is likely due to the increased number of contigs that are mapped to the reference genome (Figure 3.8A).

The effect of polishing on the frequency of INDELs was more apparent. The ONT assemblies had a higher number of INDELs than PacBio assemblies prior to polishing (Figure 3.8B). For ONT reads assembled with Unicycler (miniasm + racon), which had the highest frequency of INDELs initially, there was a 55% decrease in INDELs post polishing (Figure 3.8B). For ONT reads assembled with wtdbg2 and Flye, there was a ~34% decrease in the number of INDELs per 100 kb (Figure 3.8B). For PacBio assemblies the starting frequency of INDELs was lower than ONT prior to polishing but polishing with Illumina reads still resulted in a 21% and 4.5% decrease in INDEL frequency for wtdbg2 and Flye assemblies respectively (Figure 3.8B).

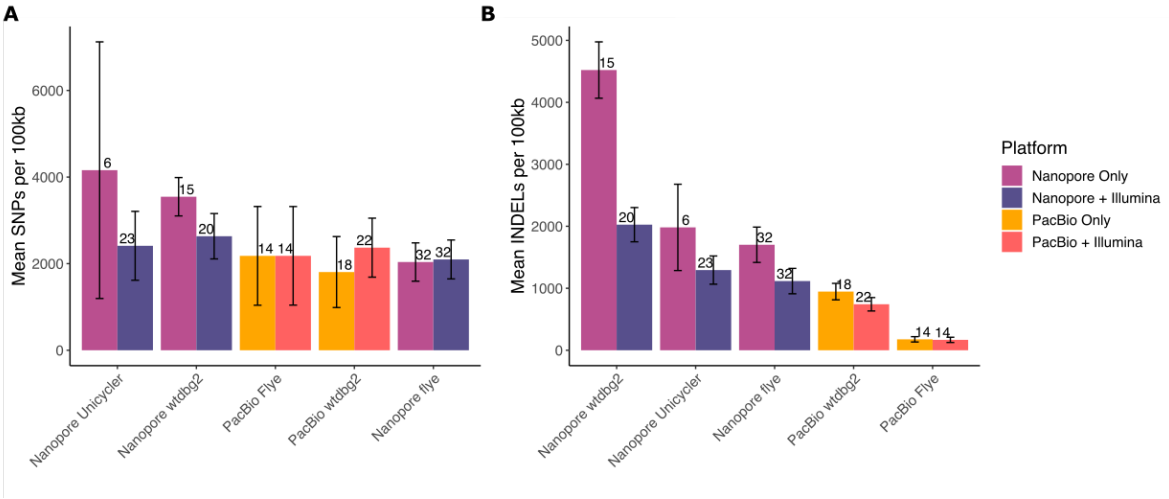


Figure 3.8 Effect of polishing on error rate

Plots showing the frequency of SNPs (**A**) and INDELs (**B**) per 100kb of reference genome before and after polishing with Illumina reads, where at least 50% of the reference genome was recovered by contigs. Error bars show standard error of mean and the number above the bar indicates the number of genomes included in mean calculation (from a total possible maximum of 60 (15 genomes, 4 assemblies)).

As assembly errors can have an effect on ORF prediction and functional annotation (Watson and Warr, 2019), we investigated the number and length of predicted ORFs on contigs which mapped to reference genomes before and after polishing. Polishing with short reads had the greatest effect on ONT data regardless of the assembler used, with mean ORF length increasing for all assemblies. Both Unicycler and wtdgb2 observed mean ORF length increases of ~66%, with a ~24% increase for Flye (Figure 3.9). For PacBio assemblies, the increases in mean ORF length were more modest at ~11% for wtgb2 assemblies and ~0.2% for Flye assemblies (Figure 3.9). While there was an increase in mean ORF length for all combinations of reads and assemblers post-polishing, all combinations were still smaller than the value obtained for the 15 reference genomes (Figure 3.9).

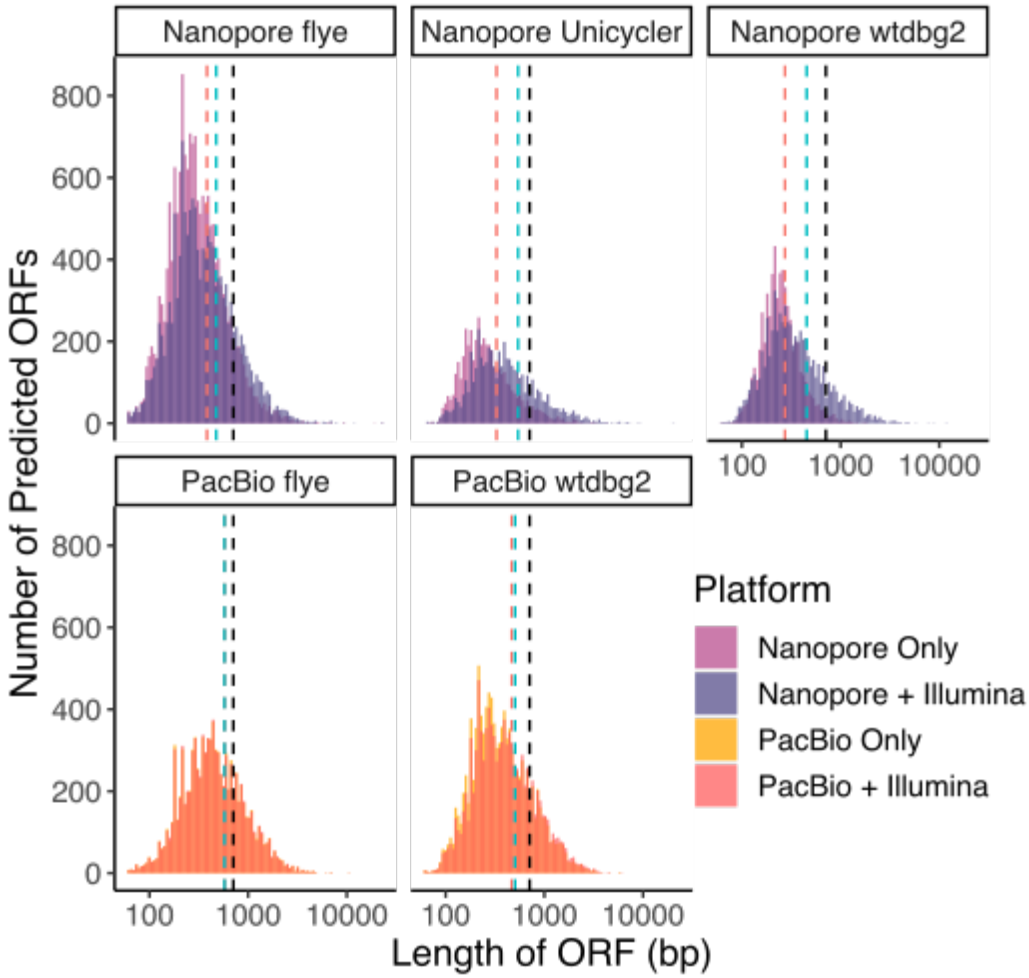


Figure 3.9 The effect of polishing long-read assemblies on predicted ORF lengths

Boxplots showing the distribution of predicted ORF length per assembly in base pairs with dashed vertical lines show the mean value before (red) and after (blue) polishing, as well as the expected value that was obtained from the reference genomes (709 bp; black).

3.5.9 Effect of Polishing Long-Read Assemblies on Viral Prediction

Many viral prediction programs use similarity of predicted proteins to known hallmark proteins for virus prediction. Thus, truncated proteins may alter the ability to predict viral contigs from viromes and metagenomes. To test if truncated proteins affect virus prediction, we compared VIBRANT (Kieft, Zhou and Anantharaman, 2020) which in part uses predicted proteins, and DeepVirFinder (Ren *et al.*, 2020) a K-mer based prediction system on all assembled contigs. Although we utilised purified phage isolates to create the mock community, up to 20% of the reads from Illumina libraries did not map to the reference genomes. Therefore, we utilised this unfortunate level of contaminating host bacterial DNA for benchmarking viral prediction. To determine how many predictions represented “true” viral predictions, we mapped the predicted vOTUs against the reference genomes.

For DeepVirFinder predictions, there were minimal differences in the number of predicted viral contigs (vOTUs) before and after polishing for all assemblies. The largest difference was observed for ONT reads assembled using Flye (61 before, 52 after) (Figure 3.10; Table 3.1). However, there was a marked increase in the number of vOTUs that could be verified as phage. For Flye, the number that could be verified as phage increased from 82% to 96% after polishing, wtdbg2 assemblies increased from 83% to 98%, and Unicycler assemblies increased from 93% to 99%. Thus, polishing ONT assemblies with Illumina reads led to an overall decrease in the number of erroneous viral predictions when using DVF (Figure 3.10; Table 3.1). For the PacBio assemblies, there was no difference in the number of predicted vOTUs and those that could be verified as phage when using DVF (Figure 3.10; Table 3.1).

When using VIBRANT for prediction, polishing of PacBio assemblies had no or minimal effect on the number of predictions or the number of verified predictions (Figure 3.10; Table 3.1). However, the polishing of ONT assemblies led to vastly different numbers of predicted vOTUs, and this varied with assembler used. The largest difference was for the ONT wtdbg2 assembly, decreasing from 199 to 133 predicted vOTUs, and the proportion of verified phages increased for all ONT assemblies after polishing. For Flye, the number of verified phages increased from 75% to 81%, Unicycler increased from 72% to 96%, and wtdbg2 increased from 51% to 87% (Figure 3.10; Table 3.1).

Thus, when using DeepVirFinder there was minimal impact of polishing on the prediction of vOTUs from either PacBio or ONT assemblies. However, there were clear benefits to the polishing of ONT assemblies when using VIBRANT for vOTU prediction, as the percentage of vOTUs that could be verified to be phage increased post polishing.

Table 3.1 Effect of polishing on vOTU predictions

Platform	Software	Assembler	Polish	Predicted	Mapped	Mapped / Predicted (%)
Nanopore	DVF	Flye	No	61	50	82.0%
			Yes	52	50	96.2%
		Unicycler	No	95	88	92.6%
			Yes	100	99	99.0%
		wtdbg2	No	110	91	82.7%
			Yes	109	107	98.2%
	VIBRANT	Flye	No	85	64	75.3%
			Yes	82	66	80.5%
		Unicycler	No	130	93	71.5%
			Yes	104	100	96.2%
		wtdbg2	No	199	101	50.8%
			Yes	133	115	86.5%
PacBio	DVF	Flye	No	173	146	84.4%
			Yes	173	146	84.4%
		wtdbg2	No	206	197	95.6%
			Yes	205	196	95.6%
	VIBRANT	Flye	No	178	159	89.3%
			Yes	179	159	88.8%
		wtdbg2	No	237	212	89.5%
			Yes	238	212	89.1%

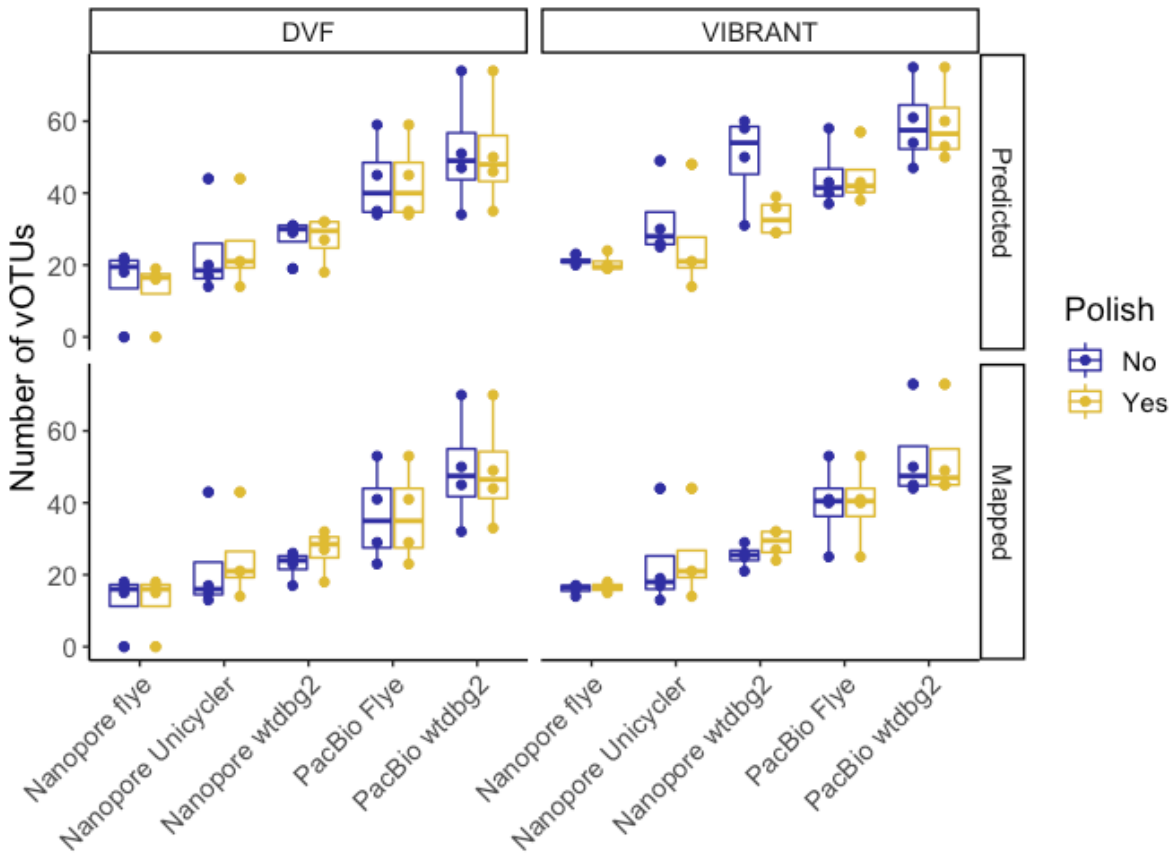


Figure 3.10 The effect of polishing long-read assemblies on viral prediction

Boxplots showing the number of predicted contigs for the five different long-read assemblies before and after polishing, with the lower two panels showing the number of contigs which mapped to the reference genomes. The left two panels show vOTUs predicted with DeepVirFinder, and the right two panels show predictions from VIBRANT. The individual boxes contain values from three individual libraries and a pooled library.

3.5.10 Effect of sequencing technology on predicted virome diversity

Having established DeepVirFinder generally performed better for all sequencing technologies, we utilised the output of DeepVirFinder predictions to assess how diversity statistics of the mock community varied with sequencing technology and assembly.

Overall, there were two clear trends in estimating alpha diversity of the mock community. When using long read assemblies for vOTU prediction, there was an overestimation in the alpha diversity. In contrast, when using Illumina and Illumina + ONT/PacBio hybrid vOTUs, there was an underestimation of alpha diversity. Within these two general trends there was also variation with the assembler used. For any assembly including short reads, there were relatively small differences in the predicted Shannon's diversity ranging from 1.1 for ONT + Illumina with Unicycler (miniasm + racon) to 1.5 for PacBio + Illumina using Unicycler (Figure 3.11). The PacBio + Illumina Unicycler assembly obtained the most accurate prediction of diversity based on Shannon's diversity index, compared to the known value of 2.05 (or 1.8, if only including those that could be detected by read mapping).

In contrast, long-read only assemblies predicted more diverse communities, with predictions ranging from 2.3 for ONT reads assembled with Unicycler (miniasm + racon) to 3.6 for PacBio reads assembled with wtdgb2 (Figure 3.11). When assessing the diversity based purely on the number of predicted vOTUs, long-read only assemblies generally overestimate the number of vOTUs within the sample (Figure 3.11). The ONT reads assembled with Unicycler (miniasm + racon) were the exception

to this, and most closely reflected the true number of vOTUs within the mock community, however this assembly still over-estimated the number of vOTUs.

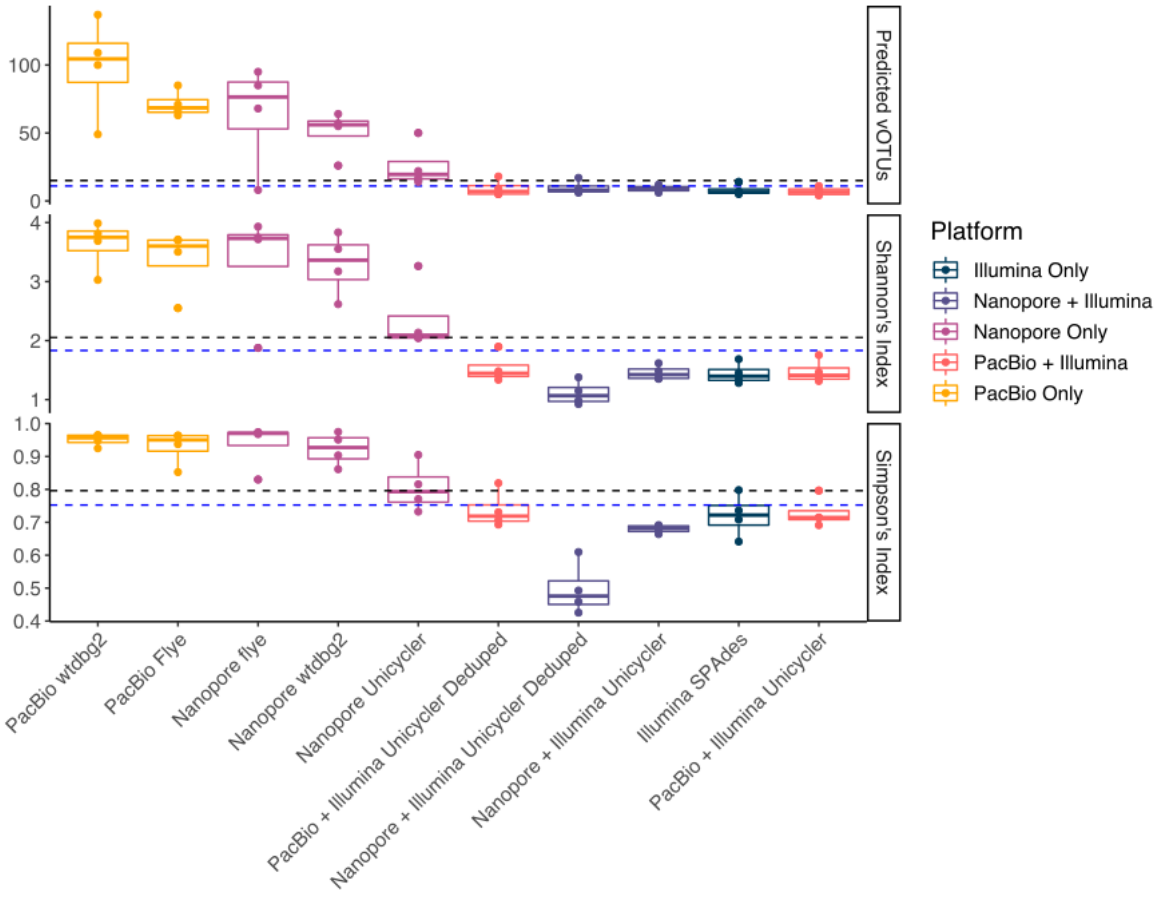


Figure 3.11 The effect of sequencing platform and assembler on diversity estimates

Boxplots showing the number of predicted vOTUs for mock virome analysis (top), and Shannon's index (middle) and Simpson's index (bottom) alpha diversity measures. Black dashed lines indicate true values for mock virome input, and blue dashed lines indicate true values excluding genomes that were not detected by read mapping in any library.

3.6 Discussion

The use of long read sequencing technologies is becoming increasingly common for the sequencing of metagenomic samples, in particular those that focus on the bacterial community. A number of studies have demonstrated the advantage of long-reads in assembling complete genomes from a variety of samples (Xie *et al.*, 2020; Arumugam *et al.*, 2021; Cuscó *et al.*, 2021; Yahara *et al.*, 2021). There have also been a number of studies benchmarking the assembly and/or recovery of bacteria from mock communities using long-reads (Nicholls *et al.*, 2019; Leidenfrost *et al.*, 2020), along with benchmarking of assembly algorithms for prokaryotic genomes (excluding phages) (Wick and Holt, 2021; Hackl, Harbig and Nieselt, 2022). However, there are no such comprehensive studies that have directly compared Illumina, ONT, and PacBio sequencing technologies for the study of viromes.

Previous benchmarking of short-read assemblers has demonstrated minimal differences in genome recovery of phage genomes when comparing multiple assemblers on a mock viral community (Roux *et al.*, 2017). For this reason, we chose only one short-read assembly algorithm: SPAdes. For long-read assembly, we chose three frequently used approaches of Unicycler (miniasm + racon), Flye, and wtdgb2 as well as using Unicycler for a direct hybrid assembly. For long read sequencing alone, we were unable to obtain assemblies from PacBio reads alone with Unicycler, even when combining all three samples suggesting it was not due to a lack of sequence coverage.

When using a single sequencing technology, only Illumina reads resulted in the complete assembly of a phage genome within any sample. Utilising a hybrid approach

increased the number of genomes that could be assembled, with ONT + Illumina reads assembled with Unicycler (minimisam + racon) recovering the largest number of genomes, whereas the addition of PacBio reads did not result in the same increased recovery of genomes. However, this may well be due to reduced yield of PacBio reads compared to ONT reads, thus increased yield of PacBio data might improve this metric.

The combination of long and short reads improving recovery of assembled genomes is consistent with previous benchmarking of a mock viral community using a virION approach (Warwick-Dugdale *et al.*, 2019). Unlike the virION approach, we were only able to assemble a complete genome with just long-reads after downsampling to lower read depths prior to assembly. However, direct comparison between the studies is difficult given the different phages used in each mock community. Furthermore, the reasons for improved assembly after down-sampling remain unclear; it is possible that the higher frequency of errors associated with long-reads is compounded as more reads are added, leading to a highly fragmented assembly when high read depths are used. Here, we utilised MDA application to provide sufficient material for long-read sequencing, whereas the virION utilises PCR to provide sufficient material (Warwick-Dugdale *et al.*, 2019; Zablocki *et al.*, 2021). The virION approach has comprehensively demonstrated relative abundance of phages are maintained due to the LASL-PCR approach (Warwick-Dugdale *et al.*, 2019; Zablocki *et al.*, 2021). Here, we observed a strong correlation in the abundance of phages in the un-amplified Illumina viromes and amplified long-read viromes. However, we are cautious in the interpretation of this data. The DNA from a ssDNA phage (ΦX174) was spiked into our mock community at a deliberately low level, as we wanted to avoid flooding our amplified DNA with

ssDNA given known biases of MDA. However, given the lack of detection of ΦX174 in any samples, we may have been overly cautious in the amount added. Thus, when ssDNA phages are present in a community, it is likely the biases observed previously are still likely to hold true (Yilmaz, Allgaier and Hugenholtz, 2010; Kim and Bae, 2011; Marine *et al.*, 2014).

When assessing any individual sequencing technology alone, the lowest number of SNPs or indels obtained was unsurprisingly observed when using Illumina reads. With ONT assemblies having a larger number of SNPs, and in particular INDELs, compared to PacBio assemblies. Both INDELs and SNPs were also affected by the method used for assembly. For ONT reads, Flye produced assemblies with the lowest number of INDELs or SNPs compared to wtdgb2 and Unicycler (miniasm + racon). It is likely for ONT data that the number of SNPs and indels will further decrease with improvements in accuracy reported for both R10 flow cells and the latest base calling algorithms that have been developed since this data was collected, as this data was generated with R9 flow cells. In contrast, Flye assemblies of PacBio reads had the lowest number of SNPs, but highest number of INDELs. Thus, the choice of assembly method should be adjusted for the type of long-reads being used. The addition of short reads to polish the long read assemblies resulted in a reduction of both SNPs and indels, as has been observed in other studies (Warwick-Dugdale *et al.*, 2019; Cook, Hooton, *et al.*, 2021; Zaragoza-Solas *et al.*, 2022).

While the combination of both short Illumina reads with long reads resulted in the “best” overall assemblies, it may well not be feasible to sequence samples with both technologies. Therefore, we treated the assemblies from multiple approaches to

assess how the different approaches affected the predicted diversity of the sample. Although polishing long read assemblies had a significant impact on reducing the number of SNPs and INDELs, there was minimal effect on the number of predicted contigs that were viral when using DVF for prediction. However, VIBRANT, which in part utilises the identification of hall-mark phage genes and was more sensitive to the higher error rates of un-polished long-read assemblies and obtained far fewer erroneous viral predictions post-polishing. Thus, choice of sequencing technology may have ramifications for downstream choices in viral prediction software.

3.7 Conclusions

We have benchmarked Illumina, ONT, and PacBio sequencing platforms for virome analysis using a number of read and assembler combinations and offer recommendations for the community: (i) if only using one sequencing platform, Illumina performs best at genome recovery and has the lowest error rates; (ii) the addition of long-reads to Illumina reads improves the assembly of lowly abundant genomes, particularly ONT; (iii) whilst long read assemblies, particularly ONT, have higher error frequencies, polishing with Illumina reads can reduce these errors to levels comparable with Illumina-only assemblies; (iv) down-sampling of long reads may aid assembly; and (v) the choice of sequencing platform should be considered when making downstream analyses decisions, such as assembler algorithm and viral prediction software.

3.8 Supplementary Figures

Below are supplementary figures from the manuscript ‘Comparison of Illumina, Nanopore and PacBio sequencing for virome analysis. Cook, R. et al (2022).’

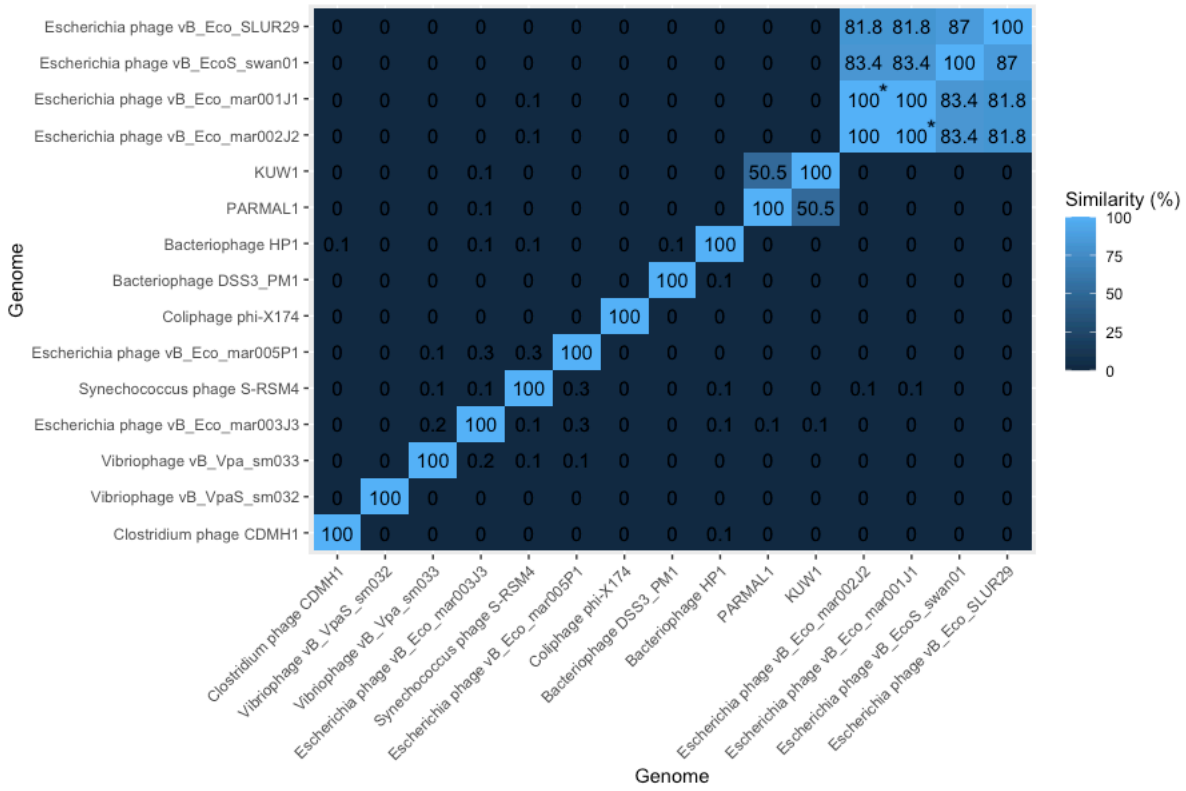


Figure 3.12 Relatedness of phages in mock community

Heatmap showing ANI (%) of phages in mock community. *Denotes phages are not quite 100% similar.

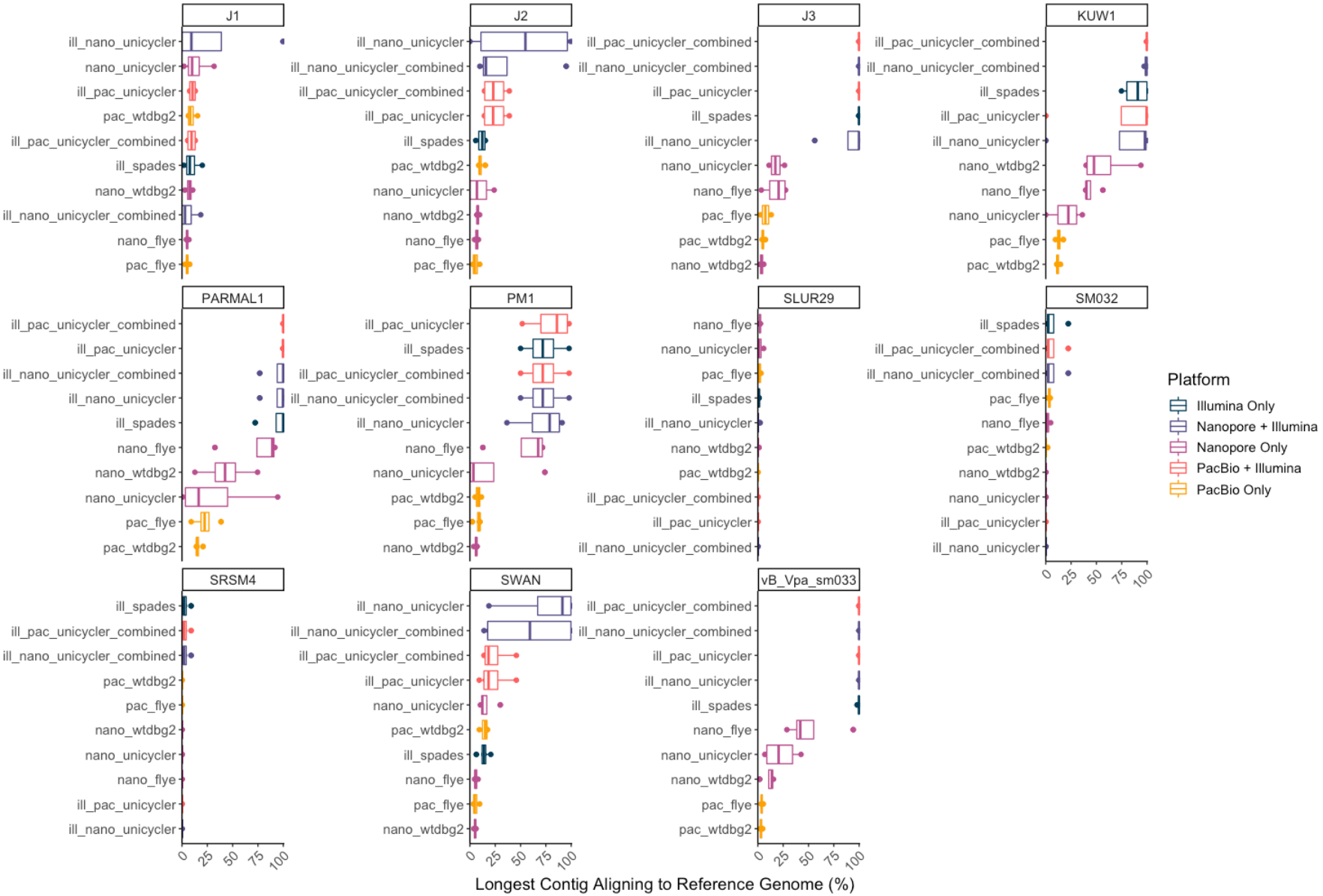

J1
ill_nano_unicycler
nano_unicycler
ill_pac_unicycler
pac_wtdbg2
ill_pac_unicycler_combined
ill_spades
nano_wtdbg2
ill_nano_unicycler_combined
nano_flye
pac_flye
J2
ill_nano_unicycler
ill_nano_unicycler_combined
ill_pac_unicycler_combined
ill_pac_unicycler
ill_spades
pac_wtdbg2
nano_unicycler
nano_wtdbg2
nano_flye
pac_flye
J3
ill_pac_unicycler_combined
ill_nano_unicycler_combined
ill_pac_unicycler
ill_spades
ill_nano_unicycler
nano_unicycler
nano_flye
pac_flye
pac_wtdbg2
nano_wtdbg2
KUW1
ill_pac_unicycler_combined
ill_nano_unicycler_combined
ill_spades
ill_pac_unicycler
ill_nano_unicycler
nano_wtdbg2
nano_flye
nano_unicycler
pac_flye
pac_wtdbg2
PARMAL1
ill_pac_unicycler_combined
ill_pac_unicycler
ill_nano_unicycler_combined
ill_nano_unicycler
ill_spades
nano_flye
nano_wtdbg2
nano_unicycler
pac_flye
pac_wtdbg2
PM1
ill_pac_unicycler
ill_spades
ill_pac_unicycler_combined
ill_nano_unicycler_combined
ill_nano_unicycler
nano_flye
nano_unicycler
pac_wtdbg2
pac_flye
nano_wtdbg2
SLUR29
nano_flye
nano_unicycler
pac_flye
ill_spades
ill_nano_unicycler
nano_wtdbg2
pac_wtdbg2
ill_pac_unicycler_combined
ill_pac_unicycler
ill_nano_unicycler_combined
SM032
ill_spades
ill_pac_unicycler_combined
ill_nano_unicycler_combined
pac_flye
nano_flye
pac_wtdbg2
nano_wtdbg2
nano_unicycler
ill_pac_unicycler
ill_nano_unicycler
SRSM4
ill_spades
ill_pac_unicycler_combined
ill_nano_unicycler_combined
pac_wtdbg2
pac_flye
nano_wtdbg2
nano_unicycler
nano_flye
ill_pac_unicycler
ill_nano_unicycler
SWAN
ill_nano_unicycler
ill_nano_unicycler_combined
ill_pac_unicycler_combined
ill_pac_unicycler
nano_unicycler
pac_wtdbg2
ill_spades
nano_flye
pac_flye
nano_wtdbg2
vB_Vpa_sm033
ill_pac_unicycler_combined
ill_nano_unicycler_combined
ill_pac_unicycler
ill_nano_unicycler
ill_spades
nano_flye
nano_unicycler
nano_wtdbg2
pac_flye
pac_wtdbg2
0
25
50
75
100
Longest Contig Aligning to Reference Genome (%)
Platform
Illumina Only
Nanopore + Illumina
Nanopore Only
PacBio + Illumina
PacBio Only

Figure 3.13 Genome by genome breakdown of assembly completeness

Boxplots showing the longest contig obtained per assembly per genome.

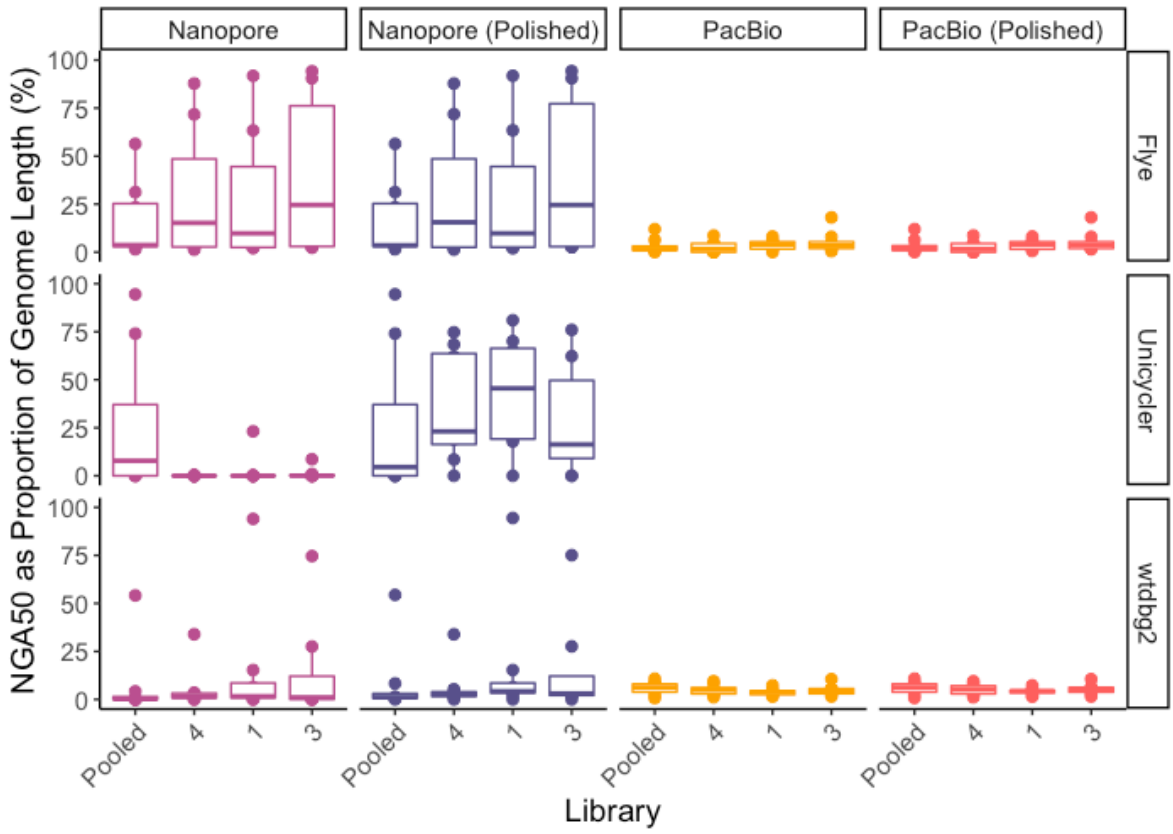


Figure 3.14 Averaged NGA50 for long-read assemblies

Boxplots show the NGA50 statistic averaged across genomes for long-read assemblies and their polished counterparts.

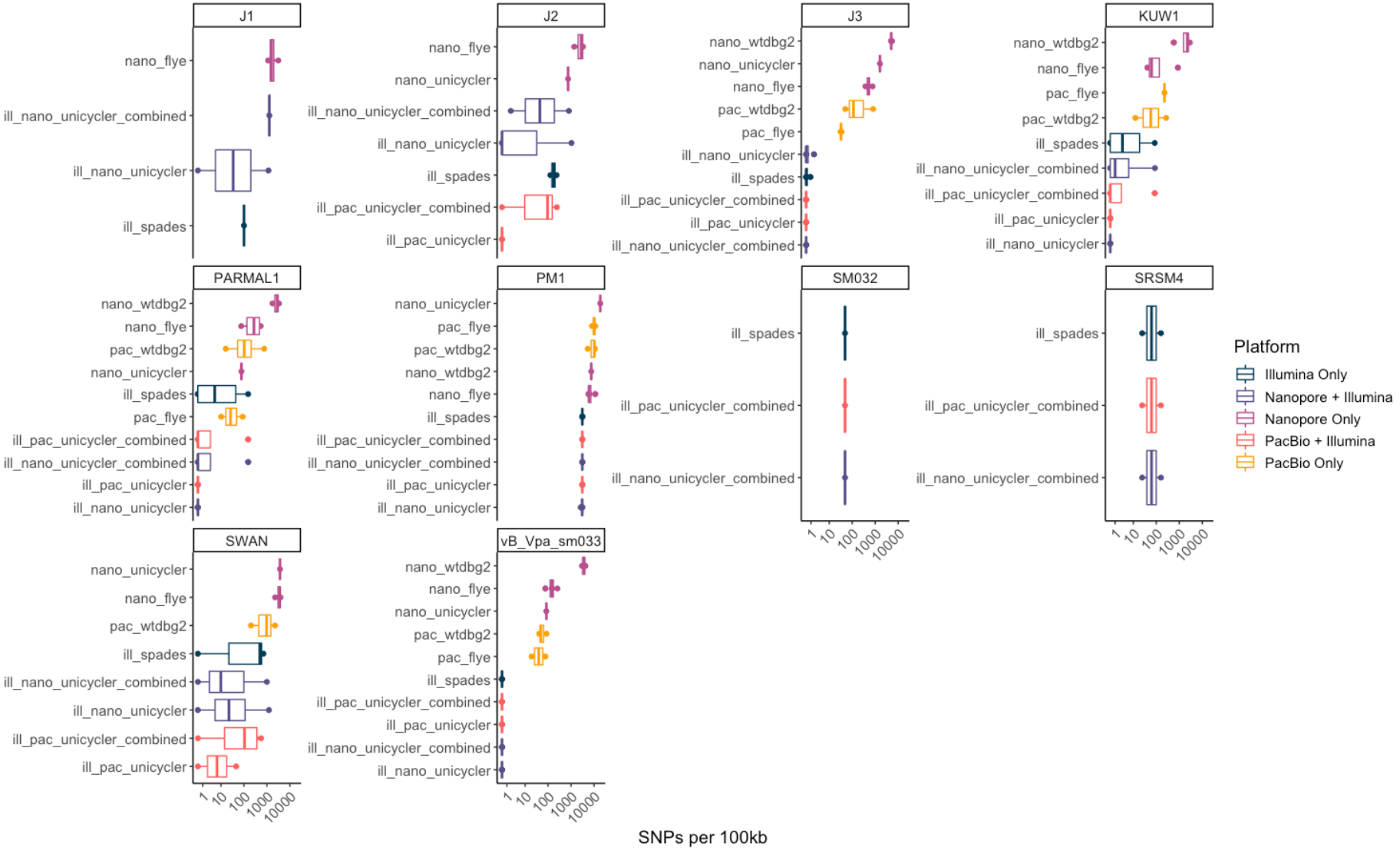

J1
nano_flye
ill_nano_unicycler_combined
ill_nano_unicycler
ill_spades
J2
nano_flye
nano_unicycler
ill_nano_unicycler_combined
ill_nano_unicycler
ill_spades
ill_pac_unicycler_combined
ill_pac_unicycler
J3
nano_wtdbg2
nano_unicycler
nano_flye
pac_wtdbg2
pac_flye
ill_nano_unicycler
ill_spades
ill_pac_unicycler_combined
ill_pac_unicycler
ill_nano_unicycler_combined
KUW1
nano_wtdbg2
nano_flye
pac_flye
pac_wtdbg2
ill_spades
ill_nano_unicycler_combined
ill_pac_unicycler_combined
ill_pac_unicycler
ill_nano_unicycler
PARMAL1
nano_wtdbg2
nano_flye
pac_wtdbg2
nano_unicycler
ill_spades
pac_flye
ill_pac_unicycler_combined
ill_nano_unicycler_combined
ill_pac_unicycler
ill_nano_unicycler
PM1
nano_unicycler
pac_flye
pac_wtdbg2
nano_wtdbg2
nano_flye
ill_spades
ill_pac_unicycler_combined
ill_nano_unicycler_combined
ill_pac_unicycler
ill_nano_unicycler
SM032
ill_spades
ill_pac_unicycler_combined
ill_nano_unicycler_combined
SRSM4
ill_spades
ill_pac_unicycler_combined
ill_nano_unicycler_combined
SWAN
nano_unicycler
nano_flye
pac_wtdbg2
ill_spades
ill_nano_unicycler_combined
ill_nano_unicycler
ill_pac_unicycler_combined
ill_pac_unicycler
vB_Vpa_sm033
nano_wtdbg2
nano_flye
nano_unicycler
pac_wtdbg2
pac_flye
ill_spades
ill_pac_unicycler_combined
ill_pac_unicycler
ill_nano_unicycler_combined
ill_nano_unicycler
1
10
100
1000
10000
SNPs per 100kb
Platform
Illumina Only
Nanopore + Illumina
Nanopore Only
PacBio + Illumina
PacBio Only

Figure 3.15 Genome by genome breakdown of SNPs per assembly

Boxplots showing number of SNPs per 100kb by assembly, by genome.

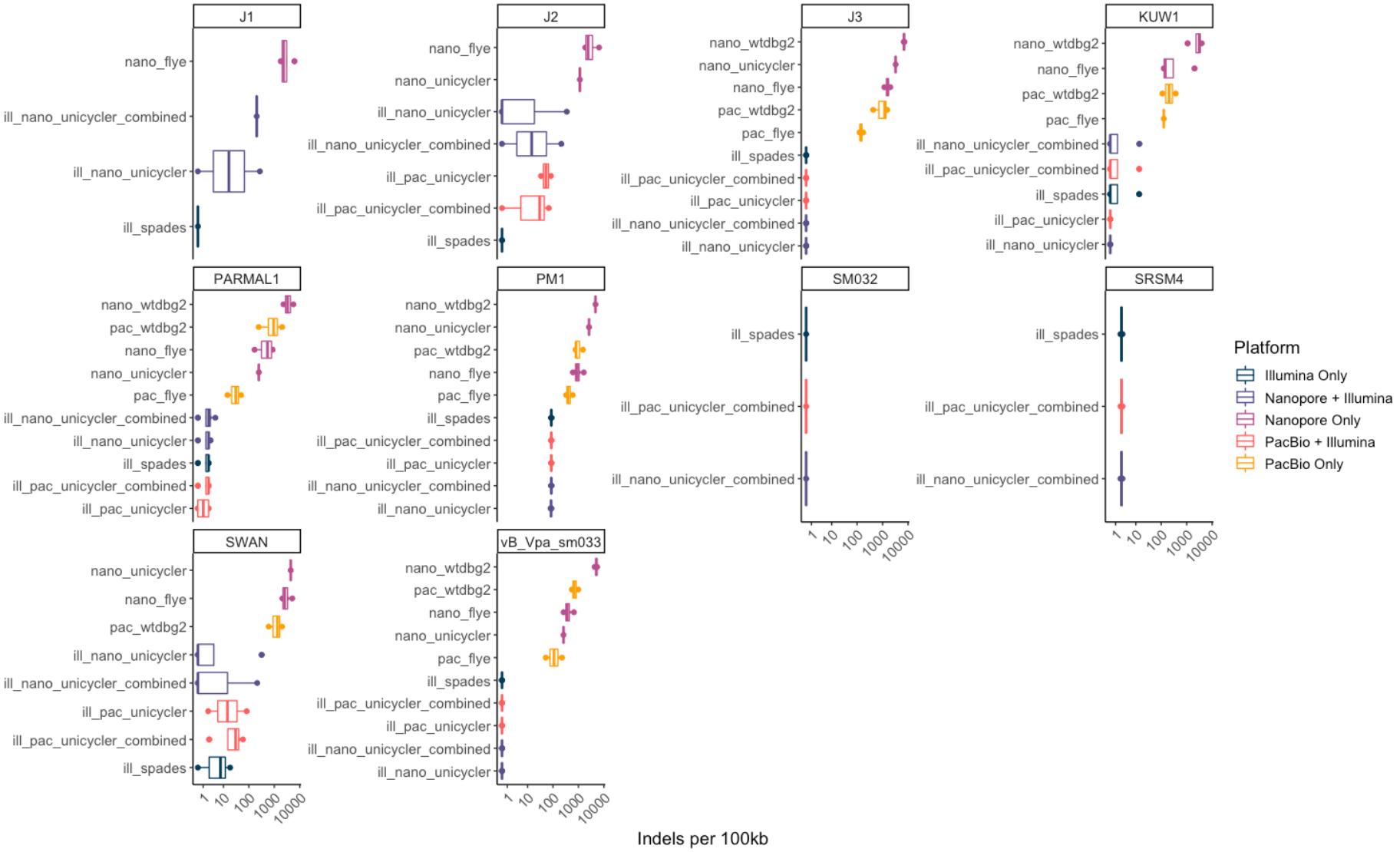

J1
nano_flye
ill_nano_unicycler_combined
ill_nano_unicycler
ill_spades
J2
nano_flye
nano_unicycler
ill_nano_unicycler
ill_nano_unicycler_combined
ill_pac_unicycler
ill_pac_unicycler_combined
ill_spades
J3
nano_wtdbg2
nano_unicycler
nano_flye
pac_wtdbg2
pac_flye
ill_spades
ill_pac_unicycler_combined
ill_pac_unicycler
ill_nano_unicycler_combined
ill_nano_unicycler
KUW1
nano_wtdbg2
nano_flye
pac_wtdbg2
pac_flye
ill_nano_unicycler_combined
ill_pac_unicycler_combined
ill_spades
ill_pac_unicycler
ill_nano_unicycler
PARMAL1
nano_wtdbg2
pac_wtdbg2
nano_flye
nano_unicycler
pac_flye
ill_nano_unicycler_combined
ill_nano_unicycler
ill_spades
ill_pac_unicycler_combined
ill_pac_unicycler
PM1
nano_wtdbg2
nano_unicycler
pac_wtdbg2
nano_flye
pac_flye
ill_spades
ill_pac_unicycler_combined
ill_pac_unicycler
ill_nano_unicycler_combined
ill_nano_unicycler
SM032
ill_spades
ill_pac_unicycler_combined
ill_nano_unicycler_combined
SRSM4
ill_spades
ill_pac_unicycler_combined
ill_nano_unicycler_combined
SWAN
nano_unicycler
nano_flye
pac_wtdbg2
ill_nano_unicycler
ill_nano_unicycler_combined
ill_pac_unicycler
ill_pac_unicycler_combined
ill_spades
vB_Vpa_sm033
nano_wtdbg2
pac_wtdbg2
nano_flye
nano_unicycler
pac_flye
ill_spades
ill_pac_unicycler_combined
ill_pac_unicycler
ill_nano_unicycler_combined
ill_nano_unicycler
1
10
100
1000
10000
Indels per 100kb
Platform
Illumina Only
Nanopore + Illumina
Nanopore Only
PacBio + Illumina
PacBio Only

Figure 3.16 Genome by genome breakdown of INDELs per assembly

Boxplots showing number of INDELs per 100kb by assembly, by genome.

Chapter 4 Exploring Phages within Dairy Farm Slurry

4.1 Chapter Preface

This chapter presents the work previously published in a paper format 'Hybrid assembly of an agricultural slurry virome reveals a diverse and stable community with the potential to alter the metabolism and virulence of veterinary pathogens. Cook, R. et al (2021) Microbiome.' https://doi.org/10.1186/s40168-021-01010-3. The text and figures from the published paper have been inserted into this chapter *verbatim*. As this work is not mine alone, the contribution of other authors is outlined below.

4.1.1 Author Contributions

Study design, sample collection, and sequencing were performed as part of the wider EVAL-FARMS consortium, prior to commencement of this PhD project (Baker *et al.*, 2022). Michael Jones, Andrew Millard, Jon Hobman, Christine Dodd and Dov Stekel conceived the study. Steven Hooton and Liz King collected and processed the samples. Ryan Cook, Steven Hooton, Urmi Trivedi and Andrew Millard carried out the bioinformatic analysis. Ryan Cook, Michael Jones and Andrew Millard drafted the manuscript. All authors approved and contributed to the final manuscript.

4.1.2 Chapter Objectives

The aim of this work was to characterise the viral community of agricultural slurry over time, and to determine if long read sequencing would uncover more viruses in a natural community. Therefore, the objectives were to:

1. To perform a comparison of Illumina, Nanopore, and hybrid approaches for sequencing a natural viral community
2. To characterise the viruses present in slurry and investigate their community dynamics over time

3. To determine if the viruses present may augment the metabolism of their bacterial hosts in the wider environment

4.2 Abstract

Background

Viruses are the most abundant biological entities on Earth, known to be crucial components of microbial ecosystems. However, there is little information on the viral community within agricultural waste. There are currently ~ 2.7 million dairy cattle in the UK producing 7–8% of their own bodyweight in manure daily, and 28 million tonnes annually. To avoid pollution of UK freshwaters, manure must be stored and spread in accordance with guidelines set by DEFRA. Manures are used as fertiliser, and widely spread over crop fields, yet little is known about their microbial composition. We analysed the virome of agricultural slurry over a 5-month period using short and long-read sequencing.

Results

Hybrid sequencing uncovered more high-quality viral genomes than long or short-reads alone; yielding 7682 vOTUs, 174 of which were complete viral genomes. The slurry virome was highly diverse and dominated by lytic bacteriophage, the majority of which represent novel genera (~ 98%). Despite constant influx and efflux of slurry, the composition and diversity of the slurry virome was extremely stable over time, with 55% of vOTUs detected in all samples over a 5-month period. Functional annotation revealed a diverse and abundant range of auxiliary metabolic genes and novel features present in the community, including the agriculturally relevant virulence factor VapE, which was widely distributed across different phage genera that were predicted to infect several hosts. Furthermore, we identified an abundance of phage-encoded diversity-generating retroelements, which were previously thought to be rare on lytic

viral genomes. Additionally, we identified a group of crAssphages, including lineages that were previously thought only to be found in the human gut.

Conclusions

The cattle slurry virome is complex, diverse and dominated by novel genera, many of which are not recovered using long or short-reads alone. Phages were found to encode a wide range of AMGs that are not constrained to particular groups or predicted hosts, including virulence determinants and putative ARGs. The application of agricultural slurry to land may therefore be a driver of bacterial virulence and antimicrobial resistance in the environment.

4.3 Introduction

Bacteriophages, or simply phages are recognised as the most abundant biological entities on the planet (Cobián Güemes *et al.*, 2016) and drive bacterial evolution through predator-prey dynamics (Bohannan and Lenski, 2000; Buckling and Rainey, 2002), and horizontal gene transfer (Canchaya *et al.*, 2003). In all systems where phages have been studied in detail, they have significant ecological roles (Clokie *et al.*, 2011; Breitbart *et al.*, 2018; Sutton and Hill, 2019). The contribution of phages to microbial communities has arguably been most extensively studied in the oceans (Yooseph *et al.*, 2007; Hurwitz and U'Ren, 2016; Paez-Espino *et al.*, 2016; Roux, Brum, *et al.*, 2016; Gregory *et al.*, 2019) where, in addition to releasing large quantities of organic carbon and other nutrients through lysing bacteria, marine phages are thought to contribute to biogeochemical cycles by augmenting host metabolism with auxiliary metabolic genes (AMGs) (Anantharaman *et al.*, 2014; Zhang, Wei and Cai, 2014; Roux, Brum, *et al.*, 2016; York, 2017). Since their initial discovery, AMGs have been identified in diverse environments, including the ocean and soils (Hurwitz and U'Ren, 2016; Jin *et al.*, 2019). The putative functions of AMGs are wide-ranging with the potential to alter photosynthesis, carbon metabolism, sulphur metabolism, nitrogen uptake and complex carbohydrate metabolism (Yooseph *et al.*, 2007; Dinsdale *et al.*, 2008; Sharon *et al.*, 2011; Hurwitz, Hallam and Sullivan, 2013; Anantharaman *et al.*, 2014; Hurwitz, Brum and Sullivan, 2015; Roux, Brum, *et al.*, 2016; Monier *et al.*, 2017; Jin *et al.*, 2019).

In addition to augmenting host metabolism, phages can contribute to bacterial virulence through phage conversion via the carriage of virulence factors and toxins (Freeman, 1951; Eklund *et al.*, 1974; Waldor and Mekalanos, 1996; Wagner *et al.*,

2002; Fortier and Sekulovic, 2013; Khalil *et al.*, 2016). Phages have also been implicated in the transfer of antimicrobial resistance genes (ARGs) (Balcázar, 2020); however, the study into the importance of phages in the transfer of ARGs has reached polarising conclusions (Enault *et al.*, 2017; Debroas and Siguret, 2019). Despite the vital and complex contributions of phages to microbial ecology, there is a lack of knowledge about their roles in agricultural slurry.

Manure is an unavoidable by-product from the farming of livestock. There are ~ 2.7 million dairy cattle in the UK, with ~ 1.8 million in milking herds (AHDB, 2018). A fully grown milking cow produces 7–8% of their own bodyweight as manure per day (Font-Palma, 2019), leading to an estimated 28.31 million tonnes of manure produced by UK dairy cattle in 2010 alone (Smith and Williams, 2016). These wastes are rich in nitrates and phosphates, making them valuable as a source of organic fertiliser, with an average value of £78 per cow per year (AHDB, no date b). However, agricultural wastes can be an environmental pollutant. Inadequate storage and agricultural run-off may lead to an increased biological oxygen demand (BOD) of freshwaters, leading to algal blooms and eutrophication (Sandars *et al.*, 2003; Thomassen *et al.*, 2008; Prapaspongsa *et al.*, 2010; De Vries, Groenestein and De Boer, 2012). Areas particularly at risk of nitrate pollution of ground or surface waters are classified as nitrate vulnerable zones (NVZs), and these constitute 55% of land in England (UK Government, 2013). For this reason, the application of organic fertilisers to fields in the UK is strictly controlled and can only be applied during certain times of the year (UK Government, no date). Thus, there is the requirement to store vast volumes of slurry for several months.

To produce slurry, solids are separated from manure using apparatus such as a screw press. The liquid fraction forms the basis of slurry, which is stored in a tank or lagoon, where it is mixed with water and other agricultural wastes before its application as fertiliser. Despite being widely used as a fertiliser, the composition of the virome within slurry is poorly studied. Culture-based approaches have been used to study phages infecting specific bacteria such as *Escherichia coli* (Smith *et al.*, 2015; Sazinas *et al.*, 2018; Besler *et al.*, 2020), but total viral diversity within cattle slurry remains largely unexplored.

Short-read viromics has transformed our understanding of phages in other systems, allowing an overview of the abundance and diversity of phages (Brum *et al.*, 2015; Paez-Espino *et al.*, 2016; Roux, Brum, *et al.*, 2016; Gregory *et al.*, 2019) and AMGs found within their genomes (Anantharaman *et al.*, 2014; Roux, Brum, *et al.*, 2016; Jin *et al.*, 2019). The power of viromics is exemplified by the study of crAssphage, which was first discovered in viromes in 2014 (Dutilh *et al.*, 2014) and has subsequently been found to be the most abundant phage in the human gut and has recently been brought into culture (Dutilh *et al.*, 2014; Guerin *et al.*, 2018; Shkoporov *et al.*, 2018). However, the use of short-reads is not without limitations. Phages that contain genomic islands and/or have high micro-diversity, such as phages of the ubiquitous *Pelagibacterales* (Zhao *et al.*, 2013; Martinez-Hernandez *et al.*, 2019), can cause genome fragmentation during assembly (Temperton and Giovannoni, 2012; Mizuno, Ghai and Rodriguez-Valera, 2014; Roux *et al.*, 2017; Olson *et al.*, 2019). The development of long-read sequencing technologies—most notably Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT)—offer a solution to such issues. The longer reads are potentially able to span the length of entire phage genomes, overcoming

assembly issues resulting from repeat regions and low coverage (Temperton and Giovannoni, 2012; Mizuno, Ghai and Rodriguez-Valera, 2014; Olson *et al.*, 2019). The cost of longer reads is a higher error rate, which can lead to inaccurate CDS prediction (Buck *et al.*, 2017; Watson and Warr, 2019).

Recently, a Long-Read Linker-Amplified Shotgun Library (LASL) approach was developed that combines LASL library preparation with ONT MinION sequencing (Warwick-Dugdale *et al.*, 2019). This approach overcame both the requirement for high DNA input for MinION sequencing and associated assembly issues with short-read sequencing. The resulting assembly increased both the number and completeness of phage genomes compared to short-read assemblies (Warwick-Dugdale *et al.*, 2019). An alternative approach that has utilised long-read sequencing used the ONT GridION platform to obtain entire phage genomes using an amplification-free approach on high molecular weight DNA (Beaulaurier *et al.*, 2020). While this approach recovered over 1000 high-quality viral genomes that could not be recovered from short-reads alone, it requires large amounts of input DNA (Beaulaurier *et al.*, 2020), that may be a limiting factor of many environments.

The aim of this work was to utilise viral metagenomics to investigate the diversity, community structure and ecological roles of viruses within dairy cattle slurry that is spread on agricultural land as an organic fertiliser.

4.4 Materials and Methods

4.4.1 DNA extraction and sequencing

DNA from the viral fraction was extracted from 10 ml of slurry as previously described (Sazinas *et al.*, 2019). Briefly, slurry was mixed with PBS buffer and centrifuged, prior to filtration to remove bacteria. Viral particles were concentrated using an Amicon column (Sigma-Aldrich) and DNA was extracted using a standard phenol-chloroform extraction. For short-read sequencing on un-amplified DNA, Illumina sequencing was carried out on NovaSeq using 2 × 150 library. For long read sequencing, DNA from four viral samples was pooled and subject to amplification with Illustra Ready-To-Go Genomphi V3 DNA amplification kit (GE, Healthcare) following the manufacturer's instructions. Post amplification DNA was de-branched with S1 nuclease (Thermo Fisher Scientific), following the manufacturer's instructions and cleaned using a DNA Clean and Concentrator column (Zymo Research). Sequencing was carried out by Edinburgh Genomics, with size selection of DNA to remove DNA < 5 kb prior to running on single PromethION flow cell. Reads were based called with guppy v2.3.35.

4.4.2 Assembly and quality control

Illumina virome reads were trimmed with Trimmomatic v0.36 (Bolger, Lohse and Usadel, 2014) using the following settings; PE illuminaclip, 2:30:10 leading:15 trailing:15 slidingwindow:4:20 minlen:50. Reads from the five samples were co-assembled with MEGAHIT v1.1.2 (Li *et al.*, 2016) using the settings; --k-min 21 --k-max 149 --k-step 24. Long-reads were assembled with flye v2.6-g0d65569, reads were mapped back against the assembly with Minimap2 v2.14-r892-dirty (Li, 2018) to produce BAM files and initially corrected with marginPolish v1.0.0 with

'allParams.np.ecoli.json'. Bacterial contamination and virus-like particle (VLP) enrichment was assessed with ViromeQC v1.0 (Zolfo *et al.*, 2019).

4.4.3 Identifying viral operational taxonomic units

To identify viral contigs, a number of filtering steps were applied. All contigs ≥ 10 kb and circular contigs < 10 kb (Roux *et al.*, 2017) were processed using MASH v2.0 (Ondov *et al.*, 2016) separately against the RefSeq70 database (O'Leary *et al.*, 2016) and a publicly available database of phage genomes (March 2020; P = 0.01). If the closest RefSeq70 hit was to a phage/virus, the contig was included as a viral operational taxonomic unit (vOTU). Failing this, if the contig obtained a closer hit to the phage database than RefSeq70, the contig was included as a vOTU. Remaining contigs were included as vOTUs if they satisfied at least two of the following criteria; 1: VIBRANT v1.0.1 indicated sequence is viral (Kieft, Zhou and Anantharaman, 2020), 2: obtained adjusted p value ≤ 0.05 from DeepVirFinder v1.0 (Ren *et al.*, 2020), 3: 30% of ORFs on the contig obtained a hit to a prokaryotic virus orthologous group (pVOG) (Grazziotin, Koonin and Kristensen, 2017) using Hmmscan v3.1b2 (-E 0.001) ('HMMER', no date). However, circular contigs < 10 kb only had to satisfy either criteria 1 or 3, as DeepVirFinder scores for these contigs were inconsistent.

4.4.4 Prophage analysis

A set of prophage sequences was identified from bacterial metagenomes from the same tank were included. These were filtered as above, however contigs < 10 kb were not included even if circular. To determine which prophage vOTUs could be detected in the free viral fraction, Illumina virome reads were mapped to vOTUs using Bbmap v38.69 (Bushnell, 2013) at 90% minimum ID and the ambiguous=all flag, and

PromethION reads were mapped to prophage vOTUs using Minimap2 v2.14-r892-dirty (Li, 2018) with parameters '-a -x map-ont'. vOTUs were deemed as present in the free viral fraction if they obtained ≥ 1x coverage across ≥ 75% of contig length in at least one sample (Roux *et al.*, 2017). To determine the ends of prophages, differential coverage obtained by mapping the Illumina virome reads was investigated. Median coverage of the whole prophage was calculated and compared to median coverage across a 500 bp sliding window (Supplementary Tables S4.6 & S4.7). If the 500 bp window had a depth of coverage ≥ 2x standard deviations lower than the median coverage of the whole prophage, this was considered a break in coverage and used to infer the ends of the prophage. An example is provided in Figure 4.7.

4.4.5 Hybrid assembly composition

Illumina reads were mapped to PromethION vOTUs using Minimap2 v2.14-r892-dirty (Li, 2018) and the contigs were polished using Pilon v1.22 (Walker *et al.*, 2014). The PromethION vOTUs underwent multiple rounds of polishing until changes to the sequence were no longer made, or the same change was swapped back and forth between rounds of polishing. The Illumina vOTUs, hybrid vOTUs and prophage vOTUs (only those detected in the viral fraction) were de-replicated at 95% average nucleotide identity (ANI) over 80% genome length using ClusterGenomes v5.1 (*GitHub - simroux/ClusterGenomes: Archive for ClusterGenomes scripts*, no date) to produce a final set of vOTUs, hereby referred to as the Final Virome. To determine assembly quality, CheckV v0.5.0 (Nayfach *et al.*, 2020) was used. As this pipeline was released after the analysis in this work was performed, this was performed post-analysis.

4.4.6 Alpha diversity and population dynamics

To estimate relative abundance, Illumina reads were mapped to vOTUs using Bbmap v38.69 (Bushnell, 2013) at 90% minimum ID and the ambiguous=all flag. vOTUs were deemed as present in a sample if they obtained ≥ 1x coverage across ≥ 75% of contig length (Roux *et al.*, 2017). The number of reads mapped to present vOTUs were normalised to reads mapped per million. Relative abundance values were analysed using Phyloseq v1.26.1 (McMurdie and Holmes, 2013) in R v3.5.1 (Team, 2018) to calculate diversity statistics.

Statistical testing of similarity of vOTU profiles between samples was carried out using DirtyGenes (Shaw *et al.*, 2019). We used the randomization option with 5000 simulations rather than chi-squared because of the small number of samples, but resampling from the null hypothesis Dirichlet distribution because there are no replicated libraries; the updated code has been uploaded to GitHub (https://github.com/LMShaw/DirtyGenes). The analysis was repeated using both the preferred cut-off of minimum 1% abundance in at least one sample and also with minimum abundance at 0.5% in at least one sample. This is because with a 1% cut-off only seven vOTUs were included (plus an 'other' category binning all remaining lower abundance vOTUs) which we did not consider to be sufficiently representative; with 0.5%, 22 vOTUs were included (plus an 'other' category).

4.4.7 Functional annotation

Final Virome vOTUs were annotated using Prokka v1.12 (Seemann, 2014) with a custom database created from phage genomes downloaded at the time (March, 2020) (Michniewski *et al.*, 2019), and ORFs were compared to profile HMMs of pVOGs

(Grazziotin, Koonin and Kristensen, 2017) using Hmmscan v3.1b2 (-E 0.001) ('HMMER', no date). Final Virome vOTU ORFs were clustered at 90% ID over 90% contig length using CD-HIT v4.6 (Fu *et al.*, 2012) to reduce redundancy. The resultant proteins were submitted to eggNOG-mapper v2.0 (Huerta-Cepas *et al.*, 2018) with default parameters, and the output was manually inspected to identify AMGs of interest. Translated ORFs identified as carbohydrate-active enzymes (CAZYmes) by eggNOG were submitted to the dbCAN2 meta-server for CAZYme identification using the HMMER method to confirm their identity (Zhang *et al.*, 2018; Jin *et al.*, 2019).

4.4.8 Diversity-generating retroelement analysis

vOTUs found to encode a putative reverse transcriptase were processed using MetaCCST (Yan *et al.*, 2019) to identify potential diversity-generating retroelements (DGRs). To identify hypervariable regions in the target gene of DGRs, reads from each sample were individually mapped to vOTUs using Bbmap v38.69 (Bushnell, 2013) at 95% minimum ID with the ambiguous=all flag. Resultant bam files were processed with Samtools v1.10 (Li *et al.*, 2009) to produce a mpileup file. Variants were called using VarScan v2.3 (Koboldt *et al.*, 2012) mpileup2snp command with parameters '--min-coverage 10 --min-avg-qual-30'. The percentage of SNP sites per gene were calculated for both DGR target gene(s) and all other genes on the vOTU, in order to identify if the DGR target gene(s) contained more SNP sites than on average across the vOTU.

4.4.9 Taxonomy and predicted host

Final Virome vOTUs were clustered using vConTACT2 v0.9.13 (Bin Jang *et al.*, 2019) with parameters; --db 'ProkaryoticViralRefSeq85-Merged' --pcs-mode MCL --vcs-

mode ClusterONE. A set of publicly available phage genome sequences (7527), that had been deduplicated at 95% identity with dedupe.sh v36.20 (Bushnell, 2013), were included. The resultant network was visualised using Cytoscape v3.7.1 (Shannon *et al.*, 2003). This method clusters vOTUs based upon shared proteins, with vOTUs belonging to the same cluster likely belonging to the same genus/sub-family. Although not precise enough for robust species/genus level classification, this method allows users to rapidly classify large numbers of vOTUs at higher taxonomic ranks.

To determine if any previously known phage genomes were present in slurry viromes, reads were mapped to a dataset of publicly a set of publicly available phage genome sequences (March, 2020; 11,030), that had been deduplicated at 95% identity with dedupe.sh v36.20 (Bushnell, 2013). Illumina reads were mapped using Bbmap v38.69 (Bushnell, 2013) at 90% minimum ID (Roux *et al.*, 2017) and the ambiguous=all flag. PromethION reads were mapped using Minimap2 v2.14-r892-dirty (Li, 2018) with parameters '-a -x map-ont'. Phages were deemed as present if they obtained ≥ 1x coverage across ≥ 75% of sequence length (Roux *et al.*, 2017).

Putative hosts for viral vOTUs were predicted with WiSH v1.0 (Galiez *et al.*, 2017) using a database of 9620 bacterial genomes. A p value cut-off of 0.05 was used. Taxonomy for the predicted hosts was obtained using the R (Team, 2018) package Taxonomizr v0.5.3 (Sherrill-Mix, 2018).

4.4.10 Lifestyle prediction

To determine which Final Virome vOTUs were temperate, ORFs were compared to a custom set of 29 profile HMMs for transposase, integrase, excisionase, resolvase and

recombinase proteins downloaded from Pfam (PF07508, PF00589, PF01609, PF03184, PF02914, PF01797, PF04986, PF00665, PF07825, PF00239, PF13009, PF16795, PF01526, PF03400, PF01610, PF03050, PF04693, PF07592, PF12762, PF13359, PF13586, PF13610, PF13612, PF13701, PF13737, PF13751, PF13808, PF13843 and PF13358) (El-Gebali *et al.*, 2019) using Hmmscan v3.1b2 ('HMMER', no date) with the --cut_ga flag. Any vOTUs with an ORF which obtained a hit were classified as temperate.

4.4.11 Positive selection analysis

Final Virome vOTUs which obtained ≥ 15x median coverage across ≥ 75% of contig length in every sample (excluding PHI75) were included in variant analysis. Briefly, reads were mapped onto the contigs using Bbmap v38.69 (Bushnell, 2013) at 95% minimum ID with the ambiguous=all flag, and a sorted indexed BAM file was produced. Snippy v4.4.5 (Seemann, no date b) was used to call variants with parameters '--mapqual 0 --mincov 10'. For genes which contained at least one single nucleotide polymorphism (SNP) or multiple nucleotide polymorphism (MNP), natural selection (pN/pS) was calculated using a method adapted from Gregory et al. (Gregory *et al.*, 2019). In this method, adjacent SNPs were linked as MNPs by Snippy.

4.5 Results

The farm in this study is a high-performance dairy farm in the East Midlands, UK with ~ 200 milking cattle. It houses a three million litre capacity slurry tank and an additional seven million litre lagoon to house overflow from the tank. The tank receives daily influent from the dairy farm including faeces, urine, washwater, footbath and waste milk through a slurry handling and general farm drainage system. Slurry solids are separated using a bed-press and solids are stored in a muck heap. The slurry tank and muck heap are open to the elements and the slurry tank also receives further influent from rainwater, muck heap run-off, and potentially from wildlife. The tank is emptied to ~ 10% of its maximum volume every ~ 6 weeks and the slurry is applied on fields as fertiliser.

4.5.1 Comparison of short- and long-read assemblies

Five samples were collected from the slurry tank over a five-month period (07/06/2017–10/10/2017) (Supplementary Table S4.1) with Illumina libraries prepared from each sample. Initial analysis of the five samples sequencing data using viromeQC (Zolfo *et al.*, 2019) indicated that one sample (PHI75) had high levels of bacterial contamination (Supplementary Table S4.1). Sample PHI75 was excluded from further analysis, with remaining DNA from the other four samples pooled, amplified and sequenced by PromethION sequencing.

Assembly was carried out with just Illumina or PromethION reads, resulting in 1844 and 4954 vOTUs ≥ 10 kb respectively. The PromethION assembly resulted in an increase in the median contig size from 12,648 to 14,658 compared to the Illumina only assembly (Figure 4.1A). The number of predicted genes per kb was also higher

in the PromethION assembly. The increased error rate of Nanopore sequencing compared to Illumina sequencing is known to result in truncated gene calls (Buck *et al.*, 2017; Watson and Warr, 2019). To alleviate this, PromethION contigs were polished with Illumina reads, creating a hybrid assembly and resulting in a decrease in the number of genes per kb from 2.059 (median length: 85 aa) to 1.706 (median length: 103 aa; Figure 4.1B).

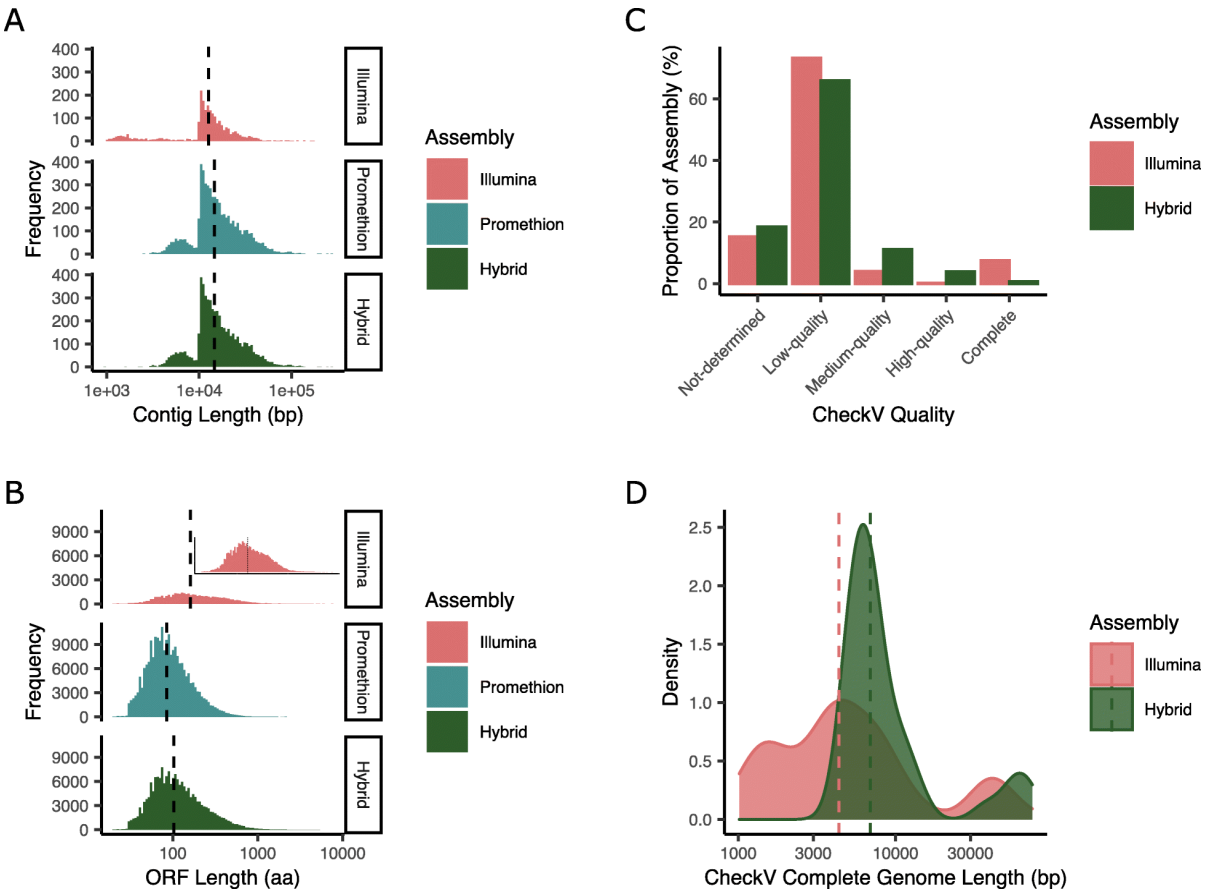


Figure 4.1 Overview of the effect of polishing PromethION vOTUs with Illumina reads

(A) Distribution of the length of vOTUs obtained from Illumina, PromethION and Hybrid assemblies. **(B)** Distribution of predicted ORF lengths obtained from Illumina, PromethION and Hybrid assemblies. **(C)** Quality assessment of vOTUs obtained from Illumina, PromethION and Hybrid assemblies from checkV analysis. **(D)** Genome completeness assessed by CheckV for the Illumina and Hybrid assemblies. The dashed lines in plots **A**, **B** and **D** indicate median values.

As whole genome amplification was used to gain sufficient material for PromethION sequencing, all diversity statistics and relative abundance data was determined from Illumina reads only. The percentage of reads that could be recruited to each different assembly was assessed. Both the PromethION (32.663%) and hybrid (33.976%) assemblies recruited more reads than the Illumina assembly (9.048%; Figure 4.2B). The median number of observed vOTUs per sample was higher in the PromethION (3,483) and hybrid (3,532) assemblies than that of the Illumina assembly (2028; Figure 4.2A). The predicted Shannon and Simpson diversity indices increased in the hybrid (Shannon: 6.909; Simpson: 0.997) and PromethION (Shannon: 6.867; Simpson: 0.997) assemblies compared to the Illumina assembly (Shannon: 5.557; Simpson: 0.972; Figure 4.2C, D).

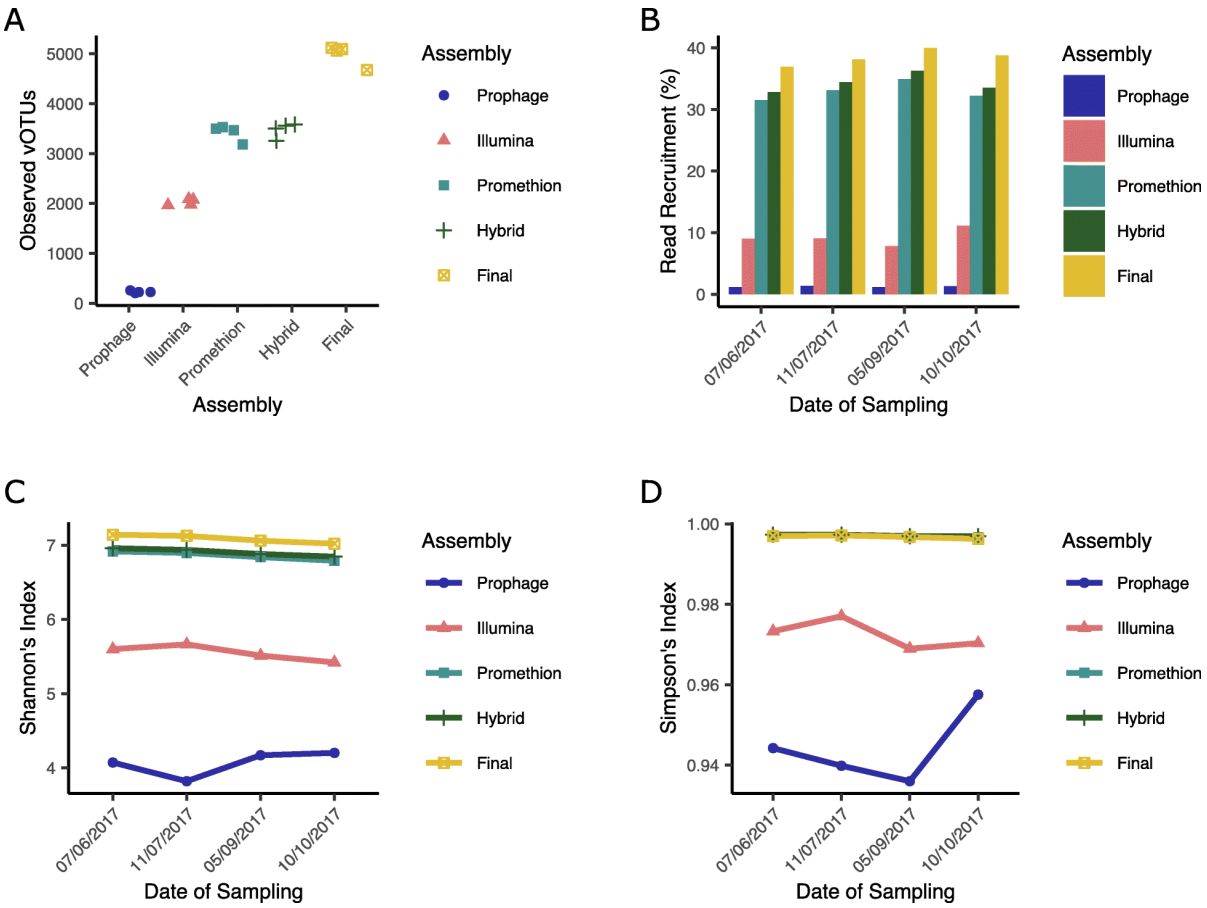


Figure 4.2 Abundance and diversity of vOTUs in different assemblies

(A) Number of vOTUs observed in each sample obtained from normalised read counts. The hybrid assembly is the combination of both Illumina and PromethION reads. Prophage were predicted from a bacterial metagenome from the same sample. Final assembly was combination of Illumina, hybrid and identified active prophage where were dereplicated at 95% ANI. **(B)** Read recruitment over time for the different assemblies. **(C)** Shanon's α-diversity from different assemblies for each sampling point. **(D)** Simpson's α-diversity assemblies for each sampling point.

To determine the completeness and quality of the identified viral contigs, CheckV (Nayfach *et al.*, 2020) was used. The hybrid assembly contained a lower proportion of low-quality genomes (65.886%), and a higher proportion of medium and high-quality (15.015%) genomes than the Illumina assembly (low-quality: 73.217%; medium and high-quality: 4.083%; Figure 4.1C). Conversely, the Illumina assembly contained more predicted complete genomes than the hybrid assembly (Illumina: 167; hybrid: 40). This may be due to the size selection of PromethION sequencing for longer reads, reflected in the longer average length of the complete genomes obtained from hybrid assembly (Figure 4.1D).

To fully understand the diversity of phages within the slurry tank, we also investigated the presence of prophage elements in the bacterial fraction. A total of 2892 putative prophages were predicted, of which only 407 could be detected in the free phage fraction by read mapping. We combined the predicted 407 active prophages, with the Illumina and hybrid assemblies. Redundancy was removed using cluster_phages_genomes.pl (*GitHub - simroux/ClusterGenomes: Archive for ClusterGenomes scripts*, no date), resulting in 7682 vOTUs. Having established the most comprehensive DNA virome possible, the data was further analysed.

4.5.2 Characterisation of the slurry virome

The percentage of reads that could be recruited from each sample varied from 36.943% (PHI73; 07/06/2017; Figure 4.2B) to 39.996% (PHI76; 05/09/2017; Figure 4.2B). Across the five-month sampling period, the Shannon's index alpha diversity estimates only varied from 7.02 (PHI77; 10/10/2017) to 7.141 (PHI73; 07/06/2017), suggesting a stable and diverse virome across seasons (Figure 4.2C, D). Although

diverse, the virome remained stable across all sampling points with 55% (4,256) of 7682 vOTUs found in all samples, and only 477 (~ 6%) of vOTUs unique to any one sampling point. Furthermore, testing with DirtyGenes (Shaw *et al.*, 2019) found no significant difference between the vOTU abundance profiles of the samples ($p = 0.1142$ with 1% cut-off; $p = 0.863$ with 0.5% cut-off). To determine if the stability in macro-diversity was mirrored by changes in micro-diversity, we assessed which predicted phage genes were under positive selection ($pN/pS > 1$). Our analysis showed 1610/210,997 genes (0.763%) to be under positive selection in at least one sample (Supplementary Table S4.2). From these, putative function could be assigned to 388 translated genes. The most common predicted functions were related to phage tail (30), and phage structure (24).

To give a broader overview of the type of viruses present in the sample, pVOGs were used to infer the taxonomic classification of each vOTU. Of the vOTUs that contained proteins that matched the pVOG databases (Grazziotin, Koonin and Kristensen, 2017), 91% were associated with the order *Caudovirales*, 2.17% associated with non-tailed viruses and the remainder not classified. Approximately 10% (710) of vOTUs were identified as temperate, suggesting that the community is dominated by lytic phages of the order *Caudovirales*. The abundance of temperate vOTUs was constant across samples, ranging from 5.605% (PHI76; 05/09/2017) to 8.866% (PHI77; 10/10/2017), further demonstrating the stability of the system across time.

In order to identify the species of phages present within the slurry, all vOTUs were compared against all known phages (March, 2020) using MASH (Ondov *et al.*, 2016), with an average nucleotide identity (ANI) of > 95% as currently defined as a cut-off for

phage species (Adriaenssens and Rodney Brister, 2017). Only vOTUs ctg5042 and ctg217 with similarity to Mycoplasma bacteriophage L2 (accession BL2CG) and Streptococcus phage Javan630 (accession MK448997) respectively were detected. Furthermore, no vOTUs were similar to any phages that have previously been isolated from this system (Smith *et al.*, 2015; Sazinas *et al.*, 2018; Besler *et al.*, 2020). Thus, the vast majority of vOTUs represent novel phage species.

To gain an understanding of the composition at higher taxonomic levels, vConTACT2 (Bin Jang *et al.*, 2019) was run. Only 217 (2.825%) vOTUs clustered with a reference genome, indicating they are related at the genus level (Figure 4.3A). Notably, 18 vOTUs formed a cluster with ΦCrAss001 (accession MH675552) and phage IAS (accession KJ003983), with ctg20 appearing to be a near-complete phage genome (~ 99 kb; Figure 4.4B). The other 7465 vOTUs clustered only with other vOTUs (3369; 43.856%) or were singletons (4096; 53.319%), indicating 5242 putative new genera. These new genera comprised 98.037% of phages across all samples, suggesting this system is dominated by novel viruses (Figure 4.3B). Working on the assumption that if a vOTU within a viral cluster (VC) was identified as temperate all other vOTUs in the cluster are, the relative abundance of temperate phages was predicted. This ranged from 13.09% (PHI76; 05/09/2017) to 16.249% (PHI77; 10/10/2017), further demonstrating the dominance of lytic viruses and stability of the system over time (Figure 4.3C).

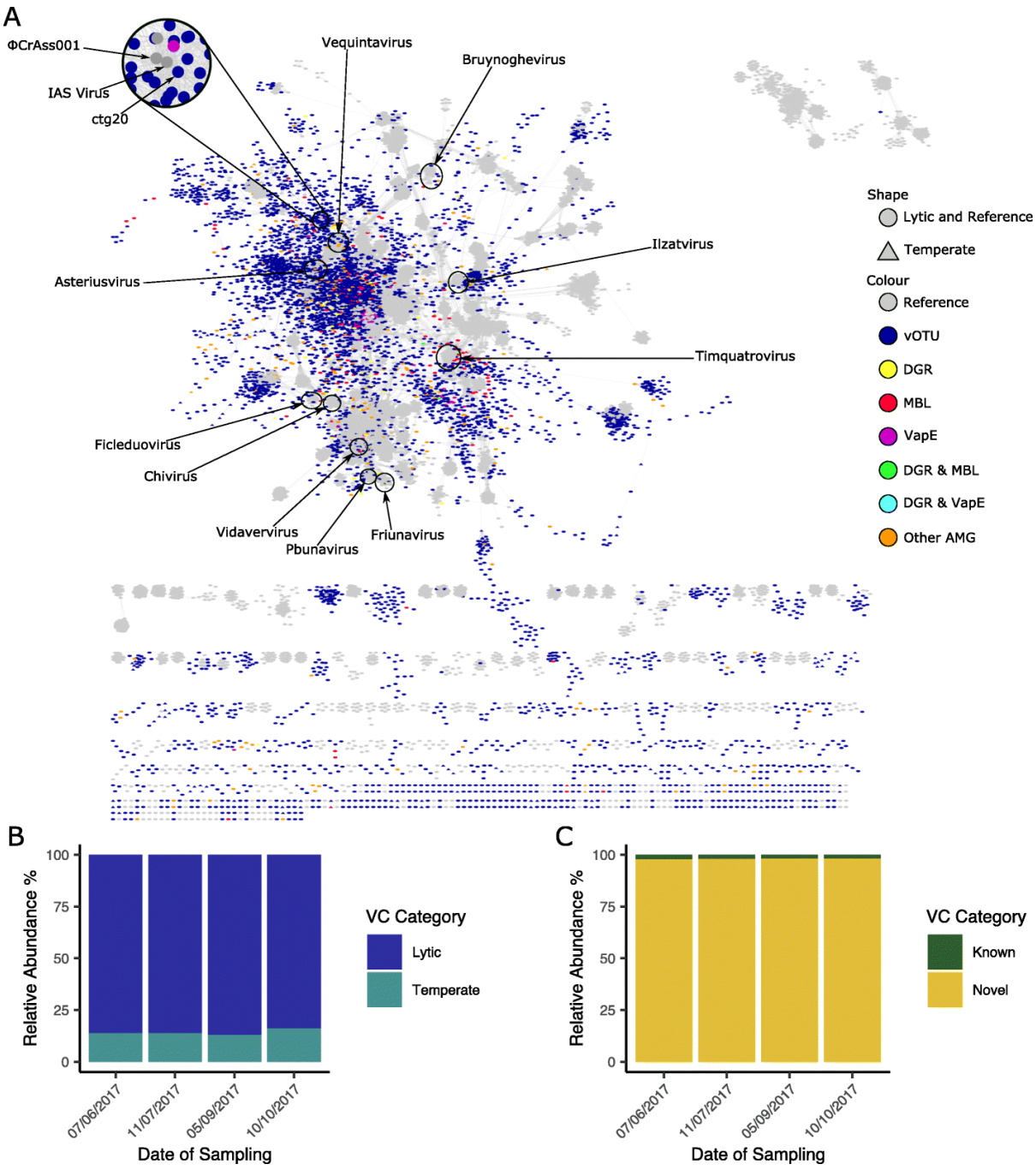


Figure 4.3 Taxonomic analysis of vOTUs

(A) vConTACT2 network analysis of vOTUs from this study and a database of phage genomes extracted from GenBank. The presence of selected viral accessory metabolic genes within viral clusters (VCs) is marked by different colours. **(B)** Abundance of viral clusters that contained ≥ 1 previously known viral genome (known) or no previously known viral genomes (novel). **(C)** Abundance of viral clusters that contained ≥ 1 vOTU predicted to be temperate (temperate) or none (lytic).

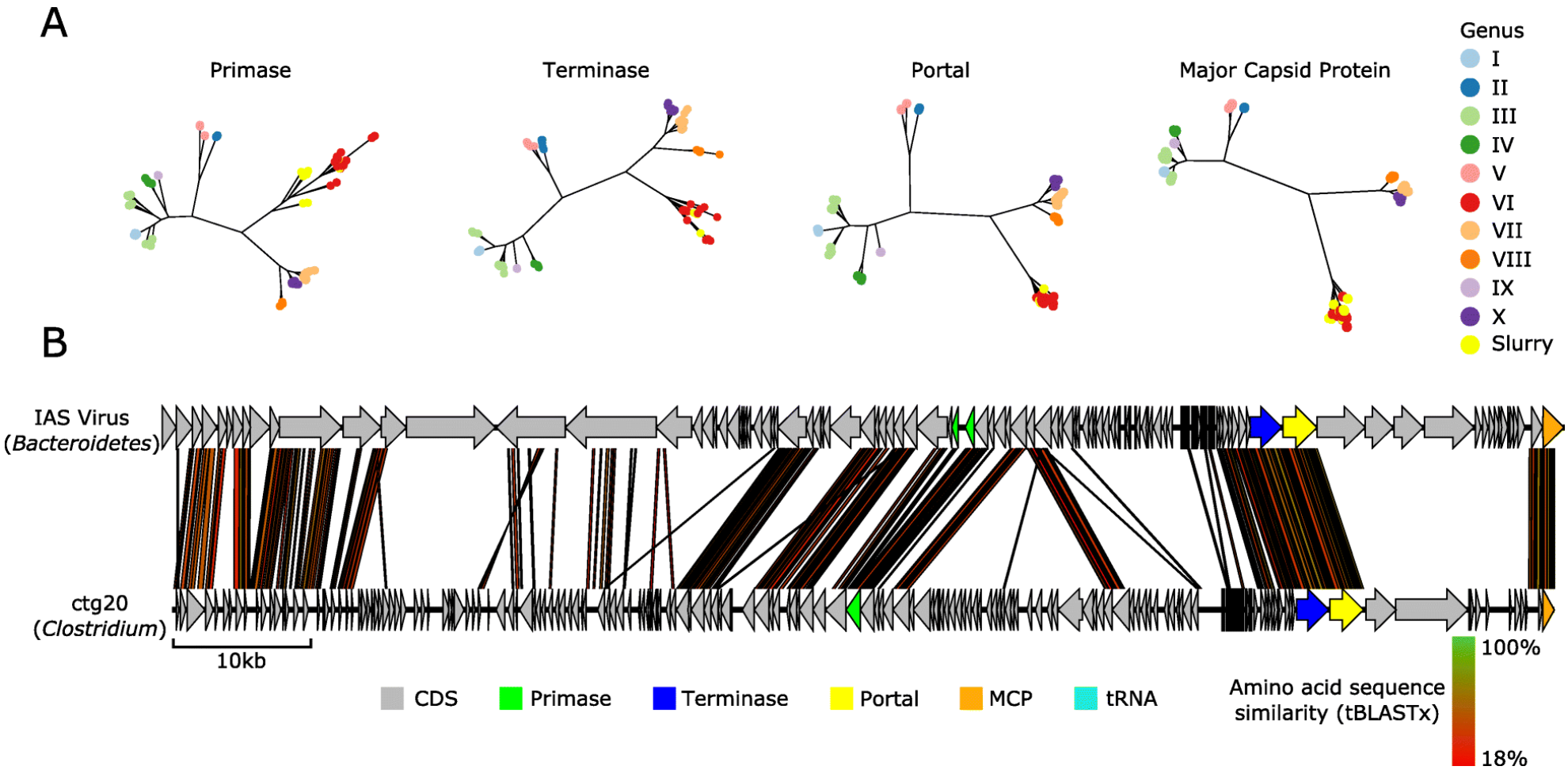

A
Primase
Terminase
Portal
Major Capsid Protein
Genus
I
II
III
IV
V
VI
VII
VIII
IX
X
Slurry
B
IAS Virus
(Bacteroidetes)
ctg20
(Clostridium)
10kb
CDS
Primase
Terminase
Portal
MCP
tRNA
Amino acid sequence
similarity (tBLASTx)
100%
18%

Figure 4.4 Phylogenetic and genomic analysis of slurry crAssphages

(A) Phylogeny of four genes that encode a primase, terminase, portal protein and major capsid protein. The analysis followed the same method as described by Guerin et al. (Guerin *et al.*, 2018), with the ten major clades as previously defined marked. **(B)** Genomic comparison between the complete genome of phage ctg20 and the IAS virus was produced using EasyFig with tBLASTx algorithm and 0.001 E value and length filter 30. Gene products with a predicted function are coloured. The predicted or known host are shown in parentheses.

Hosts were predicted for 3189 vOTUs and the system was found to be dominated by phages predicted to infect bacteria belonging to *Firmicutes* and *Bacteroidetes*, the most dominant phyla found in the cow gut (Kim and Wells, 2016; Delgado *et al.*, 2019; Li *et al.*, 2019). The proportions of host-specific abundances appeared stable across all time points (Figure 4.8).

4.5.3 Identification of CrAss-like phages in the slurry virome

The appearance of a cluster of 18 vOTUs that are similar to crAssphage was surprising given the discovery and abundance of crAssphage in human gut viromes (Dutilh *et al.*, 2014; Guerin *et al.*, 2018; Shkoporov *et al.*, 2018, 2019). To further investigate this, phylogenies based on the method of Guerin et al. were used (Guerin *et al.*, 2018) for 15 vOTUs that contained the specific marker genes. All vOTUs formed part of the previously proposed genus VI (Guerin *et al.*, 2018), including the near complete phage (ctg20; Figure 4.4A; Figure 4.9). Furthermore, the crAssphages identified from slurry did not form a single monophyletic clade. Instead, they were interspersed with human crAssphages, with some slurry crAssphages more closely related to human crAssphages than other slurry crAssphages (Figure 4.4A; Figure 4.9). Genome comparison of ctg20 and phage IAS from genus VI identified synteny in genome architecture between the phages, yet there are clearly several areas of divergence (Figure 4.4B). The predicted host of ctg20 was *Clostridium*, which contrasts to the *Bacteroides* and *Bacteroidetes* that other crAssphages have been demonstrated or predicted to infect respectively (Shkoporov *et al.*, 2018; Yutin *et al.*, 2018).

4.5.4 Abundance and diversity of auxiliary metabolic genes

In order to understand the role phages might have on the metabolic function of their hosts, function was assigned to proteins using eggNOG (Huerta-Cepas *et al.*, 2018). Out of 210,997 predicted proteins, only 48,819 (23.137%) could be assigned a putative function. The most abundant clusters of orthologous groups (COG) categories (Tatusov *et al.*, 2000) were those associated with viral lifestyle; notably replication, recombination and repair, cell wall/membrane/envelope biogenesis, transcription and nucleotide transport and metabolism (Figure 4.10).

In addition to this, a number of putative AMGs were identified, including putative ARGs, CAZYmes, assimilatory sulfate reduction (ASR) genes, MazG, VapE and Zot (Supplementary Table S4.3). These AMGs were found to be abundant and not constrained to particular set of phages or hosts they infect (Figure 4.3A; Supplementary Table S4.4). For instance, carbohydrate-active enzymes were identified on 91 vOTUs across 77 putative viral genera, with 41 vOTUs predicted to infect bacteria spanning 21 families (Supplementary Table S4.4), and genes involved in the sulphur cycle were identified on 148 vOTUs across 138 putative phage genera, with 42 vOTUs predicted to infect bacteria spanning 19 families (Supplementary Table S4.4).

4.5.5 Abundance of virulence-associated proteins

Genes encoding Zot were identified on 36 vOTUs across 33 putative genera, predicted to infect five different families of bacteria (Supplementary Table S4.4). The bacterial virulence factor VapE which is widespread in the agricultural pathogens *Streptococcus* and *Dichelobacter* was also detected (Billington, Johnston and Rood, 1996; Bloomfield

et al., 1997; Ji *et al.*, 2016). Recently, it has been demonstrated that deletions of prophage encoded *vapE* in *Streptococcus* have decreased growth rate in serum compared to wild type strains (Rezaei Javan *et al.*, 2019). VapE homologues were found on 82 vOTUs (~ 1%) across 65 clusters, including 10 high-quality genomes (Figure 4.3A). Bacterial hosts could be predicted for 17 vOTUs and spanned 10 families of bacteria (Supplementary Table S4.4). One vOTU (ctg217) shared ~ 95% ANI with the prophage Javan630 (accession MK448997) (Rezaei Javan *et al.*, 2019). Genome comparison between ctg217 and Javan630 revealed highly conserved genomes, with insertion of a gene encoding a putative methyltransferase in ctg217 being the largest single difference (Figure 4.5).

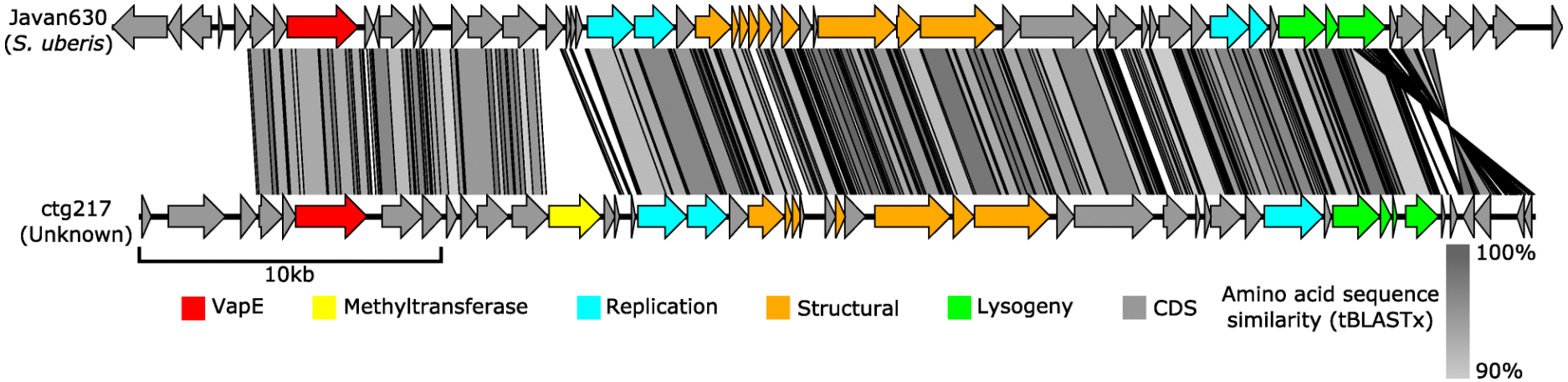


Figure 4.5 Genome comparison of Streptococcus phage Javan630 and ctg217

Genome comparison of Streptococcus phage Javan630 and ctg217 was produced using EasyFig with tBLASTx algorithm and 0.001 E value and length filter 30. The *vapE* gene that is known virulence factor is marked in red. The two genomes had genomes with an ANI > 95% across the genome. The insertion of a gene encoding a methyltransferase within the genome of ctg217 is marked in yellow.

4.5.6 Detection of putative antimicrobial resistance genes

Putative metallo-beta-lactamases (MBLs) were identified on 146 vOTUs across 116 putative genera, with 60 vOTUs predicted to infect bacterial hosts that spanned 23 families (Supplementary Table S4.4). Although low in sequence similarity, structural modelling with Phyre2 (Kelley *et al.*, 2015) found many of these sequences to have the same predicted structure as the novel *bla*$_{\text{PNGM-1}}$ beta-lactamase (100% confidence over 99% coverage) (Park *et al.*, 2018). Furthermore, these sequences contained conserved zinc-binding motifs characteristic of subclass B3 MBLs (Park *et al.*, 2018). Phylogenetic analysis of putative phage MBLs, along with representative bacterial MBLs and a known phage-encoded *bla*$_{\text{HRVM-1}}$ (Moon *et al.*, 2020), showed some clustered with previously characterised bacterial MBLs and others with a characterised phage *bla*$_{\text{HRVM-1}}$ (Figure 4.11). In addition to MBLs, two putative multidrug efflux pumps were identified on two vOTUs predicted to infect two different bacterial genera (Supplementary Table S4.4).

4.5.7 Identification of diversity-generating retroelements

In addition to AMGs, we also identified 202 vOTUs that carry genes encoding a reverse transcriptase. Although dsDNA phages are known to have genes that encode for a reverse transcriptase as part of diversity-generating retroelement (DGR) and the mechanism understood (Liu *et al.*, 2002), they are rarely reported. To determine if the identified genes encoding a reverse transcriptase were part of a DGR, MetaCCST (Yan *et al.*, 2019) was used to identify such elements. Of the 202 vOTUs carrying a reverse transcriptase gene, 82 were predicted to be part of a DGR, which accounts for ~ 1% of vOTUs in the virome. In comparison, we calculated the number of DGRs

that can be identified in publicly available phage genomes (12,354 unique genomes - March 2020) to be 0.178% (22 genomes).

For vOTUS where a complete DGR system (template repeat, variable repeat, reverse transcriptase and target gene) could be identified, the most commonly predicted function of the target gene was a tail fibre. The distribution of DGRs across 74 viral clusters and 15 families of predicted host bacteria (Supplementary Table S4.4) suggest that this is not a feature that is unique to a particular VC of phages or hosts they infect (Figure 4.3A).

DGRs were predicted to occur on four phages that were deemed high-quality complete genomes (Figure 4.6). These phage genomes varied in size from 40.3 to 52.07 kb, with two genomes containing putative integrases (k149_1459596 and k149_1764855), suggesting they are temperate, with the other two likely lytic phages (ctg154 and k149_1404499). Interestingly, phage k149_1459596 could not be detected between 07/06/2017 and 05/09/2017 but was the most abundant vOTU on 10/10/2017, representing over 3% of the viral population at that time. As vConTACT2 (Bin Jang *et al.*, 2019) analysis was unable to classify the phages, phylogenetic analysis was carried out with gene encoding TerL to identify the closest known relatives (Figure 4.12). Phage k149_1459596 closest relative was Vibrio phage Rostov 7 (accession MK575466) and member of the *Myoviridae*, whilst the closest known members of the three others phages are all members of the *Siphoviridae*.

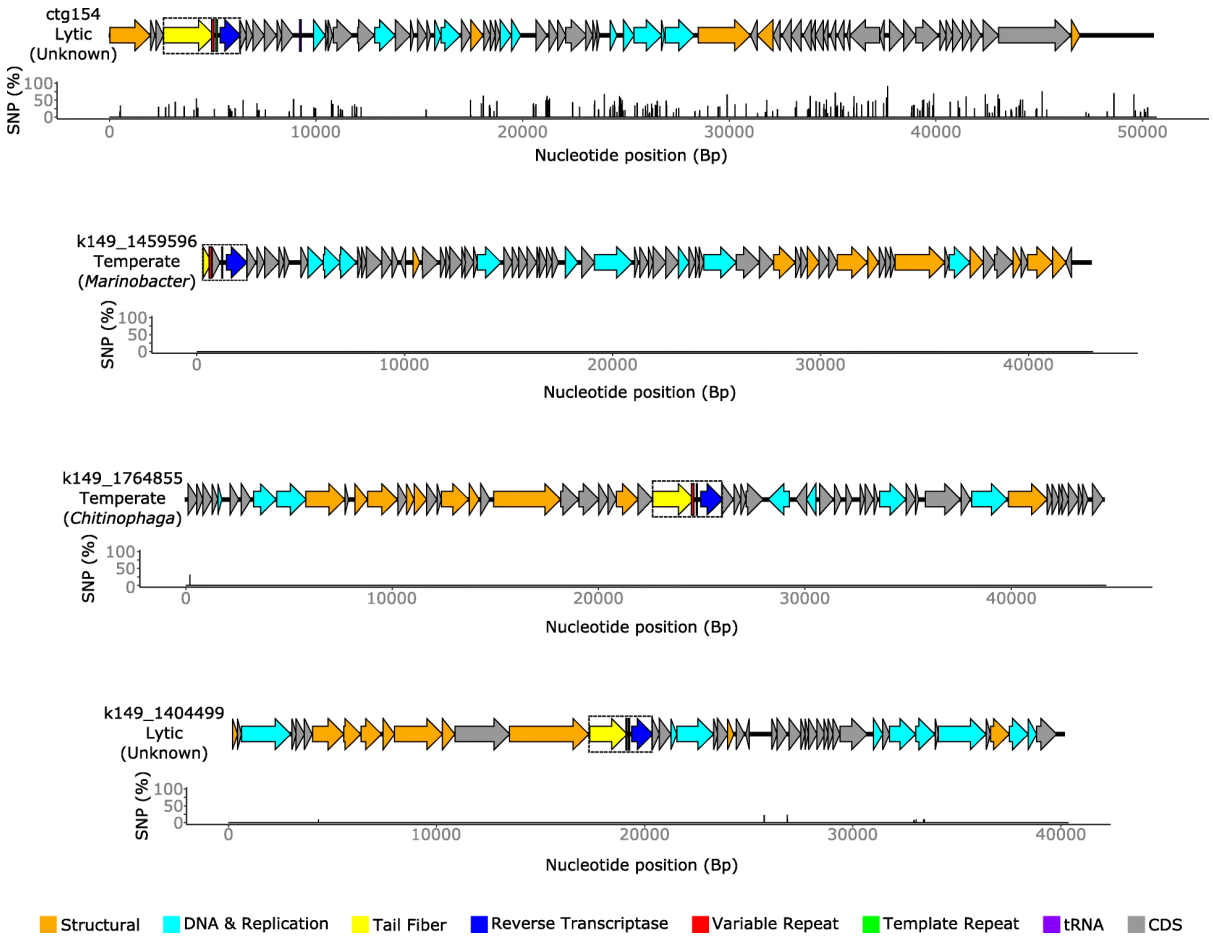


Figure 4.6 Genome maps of complete genomes containing DGRs

The four phages ctg154, k149_1459596, k149_1764855 and k149_1404499 all contain a DGR as highlighted by a dashed box. The percentage of reads that contain SNPs that map to the consensus genome was plotted below.

We hypothesised that the widespread distribution of DGRs would reflect widespread tropism switching in these phages, and that hypervariable DGR target genes could be detected. To investigate this, we examined variants per gene and calculated which genes were under positive selection. For the 69 DGR containing vOTUs in which a target gene could be identified, 22 of these contained a higher proportion of SNP sites in the DGR target gene(s) than the average proportion of SNP sites for non-DGR target genes on that given vOTU. One of which, a predicted phage tail protein (ctg187_00023), was predicted to be under positive selection. Thus, many of the DGR target genes were more variable than other genes on a given vOTU (Figure 4.6).

4.6 Discussion

4.6.1 Assembly comparison

Comparison of assemblies between both short-read and long-read based sequencing methods revealed significant differences in the distribution of viral contigs and the median gene length. As has been found previously, the use of long-reads alone causes problems in gene calling due to higher error rates (Watson and Warr, 2019). We therefore used short-reads to polish the long-read assembly and alleviate these issues (Warwick-Dugdale *et al.*, 2019). In contrast to previous methods that used LASLs combined with ONT MinION sequencing (Warwick-Dugdale *et al.*, 2019), we utilised whole genome amplification followed by size selection for PromethION sequencing.

In using MDA for production of PromethION libraries, a bias in the amplification of ssDNA phage most likely occurred due to well established preference for ssDNA using this method (Roux, Solonenko, *et al.*, 2016). A size selection of fragments was applied prior to PromethION sequencing that would likely remove some of these smaller ssDNA genomes. However, there was a peak in contigs of 4–5 kb length in the PromethION assembly, indicative of ssDNA genomes. Given the known MDA bias, we only utilised Illumina libraries (no MDA amplification) for determining the abundance of contigs and estimates of diversity. Comparison of diversity statistics on Illumina, PromethION and hybrid assemblies suggest Illumina only assemblies may underestimate the diversity within a sample, whereas diversity estimates even on un-corrected PromethION assemblies is closer to that of hybrid assemblies. We also observed a number of smaller genomes that were obtained from Illumina only assemblies and were not present in the PromethION assembly. This likely results as

part of the selection process for high molecular weight DNA (HMW) for PromethION sequencing that would exclude some small phage genomes. Therefore, whilst long-reads improved assembly statistics, the use of long-reads alone may result in exclusion of smaller phage genomes if size selection is included (as we did) and may introduce a bias of increased ssDNA genomes.

To provide the most comprehensive set of viral contigs, we included 230 predicted prophages derived from bacterial metagenomes that could be detected in the free viral fraction but were not assembled from virome reads, thus providing a more comprehensive set of viral contigs.

4.6.2 Virome composition

Comparison of diversity across the period of five months revealed a highly diverse and stable virome across time. Initially, this may be somewhat surprising given the dynamics of the slurry tank, which has constant inflow from animal waste, farm effluent and rainwater, and is emptied leaving only ~ 10% of the tank volume every ~ 6 weeks. We reason that most viruses in the slurry tank will originate from cow faeces, as this is the most dominant input of the tank. Host prediction suggested the virome was dominated by viruses predicted to infect bacteria belonging to *Firmicutes* and *Bacteroidetes*, which are the two most abundant bacterial phyla in the cow rumen and gut (Kim and Wells, 2016; Delgado *et al.*, 2019; Li *et al.*, 2019). To date, there has been limited study into the dairy cow gut virome and its dynamics over time. However, there is a parallel with the human gut virome which is known to be temporally stable despite constant influx and efflux (Reyes *et al.*, 2010; Minot *et al.*, 2013; Garmaeva *et al.*, 2019), and its composition influenced by environmental factors including diet

(Minot *et al.*, 2011; Lim *et al.*, 2015; Moreno-Gallego *et al.*, 2019). Assuming most viruses in the slurry tank are derived from cow faeces, the controlled environment and diet of dairy cattle results in a temporally stable virome.

Our positive selection analyses found the most common genes to be under positive selection were those involved in bacterial attachment and adsorption. We reasoned that these findings, in conjunction with the extreme stability in macro-diversity, fit with the Royal Family model of phage-host dynamics (Breitbart *et al.*, 2018). This model suggests that dominant phages are optimised to their specific ecological niche, and in the event of bacterial resistance to infection, a highly similar phage will fill that niche. Changes in community composition over time would therefore be reflected in fine-scale diversity changes, and macro-diversity would be relatively unchanged (Breitbart *et al.*, 2018). Instead of population crashes, phages may overcome bacterial resistance through positive selection of genes involved in attachment and adsorption, and are potentially accelerating the variation of these genes with DGRs.

4.6.3 Diversity-generating retroelements

DGRs were first discovered in the phage BPP-1 (accession AY029185) where the reverse transcriptase, in combination with terminal repeat, produces an error-prone cDNA that is then stably incorporated into the tail fibre (Liu *et al.*, 2002). This hypervariable region mediates the host switching of BPP-1 across different *Bordetella* species (Liu *et al.*, 2002). Very few DGRs have been found in cultured phage isolates since, with only two DGRs found in two temperate vibriophages (Benler *et al.*, 2018; Wu *et al.*, 2018). We expanded this to 22 phages (0.178%) by searching publicly available phage genomes. Whilst not common in phage genomes, DGRs have been

identified in bacterial genomes, with phage associated genes often localised next to the DGRs (Wu *et al.*, 2018). A recent analysis of ~32,000 prophages was able to identify a further 74 DGRs in what are thought to be active prophages from diverse bacterial phyla (Benler *et al.*, 2018). Within this study, we were able to predict a further 82 DGRs on phage genomes, four of which are thought to be complete. Two of these complete phage genomes are thought to be lytic. In fact, the majority of DGR-containing contigs in this study are thought to be lytic, thus demonstrating that DGRs on phage are far more common than previously found and also observed widely on lytic phages, which has not previously been observed.

Given the prevalence of DGRs, we expected to find evidence of widespread phage tropism switching by occurrence of SNPs in DGR target genes as others have done (Benler *et al.*, 2018). Whilst SNPs could be identified in DGR target genes supporting this, many other areas in the same phage genome contained similar levels of variation. This is likely a result of multiple evolutionary pressures and mechanisms that are exerted on a phage genome, with DGRs only one such mechanism of creating variation.

4.6.4 CrAss-like phages

Currently, crAss-like phages are classified into four subfamilies and ten genera (Guerin *et al.*, 2018), and found in a variety of environments including human waste (Dutilh *et al.*, 2014; Guerin *et al.*, 2018; Shkoporov *et al.*, 2018), primate faeces (Edwards *et al.*, 2019), dog faeces (Cuscó *et al.*, 2019) and termite guts (Yutin *et al.*, 2018). Here, we identified a further 18 crAss-like phages, including a near complete genome that belongs to the proposed genus VI (Guerin *et al.*, 2018). Genus VI is part

of the *Betacrassvirinae* subfamily and currently only includes other crAss-like phages occurring within the human gut, including IAS virus that is highly abundant in HIV-1 infected individuals (Oude Munnink *et al.*, 2014). Thus, we have expanded the environments genus VI crAss-like phages are found in to include non-human hosts. The exact source of these phages is unknown due to the number of possible inputs of the slurry tank. However, the most likely reservoir is from cows, as this is the most abundant input. Unlike its human counterpart IAS virus, which can account for 90% of viral DNA in human faeces (Dutilh *et al.*, 2014), crAss-like phages in the slurry tank were only found at low levels (~ 0.065%).

Phylogenetic analysis clearly demonstrated that human and slurry tank crAss-like phages share a common ancestor, with genetic exchange between them. The direction and route of this exchange is unclear. It may be linked to modern practices of using slurry on arable land used to produce product consumed by humans. Alternatively, it may be transferred from humans to cows via the use of biosolids derived from human waste that are applied to crops that serve as animal feed (Biosolids Assurance Scheme, 2020).

4.6.5 Auxiliary metabolic genes

We identified a vast array of diverse and abundant AMGs in dairy farm slurry including putative ARGs, CAZYmes, ASR genes, MazG, VapE and Zot. Whilst these have all been identified before in viromes from different environments (Romero *et al.*, 2009; Liu *et al.*, 2016; Enault *et al.*, 2017; Castillo *et al.*, 2018; Debroas and Siguret, 2019; Jin *et al.*, 2019; Rezaei Javan *et al.*, 2019; Rihtman *et al.*, 2019; Gao *et al.*, 2020), this is the first time they have been identified in slurry. The presence of different AMGs is

likely a reflection of the unique composition of slurry that has a very high water content combined with organic matter. CAZYmes were detected, which have previously been identified in viromes from mangrove soils and the cow rumen where they are thought to participate in the decomposition of organic carbon and boost host energy production during phage infection (Anderson, Sullivan and Fernando, 2017; Jin *et al.*, 2019). Given the high cellulose and hemicellulose content of slurry (Chen *et al.*, 2003), they likely act in a similar manner within slurry to boost energy for phage replication. As well as involvement in the cycling of carbon, it also appears phage derived genes are involved in sulphur cycling within slurry. Sulfate-reducing bacteria (SRB) are active in animal wastes (Cook *et al.*, 2008; St-Pierre and Wright, 2017), and sulfate may therefore be limiting within the tank. The ASR pathway makes sulphur available for incorporation into newly synthesised molecules, such as L-cysteine and L-methionine (Rückert, 2016), so the presence of phage encoded ASR genes on both lytic and temperate phages may overcome a metabolic bottleneck in amino acid synthesis. Alternatively, the newly synthesised ASR pathway products may be degraded for energy via the TCA cycle (Howard-Varona *et al.*, 2020).

The AMG *mazG*, that is widespread within marine phages, in particular cyanophages (Millard *et al.*, 2009; Sullivan *et al.*, 2010; Rihtman *et al.*, 2019), was also found to be abundant. The cyanophage MazG protein was originally hypothesised as a modulator of the host stringent response by altering intracellular levels of (p)ppGpp (Clokie and Mann, 2006; Clokie, Millard and Mann, 2010). However, more recent work found this not to be the case (Rihtman *et al.*, 2019). The identification in a slurry tank suggests this gene is not limited to marine environments and is widespread in different phage types, although its precise role remains to be elucidated.

4.6.6 Antibiotic resistance genes

There is ongoing debate as to the importance of phages in the transfer of ARGs (Enault *et al.*, 2017; Debroas and Siguret, 2019). We identified ARGs on ~ 2% of vOTUs; accounting for ~ 0.082% of total predicted phage genes from assembled viral contigs. The predicted ARGs were dominated by putative MBLs that contain core motifs and structural similarity with the known bacterial and phage MBLs $bla_{\text{PNGM-1}}$ (Park *et al.*, 2018) and $bla_{\text{HRVM-1}}$ (Moon *et al.*, 2020) respectively. Thus, are likely functionally active, although this remains to be proven. Our estimate of the abundance of ARGs in slurry is lower than earlier reports from other environments that predict an upper estimate of ~ 0.45% of genes in viromes are ARGs (Balcazar, 2014; Lekunberri *et al.*, 2017). However, some of these studies have used unassembled reads to estimate abundance (Balcazar, 2014; Lekunberri *et al.*, 2017), whereas we only counted ARGs on contigs that had passed stringent filtering. Our prediction of ~ 0.082% is similar to more recent estimates of 0.001% to 0.1% in viromes from six different environments that also used assembled viromes (Debroas and Siguret, 2019), suggesting that phages might be an important reservoir of ARGs in slurry.

4.6.7 Virulence-associated proteins

The virulence genes *zot* and *vapE* were found to abundant and carried by several vOTUs that were predicted to infect a range of bacterial hosts. The role of *zot* has been well studied in *Vibrio cholerae* and has previously been reported in a range of *Vibrio* and *Campylobacter* prophages (Koonin, 1992; Schmidt, Kelly and van der Walle, 2007; Liu *et al.*, 2016; Castillo *et al.*, 2018). Here, we found *zot* homologues in phages with predicted hosts other than *Vibrio* and *Camplyobacter*, further expanding the diversity of phages that carry these genes.

A similar observation was found for the virulence factor *vapE*, which has previously been found in several agricultural pathogens including *Streptococcus* and *Dichelobacter* (Billington, Johnston and Rood, 1996; Bloomfield *et al.*, 1997; Ji *et al.*, 2016). VapE encoded on prophage elements is known to enhance the virulence of *Streptococcus* and is widespread on *Streptococcus* prophages (Rezaei Javan *et al.*, 2019). Whilst the role of *vapE* in virulence has been established, previous work did not demonstrate the mobility of these prophage-like elements. Here, we identified a high quality near-complete phage genome (ctg217) which was remarkably similar to the *vapE* encoding prophage Javan630. Phage Javan630 was originally identified as a prophage within a mastitis causing strain of *Streptococcus uberis* isolated from a dairy cow some 15 years earlier on a dairy farm ~ 100 mi away (Rezaei Javan *et al.*, 2019). The identification of ctg217 in the free viral fraction indicates that a close relative of phage Javan630 is an active prophage. Along with the numerous other phages encoding *vapE* found in the free virome, it suggests that phage is active in mediating the transfer of *vapE*. The horizontal transfer of *vapE* is of particular concern in the dairy environment where mastitis causing pathogens *Strep. uberis*, *Strep. agalactiae* and *Strep. dysgalactiaea* are found (Keefe, 1997; Whist, Østerås and Sølverød, 2007; Zadoks *et al.*, 2011). Any increase in virulence of these pathogens is detrimental to the dairy industry as it affects both animal welfare and economic viability (Ruegg and Petersson-Wolfe, 2018). *Streptococcus* infections result in mastitic milk, which cannot be sold and is often disposed of into slurry tanks. The continual detection of phages containing *vapE* in slurry suggests a likely continual input, given the regular emptying of the tank. The exact source of phages containing *vapE* cannot be ascertained but is likely cow faeces or mastitic milk. It remains to be determined if the

use of slurry as an organic fertiliser contributes to the spread of phage encoded virulence factors and toxins. However, their abundance and presence suggests it is worthy of further investigation.

4.7 Conclusions

We have demonstrated that using a hybrid approach produces a more complete virome assembly than using short or long-reads alone. Whilst short-reads may underestimate the total viral diversity of a given environment, estimates from long-reads alone were far closer to the hybrid values than short-reads. The use of low input amplified genomic DNA allows the technique to be applied to previously sequenced metagenomes without need for further DNA extraction. We provide a comprehensive analysis of the slurry virome, demonstrating that the virome contains a diverse and stable viral community dominated by lytic viruses of novel genera. Functional annotation revealed a diverse and abundant range of AMGs including virulence factors, toxins and antibiotic resistance genes, suggesting that phages may play a significant role in mediating the transfer of these genes and augmenting both the virulence and antibiotic resistance of their hosts.

4.8 Supplementary Figures

Below are supplementary figures from the publication ‘Hybrid assembly of an agricultural slurry virome reveals a diverse and stable community with the potential to alter the metabolism and virulence of veterinary pathogens. Cook, R. et al (2021) Microbiome.’ https://doi.org/10.1186/s40168-021-01010-3.

Supplementary figure 1

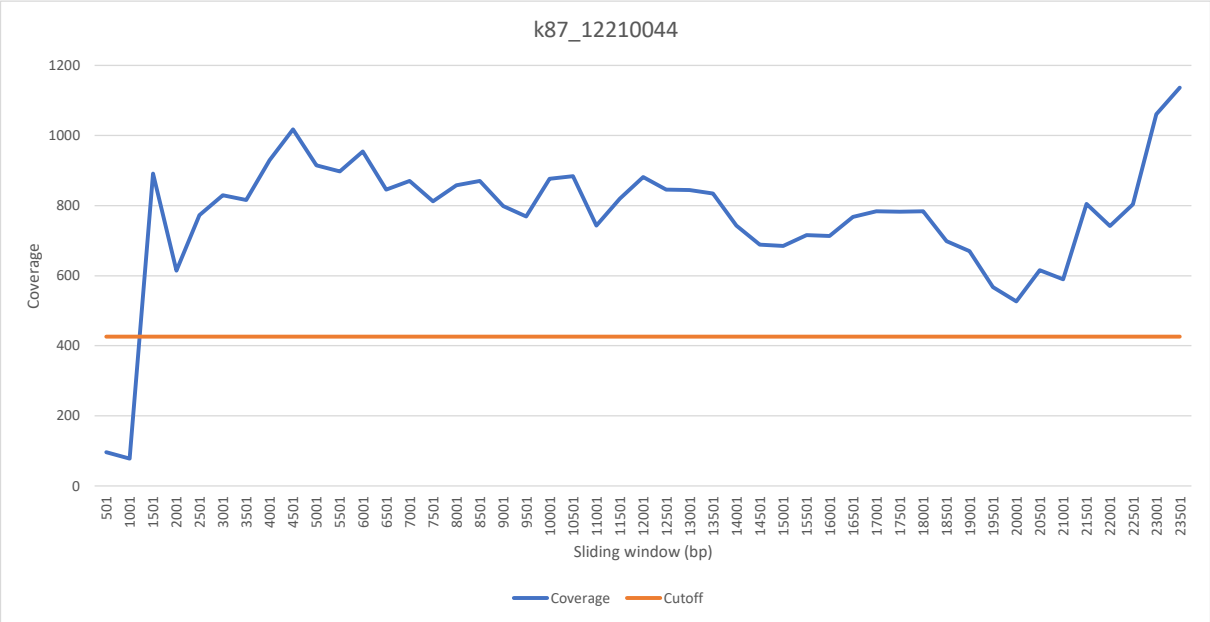


Figure 4.7 Representative figure for the identification of prophage ends

Reads were mapped against vOTU k87_12210044 at 95 % identity threshold, the median coverage was calculated for 500 bp windows with the cutoff value calculated as median coverage minus (2 * standard deviations of median coverage) and plotted in orange. In this particular example, only one end was predicted.

Supplementary figure 2

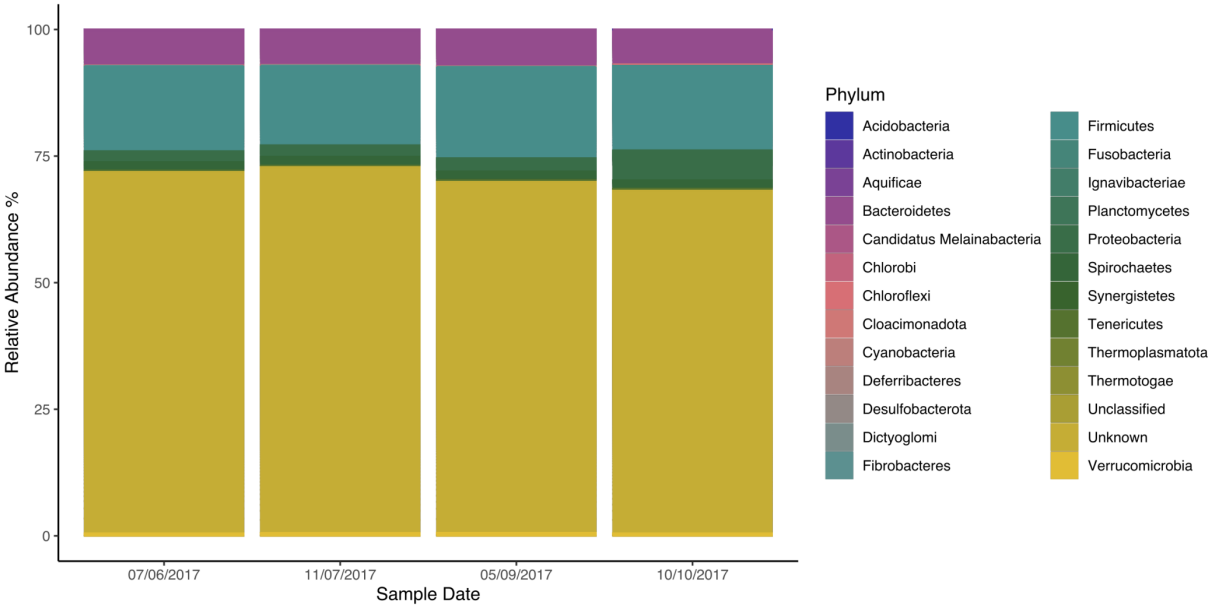
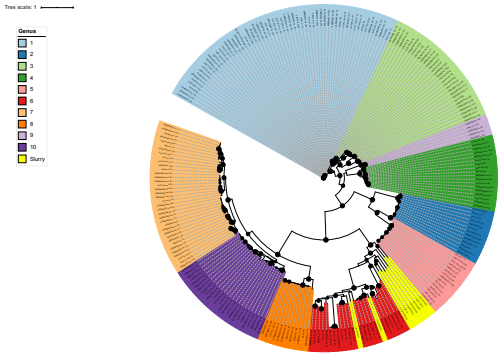


Figure 4.8 Predicted hosts of viral contigs at the phylum level

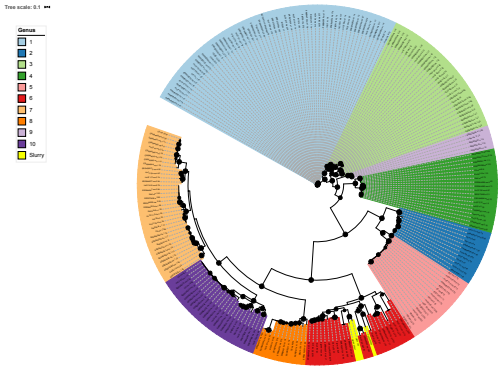
Predicted hosts were obtained using WiSH. The relative abundance of phages predicted to infect different hosts was calculated by stringent mapping of reads to each viral contig as normalising for contig length and sequencing depth as described in materials and methods.

Supplementary figure 3

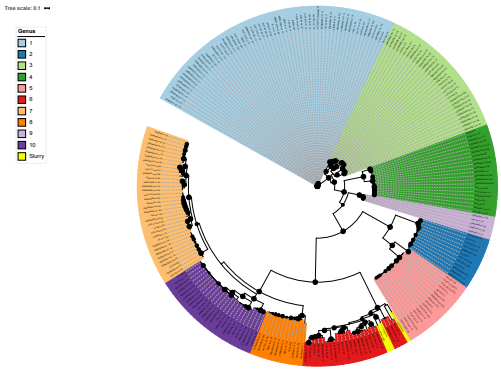
Primase



Terminase



Portal



Major Capsid Protein

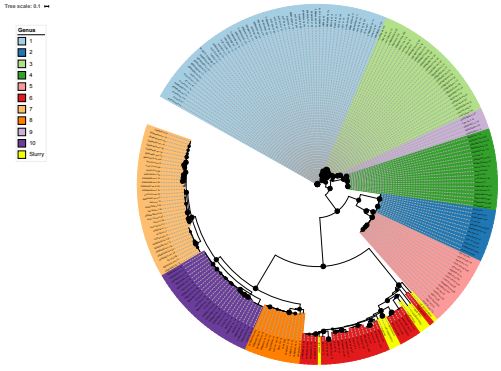


Figure 4.9 Phylogeny of crAss-like vOTUs based upon the method of Guerin et al.

Phylogeny of four genes that encode a primase, terminase, portal protein and major capsid protein. The analysis followed the same method as described by Guerin et al., with the ten major clades as previously defined marked. Bootstrap values >70% are marked by a circle.

Supplementary figure 4

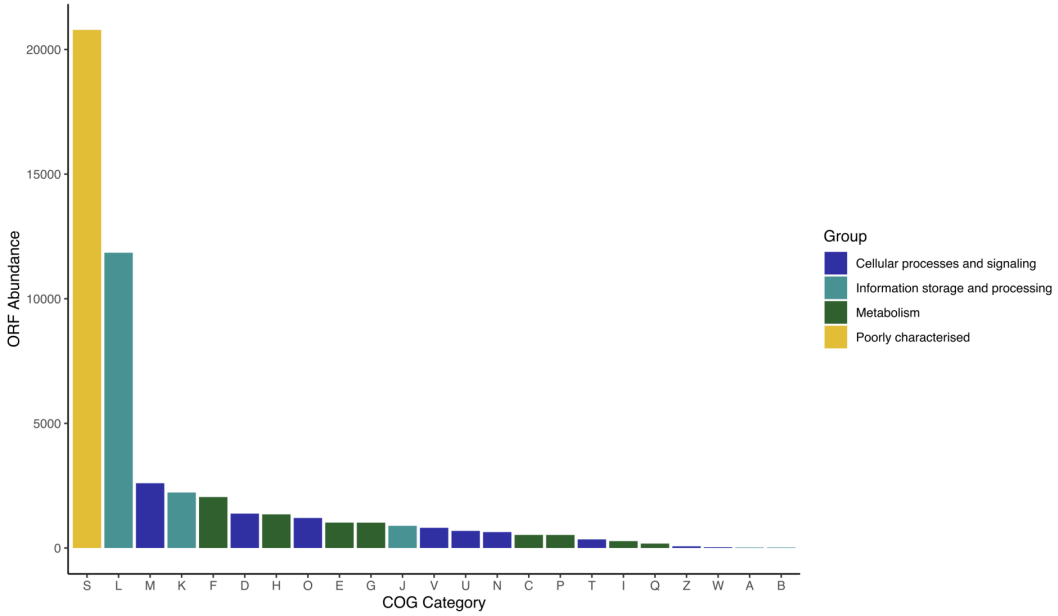


Figure 4.10 Functional classification of viral proteins into COG categories by eggNOG mapping

Abundance of COG categories for vOTU predicted proteins.

Supplementary figure 5

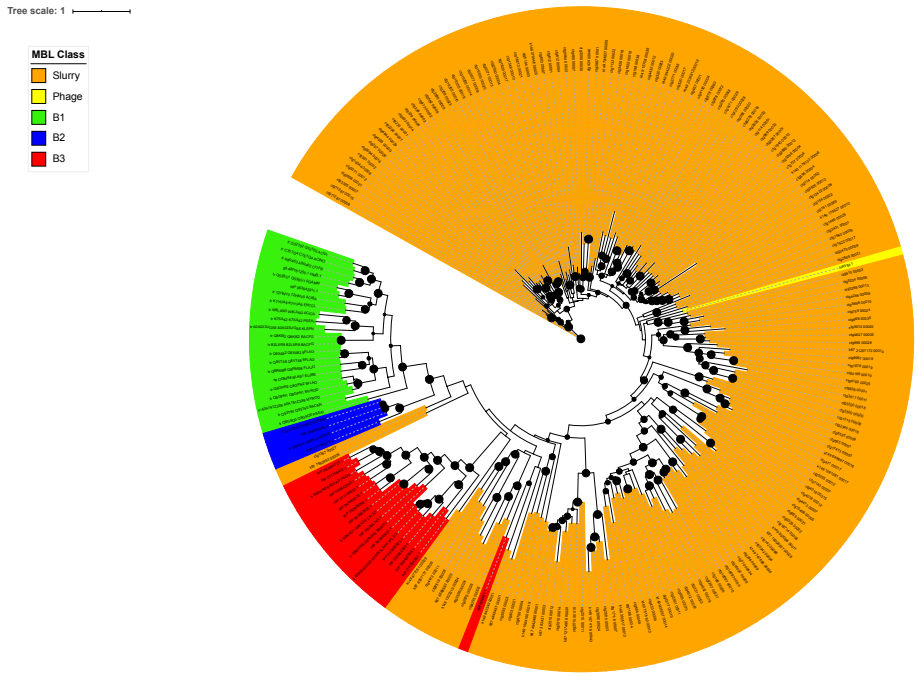


Figure 4.11 Phylogeny of putative metallo-β-lactamases

The phylogeny was built on the alignment of the amino acid sequences that were aligned by MAFFT. A WAG model of evolution was used in IQ-TREE with 1000 boostraps. Putative MBLs identified in the slurry tank are marked in orange, along with a previously experimentally validated phage-encoded MBL (yellow). Bacterial subclass B1 (green), B2 (blue), B3 (red) MBLs are also marked. Bootstrap values >70% are marked by a circle. Tree is rooted at the mid-point.

Supplementary figure 6

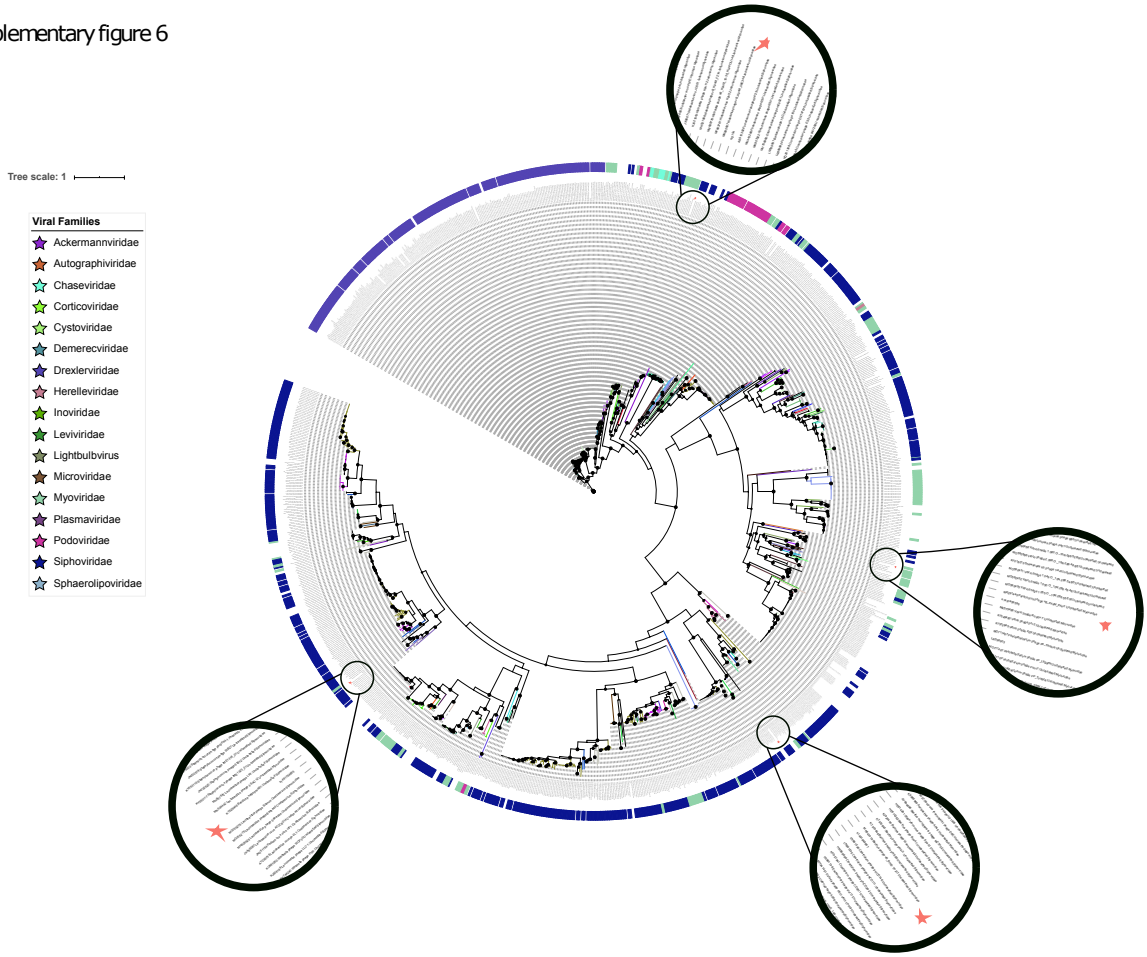


Figure 4.12 Phylogeny of phage genomes that contain a complete DGR

Phylogeny was constructed from the amino acid sequence of TerL protein that were aligned in mafft and phylogeny constructed with IQTREE with a WAG model of evolution and 1000 bootstraps. Bootstrap values >70% are marked by a circle. Different viral families are differentiated by the coloured ring around the outside of the tree. Tree is rooted at the mid-point.

Chapter 5 Determining the Effect of Antimicrobials on Modelled Slurry Tank Viromes

5.1 Introduction

Antimicrobial resistance is a growing global concern. The widespread use of antimicrobials in the rearing of livestock has been implicated in the emergence of drug-resistant infections in humans and animals (Aarestrup *et al.*, 2000; O'Neill, 2015; Van Boeckel *et al.*, 2017).

In the UK, dairy cattle are routinely treated with antibiotics for common illnesses including mastitis and respiratory illnesses (Oliver, Murinda and Jayarao, 2011). Furthermore, lameness—the costliest disease to UK dairy cattle (CHAWG, 2020)—is typically prevented by treatment with footbaths that contain antimicrobial metals (e.g., copper and zinc) and/or other chemicals (e.g., formalin and glutaraldehyde) that are known to co-select for AMR (Pal *et al.*, 2015; Griffiths, White and Oikonomou, 2018; Davies and Wales, 2019). Therefore, dairy cattle slurries may contain selective and co-selective pressures for the transmission of AMR.

Phages are known to encode a plethora of diverse genes that confer an advantage to the fitness of their host, with the potential to augment nutrient acquisition and metabolism (Yooseph *et al.*, 2007; Dinsdale *et al.*, 2008; Sharon *et al.*, 2011; Hurwitz, Hallam and Sullivan, 2013; Anantharaman *et al.*, 2014; Zhang, Wei and Cai, 2014; Hurwitz, Brum and Sullivan, 2015; Hurwitz and U'Ren, 2016; Roux, Brum, *et al.*, 2016; York, 2017; Monier *et al.*, 2017; Jin *et al.*, 2019), as well as virulence (Freeman, 1951; Eklund *et al.*, 1974; Waldor and Mekalanos, 1996; Wagner *et al.*, 2002; Fortier and Sekulovic, 2013; Khalil *et al.*, 2016). However, the carriage of ARGs in phage genomes is seemingly a rare event (Enault *et al.*, 2017; Cook, Brown, *et al.*, 2021). The paucity of reported phage-encoded ARGs may be a true reflection of their rarity,

however, phages and viromes are commonly isolated from environments where the concentration of antibiotics that may not be high enough to have a selective pressure for ARG carriage.

Prior to commencement of this PhD project, a study was designed to determine the effect of agricultural antimicrobials, including foot-wash, on the microbial ecology of agricultural slurry. Miniaturised versions of the slurry tank described in Chapter 4 were devised to assess the impact of storing slurry, and the impacts of particular antimicrobial additions, as described in Baker *et al.*, (2022). The twelve “mini-tanks” were buckets containing 10 L of slurry taken from the tank described in Chapter 4, stored at ambient temperature on the farm for a duration of seven weeks. The mini-tanks were protected from rain and direct sunlight, and unlike the main slurry tank, the mini-tanks did not receive further influent after initial setup. Six different conditions were tested in duplicate (Table 5.1), with samples being taken at the point of setup (T=0) and seven weeks later (T=7). Viral fractions were taken from the samples and sequenced. Study design, sample collection, and sequencing were performed as part of the wider EVAL-FARMS consortium, prior to commencement of this PhD project. For the work described in this chapter, I started with the existing raw virome datasets.

The aim of this work was to determine the impact of agricultural antimicrobial compounds on the diversity and community composition of bacteriophages within agricultural slurry, as well as the phage carriage of ARGs. Therefore, the objectives were to:

1. To describe the viromes for model slurry mini-tanks

2. To determine the selective effect of the agricultural antimicrobials footwash and cefquinome on the composition and structure of viral communities in agricultural slurry
3. To determine if the exposure of agricultural antimicrobials increases the frequency of phage-encoded ARGs

5.2 Materials and Methods

Study design, sample collection, and sequencing were performed as part of the wider EVAL-FARMS consortium, prior to commencement of this PhD project (Baker *et al.*, 2022). In brief, twelve mock slurry tanks containing 10L samples of slurry from the surface of the main slurry tank were positioned on the farm for a seven-week period at ambient temperature (mean 24 h temperature in liquid ranged between 7° to 17°) and protected from rain and direct sunlight. Six different conditions were tested in duplicate (all amounts per litre): control; + SSD (SSD being 0.2 mL of slurry solids homogenised by stomacher, including 67 CFU of CTX-resistant *E. coli*); + SSD + 3 µg cefquinome weekly addition; + SSD + 40 µg cefalexin weekly addition; + SSD + 16.8 g of footbath mix (Cu + Zn); + SSD + footbath + cefquinome). Mini-tanks were sampled four times as part of the main study (0, 2, 4 and 7 weeks after initial filling), and twice for virome sequencing (0 and 7 weeks after initial filling). The experimental conditions and timepoints of the samples used for virome sequencing are shown in Table 5.1.

Table 5.1 Minitank conditions and timepoints

Sample	Tank	Timepoint (Weeks)	Condition
MiniT0Phi49	1	0	Control
MiniT7Phi61		7	
MiniT0Phi55	7	0	
MiniT7Phi67		7	
MiniT0Phi50	2	0	SSD (0.2 mL of slurry solids homogenised by stomacher, including 67 CFU of CTX-resistant *E. coli*)
MiniT7Phi62		7	
MiniT0Phi56	8	0	
MiniT7Phi68		7	
MiniT0Phi51	3	0	SSD + 16.8 g of footbath mix (Cu + Zn)
MiniT7Phi63		7	
MiniT0Phi57	9	0	
MiniT7Phi69		7	
MiniT0Phi52	4	0	SSD + 3 µg cefquinome weekly addition
MiniT7Phi64		7	
MiniT0Phi58	10	0	
MiniT7Phi70		7	
MiniT0Phi53	5	0	SSD + footbath + cefquinome
MiniT7Phi65		7	
MiniT0Phi59	11	0	
MiniT7Phi71		7	
MiniT0Phi54	6	0	SSD + 40 µg cefalexin weekly addition
MiniT7Phi66		7	
MiniT0Phi60	12	0	
MiniT7Phi72		7	

5.2.1 Virome Preparation, Sequencing and Assembly

The preparation of viromes and sequencing (Section 4.4.1), quality control and assembly (Section 4.4.2), and filtering of vOTUs (Section 4.4.3) was the same as described in Chapter 4. After filtering, the mini-tank vOTUs were de-replicated alongside the main-tank vOTUs at 95% average nucleotide identity (ANI) over 80% genome length using ClusterGenomes v5.1 (*GitHub - simroux/ClusterGenomes: Archive for ClusterGenomes scripts*, no date) to produce a combined set of slurry vOTUs. The de-replicated vOTUs were processed using CheckV v0.9.0 (Nayfach *et al.*, 2020), and those with the ”no viral genes” warning, < 3 total genes, or ≥ 25% “host” genes (and not identified as a prophage) were excluded. For those identified as prophages, the CheckV trimmed versions were used in downstream analyses (Nayfach *et al.*, 2020).

The detection of known phages, and functional annotation, lifestyle prediction, and taxonomic analysis of the new vOTUs were performed as described earlier (Sections 4.4.7, 4.4.9, and 4.4.10). Host prediction was performed using iPHoP v0.9beta (Roux *et al.*, 2022); a pipeline that combines RaFAH (Coutinho *et al.*, 2021), WIsH (Galiez *et al.*, 2017), oligonucleotide frequencies (Ahlgren *et al.*, 2017), PHP (Lu *et al.*, 2021), and BLAST (Altschul *et al.*, 1990).

5.2.2 Population Dynamics

Reads from each sample were separately mapped to the vOTUs using Bowtie 2 v2.3.4.3 with --non-deterministic --maxins 2000 (Langmead and Salzberg, 2012), as described in the MetaPop paper (Gregory *et al.*, 2022). MetaPop was performed with --genome_detection_cutoff 75 --no_viz (Roux *et al.*, 2017; Gregory *et al.*, 2022). To

allow previously predicted genes to be used as input for MetaPop, they were modified with an in-house script (Supplementary File 1). The main-tank samples described in Chapter 4 were included in this analysis as a point of orientation. Pairwise comparisons of beta-diversity between groups were performed by PERMANOVA with 1,000 permutations using adonis as part of Vegan (Oksanen *et al.*, 2020), and p-values were adjusted for multiple comparisons using the Benjamini-Hochberg correction (Benjamini and Hochberg, 1995). Pairwise comparison of means (for Shannon's index, observed vOTUs, microdiversity, and abundance of temperate and novel genera) were performed using the T-test with all groups compared to the T=0 samples.

5.3 Results

Illumina sequencing of 24 viromes produced from agricultural slurry from 12 “mini-tanks” over two sampling points (T=0 at initial setup, and T=7 seven weeks later) yielded 419.2 Gb of sequence data. Individual viromes ranged from 9.6 – 29.7 Gb with a mean of 17.5 (± 5.6 standard deviation) (Figure 5.1B). ViromeQC enrichment scores ranged from 3.1 – 19.1 with a mean of 11.4 (± 3.4 standard deviation) (Figure 5.1A).

Co-assembly of mini-tank viromes, followed by viral filtering, and de-replication with the slurry main-tank vOTUs (Chapter 4) resulted in 12,566 vOTUs with mean and median lengths of 18,107 and 13,962 bp respectively (Figure 5.1C). Prediction of vOTU completeness using CheckV estimated 107 vOTUs to represent complete genomes, with a further 372 estimated high-quality (≥ 90% complete; Figure 5.1D).

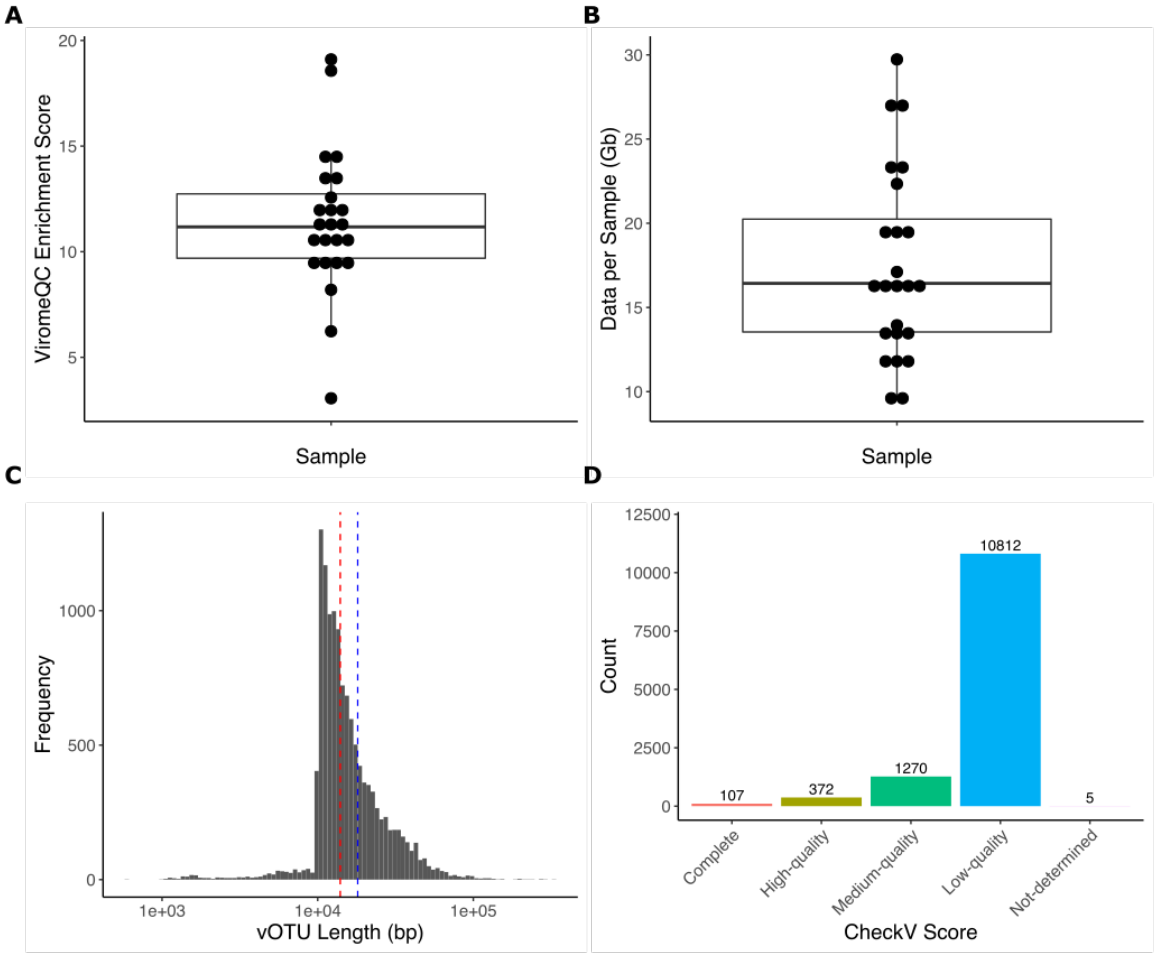


Figure 5.1 Mini-tanks data summary

Summary of data obtained from the mini-tank datasets, showing (**A**) ViromeQC enrichment scores, (**B**) amount of sequence data generated per sample in gigabases, (**C**) distribution of vOTU lengths with mean (red dashed line) and median (blue dashed line) values shown, and (**D**) CheckV quality estimates for filtered vOTUs.

5.3.1 Effect of footwash and cefquinome on beta-diversity

Comparison of beta-diversity using Bray-Curtis dissimilarity (i.e., the distance between communities) demonstrated that there was little difference between the viral communities at the start of the experiment (T=0), although there was one obvious outlier (Effect of footwash on viral community composition). The T=0 samples were most similar to the viromes taken from the main slurry tank described in Chapter 4 (Effect of footwash on viral community composition). However, the viral communities were varied at the end of the seven-week experiment (T=7), with many T=7 samples were substantially different to those at T=0, although others remained similar (Effect of footwash on viral community compositionA).

The original experimental conditions (Table 5.1) were performed in duplicate. To increase the statistical power of the experiment, I grouped conditions that shared an addition. This led to two conditions being investigated: the addition of footwash (16.8 g of footbath mix (Cu + Zn) added at setup) and the addition of cefquinome (3 µg weekly addition). This increased the number of samples in each condition from two to four.

The T=7 samples that had received footwash remained similar to the T=0 and main-tank samples (Effect of footwash on viral community compositionB). PERMANOVA analysis showed the difference between T=7 samples with and without footwash was significant ($p = 0.04$, adjusted using BH for multiple comparisons). Suggesting that the inclusion/exclusion of footwash has an influence of the viral community composition. Conversely, the addition of cefquinome seemingly had no effect on the composition of the virome (Figure 5.3C).

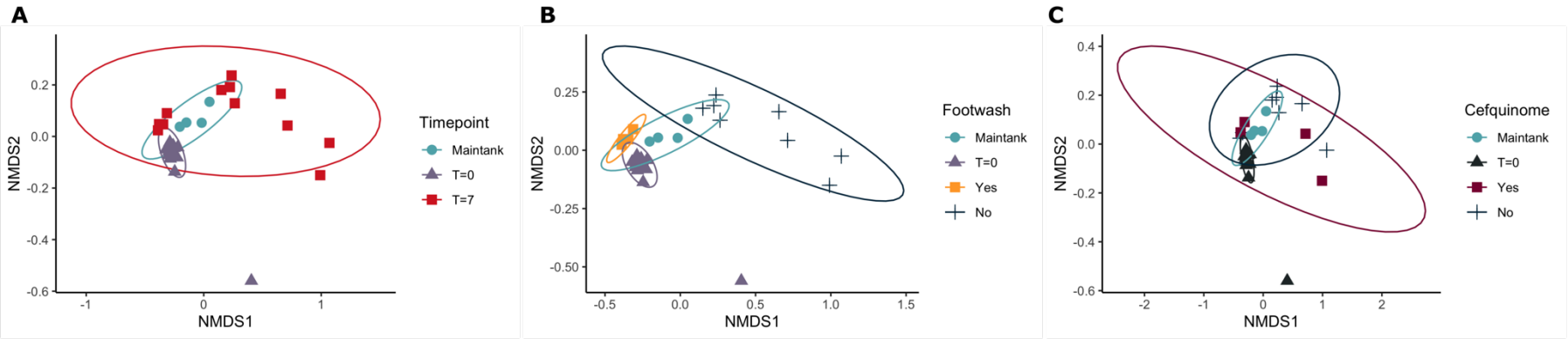


Figure 5.2 The effect of footwash and cefquinome on beta-diversity

Bray-cutis dissimilarity showing beta-diversity of mini-tank samples (**A**) at the beginning and end of the experiment, (**B**) with and without footwash, and (**C**) with and without cefquinome. All plots include the main-tank samples as a point of reference, and ellipses show normal distribution of data.

5.3.2 Effect of footwash on viral community composition

The samples that received footwash maintained a high species richness (observed vOTUs) and alpha diversity (Shannon's index) that was comparable to the T=0 samples (Figure 5.3). Conversely, those that did not receive footwash observed a marked decrease in species richness and alpha diversity (Figure 5.3). Whilst only the T=7 samples without footwash had a decrease in macro-diversity, all T=7 samples obtained a significant decrease in micro-diversity (nucleotide diversity π) regardless of inclusion/exclusion of footwash (Figure 5.3).

The exclusion of footwash led to a significant increase in the proportion of novel and lytic phages (Figure 5.3). Although the proportion of novel genera seemed to decrease at T=7 with the inclusion of footwash, this difference was not significant (Figure 5.3). Furthermore, the inclusion of footwash led to no changes in the proportion of putatively lytic genera over the course of the experiment (Figure 5.3).

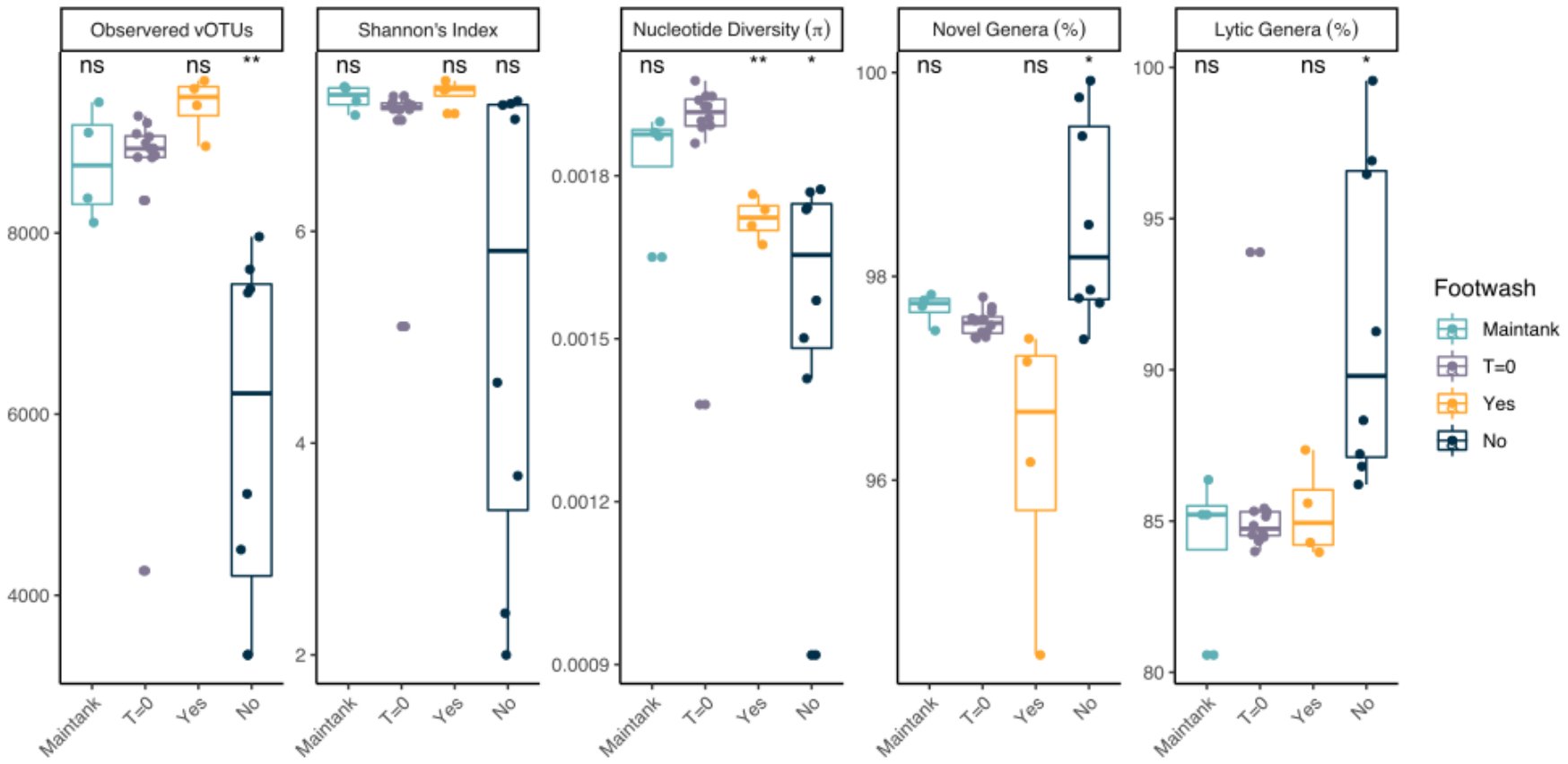


Figure 5.3 Influence of footwash on virome composition

Boxplots showing the species richness, Shannon's index, nucleotide diversity, and proportions of novel and lytic genera for T=7 samples with and without the addition of footwash. Significance was tested using the T-test with the T=0 samples as a reference group. P-values were adjusted for multiple comparisons (*0.05, **0.01).

5.4 Discussion

Building upon previous work that characterised the virome of agricultural slurry (Chapter 4), I analysed previously generated viromes from agricultural slurry of mock slurry “mini-tanks” to determine the effect of footwash and cefquinome on the composition of the virome. Whilst there is limited study into the effect of agricultural antimicrobials on the bacterial fraction of slurry (Baker *et al.*, 2022), there is no such work on the viral fraction.

Previous work demonstrated that the agricultural slurry virome was stable over time, despite constant influx and efflux (4.5.2 (Cook, Hooton, *et al.*, 2021)). This work has shown that footwash may be an important component for the maintenance of the viral community, suggesting that the constant addition of footwash to the slurry tank via farm waste may be a factor in the stability of the slurry virome described earlier (4.5.2 (Cook, Hooton, *et al.*, 2021)). The footwash mix, used to control lameness, contains copper and zinc. As these metals are antimicrobial, it is possible that the footwash selects for a particular bacterial community composition which is mirrored in the viral fraction. It may be that footwash prevents the growth of metal-sensitive bacteria that may otherwise proliferate, hence the divergence of mini-tank viromes that did not contain footwash. The same effect was not observed for cefquinome. Conversely, it may be that the metal ions have a more direct effect on the viral community. Heavy metals such as copper are known to induce prophages (Lee *et al.*, 2006; Guo *et al.*, 2017), and chromium-contaminated soil viromes have been found to be enriched for temperate phages (Huang *et al.*, 2021). Therefore, the higher proportion of temperate phages in mini-tanks with footwash than those without footwash may be due to higher levels of prophage induction caused by the presence of copper ions.

Whilst there are preliminary results, the work described in this chapter is largely incomplete. Metal ions, such as those used in footwash, are known to co-select for the carriage of ARGs on mobile genetic elements such as plasmids and transposons (Pal *et al.*, 2015; Griffiths, White and Oikonomou, 2018; Davies and Wales, 2019). However, I did not investigate whether the use of footwash or cefquinome impacted the carriage of phage-encoded ARGs. As phages commonly carry niche-specific genes that confer fitness advantages to their hosts, it is possible that the continued selective pressure of agricultural antimicrobials will select for ARGs to be encoded on phages. Furthermore, metal resistance genes are known to be prevalent in farmed environments that contain high levels of heavy metal ions (Li *et al.*, 2022). The carriage of metal resistance genes on vOTUs could be determined using MEGARes and BacMet (Pal *et al.*, 2014; Doster *et al.*, 2020). It may also be possible to use a read-based approach for the quantification of ARGs in the viral fractions to assess whether ARGs are packaged into virions more commonly in the presence of agricultural antimicrobials (i.e., generalised transduction), however, read-based approaches should always be used with caution as it is impossible to differentiate ARGs found within viral particles from contaminating bacterial DNA (Enault *et al.*, 2017). Beyond ARGs, it would also be of note to determine the abundance and distribution of other AMGs within the mini-tank viromes, as phages are known to have diverse impacts on the metabolism of their hosts (discussed in Section 1.7).

It is likely that the differences observed in the viral fraction with and without footwash are mirroring changes in the bacterial fraction. It would therefore be of interest to determine the changes to bacterial taxa and see how this corresponds to what was observed in the viral fraction. As part of the wider EVAL-FARMS research consortium,

there are high quality bacterial metagenomes derived from the same samples as the viral fractions. The metagenomes are briefly described in Baker *et al.*, (2022), although no in-depth analysis of bacterial community structure is performed. Additionally, I have predicted the hosts of the vOTUs using iPHoP (Roux *et al.*, 2022). Future work could determine if the abundance of vOTUs predicted to infect bacterial taxa correlates with the abundance of said bacterial taxa, as determined from the metagenomes using tools such as Kraken 2 (Wood, Lu and Langmead, 2019). Additional analyses using the bacterial fraction could predict prophages (as done in Section 4.4.4). Furthermore, the use of viral and bacterial fractions together could elucidate whether prophages are being induced via tools such as PropagAtE (Kieft and Anantharaman, 2022). These future analyses could shed light on the apparent shift from temperate to lytic phages in the absence of footwash.

Chapter 6 Characterising the Dairy Cow Gut Virome Across Life Stages

6.1 Introduction

The farming of cattle constitutes 50% of global Livestock Standard Units (FAO, 2020), with an estimated 265 million dairy cows globally (AHDB, 2020). Consequently, the dairy industry has significant impacts on the health and welfare of enormous numbers of cattle, global food production, agricultural economics, and the wider environment (Peterson and Mitloehner, 2021).

The life of a UK dairy cow can be split into several distinct stages, during which the cows are housed separately from other groups and given a diet specific for their requirements at the time. The infant cow is referred to as a calf. Although timings will differ, dairy calves are fed a liquid diet of either milk or milk replacer (i.e., formulated milk) for the first few months of life (Khan *et al.*, 2016). Although there is no defined cut-off between calves and heifers, a sexually mature female dairy cow that has not yet calved is commonly referred to as a heifer (Sakaguchi, 2011).

As mammals, dairy cattle need to calve to produce milk (Sakaguchi, 2011). After their first calving, the cow will enter the milking herd. For optimum dairy production, ideally a dairy cow will calve every 12 months, with the average UK dairy cow yielding over 8,000 litres of milk per year (AHDB, 2022b). To optimise milk production after calving, the lactation cycle and diets of the cows is tightly monitored and controlled. A dairy cow is only able to produce milk for ~ten months of the year, and the final two months of pregnancy are commonly referred to as the “drying off” period (AHDB, no date a). During the drying off period, the cows are housed separately to the milking herd, and given a diet designed to optimise milk production post-calving (AHDB, 2022a). A summary of the lactation cycle is shown in Figure 6.1.

Previously, I characterised the virome of agricultural slurry that is primarily derived from dairy cattle faeces (see Chapter 4 (Cook, Hooton, *et al.*, 2021)), however study into the cattle virome remains limited. Whilst there has been a handful of studies that have investigated the rumen virome (Berg Miller *et al.*, 2012; Ross *et al.*, 2013; Anderson, Sullivan and Fernando, 2017), the only recent dairy cow gut virome study focussed on the bacterial fraction and its virome analysis was limited (Park and Kim, 2019). Conversely, there has been extensive study into the human gut virome, showing the human gut is sterile at birth and the virome develops in multiple stages of ecological succession (Beller *et al.*, 2022). Once developed, the human gut virome is temporally stable, with 80% of vOTUs being maintained over a 2.5 year period (Minot *et al.*, 2013), and likely shaped by environmental factors such as diet (Minot *et al.*, 2011; Edwards *et al.*, 2019; Shkoporov *et al.*, 2019). Whilst the impacts of the gut virome on human health are yet to be elucidated, there is a growing body of evidence that the human gut virome is altered in certain disease states, such as Crohn's disease and ulcerative colitis (Norman *et al.*, 2015; Clooney *et al.*, 2019). The gut virome of other animals may therefore have roles in health and disease.

The aim of this work was to determine the diversity and ecological roles of bacteriophages within the dairy cow gut, and to elucidate how this community differs across life stages. Therefore, the objectives were to:

1. To isolate and sequence the viral fraction of dairy cattle across different life stages
2. To compare the composition and structure of viral communities between sampling groups

3. To compare the dairy cow gut virome with that of the slurry tank and human gut, to determine if the dairy cow shares properties with more characterised systems

6.2 Materials and Methods

6.2.1 Sample Collection and Processing

Cow faeces was collected from a UK dairy farm using a drop catch method (i.e., catching the sample in a sterile tube before it can contact the ground). Samples (n = 20) were collected from five each of pre-weaning calves (< 30 days old), heifers, dry adults, and milking adults (Figure 6.1). Samples were kept on ice and processed the same day. Sample collection for this project was reviewed and approved by the SVMS ethics committee on the 14th of November 2017 with approval number 2132 171010.

Viral-like particle (VLP) enrichment was performed as described previously (Chapter 4), based on the method of Sazinas *et al.*, (2019).

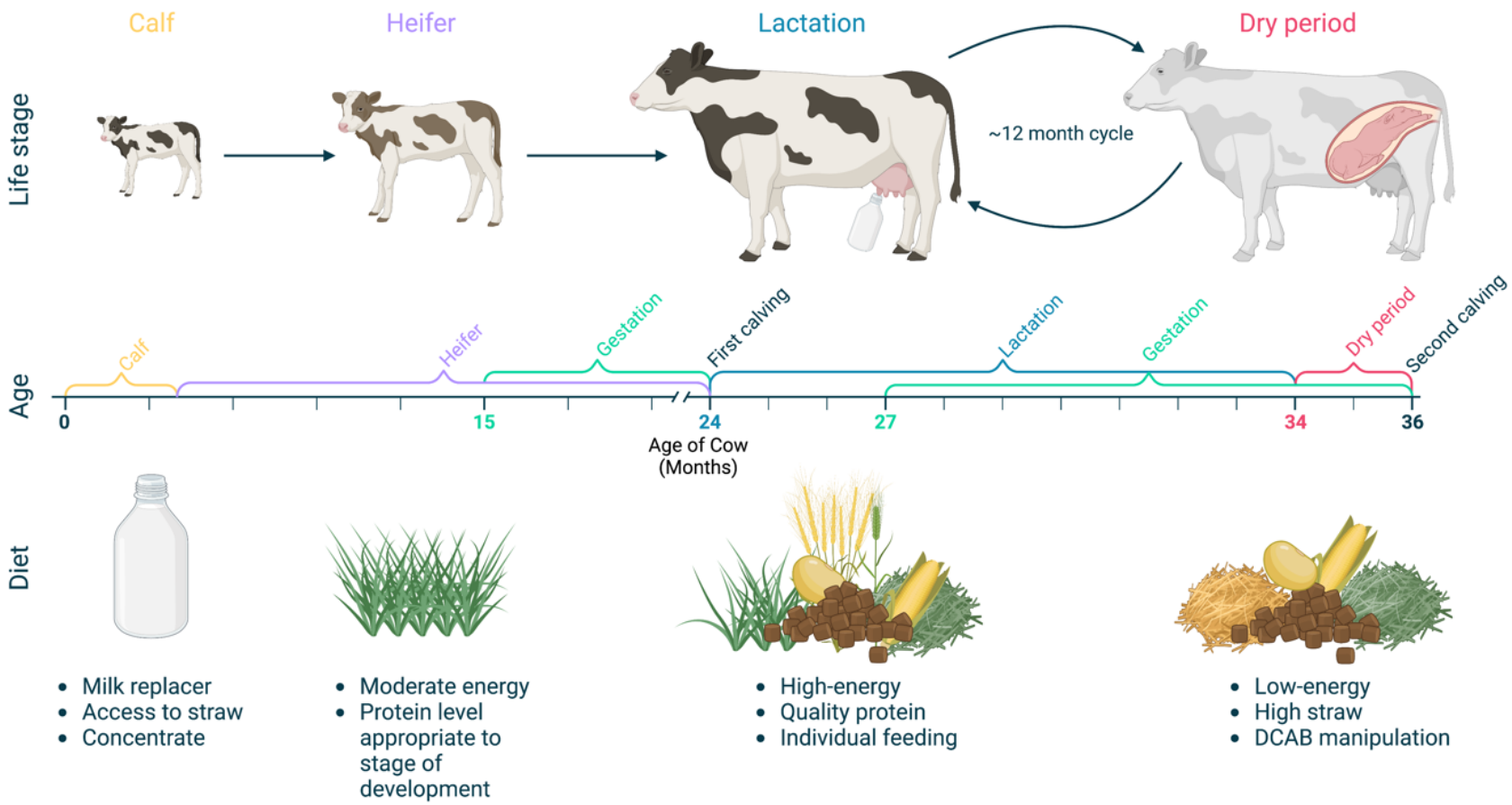


Figure 6.1 Longitudinal overview of the dairy cow lactation cycle

A longitudinal overview of life stages of the dairy cow, highlighting the ages and diets of calves, heifers, milking adults and dry adults. Note that beyond 36 months, the cow enters a ~12 month cycle that repeats months 24-36 birthing one calf per year.

6.2.2 Short Read Sequencing

All 20 individual samples were used as template for short-read sequencing by NUomics at Northumbria University. DNA was quantified using Qubit high sensitivity and normalised to 2ng per library starting concentration. The libraries were prepared using DNAprep (M) (Illumina, San Diego, CA, USA) with unique dual indexes as per manufacturer's instructions. The library was checked using BioAnalyzer (Agilent, Santa Clara, CA, USA) and Qubit high sensitivity and normalised to 30 nM. The libraries were pooled and ran on a MiSeq V2 300 cycle nano kit prior to sequencing on the Novaseq 6000 300 cycle SP kit.

6.2.3 Long Read Sequencing

DNA from the 20 samples was pooled and amplified separately using either the REPLI-g Mini kit (Qiagen, Valencia, CA, USA) or GenomiPhi V3 (GE Healthcare, Chicago, IL, USA) to gain sufficient material for ONT sequencing. As yields for GenomiPhi were comparatively low, I proceeded with REPLI-g amplifications only. Amplified DNA was de-branched using S1 Nuclease (Promega) at 10 units per µg of DNA (quantified using a Qubit) to minimise chimeras introduced during amplification (Lasken and Stockwell, 2007), followed by passage through a Zymo Clean & Concentrator-25 column. To enrich for high molecular weight DNA, a 10 kb short read exclusion kit (Circulomics, Baltimore, MD, USA) was used following manufacturer's instructions, with the following modifications. The DNA pellet was re-suspended in 50 µl of nuclease-free water rather than the provided buffer. Libraries were prepared using the SQK LSK-110 ligation sequencing kit (ONT, Oxford, UK) prior to sequencing on ten MinION flow cells (six r9.4.1 and four r10.3), with four out of ten being loaded

with DNA that had been processed using the short read exclusion kit prior to library preparation.

6.2.4 Quality Control and Assembly

Adapters were trimmed from short reads using bbduk.sh v38.84 with ktrim=r minlen=40 minlenfraction=0.6 mink=11 tbo tpe k=23 hdist=1 hdist2=1 ftm=5 ref=/bbmap/resources/adapters.fa (Bushnell, 2013), followed by quality trimming with maq=8 maxns=1 minlen=40 minlenfraction=0.6 k=27 hdist=1 trimq=12 qtrim=rl (Bushnell, 2013). Tadpole.sh v38.84 was used to correct sequencing errors with mode=correct ecc=t prefilter=2 (Bushnell, 2013). Trimmed reads were mapped to the *Bos taurus* reference genome (NKLS00000000) using bbmap.sh v38.84 with local=t minid=0.95 maxindel=6 tipsearch=4 bandwidth=18 bandwidthratio=0.18 usemodulo=t printunmappedcount=t idtag=t minhits=1, and unmapped reads were split back into paired end files using reformat.sh v38.84 (Bushnell, 2013). VLP enrichment of samples was estimated using ViromeQC v1.0 (Zolfo *et al.*, 2019). Assembly was performed using MEGAHIT v1.1.1-2-g02102e1 with `--k-min 21 --k-max 149 --k-step 24` and contigs ≥ 1.5 kb were retained (Li *et al.*, 2016).

Long reads were pooled, and low-quality reads removed using Filtlong v0.2.1 with --min_length 1000 --keep_percent 95 (https://github.com/rrwick/Filtlong). Assembly was performed using Flye v2.8.1-b1676 with --meta --min-overlap 1000 (Kolmogorov *et al.*, 2020). Long read polishing was performed with Medaka v1.6.0 with -b 50 (https://github.com/nanoporetech/medaka) in two rounds, first with reads obtained from r9.4.1 flow cells followed by reads obtained from r10.3 flow cells. Illumina reads were pooled, and forward and reverse reads were mapped separately to the medaka-

polished assembly using bwa v0.7.12-r1039 to generate SAM files (e.g., bwa mem -a medaka_polished.fa pooled_R1.fq.gz > alignments_1.sam and bwa mem -a medaka_polished.fa pooled_R2.fq.gz > alignments_2.sam) (Li and Durbin, 2009). Alignments were filtered using polypolish_insert_filter.py and the medaka-polished contigs were polished using Polypolish v0.5.0 (Wick and Holt, 2022). The polished ONT contigs were processed using CheckV v0.9.0 and any with a kmer frequency of ≥ 1.5 were excluded to remove potential chimeras (Nayfach *et al.*, 2020).

6.2.5 Filtering vMAGs and vOTUs

The 20 samples of Illumina reads were mapped separately to the Illumina assembly using minimap2 v2.17-r941 with -ax sr and sorted BAM files were produced using samtools v1.9 (Li *et al.*, 2009; Li, 2018). The Illumina assembly and BAM files were used as input for vRhyme v1.1.0 to produce vMAGs (Kieft *et al.*, 2022). vMAGs from the "best bins" were concatenated into single contigs padded with N's using concatenate.sh v38.84 and processed using CheckV v0.9.0 (Bushnell, 2013; Nayfach *et al.*, 2020). Bins that obtained a CheckV quality estimate of "low-quality" or "not-determined", a protein redundancy > 1, a contamination estimate > 10%, the warning flag "no viral genes", or ≥ 25% "host" genes (and not identified as a prophage) were excluded from filtering. Bins ≥ 10 kb (and one < 10 kb bin that was estimated to be complete due to a high confidence DTR) that satisfied at least one of the following conditions were included: (1) predicted viral by VIBRANT v1.2.0 (Kieft, Zhou and Anantharaman, 2020), (2) obtained an adjusted P-value from DeepVirFinder of ≤ 0.05, or (3) had a significant (-E 0.001) to either the viral RefSeq or INPHARED databases using MASH v2.0 (July 2022) (O'Leary *et al.*, 2016; Ondov *et al.*, 2016; Cook, Brown, *et al.*, 2021). For any tool used to process the concatenated vMAGs that uses Prodigal

to predict open reading frames (ORFs), the -m flag was manually added to their code so ORFs were not predicted over ambiguous bases (Ns) (Hyatt *et al.*, 2010).

Illumina and polished ONT contigs ≥ 10kb and those predicted to be circular (determined using apc.pl (https://github.com/jfass/apc)) were de-replicated using the MIUVIG recommended parameters (95% ANI over 85% length of the shorter sequence) with blast and CheckV scripts as described in the CheckV documentation (https://bitbucket.org/berkeleylab/checkv/src/master/) (Altschul *et al.*, 1990; Nayfach *et al.*, 2020). Contig clusters that belonged to an included vMAG were excluded from further analysis. The remaining clustered contigs were filtered using the same three criteria as the vMAGs. The filtered contigs were processed using CheckV v0.9.0 and those with the ”no viral genes” warning, < 3 total genes, or ≥ 25% “host” genes (and not identified as a prophage) were excluded (Nayfach *et al.*, 2020). For those identified as prophages, the CheckV trimmed versions were used in downstream analyses (Nayfach *et al.*, 2020). The contigs and vMAGs that passed filtering formed the 30,321 vOTUs included in this analysis.

6.2.6 Functional Annotation and AMG Analysis

vOTUs were annotated using Prokka v1.14.6 with a publicly available set of HMMs derived from PHROGs (http://s3.climb.ac.uk/ADM_share/all_phrogs.hmm.gz) (Seemann, 2014; Terzian *et al.*, 2021). Translated ORFs were processed using METABOLIC v4.0 (Zhou *et al.*, 2022), and submitted to eggNOG for additional annotation and AMG prediction (Huerta-Cepas *et al.*, 2018). Translated ORFs on predicted complete vOTUs with a hit to a CAZYme from eggNOG were submitted to

Phyre2 to predict their structure (Kelley *et al.*, 2015). Diversity-generating retroelements (DGRs) were predicted using MetaCSST (Yan *et al.*, 2019).

6.2.7 Taxonomy

The vOTUs were processed alongside the INPHARED database (August 2022) using vConTACT2 with --rel-mode 'Diamond' --db 'None' --pcs-mode MCL --vcs-mode ClusterONE --min-size 1 (Bin Jang *et al.*, 2019; Cook, Brown, *et al.*, 2021). If a viral cluster (VC) contained a reference genome belonging to *Crassvirales*, the VC was considered to be crAss-like.

6.2.8 Lifestyle and Host Prediction

Phages that may be able to access a lysogenic lifestyle (temperate phages) were identified with PhageLeads (Yukgehnaish *et al.*, 2022) and BACPHLIP (≥ 95% probability only) (Hockenberry and Wilke, 2021). If a temperate vOTU was identified, all vOTUs within its vConTACT2 VC were also classified as temperate. Hosts were predicted for the vOTUs using iPHoP v0.9beta (Roux *et al.*, 2022); a pipeline that combines RaFAH (Coutinho *et al.*, 2021), WIsH (Galiez *et al.*, 2017), oligonucleotide frequencies (Ahlgren *et al.*, 2017), PHP (Lu *et al.*, 2021), and BLAST (Altschul *et al.*, 1990).

6.2.9 Micro- and Macro-Diversity Statistics

Each read set was randomly down-sampled to the size of the smallest sample in which ≥1 could be detected by read mapping (Calf 3: 481,471 x 2 paired end reads) using seqtk with -s 100 (https://github.com/lh3/seqtk). Rarefied reads were mapped to the vOTUs using Bowtie 2 v2.3.4.3 with --non-deterministic --maxins 2000 (Langmead and

Salzberg, 2012), as described in the MetaPop paper (Gregory *et al.*, 2022). MetaPop was performed with --genome_detection_cutoff 75 --no_viz (Roux *et al.*, 2017; Gregory *et al.*, 2022). To allow our previously predicted genes to be used as input for MetaPop, they were modified with an in-house script (Supplementary File 1). Mapping and MetaPop analyses were subsequently re-performed using the full read sets (i.e., not rarefied). Pairwise comparisons of beta-diversity between groups were performed by PERMANOVA with 1,000 permutations using adonis as part of Vegan (Oksanen *et al.*, 2020), and p-values were adjusted for multiple comparisons using the Benjamini-Hochberg correction (Benjamini and Hochberg, 1995).

6.2.10 Detection of Previously Characterised Phages, Human Gut Phages and Slurry vOTUs

Reads were mapped separately to the INPHARED database (August 2022) (Cook, Brown, *et al.*, 2021), a set of vOTUs produced from a virome analysis of a dairy cattle slurry tank on the same farm (Cook, Hooton, *et al.*, 2021), viral RefSeq (August 2022) (O'Leary *et al.*, 2016), and a human gut phage database (Unterer, Khan Mirzaei and Deng, 2021) using bbmap.sh v38.84 with minid=0.90 ambiguous=all (Bushnell, 2013). A sequence was determined as present if it obtained ≥ 1x coverage over ≥ 75% sequence length (Roux *et al.*, 2017).

6.2.11 Curation of Predicted Complete Genomes

The annotations of vOTUs predicted complete by CheckV (n = 1,338) were manually inspected to determine if the vOTU was demonstrably viral (e.g., presence of viral signature genes (such as terminase, portal, tail, capsid etc), a high number of hypothetical proteins, and few genes typically associated with bacteria). The complete

vOTUs were pooled with ICTV classified dsDNA phage genomes (n = 3,652) and used as input for ViPTree v1.1.2 (Nishimura *et al.*, 2017). The terminase large subunit (*terL*) could be readily identified on 1,109 vOTUs. For those 1,109, the translated *terL* sequence was aligned with the *terL* of 3,451 ICTV classified phage genomes using MAFFT (Nakamura *et al.*, 2018). The resultant alignment was used as input for IQ-Tree (Nguyen *et al.*, 2015), and visualised using IToL (Letunic and Bork, 2019). The ICTV classified genomes and *terL* sequences are publicly available at http://millardlab.org/2022/08/04/ictv-bacteriophage-genera/.

6.3 Results

The farm in this study is the same high-performance dairy farm in the East Midlands, UK with ~ 200 milking cattle described in chapter four (Exploring Phages within Dairy Farm Slurry). The dairy cattle sampled are Holstein-Friesian, a high yielding breed that is commonly farmed for dairy. A summary of the four sampling groups (calves, heifers, milking adults, and dry adults) and a timeline of the dairy lactation cycle is shown in Figure 6.1.

6.3.1 Sequencing and Assembly Statistics

The twenty Illumina viromes (one for each cow/sample) and the pooled Nanopore virome produced ~277 and ~104 Gb of data, respectively. The four Nanopore flow cells loaded with DNA that had been processed with a short read exclusion kit obtained median read lengths of 6.9, 6.0, 6.9, and 6.2 kb whereas those that had not used short read exclusion obtained 2.6, 1.9, 1.8, 3.0, 1.9, and 1.7 kb (Figure 6.2). Furthermore, the short read exclusion did not reduce the total output of data produced (Figure 6.2).

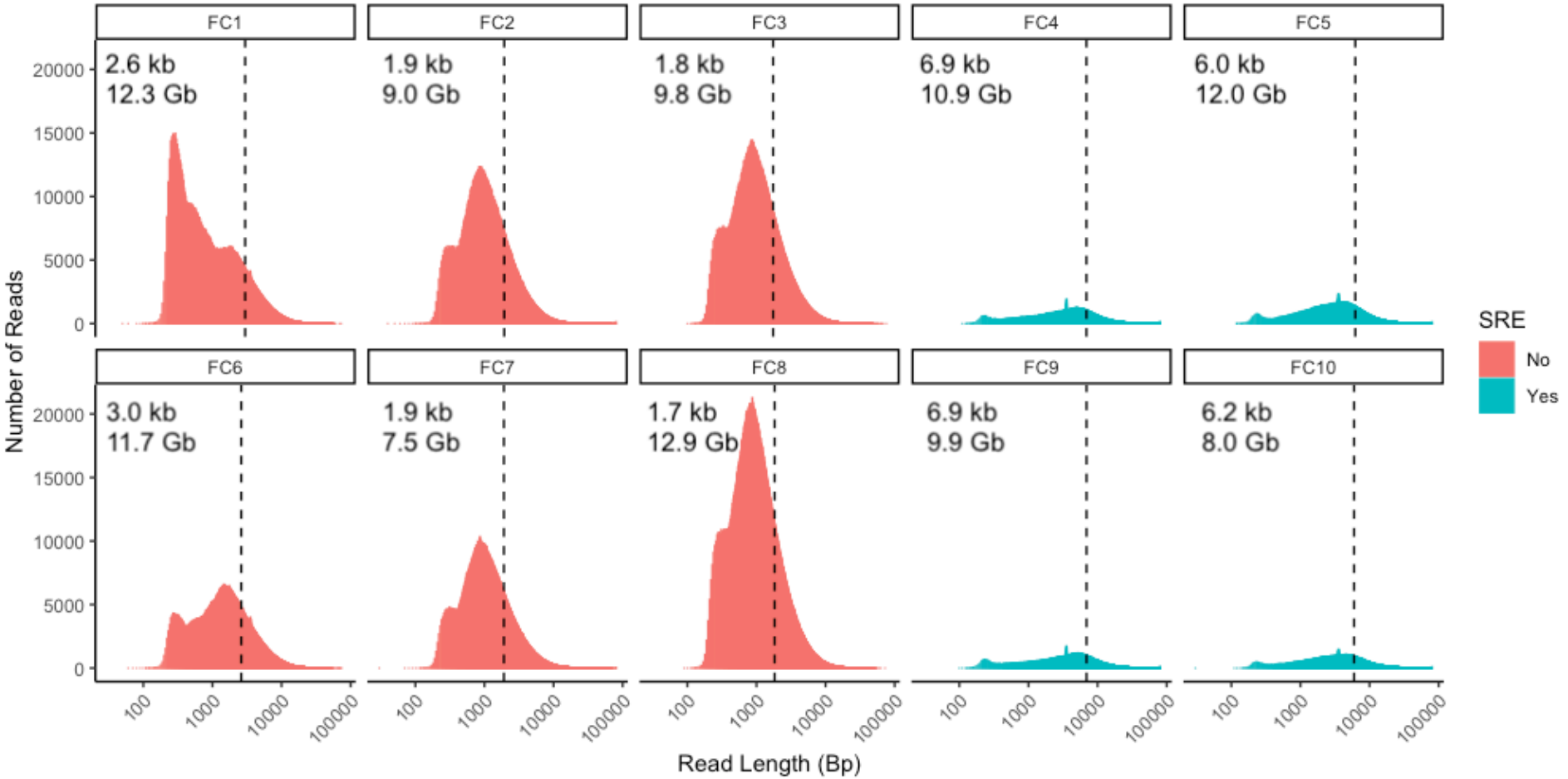


Figure 6.2 Effect of short read exclusion

Histograms showing read lengths obtained from the ten Nanopore flow cells with colour indicating the use of short read exclusion. Dashed lines show median read length. In panel labels indicate the median read length (kb) and total amount of data produced (Gb).

The ViromeQC predicted VLP enrichment (i.e., how enriched for viruses the sample is compared to the expected level obtained from a metagenome, with a score of 5 indicating a 5x enrichment) for the samples typically grouped within samples with mean values of 1.41, 7.30, 11.28, and 11.58, for calves, heifers, dry adults, and milking adults respectively (Figure 6.3A, B). The difference in predicted VLP enrichment may be due to differences in the viral community between these groups, or through differences in sample consistency/heterogeneity leading to differences in VLP extraction from the sample. Further to this, the calf sample with the highest VLP enrichment score had a different colour and consistency to other calf samples that was more similar to the adult samples (Sample C1; Figure 6.4). Whilst four of the calf samples were a viscous off-white liquid, one of them was semi-solid and brown (Sample C1; Figure 6.4).

The final filtered virome comprised 30,321 unique vOTUs ranging from 1.5 kb to 1.2 Mb, with mean and median lengths of 25.3 kb and 16.9 kb respectively (Figure 6.3C). CheckV estimated completeness scores of 1,338 complete, 3,898 high-quality, 6,497 medium-quality, 18,587 low-quality vOTUs, with one not-determined (Figure 6.3D).

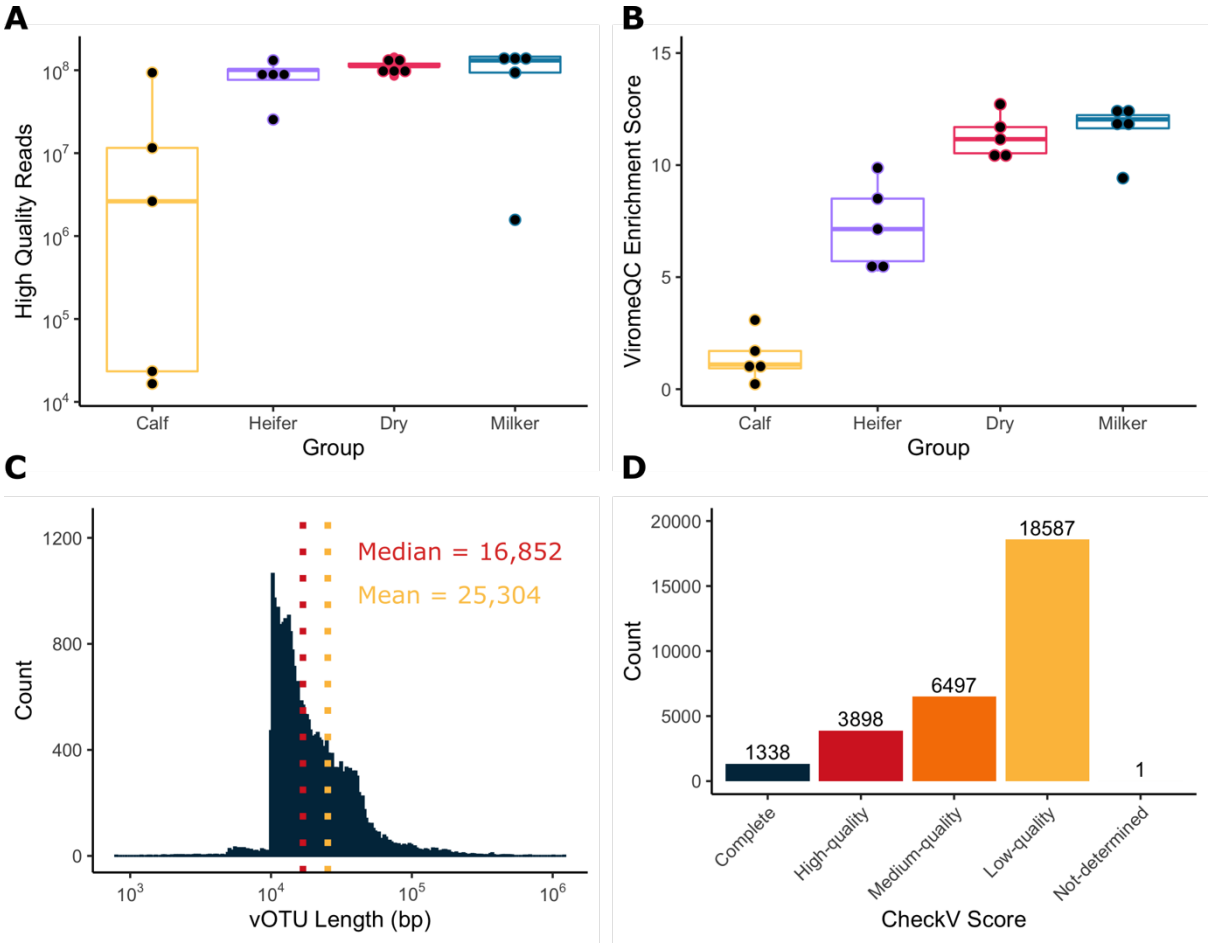


Figure 6.3 Virome summary statistics

A summary of the cows virome dataset including **(A)** number of high-quality reads and **(B)** viral enrichment as determined by ViromeQC (Zolfo *et al.*, 2019), **(C)** length of predicted vOTUs, and **(D)** completeness of vOTUs as determined by CheckV (Nayfach *et al.*, 2020).

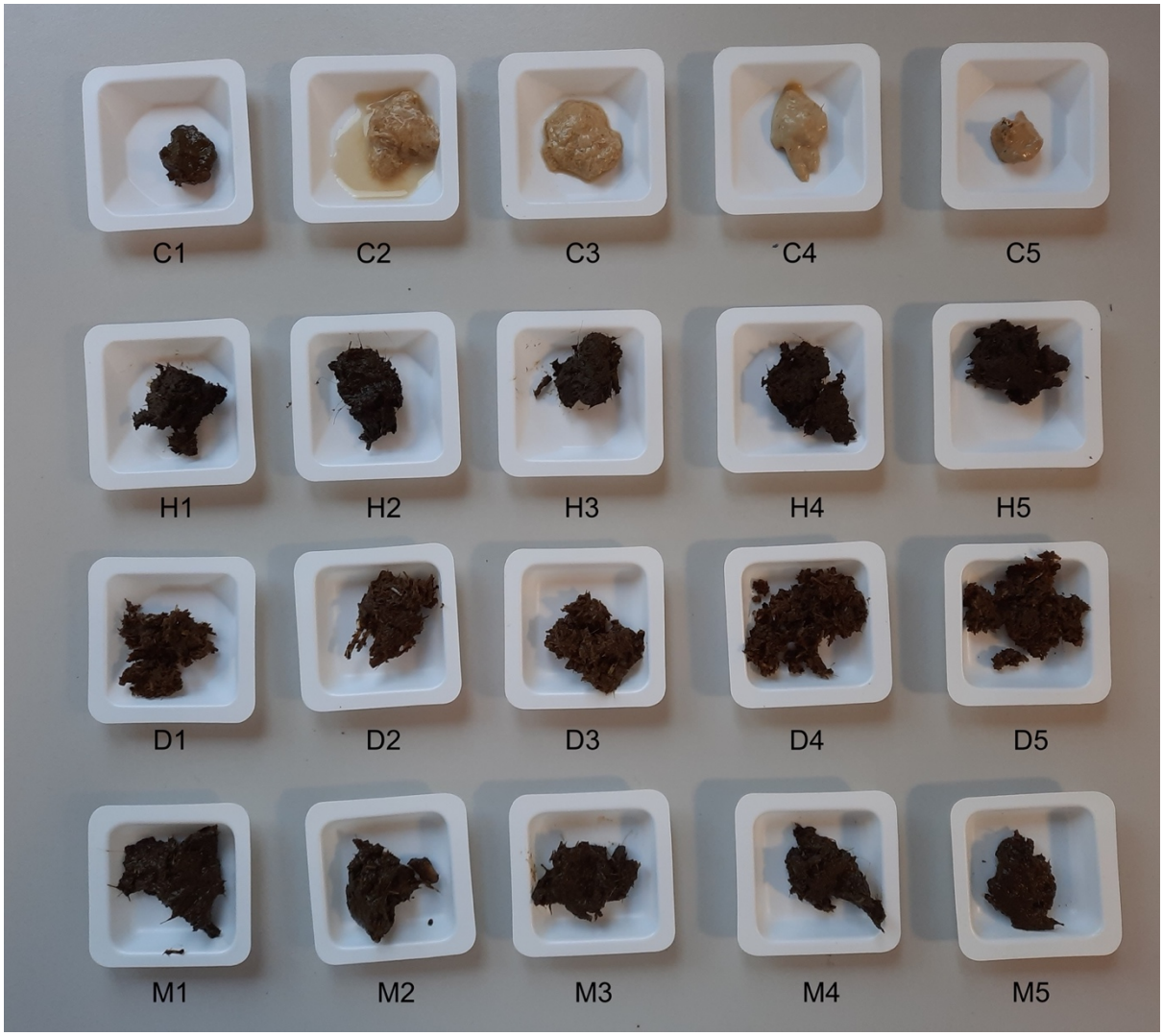


Figure 6.4 Stool samples from cows

Photographs showing individual cow faeces samples with sample ID beneath each sample. Rows from top to bottom show calf (C), heifer (H), dry adult (D), and milking adult samples respectively.

6.3.2 Virome composition

To determine the taxa of dairy cow gut viruses, vConTACT2 was used alongside all publicly available complete phage genomes (Bin Jang *et al.*, 2019; Cook, Brown, *et al.*, 2021). Of the 30,321 unique vOTUs, only 587 clustered with a known viral genome at the level of genus, and a further 78 shared the same overlap space (i.e., they overlapped with the same two viral clusters). The remaining 29,656 vOTUs were found to represent 14,506 novel genera (1,032 were singletons and 8,904 were outliers). Notably, 1,926 vOTUs were found to cluster with a member of *Crassvirales*, 234 of which estimated to be complete and a further 333 high-quality (Supplementary Table S6.1). Of the vOTUs related to a known virus at the level of genus, only three were related to a known genome at the level of species (95% ANI).

6.3.3 Comparison of the dairy cow virome across life stages

To investigate the presence/abundance of vOTUs within the 20 samples, down-sampled reads were mapped to the vOTUs, and the community structure of individual samples was investigated. No vOTUs could be detected in two of the calf samples (1x coverage over 75% vOTU length (Roux *et al.*, 2017)), and these samples were excluded from further analysis. From the remaining three calf samples, 225 vOTUs could be detected in ≥ 1 sample (35 of which were detected in all three). Of the 225 vOTUs detected in ≥ 1 calf sample, only 47 could be detected in ≥ 1 adult sample; suggesting the calf virome is vastly different from that of the adult cows.

The three adult groups each had a large “core” virome with 3,385, 3,049, and 3,254 vOTUs detected in 4/5 samples for heifers, dry cows, and milkers respectively. However, many of these vOTUs were not shared between the adult groups with a core

virome of 711 vOTUs in the adult dairy cow gut (defined by detection in ≥ 4 of all three adult groups).

With regard to alpha diversity (the diversity within a sample), the mean Shannon's index obtained for the calves (2.46), was far lower than that obtained from any of the adult groups (6.239, 6.506, and 5.296 for heifers, dry adults, and milking adults respectively; pairwise comparisons of the groups were performed using PERMANOVA with 1,000 permutations and P-values were adjusted for multiple comparisons using the Benjamini-Hochberg method, resulting in $P < 0.05$ for calves versus adult groups); suggesting the adult gut virome is more diverse than that of the calf (Figure 6.5D). This difference in macro-diversity was reflected in the micro-diversity, with the adult cow groups all obtaining high nucleotide diversity than the calf groups (Figure 6.5F). For all groups, a large proportion of detected vOTUs were found to have ≥ 1 gene under positive selection within that sample, with the dry adults obtaining the highest value (means of 26.24%, 33.03%, 38.03%, and 31.15% for calves, heifers, dry adults, and milking adults respectively) (Figure 6.5E).

Comparisons of beta-diversity using Bray-Curtis dissimilarity demonstrated the calf virome to be vastly different from that of the adult cows (Figure 6.5A), and the three adult groups to be significantly different from one another (Figure 6.5B) (Pairwise comparisons of the groups were performed using PERMANOVA with 1,000 permutations and P-values were adjusted for multiple comparisons using the Benjamini-Hochberg method, resulting in $P = 0.020979021$ for all groups).

When examining the types of phages within the groups, clear differences can be observed between the calves and adults (Figure 6.5). The calf virome contains a far higher proportion of known and temperate genera than adults, which are dominated by novel and lytic genera (Figure 6.5G, H). This finding was supported by read mapping to viral datasets, from which the calves had the highest proportion of reads mapping to known viral databases (INPHARED and Viral RefSeq) when compared to the adult groups (Figure 6.5I). Notably, the calf samples contained a large proportion of reads which mapped to the human gut phage database (mean 50.7%) (Figure 6.5I). To determine similarity of the groups to the slurry tank virome analysed in Chapter 4, cow virome reads were mapped to the vOTUs produced from the slurry virome analysis. Only a small proportion of cow virome reads mapped to the slurry vOTUs (means of 0.334684801, 1.800015517, 3.231737382, and 4.362275069 for calves, heifers, dry adults, and milking adults respectively) (Figure 6.5I).

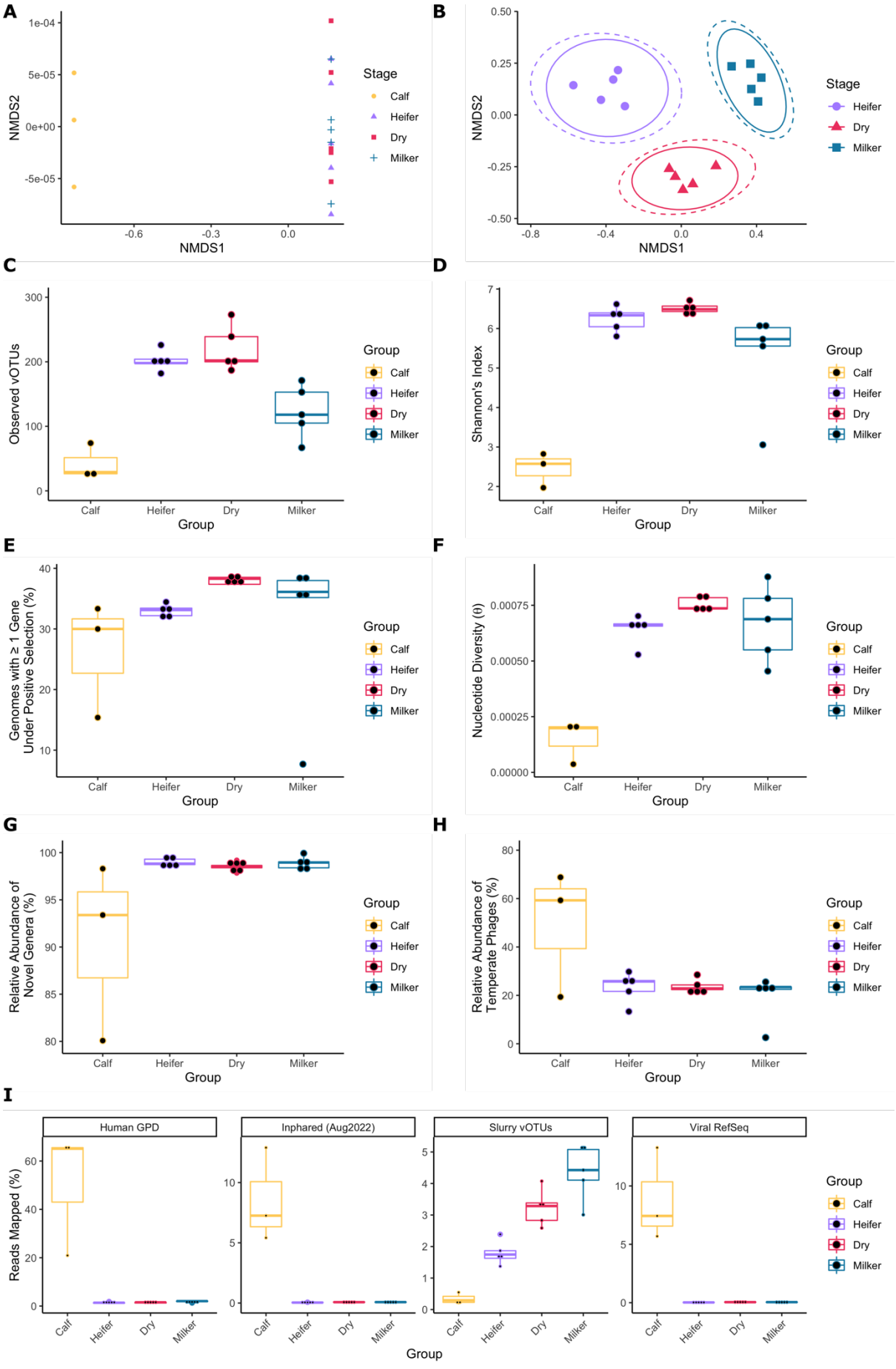

A
NMDS2
NMDS1
Stage
Calf
Heifer
Dry
Milker
B
NMDS2
NMDS1
Stage
Heifer
Dry
Milker
C
Observed vOTUs
Group
Calf
Heifer
Dry
Milker
D
Shannon's Index
Group
E
Genomes with ≥ 1 Gene Under Positive Selection (%)
Group
F
Nucleotide Diversity (θ)
Group
G
Relative Abundance of Novel Genera (%)
Group
H
Relative Abundance of Temperate Phages (%)
Group
I
Human GPD
Inphared (Aug2022)
Slurry vOTUs
Viral RefSeq
Reads Mapped (%)
Group

Figure 6.5 Comparison of abundance and diversity of viruses in different cow groups

Macro- and Micro-diversity. Beta-diversity for the four different groups with **(A)** and without calves **(B)**, with the ellipses using normal (dashed) and t (solid) distributions. Pairwise comparisons of the groups were performed using PERMANOVA with 1,000 permutations and P-values were adjusted for multiple comparisons using BH method, resulting in P = 0.020979021 for all groups. **(C)** Observed vOTUs and **(D)** Shannon's index for each library. **(E)** The percentage of vOTUs with ≥ 1 gene under positive selection and **(F)** mean microdiversity (θ) for vOTUs in each library. **(G)** Relative abundance of vOTUs which did not cluster with a known phage using vConTACT2 and **(H)** relative abundance of vOTUs predicted to be temperate. **(I)** The proportion of reads which mapped to relevant databases. Note for panels **A-D**, rarefied reads were used and for panels **E-I** the full read sets were used.

6.3.4 Predicted hosts for vOTUs reflective of gut metabolism

Host genera could be predicted for 2,416 (~8%) vOTUs with ≥ 90% confidence using iPHoP (Roux *et al.*, 2022). Of these, 121 were core. The most commonly predicted host phyla were *Bacteroidota* (n=999) and *Firmicutes* (n=919) (Figure 6.6), the two most dominant phyla in the cow gut (Kim and Wells, 2016; Delgado *et al.*, 2019; Li *et al.*, 2019). Notably, 51 of the vOTUs were predicted to infect the domain Archaea, with phyla including class I methanogens such as *Methanobacteriota* (n=21), and class II methanogens such as *Halobacteriota* (n=10), as well as *Thermoplasmatota* (n=8), *Nanoarchaeota* (n=6), *Asgardarchaeota* (n=3), *Aenigmatarchaeota* (n=2), and *Thermoproteota* (n=1) (Figure 6.6). Of the vOTUs predicted to infect Archaea, one was core (based on 4/5 of all three adult groups); predicted to infect *Asgardarchaeota*. Furthermore, 35 vOTUs were predicted to infect *Fibrobacterota* (a major component of the rumen microbiota), three were predicted to infect the sulfate-reducing phylum *Desulfobacterota*, and one high quality vOTU (predicted 95.76% complete) was predicted to infect *Fusobacterium*, an environmental bacterium associated with bovine foot rot.

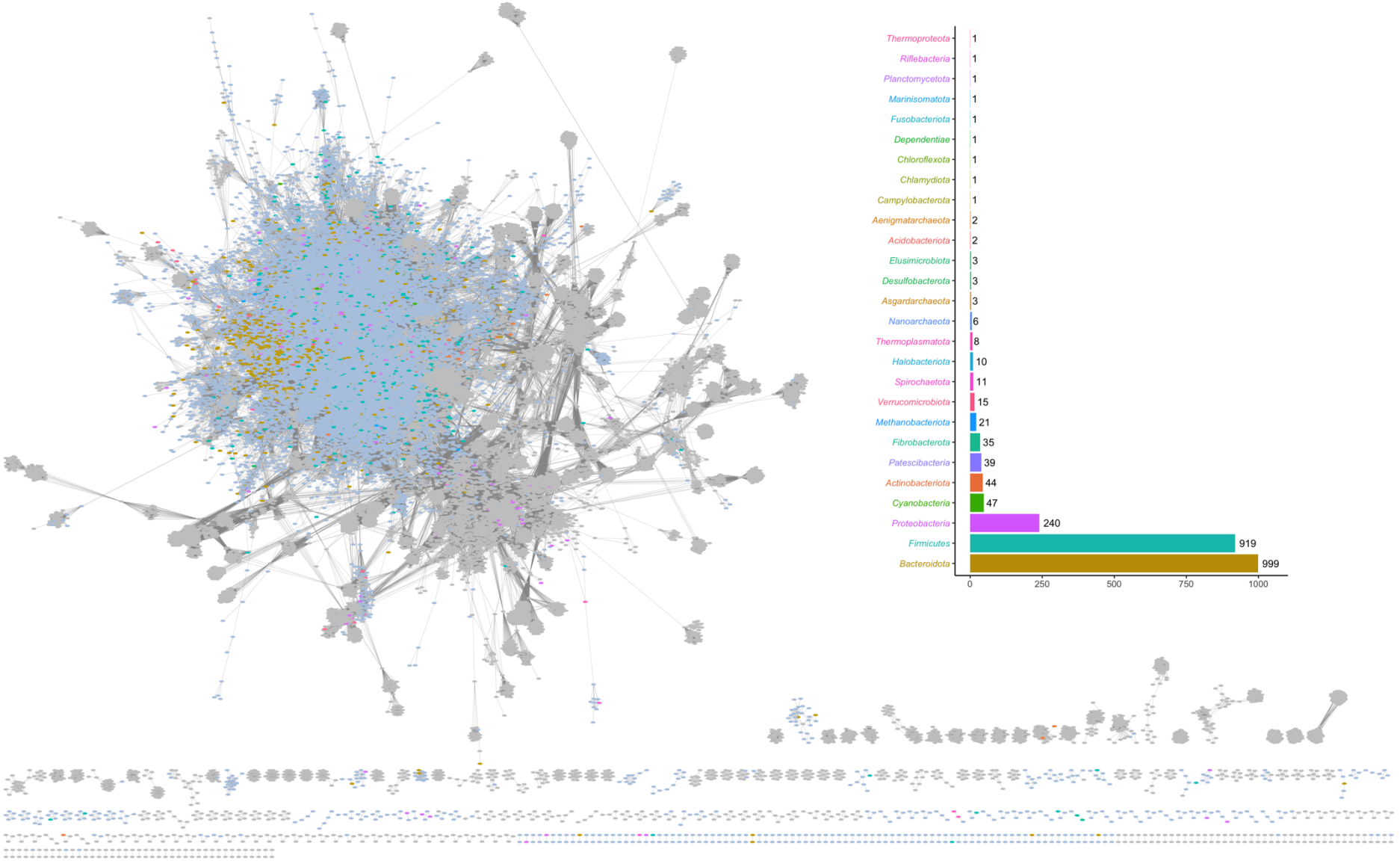

Thermoproteota 1
Riflebacteria 1
Planctomycetota 1
Marinisomatota 1
Fusobacteriota 1
Dependentiae 1
Chloroflexota 1
Chlamydiota 1
Campylobacterota 1
Aenigmatarchaeota 2
Acidobacteriota 2
Elusimicrobiota 3
Desulfobacterota 3
Asgardarchaeota 3
Nanoarchaeota 6
Thermoplasmatota 8
Halobacteriota 10
Spirochaetota 11
Verrucomicrobiota 15
Methanobacteriota 21
Fibrobacterota 35
Patescibacteria 39
Actinobacteriota 44
Cyanobacteria 47
Proteobacteria 240
Firmicutes 919
Bacteroidota 999
0
250
500
750
1000

Figure 6.6 Diversity and large-scale taxonomy of cow vOTUs by host phylum

A vConTACT2 network showing taxonomy of cow vOTUs coloured by host phylum, with bar chart indicating the number of vOTUs predicted to infect each phylum. Those in grey are reference sequences, and those in pale blue are cow vOTUs with no host predicted.

6.3.5 Compendium of complete genomes

CheckV predicted 1,338 complete viral genomes which were checked manually to identify any that were clearly not viral (Section 6.2.11), of which none were identified. To investigate the diversity of the complete cow vOTUs, ViPTree was used alongside all phage genomes classified at the level of genus by the ICTV (n=3,652) (Nishimura *et al.*, 2017) (Figure 6.7). Many cow vOTUs were interspersed with known viral familes, including *Intestiviridae* (n = 1), *Steigviridae* (n = 4), *Salasmaviridae* (n = 10), *Microviridae* (n = 84), *Autographviridae* (n = 5), and *Peduoviridae* (n = 1) (Figure 6.7). However, large numbers of cow vOTUs formed monophyletic clades which did not contain an ICTV classified genome. Notably, a deeply branching clade of 295 cow vOTUs was not represented by any currently classified phages (Figure 6.7). Furthermore, the nearest sister clades to this large clade of cow vOTUs contain reference genomes that are classified to the level of genus but currently do not belong to a family. These results suggest there are families, and possibly orders, of novel phages in the cow gut that have yet to be isolated and sequenced.

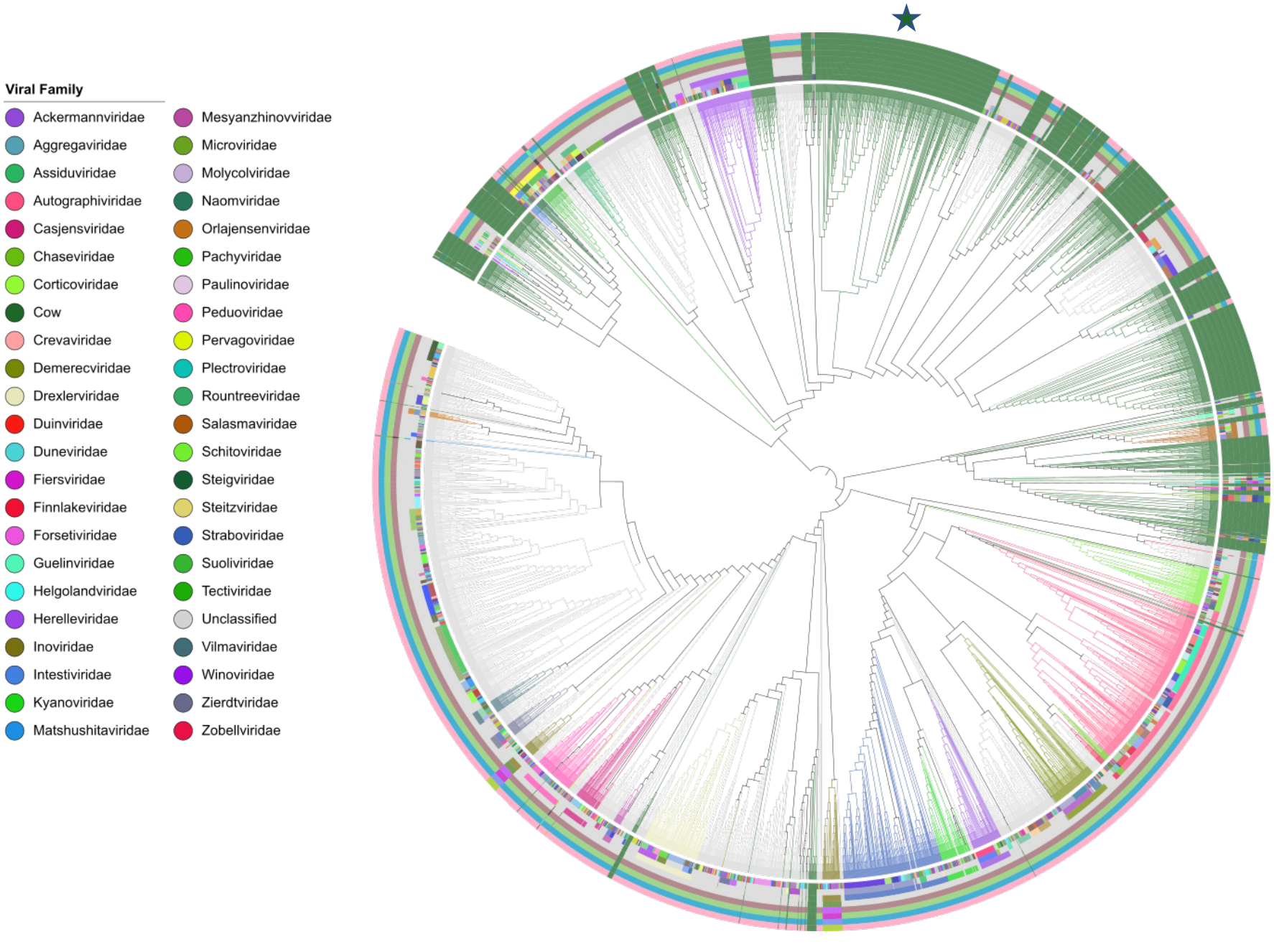

Viral Family
Ackermannviridae
Aggregaviridae
Assiduviridae
Autographiviridae
Casjensviridae
Chaseviridae
Corticoviridae
Cow
Crevaviridae
Demerecviridae
Drexlerviridae
Duinviridae
Duneviridae
Fiersviridae
Finnlakeviridae
Forsetiviridae
Guelinviridae
Helgolandviridae
Herelleviridae
Inoviridae
Intestiviridae
Kyanoviridae
Matshushitaviridae
Mesyanzhinovviridae
Microviridae
Molycolviridae
Naomviridae
Orlajensenviridae
Pachyviridae
Paulinoviridae
Peduoviridae
Pervagoviridae
Plectroviridae
Rountreeviridae
Salasmaviridae
Schitoviridae
Steigviridae
Steitzviridae
Straboviridae
Suoliviridae
Tectiviridae
Unclassified
Vilmaviridae
Winoviridae
Zierdtviridae
Zobellviridae

Figure 6.7 Phylogeny of complete genomes

VipTree phylogeny of complete cow vOTUs alongside all classified phages at level of genus. Coloured rings indicate viral taxa (inner to outer: genus, sub-family, family, order, class, phylum, kingdom, and realm). Solid green indicates a cow vOTU, and the green star indicates the large cow-only monophyletic clade described in section 6.3.5. Coloured branches indicate viral family. Tree visualised using IToL. Tree is rooted at the mid-point.

6.4 Discussion

Building upon previous work (Chapter 3 and Chapter 4), and that of others (Warwick-Dugdale *et al.*, 2019; Zablocki *et al.*, 2021), this analysis has continued to develop the methodology of hybrid viromics. The use of short read exclusion prior to library preparation obtained median read lengths ≥ 6 kb for all flow cells used, whereas a previously used LASL approach obtained median read length of ~4kb (Warwick-Dugdale *et al.*, 2019). Increasing the length of reads may help to uncover and assemble more viral genomes from mixed community samples. Furthermore, this analysis has expanded the use of hybrid viromics to a new environment, the Holstein-Friesian dairy cattle gut.

To date, there has been limited study into the dairy cattle gut virome and most dairy cattle viromics has focussed on the rumen, where most phages are novel and some are thought to augment their host metabolism to aid the breakdown of complex carbohydrates (Berg Miller *et al.*, 2012; Ross *et al.*, 2013; Anderson, Sullivan and Fernando, 2017). However, a more recent study examined the dairy cattle gut virome and compared it to that of horses found on the same farm, however the methodology of this analysis focussed on the bacterial fraction and their investigations into the viral fraction were minimal (Park and Kim, 2019). Conversely, studies into the human gut virome have revealed ecologically significant phages (Dutilh *et al.*, 2014), and enigmatic eukaryotic viruses whose role in human health remains to be elucidated (e.g. *Anelloviridae*) (Reyes *et al.*, 2015). Exploring the viromes of other animals may therefore uncover viruses with significant ecological impacts on widely reared members of livestock.

This study identified 30,321 non-redundant vOTUs from the cow gut that represent 14,506 new genera that are not currently represented in cultured phage isolates. As the ICTV currently recognises 1,652 genera of viruses that infect prokaryotes (September 2022; https://ictv.global/taxonomy), this represents an enormous volume of previously unseen viral diversity. Whilst there are no studies of a comparable size regarding the gut of a non-human animal, there have been large studies into the human gut. Most notably, a recent meta-study that mined viral genomes from 11,810 publicly available human stool metagenomes using data from 61 previously published studies was able to identify vOTUs belonging to 5,800 genera (Nayfach *et al.*, 2021). However, the two studies differ in how viral taxonomy is assigned. Furthermore, the cow vOTUs described here are estimated to represent 11,733 ≥ 50% complete phage genomes, with 1,338 predicted to be 100% complete and manually verified to be viral in origin. Whilst the human gut meta-study identified 26,030 complete phage genomes, these were not manually inspected and the data was obtained from 61 different studies (Nayfach *et al.*, 2021). The work described here therefore likely represents the single highest number of complete (or near complete) vOTUs obtained from a single study of gut viruses.

Phylogenetic analysis of the complete genomes revealed large monophyletic clades that contained no currently classified phage genomes, suggesting the presence of novel families (and potentially orders) in the dairy cow gut that are not represented in currently known viral diversity. Therefore, agricultural sites may offer a reservoir of novel viruses. However, this is not necessarily surprising as are likely far from reaching saturation of viral diversity even for commonly sampled hosts and sites (Cook, Brown, *et al.*, 2021).

The large number of novel and diverse vOTUs presented in this work may therefore offer a community resource; a compendium of complete phage genomes derived from the cow gut. This dataset contains extensive viral genomic diversity —not represented in currently classified phages —that may aid the further study of viromes from a variety of animal and environmental reservoirs. For example, this compendium may be able to provide genomic context for under-sampled groups of phages that do not currently belong to a family and aid their taxonomic classification.

Although many of the predicted complete genomes likely represented novel families, there were some that fell within clades of currently classified viral families (*Intestiviridae*, *Steigviridae*, *Salasmaviridae*, *Microviridae*, *Autographviridae*, and *Peduoviridae*). Notably *Intestiviridae* and *Steigviridae* being members of *Crassvirales*. The *Crassvirales* are an enigmatic order of phages that are known to be dominant in the human gut (Shkoporov *et al.*, 2018; Yutin *et al.*, 2018; Camarillo-Guerrero *et al.*, 2021). Whilst early studies suggested crAss-like phages (before the order *Crassvirales* was ratified) were unique to the human gut, the number of environments they are found in has expanded beyond the human gut (Yutin *et al.*, 2018; Cuscó *et al.*, 2019; Edwards *et al.*, 2019), and this work has expanded this further to include the dairy cattle gut. However, this is not necessarily surprising, as previous work found *Crassvirales* to be present in agricultural slurry that is largely derived from cattle faeces (Cook, Hooton, *et al.*, 2021). Although *Crassvirales* were found to be present in both cattle faeces and slurry, a read mapping approach showed the dairy cattle gut virome shared little similarity to the slurry tank virome from the same site (Chapter 4), suggesting that most viruses in the slurry tank are not those found in cattle faeces. As

cattle faeces is the main input of the slurry tank, this posits the question, where are the slurry phages coming from?

Comparison of the community composition of the dairy cow gut virome across life stages revealed the calf virome was significantly different from that of the adult dairy cow. The calf virome is less diverse than that of adults and contains a higher proportion of temperate phages. This draws a parallel from the more widely studied human gut virome, which is thought to be initially colonised by prophages induced from early colonising bacteria, with the virome become more complex and diverse over stages of ecological succession (Lim *et al.*, 2015; Liang *et al.*, 2020; Beller *et al.*, 2022). The calf virome may therefore follow a similar path of early seeding prophages followed by stages of ecological succession, as the adult cow virome was far more diverse than that of the calf. Moreover, the predicted VLP enrichment for calf samples was typically far lower than that of the adult samples. Whilst this may be indicative of lower levels of viral diversity in the calf gut similar to what is known of the infant human gut (Lim *et al.*, 2015; Liang *et al.*, 2020; Beller *et al.*, 2022), it may simply represent a technical challenge with extracting VLPs from this particular sample type, as the consistency of the calf faeces was different to that of the adult cows. Although the calf viromes shared similarity with human gut viruses, and the ecological succession of the dairy cow virome over time may be similar to that of humans, the adult dairy cow virome shared little with the human gut virome. Future work would compare the dairy cow gut viromes to those of other domesticated animals that share more similar environmental conditions (e.g., diet and housing) to cows than humans, such as horses and pigs (Park and Kim, 2019; Babenko *et al.*, 2020; Billaud *et al.*, 2021). It may be that

domesticated animals have a more similar virome to one another than they do to humans.

Comparison of three different groups of adult cow samples (heifers, dry adults, and milking adults) revealed the virome to be significantly different across groups. These adult cows are at different stages of the lactic cycle, with different housing and different diets. The human gut virome is known to be influenced by environmental factors, such as diet (Minot *et al.*, 2011; Edwards *et al.*, 2019), and so this finding in dairy cattle is not necessarily surprising. Notably, the cows in the drying off period contained the highest proportion of vOTUs under positive selection. The drying off period sees dairy cattle transition to a different diet for a ~2 month period. Therefore, the sudden change in diet may alter the bacterial composition of the dairy cattle gut, which is reflected in increased selection pressures on the viral community. However, a more robust longitudinal study that follows individual cows over time may be better positioned to investigate these differences.

The host-prediction analysis revealed vOTUs predicted to infect a diverse range of bacterial and archaeal phyla. The two most commonly predicted host phyla were *Bacteroidota* and *Firmicutes*, which is unsurprising as these are the two most dominant phyla in the cow gut (Kim and Wells, 2016; Delgado *et al.*, 2019; Li *et al.*, 2019). Other commonly predicted bacterial host phyla included *Fibrobacterota*, which is known to be an important member of the rumen microbiota involved in lignocellulose degradation (Xie *et al.*, 2018), and *Desulfobacterota* that are thought to contribute to biogas production in animal wastes through H_2S production (St-Pierre and Wright, 2017). Moreover, 51 vOTUs were predicted to infect the domain Archaea, including

class I methanogens (e.g., *Methanobacteriota*) and class II methanogens (e.g., *Halobacteriota*). The viral community of dairy cattle may therefore have significant impacts on the microbial community that contributes to greenhouse gas emissions resulting from the rearing of dairy cattle (Lahart *et al.*, 2021). Furthermore, one vOTU was predicted to infect *Fusobacterium*, an environmental bacterium associated with bovine foot rot; the costliest disease to UK dairy cattle (Van Metre, 2017; CHAWG, 2020). The recent interest in the development and application of phage therapy may be suitable for the treatment of relevant agricultural pathogens (Luong, Salabarria and Roach, 2020), such as foot rot, although those against *Dichelobacter* may have more clinical relevance as *D. nodosus* is thought to be the main aetiological agent of foot rot (Prosser *et al.*, 2020).

Although there are exciting initial results described in this chapter, there is much analysis yet to be performed. Whilst previous work investigated the diversity and abundance of AMGs within agricultural slurry (Section 4.5.4), I have not yet done so for the cow viromes. Phages in the rumen are thought to contribute to complex carbohydrate metabolism through the presence of AMGs (Ross *et al.*, 2013; Anderson, Sullivan and Fernando, 2017), and there may be a similar pattern of AMGs present in the dairy cattle gut. Furthermore, it would be of note to determine if the pattern of AMGs varies with the life stages used in this study. Additionally, the differences in the virome of cows across life stages may be reflected by changes in the bacterial flora. As part of a linked study, bacterial metagenomes from the same 20 cow samples have been sequenced. These may be incorporated into this study with analyses including the prediction of bacterial community composition and detecting prophages.

Chapter 7 General Discussion

7.1 Conclusions

In this thesis, I have compared how viromic approaches are currently used to investigate viral communities and implemented viromics to explore the diversity and community composition of viruses in the dairy farm environment.

In Chapter 2, I explored the current extent of sequencing for complete phage genomes obtained from cultured isolates. As reference databases are vital for numerous viromics analyses (e.g., virus prediction and phylogenetic analyses), a deeper understanding of complete phage genomes will likely aid the field of viromics. I demonstrated that, while the number of complete genomes is continuing to increase at a rapid rate, there are biases within the current collection of available genomes, likely due to common sampling hosts and sites. However, even for hosts most commonly sampled, we are far from reaching saturation of viral diversity in nature. Therefore, to uncover more viral diversity, we need to isolate phages using a wider range of hosts from a wider range of environments.

In Chapter 3, I benchmarked long, short and hybrid sequencing approaches using a number of different assembly algorithms for the recovery of viral genomes from a mock viral community. The continual improvement of sequencing technologies, such as the move away from cloning-based approaches, has allowed for a deeper understanding of global viral diversity. This upward trend can only continue as a result of continuous improvements to sequencing platforms and assemblers, and benchmarking these approaches allows for the community to use optimised work-through for their analyses (Roux *et al.*, 2017; Fung *et al.*, 2022). Building upon the work of others, I demonstrated that the addition of long-reads to short-reads is able to aid the recovery of viral

genomes from a mixed viral community (Warwick-Dugdale *et al.*, 2019). However, this work was the first to comprehensively benchmark different sequencing approaches using a mock viral community and was the first use of PacBio sequencing on a VLP-enriched virome. The use of multiple sequencing platforms, and continual optimisation of viromic work-throughs will expand our understanding of viral diversity. Future work should seek to enhance viral nucleic extraction for sequencing with long read platforms.

In Chapter 4, I analysed the virome of dairy cattle slurry in a longitudinal study with Illumina and ONT sequencing. Building upon previous work (Chapter 3), and that of others (Warwick-Dugdale *et al.*, 2019), this work demonstrated that the addition of long-reads to short-reads aids the recovery of viral genomes. I characterised the viral community of a unique agricultural environment that had not yet been explored, despite the widespread application of slurry to land. This work demonstrated a diverse and stable community of novel viruses that may impact on the metabolism of their hosts, notably through widespread virulence determinants that are associated with relevant agricultural pathogens. Subsequent analyses of viromes from modelled slurry "mini-tanks" suggested the stable composition of this community may be due to the continual addition of footwash, used for the prevention of lameness (Chapter 5). Furthermore, I uncovered a large number of putative phage-encoded metallo-beta-lactamases, although subsequent experimental work suggested these may not be functional (**Error! Reference source not found.**). Whilst the putative ARGs in this work are likely non-functional, more phenotypic screening of putative AMGs (not just ARGs) is required to better understand the role of phages in augmenting the metabolism of their hosts in the wider environment.

In Chapter 6, I characterised the gut virome of dairy cattle across life stages. This work continued the improvement of viromic sequencing, by demonstrating the use of a short read exclusion kit prior to ONT sequencing increases the median read length obtained from VLP-enriched viromes. Furthermore, this work expanded the use of hybrid viromics to a new environment, for which there was a paucity of knowledge: the cow gut. I showed that the cow gut virome differs across stages of the lactic cycle, likely caused by age, diet and communal housing. Additionally, this work uncovered the highest number of predicted complete phage genomes obtained from a single study that I am currently aware of.

7.2 Next steps

This study has comprehensively characterised the virome of agricultural slurry, however, there are still many questions to be addressed. Due to time constraints, the analysis of the slurry "mini-tanks" is still largely incomplete (Chapter 5). The next logical steps would be to further characterise the viral communities and to link these analyses with bacterial metagenomes obtained from the same samples. This would allow for a deeper understanding of the role footwash plays in shaping the microbial ecology of agricultural slurry.

Similarly, the analysis of the cow gut viromes remains incomplete (Chapter 6). The diversity and phylogeny of the predicted complete genomes can be analysed further, to gain a deeper understanding of the novel families that likely reside within cows. Additionally, I prepared bacterial DNA fractions from the same cow samples, that were sequenced for a fellow PhD student's project. Whilst this linked project has performed analyses of the bacterial fraction, no linked analyses between viral and bacterial fractions have been performed thus far. Creating synergy between the two datasets would likely maximise the understanding of microbial ecology within the cow gut across the life stages sampled in this experiment. Furthermore, whilst these analyses show clear differences between the groups that we infer are due to age and diet, a longitudinal study of the same cows over time would demonstrate this further.

Although I appear to have successfully cloned at least two of the putative MBLs identified on phage contigs (**Error! Reference source not found.**), this work was still in its infancy at the end of the PhD project. Future work would perform a more robust antimicrobial susceptibility assay to determine if the successfully cloned inserts are

indeed functional. This would involve using a larger range of beta-lactam antibiotics at a larger range of concentrations, as well as using a positive control such as the phage-encoded MBL characterised previously (Moon *et al.*, 2020). With regard to the inserts I suggest may be toxic to *E. coli*, future work could clone these into a tightly regulated inducible expression system to determine if they are indeed toxic.

As mentioned previously, I characterised the virome of agricultural slurry that is applied to land as fertiliser. However, the consequences of the application to land remain unknown. Future experiments could study the virome of soil to which the slurry is applied, compared against those that do not receive slurry. Whilst we identified a number of putative virulence factors and ARGs in agricultural slurry, it is not clear if these persist in the wider environment after the slurry is applied to land.

Bibliography

Aarestrup, F. M. *et al.* (2000) 'Associations between the use of antimicrobial agents for growth promotion and the occurrence of resistance among Enterococcus faecium from broilers and pigs in Denmark, Finland, and Norway', *Microbial drug resistance (Larchmont, N.Y.)*. Microb Drug Resist, 6(1), pp. 63–70. doi: 10.1089/MDR.2000.6.63.

Ackermann, H. W. (2009) 'Phage classification and characterization', *Methods in molecular biology (Clifton, N.J.)*. Methods Mol Biol, 501, pp. 127–140. doi: 10.1007/978-1-60327-164-6_13.

Addy, H. S. *et al.* (2012) 'Loss of Virulence of the Phytopathogen Ralstonia solanacearum Through Infection by φRSM Filamentous Phages', *http://dx.doi.org/10.1094/PHYTO-11-11-0319-R*. The American Phytopathological Society, 102(5), pp. 469–477. doi: 10.1094/PHYTO-11-11-0319-R.

Adriaenssens, E. M. *et al.* (2015) 'Integration of genomic and proteomic analyses in the classification of the Siphoviridae family', *Virology*. Virology, 477, pp. 144–154. doi: 10.1016/J.VIROL.2014.10.016.

Adriaenssens, E. M. *et al.* (2018) 'Taxonomy of prokaryotic viruses: 2017 update from the ICTV Bacterial and Archaeal Viruses Subcommittee', *Archives of Virology*. Springer Vienna, 163(4), pp. 1125–1129. doi: 10.1007/s00705-018-3723-z.

Adriaenssens, E. M. and Rodney Brister, J. (2017) 'How to name and classify your phage: An informal guide', *Viruses*, 9(4), pp. 1–9. doi: 10.3390/v9040070.

AHDB (2018) *UK and EU cow numbers*. Available at: https://ahdb.org.uk/dairy/uk-and-eu-cow-numbers (Accessed: 19 June 2020).

AHDB (2020) *UK milk productivity: The global context | AHDB*. Available at: https://ahdb.org.uk/news/uk-milk-productivity-the-global-context (Accessed: 20

September 2022).

AHDB (2022a) *Drying off dairy cows | AHDB*. Available at: https://ahdb.org.uk/knowledge-library/drying-off-dairy-cows (Accessed: 20 September 2022).

AHDB (2022b) *UK milk yield | AHDB*. Available at: https://ahdb.org.uk/dairy/uk-milk-yield (Accessed: 20 September 2022).

AHDB (no date a) *Considerations when drying off dairy cows | AHDB*. Available at: https://ahdb.org.uk/knowledge-library/considerations-when-drying-off-dairy-cows (Accessed: 25 October 2022).

AHDB (no date b) *Cost effective slurry storage strategies*. Available at: https://dairy.ahdb.org.uk/resources-library/technical-information/health-welfare/cost-effective-slurry-storage-strategies/#.XvCQompKjwd.

Ahlgren, N. A. *et al.* (2017) 'Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences', *Nucleic Acids Research*. Oxford Academic, 45(1), pp. 39–53. doi: 10.1093/NAR/GKW1002.

Aiewsakun, P. *et al.* (2018) 'Evaluation of the genomic diversity of viruses infecting bacteria, archaea and eukaryotes using a common bioinformatic platform: steps towards a unified taxonomy', *The Journal of General Virology*. Microbiology Society, 99(9), p. 1331. doi: 10.1099/JGV.0.001110.

Akhter, S., Aziz, R. K. and Edwards, R. a. (2012) 'PhiSpy: A novel algorithm for finding prophages in bacterial genomes that combines similarity-and composition-based strategies', *Nucleic Acids Research*, 40(16), pp. 1–13. doi: 10.1093/nar/gks406.

Akhwale, J. K. *et al.* (2019) 'Comparative genomic analysis of eight novel

haloalkaliphilic bacteriophages from Lake Elmenteita, Kenya', *PLOS ONE*. Public Library of Science, 14(2), p. e0212102. doi: 10.1371/JOURNAL.PONE.0212102.

Ali, Y. *et al.* (2014) 'Temperate Streptococcus thermophilus phages expressing superinfection exclusion proteins of the Ltp type', *Frontiers in Microbiology*, 5. doi: 10.3389/fmicb.2014.00098.

Altschul, S. F. *et al.* (1990) 'Basic local alignment search tool', *Journal of Molecular Biology*. Academic Press, 215(3), pp. 403–410. doi: 10.1016/S0022-2836(05)80360-2.

Anantharaman, K. *et al.* (2014) 'Sulfur oxidation genes in diverse deep-sea viruses', *Science*. American Association for the Advancement of Science, 344(6185), pp. 757–760. doi: 10.1126/science.1252229.

Anderson, C. L., Sullivan, M. B. and Fernando, S. C. (2017) 'Dietary energy drives the dynamic response of bovine rumen viral communities', *Microbiome*. BioMed Central, 5(1), p. 155. doi: 10.1186/s40168-017-0374-3.

Angly, F. E. *et al.* (2006) 'The Marine Viromes of Four Oceanic Regions', *PLoS Biology*. Public Library of Science, 4(11), pp. 2121–2131. doi: 10.1371/JOURNAL.PBIO.0040368.

Antipov, D. *et al.* (2022) 'viralFlye: assembling viruses and identifying their hosts from long-read metagenomics data', *Genome Biology 2021 23:1*. BioMed Central, 23(1), pp. 1–21. doi: 10.1186/S13059-021-02566-X.

Arndt, D. *et al.* (2017) 'PHAST, PHASTER and PHASTEST: Tools for finding prophage in bacterial genomes', *Briefings in Bioinformatics*, (May), pp. 1–8. doi: 10.1093/bib/bbx121.

Arnold, B. J., Huang, I. T. and Hanage, W. P. (2021) 'Horizontal gene transfer and adaptive evolution in bacteria', *Nature Reviews Microbiology 2021 20:4*. Nature

Publishing Group, 20(4), pp. 206–218. doi: 10.1038/s41579-021-00650-4.

Arumugam, K. *et al.* (2021) 'Recovery of complete genomes and non-chromosomal replicons from activated sludge enrichment microbial communities with long read metagenome sequencing', *npj Biofilms and Microbiomes*. Nature Publishing Group, 7(1), pp. 1–13. doi: 10.1038/s41522-021-00196-6.

Babenko, V. V. *et al.* (2020) 'The ecogenomics of dsDNA bacteriophages in feces of stabled and feral horses', *Computational and Structural Biotechnology Journal*. Elsevier, 18, pp. 3457–3467. doi: 10.1016/J.CSBJ.2020.10.036.

Bächi, B. and Arber, W. (1977) 'Physical mapping of BglII, BamHI, EcoRI, HindIII and PstI Restriction fragments of bacteriophage P1 DNA', *Molecular and General Genetics MGG*, 153(3), pp. 311–324. doi: 10.1007/BF00431596.

Bailly-Bechet, M., Vergassola, M. and Rocha, E. (2007) 'Causes for the intriguing presence of tRNAs in phages', *Genome research*. 2007/09/04. Cold Spring Harbor Laboratory Press, 17(10), pp. 1486–1495. doi: 10.1101/gr.6649807.

Baker, M. *et al.* (2022) 'Antimicrobial Resistance in Dairy Slurry Tanks: A Critical Point for Measurement and Control', *SSRN Electronic Journal*. Cold Spring Harbor Laboratory, p. 2022.02.22.481441. doi: 10.2139/ssrn.4079732.

Balcazar, J. L. (2014) 'Bacteriophages as vehicles for antibiotic resistance genes in the environment', *PLoS pathogens*. Public Library of Science, 10(7), pp. e1004219–e1004219. doi: 10.1371/journal.ppat.1004219.

Balcázar, J. L. (2020) 'Implications of bacteriophages on the acquisition and spread of antibiotic resistance in the environment', *International Microbiology*. Springer. doi: 10.1007/s10123-020-00121-5.

Banks, D. J., Lei, B. and Musser, J. M. (2003) 'Prophage Induction and Expression of Prophage-Encoded Virulence Factors in Group A Streptococcus Serotype M3

Strain MGAS315', *Infection and Immunity*. American Society for Microbiology, 71(12), pp. 7079–7086. doi: 10.1128/IAI.71.12.7079-7086.2003/ASSET/DBD7C490-F635-417B-A144-7575FF1459BC/ASSETS/GRAPHIC/II1230693007.JPEG.

Barylski, J. *et al.* (2019) 'Analysis of Spounaviruses as a Case Study for the Overdue Reclassification of Tailed Phages', *Systematic Biology*, 0(0), pp. 1–14. doi: 10.1093/sysbio/syz036.

Beaulaurier, J. *et al.* (2020) 'Assembly-free single-molecule sequencing recovers complete virus genomes from natural microbial communities', *Genome Research*. Cold Spring Harbor Laboratory Press, 30(3), pp. 437–446. doi: 10.1101/gr.251686.119.

van Belkum, A. *et al.* (2015) 'Phylogenetic Distribution of CRISPR-Cas Systems in Antibiotic-Resistant <span class="named-content genus-species" id="named-content-1">Pseudomonas aeruginosa</span&gt';, *mBio*. Edited by J. Parkhill, 6(6), pp. e01796-15. doi: 10.1128/mBio.01796-15.

Beller, L. *et al.* (2022) 'The virota and its transkingdom interactions in the healthy infant gut', *Proceedings of the National Academy of Sciences of the United States of America*. NLM (Medline), 119(13). doi: 10.1073/PNAS.2114619119/SUPPL_FILE/PNAS.2114619119.SD12.XLSX.

Bench, S. R. *et al.* (2007) 'Metagenomic Characterization of Chesapeake Bay Virioplankton', *Applied and Environmental Microbiology*. American Society for Microbiology (ASM), 73(23), p. 7629. doi: 10.1128/AEM.00938-07.

Benjamini, Y. and Hochberg, Y. (1995) 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing', *Journal of the Royal Statistical Society: Series B (Methodological)*. John Wiley & Sons, Ltd, 57(1), pp. 289–300. doi: 10.1111/J.2517-6161.1995.TB02031.X.

Benler, S. *et al.* (2018) ‘A diversity-generating retroelement encoded by a globally ubiquitous Bacteroides phage 06 Biological Sciences 0605 Microbiology’, *Microbiome*. Microbiome, 6(1), pp. 1–10. doi: 10.1186/s40168-018-0573-6.

Berg Miller, M. E. *et al.* (2012) ‘Phage–bacteria relationships and CRISPR elements revealed by a metagenomic survey of the rumen microbiome’, *Environmental Microbiology*. John Wiley & Sons, Ltd, 14(1), pp. 207–227. doi: 10.1111/J.1462-2920.2011.02593.X.

Bertrand, D. *et al.* (2019) ‘Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes’, *Nature Biotechnology 2019 37:8*. Nature Publishing Group, 37(8), pp. 937–944. doi: 10.1038/s41587-019-0191-2.

Besler, I. *et al.* (2020) ‘Genome sequence and characterization of Coliphage vB_Eco_SLUR29’, *PHAGE*. Mary Ann Liebert Inc, 1(1), pp. 38–44. doi: 10.1089/phage.2019.0009.

Billard-Pomares, T. *et al.* (2014) ‘Characterization of a P1-like bacteriophage carrying an SHV-2 extended-spectrum β-lactamase from an Escherichia coli strain’, *Antimicrobial Agents and Chemotherapy*. American Society for Microbiology, 58(11), pp. 6550–6557. doi: 10.1128/AAC.03183-14/SUPPL_FILE/ZAC011143381SO1.PDF.

Billaud, M. *et al.* (2021) ‘Analysis of viromes and microbiomes from pig fecal samples reveals that phages and prophages rarely carry antibiotic resistance genes’, *ISME Communications 2021 1:1*. Nature Publishing Group, 1(1), pp. 1–10. doi: 10.1038/s43705-021-00054-8.

Billington, S. J., Johnston, J. L. and Rood, J. I. (1996) ‘Virulence regions and virulence factors of the ovine footrot pathogen, Dichelobacter nodosus’, *FEMS*

Microbiology Letters, 145(2), pp. 147–156. doi: 10.1111/j.1574-6968.1996.tb08570.x.

Biosolids Assurance Scheme (2020) *ABOUT BIOSOLIDS : Assured biosolids*. Available at: https://assuredbiosolids.co.uk/about-biosolids/ (Accessed: 22 July 2020).

Bjornsti, M. A., Reilly, B. E. and Anderson, D. L. (1983) 'Morphogenesis of bacteriophage phi 29 of Bacillus subtilis: oriented and quantized in vitro packaging of DNA protein gp3', *Journal of virology*, 45(1), pp. 383–396. doi: 10.1128/JVI.45.1.383-396.1983.

Black, L. W. (1989) 'DNA PACKAGING IN dsDNA BACTERIOPHAGES', *Annual Review of Microbiology*. Annual Reviews, 43(1), pp. 267–292. doi: 10.1146/annurev.mi.43.100189.001411.

Bloomfield, G. A. *et al.* (1997) 'Analysis of sequences flanking the vap regions of Dichelobacter nodosus: Evidence for multiple integration events, a killer system, and a new genetic element', *Microbiology*. Microbiology Society, 143(2), pp. 553–562. doi: 10.1099/00221287-143-2-553.

Boeckaerts, D. *et al.* (2021) 'Predicting bacteriophage hosts based on sequences of annotated receptor-binding proteins', *Scientific Reports 2021 11:1*. Nature Publishing Group, 11(1), pp. 1–14. doi: 10.1038/s41598-021-81063-4.

Van Boeckel, T. P. *et al.* (2017) 'Reducing antimicrobial use in food animals', *Science*. American Association for the Advancement of Science, 357(6358), pp. 1350–1352. doi: 10.1126/SCIENCE.AAO1495/SUPPL_FILE/AAO1496-VANBOECKEL-SM.PDF.

Bohannan, B. J. M. and Lenski, R. E. (2000) 'Linking genetic change to community evolution: Insights from studies of bacteria and bacteriophage', *Ecology Letters*. John Wiley & Sons, Ltd, pp. 362–377. doi: 10.1046/j.1461-0248.2000.00161.x.

Bolduc, B. *et al.* (2017) ‘vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect *Archaea* and *Bacteria*’, *PeerJ*, 5, p. e3243. doi: 10.7717/peerj.3243.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) ‘Trimmomatic: a flexible trimmer for Illumina sequence data’, *Bioinformatics*, 30(15), pp. 2114–2120. doi: 10.1093/bioinformatics/btu170.

Borsheim, K. Y., Bratbak, G. and Heldal, M. (1990) ‘Enumeration and biomass estimation of planktonic bacteria and viruses by transmission electron microscopy.’, *Applied and Environmental Microbiology*. American Society for Microbiology (ASM), 56(2), p. 352. doi: 10.1128/aem.56.2.352-356.1990.

Breitbart, M. *et al.* (2002) ‘Genomic analysis of uncultured marine viral communities’, *Proceedings of the National Academy of Sciences of the United States of America*. Proc Natl Acad Sci U S A, 99(22), pp. 14250–14255. doi: 10.1073/PNAS.202488399.

Breitbart, M. *et al.* (2003) ‘Metagenomic Analyses of an Uncultured Viral Community from Human Feces’, *Journal of Bacteriology*. American Society for Microbiology (ASM), 185(20), p. 6220. doi: 10.1128/JB.185.20.6220-6223.2003.

Breitbart, M. *et al.* (2007) ‘Exploring the vast diversity of marine viruses’, *Oceanography*, 20(SPL.ISS. 2), pp. 135–139. doi: 10.5670/oceanog.2007.58.

Breitbart, M. *et al.* (2008) ‘Viral diversity and dynamics in an infant gut’, *Research in microbiology*. Res Microbiol, 159(5), pp. 367–373. doi: 10.1016/J.RESMIC.2008.04.006.

Breitbart, M. *et al.* (2018) ‘Phage puppet masters of the marine microbial realm’, *Nature Microbiology*. Nature Publishing Group, pp. 754–766. doi: 10.1038/s41564-018-0166-y.

Breitbart, M. and Rohwer, F. (2005) ‘Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing’, *BioTechniques*. Biotechniques, 39(5), pp. 729–736. doi: 10.2144/000112019.

Brenciani, A. *et al.* (2010) ‘Φm46.1, the main Streptococcus pyogenes element carrying mef(A) and tet(O) genes’, *Antimicrobial Agents and Chemotherapy*. American Society for Microbiology, 54(1), pp. 221–229. doi: 10.1128/AAC.00499-09/ASSET/400CAAEA-D0E7-435A-9339-124005A885E2/ASSETS/GRAPHIC/ZAC0011086780003.JPEG.

Brown, C. L. *et al.* (2021) ‘Critical evaluation of short, long, and hybrid assembly for contextual analysis of antibiotic resistance genes in complex environmental metagenomes’, *Scientific Reports 2021 11:1*. Nature Publishing Group, 11(1), pp. 1–12. doi: 10.1038/s41598-021-83081-8.

Browne, P. D. *et al.* (2020) ‘GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms’, *GigaScience*. Oxford Academic, 9(2), pp. 1–14. doi: 10.1093/gigascience/giaa008.

Brum, J. R. *et al.* (2015) ‘Patterns and ecological drivers of ocean viral communities’, *Science*. American Association for the Advancement of Science, 348(6237). doi: 10.1126/science.1261498.

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

Brüssow, H., Canchaya, C. and Hardt, W.-D. (2004) ‘Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion’, *Microbiology and molecular biology reviews : MMBR*. American Society for

Microbiology, 68(3), pp. 560–602. doi: 10.1128/MMBR.68.3.560-602.2004.

Buck, D. *et al.* (2017) 'Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis', *F1000Research*. Faculty of 1000 Ltd, 6. doi: 10.12688/f1000research.10571.2.

Buckling, A. and Rainey, P. B. (2002) 'Antagonistic coevolution between a bacterium and a bacteriophage', *Proceedings of the Royal Society B: Biological Sciences*. The Royal Society, 269(1494), pp. 931–936. doi: 10.1098/rspb.2001.1945.

Bushnell, B. (2013) *BBMap download | SourceForge.net*. Available at: https://sourceforge.net/projects/bbmap/ (Accessed: 29 May 2020).

Camarillo-Guerrero, L. F. *et al.* (2021) 'Massive expansion of human gut bacteriophage diversity', *Cell*. Cell Press, 184(4), pp. 1098-1109.e9. doi: 10.1016/J.CELL.2021.01.029.

Canchaya, C. *et al.* (2003) 'Phage as agents of lateral gene transfer', *Current Opinion in Microbiology*. Elsevier Ltd, 6(4), pp. 417–424. doi: 10.1016/S1369-5274(03)00086-9.

Cao, J. *et al.* (2020) 'Profiling of Human Gut Virome with Oxford Nanopore Technology', *Medicine in Microecology*. Elsevier, 4, p. 100012. doi: 10.1016/J.MEDMIC.2020.100012.

Casjens, S. (2003) 'Prophages and bacterial genomics: what have we learned so far?', *Molecular Microbiology*. John Wiley & Sons, Ltd, 49(2), pp. 277–300. doi: https://doi.org/10.1046/j.1365-2958.2003.03580.x.

Casjens, S. R. and Hendrix, R. W. (2015) 'Bacteriophage lambda: Early pioneer and still relevant', *Virology*. Virology, 479–480, pp. 310–330. doi: 10.1016/J.VIROL.2015.02.010.

Castillo, D. *et al.* (2018) 'Exploring the genomic traits of non-toxigenic Vibrio

parahaemolyticus strains isolated in southern Chile', *Frontiers in Microbiology*. Frontiers Media S.A., 9(FEB), p. 161. doi: 10.3389/fmicb.2018.00161.

CHAWG (2020) *GB Cattle Health & Welfare Group Fifth Report*. Available at: https://ahdb.org.uk/knowledge-library/gb-cattle-health-welfare-group-fifth-report-2020.

Chelala, C. A. and Margolin, P. (1976) 'Evidence that HT mutant strains of bacteriophage P22 retain an altered form of substrate specificity in the formation of transducing particles in Salmonella typhimurium', *Genetical research*. Genet Res, 27(2), pp. 315–322. doi: 10.1017/S0016672300016505.

Chen, J. *et al.* (2018) 'Genome hypermobility by lateral transduction', *Science (New York, N.Y.)*. Science, 362(6411), pp. 207–212. doi: 10.1126/SCIENCE.AAT5867.

Chen, L. *et al.* (2016) 'VFDB 2016: hierarchical and refined dataset for big data analysis--10 years on', *Nucleic acids research*. 2015/11/17. Oxford University Press, 44(D1), pp. D694–D697. doi: 10.1093/nar/gkv1239.

Chen, S. *et al.* (2003) 'Value-Added Chemicals from Animal Manure', *Pacific Northwest National Laboratory*, PNNL-14495(December), pp. 1–142. doi: http://dx.doi.org/10.2172/15009485.

Cheung, M. S. *et al.* (2011) 'Systematic bias in high-throughput sequencing data and its correction by BEADS', *Nucleic Acids Research*. Oxford Academic, 39(15), pp. e103–e103. doi: 10.1093/nar/gkr425.

Chiang, Y. N., Penadés, J. R. and Chen, J. (2019) 'Genetic transduction by phages and chromosomal islands: The new and noncanonical', *PLoS Pathogens*. Public Library of Science, 15(8). doi: 10.1371/JOURNAL.PPAT.1007878.

Chibani, C. M. *et al.* (2019) 'Classifying the Unclassified: A Phage Classification Method', *Viruses*, 11(2), p. 195. doi: 10.3390/v11020195.

Chung, C. T., Niemela, S. L. and Miller, R. H. (1989) ‘One-step preparation of competent Escherichia coli: transformation and storage of bacterial cells in the same solution.’, *Proceedings of the National Academy of Sciences*. Proceedings of the National Academy of Sciences, 86(7), pp. 2172–2175. doi: 10.1073/PNAS.86.7.2172.

Clokie, M. R. *et al.* (2011) ‘Phages in nature’, *Bacteriophage*. Landes Bioscience, 1(1), pp. 31–45. doi: 10.4161/bact.1.1.14942.

Clokie, M. R. J. and Mann, N. H. (2006) ‘Marine cyanophages and light’, *Environmental Microbiology*. John Wiley & Sons, Ltd, 8(12), pp. 2074–2082. doi: 10.1111/j.1462-2920.2006.01171.x.

Clokie, M. R. J., Millard, A. D. and Mann, N. H. (2010) ‘T4 genes in the marine ecosystem: studies of the T4-like cyanophages and their role in marine ecology’, *Virology journal*. BioMed Central, 7, p. 291. doi: 10.1186/1743-422X-7-291.

Clooney, A. G. *et al.* (2019) ‘Whole-Virome Analysis Sheds Light on Viral Dark Matter in Inflammatory Bowel Disease’, *Cell Host & Microbe*, 26(6), pp. 764-778.e5. doi: https://doi.org/10.1016/j.chom.2019.10.009.

Cobián Güemes, A. G. *et al.* (2016) ‘Viruses as winners in the game of life’, *Annual Review of Virology*. Annual Reviews, 3(1), pp. 197–214. doi: 10.1146/annurev-virology-100114-054952.

Comeau, A. M. *et al.* (2008) ‘Exploring the prokaryotic virosphere’, *Research in Microbiology*. Elsevier Masson, 159(5), pp. 306–313. doi: 10.1016/J.RESMIC.2008.05.001.

Cook, K. L. *et al.* (2008) ‘Evaluation of the sulfate-reducing bacterial population associated with stored swine slurry’, *Anaerobe*. Academic Press, 14(3), pp. 172–180. doi: 10.1016/j.anaerobe.2008.03.003.

Cook, R., Hooton, S., *et al.* (2021) ‘Hybrid assembly of an agricultural slurry virome reveals a diverse and stable community with the potential to alter the metabolism and virulence of veterinary pathogens’, *Microbiome*, 9(1), p. 65. doi: 10.1186/s40168-021-01010-3.

Cook, R., Brown, N., *et al.* (2021) ‘INfrastructure for a PHAge REference Database: Identification of Large-Scale Biases in the Current Collection of Cultured Phage Genomes’, *Phage*. Cold Spring Harbor Laboratory, 2(4), pp. 214–223. doi: 10.1089/phage.2021.0007.

Coutinho, F. H. *et al.* (2021) ‘RaFAH: Host prediction for viruses of Bacteria and Archaea based on protein content’, *Patterns*. Elsevier, 2(7), p. 100274. doi: 10.1016/J.PATTER.2021.100274.

Cumby, N. *et al.* (2015) ‘The phage tail tape measure protein, an inner membrane protein and a periplasmic chaperone play connected roles in the genome injection process of E. coli phage HK97’, *Molecular Microbiology*. John Wiley & Sons, Ltd, 96(3), pp. 437–447. doi: https://doi.org/10.1111/mmi.12918.

Cuscó, A. *et al.* (2019) ‘Shallow metagenomics with Nanopore sequencing in canine fecal microbiota improved bacterial taxonomy and identified an uncultured CrAssphage’, *bioRxiv*. Cold Spring Harbor Laboratory, pp. 1–12. doi: 10.1101/585067.

Cuscó, A. *et al.* (2021) ‘Long-read metagenomics retrieves complete single-contig bacterial genomes from canine feces’, *BMC Genomics*. BioMed Central Ltd, 22(1), pp. 1–15. doi: 10.1186/S12864-021-07607-0/FIGURES/5.

D’Hérelle, F. (2007) ‘On an invisible microbe antagonistic toward dysenteric bacilli: brief note by Mr. F. D’Herelle, presented by Mr. Roux’, *Research in Microbiology*, 158(7), pp. 553–554. doi: https://doi.org/10.1016/j.resmic.2007.07.005.

Daly, M. M. *et al.* (2004) ‘Characterization and prevalence of MefA, MefE, and the associated msr(D) gene in Streptococcus pneumoniae clinical isolates’, *Journal of clinical microbiology*. American Society for Microbiology, 42(8), pp. 3570–3574. doi: 10.1128/JCM.42.8.3570-3574.2004.

Davies, R. and Wales, A. (2019) ‘Antimicrobial Resistance on Farms: A Review Including Biosecurity and the Potential Role of Disinfectants in Resistance Selection’, *Comprehensive Reviews in Food Science and Food Safety*. John Wiley & Sons, Ltd, 18(3), pp. 753–774. doi: 10.1111/1541-4337.12438.

Debroas, D. and Siguret, C. (2019) ‘Viruses as key reservoirs of antibiotic resistance genes in the environment’, *ISME Journal*. Nature Publishing Group, 13(11), pp. 2856–2867. doi: 10.1038/s41396-019-0478-9.

Delgado, B. *et al.* (2019) ‘Whole rumen metagenome sequencing allows classifying and predicting feed efficiency and intake levels in cattle’, *Scientific Reports*. Nature Publishing Group, 9(1), pp. 1–13. doi: 10.1038/s41598-018-36673-w.

Devoto, A. E. *et al.* (2019) ‘Megaphages infect Prevotella and variants are widespread in gut microbiomes’, *Nature Microbiology*. doi: 10.1038/s41564-018-0338-9.

Dinsdale, E. A. *et al.* (2008) ‘Functional metagenomic profiling of nine biomes’, *Nature*. Nature Publishing Group, 452(7187), pp. 629–632. doi: 10.1038/nature06810.

Dion, M. B., Oechslin, F. and Moineau, S. (2020) ‘Phage diversity, genomics and phylogeny’, *Nature Reviews Microbiology 2020 18:3*. Nature Publishing Group, 18(3), pp. 125–138. doi: 10.1038/s41579-019-0311-5.

Doster, E. *et al.* (2020) ‘MEGARes 2.0: A database for classification of antimicrobial drug, biocide and metal resistance determinants in metagenomic sequence data’,

Nucleic Acids Research. Nucleic Acids Res, 48(D1), pp. D561–D569. doi: 10.1093/nar/gkz1010.

Dunn, J. J., Studier, F. W. and Gottesman, M. (1983) 'Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements', *Journal of Molecular Biology*, 166(4), pp. 477–535. doi: https://doi.org/10.1016/S0022-2836(83)80282-4.

Dutilh, B. E. *et al.* (2012) 'Reference-independent comparative metagenomics using cross-assembly: crAss', *Bioinformatics*. Oxford Academic, 28(24), pp. 3225–3231. doi: 10.1093/BIOINFORMATICS/BTS613.

Dutilh, B. E. *et al.* (2014) 'A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes', *Nature Communications*, 5(1), p. 4498. doi: 10.1038/ncomms5498.

Eckstein, S. *et al.* (2021) 'Isolation and characterization of lytic phage TUN1 specific for Klebsiella pneumoniae K64 clinical isolates from Tunisia', *BMC Microbiology*. BioMed Central Ltd, 21(1), pp. 1–9. doi: 10.1186/S12866-021-02251-W/FIGURES/5.

Edgar, R. C. and Bateman, A. (2010) 'Search and clustering orders of magnitude faster than BLAST', *Bioinformatics*. Oxford Academic, 26(19), pp. 2460–2461. doi: 10.1093/BIOINFORMATICS/BTQ461.

Edwards, R. A. *et al.* (2019) 'Global phylogeography and ancient evolution of the widespread human gut virus crAssphage', *Nature Microbiology*, 4(10), pp. 1727–1736. doi: 10.1038/s41564-019-0494-6.

Edwards, R. A. and Rohwer, F. (2005) 'Viral metagenomics', *Nature Reviews Microbiology*, 3(6), pp. 504–510. doi: 10.1038/nrmicro1163.

Eklund, M. W. *et al.* (1974) 'Interspecies conversion of clostridium botulinum type C to Clostridium novyi type A by bacteriophage', *Science*. Science, 186(4162), pp.

456–458. doi: 10.1126/science.186.4162.456.

El-Gebali, S. *et al.* (2019) 'The Pfam protein families database in 2019', *Nucleic acids research*. Oxford University Press, 47(D1), pp. D427–D432. doi: 10.1093/nar/gky995.

Enault, F. *et al.* (2017) 'Phages rarely encode antibiotic resistance genes: A cautionary tale for virome analyses', *ISME Journal*. Nature Publishing Group, 11(1), pp. 237–247. doi: 10.1038/ismej.2016.90.

Fancello, L. *et al.* (2011) 'Bacteriophages and diffusion of genes encoding antimicrobial resistance in cystic fibrosis sputum microbiota', *The Journal of antimicrobial chemotherapy*. J Antimicrob Chemother, 66(11), pp. 2448–2454. doi: 10.1093/JAC/DKR315.

FAO (2020) *Livestock and environment statistics: manure and greenhouse gas emissions*. Available at: https://www.fao.org/documents/card/en/c/cb1922en/.

Feiss, M. *et al.* (1983) 'Structure of the bacteriophage lambda cohesive end site: location of the sites of terminase binding (cosB) and nicking (cosN)', *Gene*, 24(2), pp. 207–218. doi: https://doi.org/10.1016/0378-1119(83)90081-1.

Fierer, N. *et al.* (2007) 'Metagenomic and Small-Subunit rRNA Analyses Reveal the Genetic Diversity of Bacteria, Archaea, Fungi, and Viruses in Soil', *Applied and Environmental Microbiology*. American Society for Microbiology (ASM), 73(21), p. 7059. doi: 10.1128/AEM.00358-07.

Fillol-Salom, A. *et al.* (2021) 'Lateral transduction is inherent to the life cycle of the archetypical Salmonella phage P22', *Nature Communications 2021 12:1*. Nature Publishing Group, 12(1), pp. 1–12. doi: 10.1038/s41467-021-26520-4.

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

10.4161/BACT.28281.

Font-Palma, C. (2019) ‘Methods for the Treatment of Cattle Manure—A Review’, *C*. MDPI AG, 5(2), p. 27. doi: 10.3390/c5020027.

Fortier, L. C. and Sekulovic, O. (2013) ‘Importance of prophages to evolution and virulence of bacterial pathogens’, *Virulence*. Taylor and Francis Inc., 4(5), pp. 354–365. doi: 10.4161/viru.24498.

Freeman, V. J. (1951) ‘Studies on the virulence of bacteriophage-infected strains of Corynebacterium diphtheriae.’, *Journal of bacteriology*. American Society for Microbiology (ASM), 61(6), pp. 675–688. doi: 10.1128/JB.61.6.675-688.1951.

Fu, L. *et al.* (2012) ‘CD-HIT: accelerated for clustering the next-generation sequencing data’, *Bioinformatics*. 2012/10/11. Oxford University Press, 28(23), pp. 3150–3152. doi: 10.1093/bioinformatics/bts565.

Fung, S. *et al.* (2022) ‘Gauge your phage: Benchmarking of bacteriophage identification tools in metagenomic sequencing data’, *bioRxiv*. Cold Spring Harbor Laboratory, p. 2021.04.12.438782. doi: 10.1101/2021.04.12.438782.

Galiez, C. *et al.* (2017) ‘WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs’, *Bioinformatics*, 33(19), pp. 3113–3114. doi: 10.1093/bioinformatics/btx383.

Gao, S. M. *et al.* (2020) ‘Depth-related variability in viral communities in highly stratified sulfidic mine tailings’, *Microbiome*. BioMed Central, 8(1), p. 89. doi: 10.1186/s40168-020-00848-3.

Garmaeva, S. *et al.* (2019) ‘Studying the gut virome in the metagenomic era: Challenges and perspectives’, *BMC Biology*. BioMed Central Ltd., 17(1), pp. 1–14. doi: 10.1186/s12915-019-0704-y.

George, M. and Bukhari, A. I. (1981) ‘Heterogeneous host DNA attached to the left

end of mature bacteriophage Mu DNA', *Nature*, 292(5819), pp. 175–176. doi: 10.1038/292175a0.

Girons, I. S. *et al.* (2000) 'The LE1 bacteriophage replicates as a plasmid within Leptospira biflexa: construction of an L. biflexa-Escherichia coli shuttle vector', *Journal of bacteriology*. J Bacteriol, 182(20), pp. 5700–5705. doi: 10.1128/JB.182.20.5700-5705.2000.

GitHub - simroux/ClusterGenomes: Archive for ClusterGenomes scripts (no date). Available at: https://github.com/simroux/ClusterGenomes (Accessed: 29 May 2020).

Grazziotin, A. L., Koonin, E. V and Kristensen, D. M. (2017) 'Prokaryotic Virus Orthologous Groups (pVOGs): A resource for comparative genomics and protein family annotation', *Nucleic Acids Research*. Oxford University Press, 45(D1), pp. D491–D498. doi: 10.1093/nar/gkw975.

Gregory, A. C. *et al.* (2016) 'Genomic differentiation among wild cyanophages despite widespread horizontal gene transfer', *BMC genomics*. BioMed Central, 17(1), p. 930. doi: 10.1186/s12864-016-3286-x.

Gregory, A. C. *et al.* (2019) 'Marine DNA viral macro- and microdiversity from Pole to Pole', *Cell*, 177(5), pp. 1109-1123.e14. doi: 10.1016/j.cell.2019.03.040.

Gregory, A. C. *et al.* (2022) 'MetaPop: a pipeline for macro- and microdiversity analyses and visualization of microbial and viral metagenome-derived populations', *Microbiome*. BioMed Central Ltd, 10(1), pp. 1–19. doi: 10.1186/S40168-022-01231-0/FIGURES/5.

Griffiths, B. E., White, D. G. and Oikonomou, G. (2018) 'A cross-sectional study into the prevalence of dairy cattle lameness and associated herd-level risk factors in England and Wales', *Frontiers in Veterinary Science*. Frontiers Media S.A., 5(APR), p. 65. doi: 10.3389/FVETS.2018.00065/BIBTEX.

Guerin, E. *et al.* (2018) ‘Biology and taxonomy of crAss-like bacteriophages, the most abundant virus in the human gut’, *Cell Host and Microbe*. Cell Press, 24(5), pp. 653-664.e6. doi: 10.1016/j.chom.2018.10.002.

Guo, J. *et al.* (2017) ‘Copper Oxide Nanoparticles Induce Lysogenic Bacteriophage and Metal-Resistance Genes in Pseudomonas aeruginosa PAO1’, *ACS Applied Materials and Interfaces*. American Chemical Society, 9(27), pp. 22298–22307. doi: 10.1021/ACSAMI.7B06433/SUPPL_FILE/AM7B06433_SI_004.XLSX.

Haaber, J. *et al.* (2016) ‘Bacterial viruses enable their host to acquire antibiotic resistance genes from neighbouring cells’, *Nature Communications 2016 7:1*. Nature Publishing Group, 7(1), pp. 1–8. doi: 10.1038/ncomms13333.

Hackl, S. T., Harbig, T. A. and Nieselt, K. (2022) ‘Technical report on best practices for hybrid and long read de novo assembly of bacterial genomes utilizing Illumina and Oxford Nanopore Technologies reads’, *bioRxiv*. Cold Spring Harbor Laboratory, p. 2022.10.25.513682. doi: 10.1101/2022.10.25.513682.

Hanauer, D. I. *et al.* (2017) ‘An inclusive Research Education Community (iREC): Impact of the SEA-PHAGES program on research outcomes and student learning’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 114(51), pp. 13531–13536. doi: 10.1073/pnas.1718188115.

Hargreaves, K. R., Kropinski, A. M. and Clokie, M. R. (2014a) ‘Bacteriophage behavioral ecology: How phages alter their bacterial host’s habits’, *Bacteriophage*. Landes Bioscience, 4, pp. e29866–e29866. doi: 10.4161/bact.29866.

Hargreaves, K. R., Kropinski, A. M. and Clokie, M. R. (2014b) ‘What does the talking?: quorum sensing signalling genes discovered in a bacteriophage genome’, *PloS one*. Public Library of Science, 9(1), pp. e85131–e85131. doi:

10.1371/journal.pone.0085131.

Hatfull, G. F. *et al.* (2006) ‘Exploring the mycobacteriophage metaproteome: Phage genomics as an educational platform’, *PLoS Genetics*. doi: 10.1371/journal.pgen.0020092.

Hatfull, G. F. (2008) ‘Bacteriophage genomics’, *Current opinion in microbiology*. 2008/10/14, 11(5), pp. 447–453. doi: 10.1016/j.mib.2008.09.004.

Haudiquet, M. *et al.* (2022) ‘Selfish, promiscuous and sometimes useful: how mobile genetic elements drive horizontal gene transfer in microbial populations’, *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1861). doi: 10.1098/rstb.2021.0234.

Hayes, S. *et al.* (2017) ‘Metagenomic Approaches to Assess Bacteriophages in Various Environmental Niches’, *Viruses*. Multidisciplinary Digital Publishing Institute (MDPI), 9(6). doi: 10.3390/V9060127.

‘HMMER’ (no date). Available at: http://hmmer.org/ (Accessed: 29 May 2020).

Hobbs, Z. and Abedon, S. T. (2016) ‘Diversity of phage infection types and associated terminology: the problem with “Lytic or lysogenic”’, *FEMS microbiology letters*. FEMS Microbiol Lett, 363(7). doi: 10.1093/FEMSLE/FNW047.

Hockenberry, A. J. and Wilke, C. O. (2021) ‘BACPHLIP: Predicting bacteriophage lifestyle from conserved protein domains’, *PeerJ*. PeerJ Inc., 9. doi: 10.7717/PEERJ.11396/SUPP-1.

Howard-Varona, C. *et al.* (2017) ‘Lysogeny in nature: mechanisms, impact and ecology of temperate phages’, *The ISME Journal*. Nature Publishing Group, 11(7), p. 1511. doi: 10.1038/ISMEJ.2017.16.

Howard-Varona, C. *et al.* (2020) ‘Phage-specific metabolic reprogramming of virocells’, *ISME Journal*. Springer Nature, 14(4), pp. 881–895. doi: 10.1038/s41396-

019-0580-z.

Huang, D. *et al.* (2021) ‘Enhanced mutualistic symbiosis between soil phages and bacteria with elevated chromium-induced environmental stress’, *Microbiome*. BioMed Central Ltd, 9(1), pp. 1–15. doi: 10.1186/s40168-021-01074-1.

Huerta-Cepas, J. *et al.* (2018) ‘eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses’, *Nucleic Acids Research*, 47(D1), pp. D309–D314. doi: 10.1093/nar/gky1085.

Hurwitz, B. L., Brum, J. R. and Sullivan, M. B. (2015) ‘Depth-stratified functional and taxonomic niche specialization in the “core” and “flexible” Pacific Ocean Virome’, *ISME Journal*. Nature Publishing Group, 9(2), pp. 472–484. doi: 10.1038/ismej.2014.143.

Hurwitz, B. L., Hallam, S. J. and Sullivan, M. B. (2013) ‘Metabolic reprogramming by viruses in the sunlit and dark ocean’, *Genome Biology*. BioMed Central, 14(11), p. R123. doi: 10.1186/gb-2013-14-11-r123.

Hurwitz, B. L. and U’Ren, J. M. (2016) ‘Viral metabolic reprogramming in marine ecosystems’, *Current Opinion in Microbiology*. Elsevier Ltd, 31, pp. 161–168. doi: 10.1016/j.mib.2016.04.002.

Hyatt, D. *et al.* (2010) ‘Prodigal: Prokaryotic gene recognition and translation initiation site identification’, *BMC Bioinformatics*. BioMed Central, 11(1), pp. 1–11. doi: 10.1186/1471-2105-11-119/TABLES/5.

Iannelli, F. *et al.* (2014) ‘Nucleotide sequence of conjugative prophage Φ1207.3 (formerly Tn1207.3) carrying the mef(A)/msr(D) genes for efflux resistance to macrolides in Streptococcus pyogenes’, *Frontiers in Microbiology*. Frontiers Media S.A., 5(DEC), p. 687. doi: 10.3389/FMICB.2014.00687/BIBTEX.

Ibrahim, D. R. *et al.* (2016) 'Multidrug resistant, extended spectrum β-lactamase (ESBL)-producing Escherichia coli isolated from a dairy farm', *FEMS Microbiology Ecology*. Oxford Academic, 92(4). doi: 10.1093/FEMSEC/FIW013.

Iranzo, J., Krupovic, M. and Koonin, E. V. (2016) 'The Double-Stranded DNA Virosphere as a Modular Hierarchical Network of Gene Sharing', *mBio*. mBio, 7(4). doi: 10.1128/MBIO.00978-16.

Iyer, L. M. *et al.* (2021) 'Jumbo Phages: A Comparative Genomic Overview of Core Functions and Adaptions for Biological Conflicts', *Viruses 2021, Vol. 13, Page 63*. Multidisciplinary Digital Publishing Institute, 13(1), p. 63. doi: 10.3390/V13010063.

Bin Jang, H. *et al.* (2019) 'Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks', *Nature Biotechnology*. Nature Publishing Group, 37(6), pp. 632–639. doi: 10.1038/s41587-019-0100-8.

Ji, X. *et al.* (2016) 'A novel virulence-associated protein, vapE, in Streptococcus suis serotype 2', *Molecular Medicine Reports*. Spandidos Publications, 13(3), pp. 2871–2877. doi: 10.3892/mmr.2016.4818.

Jin, H. *et al.* (2022) 'Hybrid, ultra-deep metagenomic sequencing enables genomic and functional characterization of low-abundance species in the human gut microbiome', *Gut Microbes*. Taylor & Francis, 14(1). doi: 10.1080/19490976.2021.2021790.

Jin, M. *et al.* (2019) 'Diversities and potential biogeochemical impacts of mangrove soil viruses', *Microbiome*. BioMed Central Ltd., 7(1), pp. 1–15. doi: 10.1186/s40168-019-0675-9.

Juhala, R. J. *et al.* (2000) 'Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages11Edited by M. Gottesman', *Journal of Molecular Biology*, 299(1), pp. 27–51. doi:

https://doi.org/10.1006/jmbi.2000.3729.

Kalinski, A. and Black, L. W. (1986) ‘End structure and mechanism of packaging of bacteriophage T4 DNA’, *Journal of virology*. J Virol, 58(3), pp. 951–954. doi: 10.1128/JVI.58.3.951-954.1986.

Kanwar, N. *et al.* (2021) ‘PacBio sequencing output increased through uniform and directional fivefold concatenation’, *Scientific Reports 2021 11:1*. Nature Publishing Group, 11(1), pp. 1–13. doi: 10.1038/s41598-021-96829-z.

Keefe, G. P. (1997) ‘Streptococcus agalactiae mastitis: A review’, *Canadian Veterinary Journal*. Canadian Veterinary Medical Association, 38(7), pp. 429–437.

Keen, E. C. *et al.* (2017) ‘Novel “Superspreader” Bacteriophages Promote Horizontal Gene Transfer by Transformation’, *mBio*. American Society for Microbiology, 8(1), pp. e02115-16. doi: 10.1128/mBio.02115-16.

Kelley, L. A. *et al.* (2015) ‘The Phyre2 web portal for protein modeling, prediction and analysis’, *Nature Protocols*. Nature Publishing Group, 10(6), pp. 845–858. doi: 10.1038/nprot.2015.053.

Kenzaka, T. *et al.* (2007) ‘High-frequency phage-mediated gene transfer among Escherichia coli cells, determined at the single-cell level’, *Applied and environmental microbiology*. Appl Environ Microbiol, 73(10), pp. 3291–3299. doi: 10.1128/AEM.02890-06.

Khalil, R. K. S. *et al.* (2016) ‘Phage-mediated Shiga toxin (Stx) horizontal gene transfer and expression in non-Shiga toxigenic Enterobacter and Escherichia coli strains’, *Pathogens and Disease*. Pathog Dis, 74(5). doi: 10.1093/femspd/ftw037.

Khan, M. A. *et al.* (2016) ‘Invited review: Transitioning from milk to solid feed in dairy heifers’, *Journal of Dairy Science*. Elsevier, 99(2), pp. 885–902. doi: 10.3168/JDS.2015-9975.

Kieft, K. *et al.* (2022) ‘vRhyme enables binning of viral genomes from metagenomes’, *Nucleic Acids Research*. Oxford University Press (OUP), (1). doi: 10.1093/NAR/GKAC341.

Kieft, K. and Anantharaman, K. (2022) ‘Deciphering Active Prophages from Metagenomes’, *mSystems*. American Society for Microbiology, 7(2). doi: 10.1128/msystems.00084-22.

Kieft, K., Zhou, Z. and Anantharaman, K. (2020) ‘VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences’, *Microbiome*. Cold Spring Harbor Laboratory, 8(1), p. 90. doi: 10.1186/s40168-020-00867-0.

Kim, K. H. *et al.* (2008) ‘Amplification of uncultured single-stranded DNA viruses from rice paddy soil’, *Applied and Environmental Microbiology*, 74(19), pp. 5975–5985. doi: 10.1128/AEM.01275-08.

Kim, K. H. and Bae, J. W. (2011) ‘Amplification Methods Bias Metagenomic Libraries of Uncultured Single-Stranded and Double-Stranded DNA Viruses’, *Applied and Environmental Microbiology*. American Society for Microbiology (ASM), 77(21), p. 7663. doi: 10.1128/AEM.00289-11.

Kim, M. S. *et al.* (2011) ‘Diversity and abundance of single-stranded DNA viruses in human feces’, *Applied and Environmental Microbiology*. American Society for Microbiology, 77(22), pp. 8062–8070. doi: 10.1128/AEM.06331-11/SUPPL_FILE/AEM6331-11_SUPPLEMENTAL.PDF.

Kim, M. and Wells, J. E. (2016) ‘A meta-analysis of bacterial diversity in the feces of cattle’, *Current Microbiology*. Springer New York LLC, 72(2), pp. 145–151. doi: 10.1007/s00284-015-0931-6.

King, A. M. Q. *et al.* (2012) *Virus Taxonomy Classification and Nomenclature of*

Viruses Ninth Report of the International Committee on Taxonomy of Viruses, *Virus Taxonomy, Ninth Report of the International Committee on Taxonomy of Viruses*. Elsevier, Amsterdam.

Kleinheinz, K. A., Joensen, K. G. and Larsen, M. V. (2014) 'Applying the ResFinder and VirulenceFinder web-services for easy identification of acquired antibiotic resistance and E. coli virulence genes in bacteriophage and prophage nucleotide sequences', *Bacteriophage*. Taylor & Francis, 4(2), p. e27943. doi: 10.4161/bact.27943.

Ko, C. C. and Hatfull, G. F. (2020) 'Identification of mycobacteriophage toxic genes reveals new features of mycobacterial physiology and morphology', *Scientific Reports 2020 10:1*. Nature Publishing Group, 10(1), pp. 1–17. doi: 10.1038/s41598-020-71588-5.

Koboldt, D. C. *et al.* (2012) 'VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing', *Genome research*. 2012/02/02. Cold Spring Harbor Laboratory Press, 22(3), pp. 568–576. doi: 10.1101/gr.129684.111.

Kolmogorov, M. *et al.* (2019) 'Assembly of long, error-prone reads using repeat graphs', *Nature Biotechnology*. Nature Publishing Group, 37(5), pp. 540–546. doi: 10.1038/s41587-019-0072-8.

Kolmogorov, M. *et al.* (2020) 'metaFlye: scalable long-read metagenome assembly using repeat graphs', *Nature methods*. Nat Methods, 17(11), pp. 1103–1110. doi: 10.1038/S41592-020-00971-X.

Koonin, E. V. (1992) 'The second cholera toxin, Zot, and its plasmid-encoded and phage-encoded homologues constitute a group of putative ATPases with an altered purine NTP-binding motif', *FEBS Letters*. No longer published by Elsevier, 312(1), pp. 3–6. doi: 10.1016/0014-5793(92)81398-6.

Koonin, E. V. *et al.* (2020) ‘Global Organization and Proposed Megataxonomy of the Virus World’, *Microbiology and Molecular Biology Reviews : MMBR*. American Society for Microbiology (ASM), 84(2). doi: 10.1128/MMBR.00061-19.

Kovach, M. E. *et al.* (1995) ‘Four new derivatives of the broad-host-range cloning vector pBBR1MCS, carrying different antibiotic-resistance cassettes’, *Gene*. Gene, 166(1), pp. 175–176. doi: 10.1016/0378-1119(95)00584-1.

Kupritz, J. *et al.* (2021) ‘Isolation and characterization of a novel bacteriophage WO from Allonemobius socius crickets in Missouri’, *PLOS ONE*. Public Library of Science, 16(7), p. e0250051. doi: 10.1371/JOURNAL.PONE.0250051.

Lahart, B. *et al.* (2021) ‘Greenhouse gas emissions and nitrogen efficiency of dairy cows of divergent economic breeding index under seasonal pasture-based management’, *Journal of Dairy Science*. Elsevier, 104(7), pp. 8039–8049. doi: 10.3168/JDS.2020-19618.

Langmead, B. and Salzberg, S. L. (2012) ‘Fast gapped-read alignment with Bowtie 2’, *Nature methods*. NIH Public Access, 9(4), p. 357. doi: 10.1038/NMETH.1923.

Lasken, R. S. and Stockwell, T. B. (2007) ‘Mechanism of chimera formation during the Multiple Displacement Amplification reaction’, *BMC Biotechnology*. BioMed Central, 7, p. 19. doi: 10.1186/1472-6750-7-19.

Lavigne, R. *et al.* (2008) ‘Unifying classical and molecular taxonomic classification: analysis of the Podoviridae using BLASTP-based tools’, *Research in microbiology*. Res Microbiol, 159(5), pp. 406–414. doi: 10.1016/J.RESMIC.2008.03.005.

Lavigne, R. *et al.* (2009) ‘Classification of Myoviridae bacteriophages using protein sequence similarity’, *BMC microbiology*. BMC Microbiol, 9. doi: 10.1186/1471-2180-9-224.

Lawrence, J. G., Hatfull, G. F. and Hendrix, R. W. (2002) ‘Imbroglios of viral

taxonomy: genetic exchange and failings of phenetic approaches', *Journal of bacteriology*. J Bacteriol, 184(17), pp. 4891–4905. doi: 10.1128/JB.184.17.4891-4905.2002.

Lee, L. H. *et al.* (2006) 'Induction of temperate cyanophage AS-1 by heavy metal - Copper', *BMC Microbiology*. BioMed Central, 6(1), pp. 1–7. doi: 10.1186/1471-2180-6-17/TABLES/2.

Leidenfrost, R. M. *et al.* (2020) 'Benchmarking the MinION: Evaluating long reads for microbial profiling', *Scientific Reports*. Nature Publishing Group, 10(1), pp. 1–10. doi: 10.1038/s41598-020-61989-x.

Leite, D. M. C. *et al.* (2018) 'Computational prediction of inter-species relationships through omics data analysis and machine learning', *BMC Bioinformatics*. BioMed Central Ltd., 19(14), pp. 151–159. doi: 10.1186/S12859-018-2388-7/TABLES/3.

Leite, D. M. C. *et al.* (2019) 'Exploration of multiclass and one-class learning methods for prediction of phage-bacteria interaction at strain level', *Proceedings - 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018*. Institute of Electrical and Electronics Engineers Inc., pp. 1818–1825. doi: 10.1109/BIBM.2018.8621433.

Lekunberri, I. *et al.* (2017) 'Exploring the contribution of bacteriophages to antibiotic resistance', *Environmental Pollution*. Elsevier Ltd, 220, pp. 981–984. doi: 10.1016/j.envpol.2016.11.059.

Letunic, I. and Bork, P. (2019) 'Interactive Tree Of Life (iTOL) v4: recent updates and new developments', *Nucleic Acids Research*, 47(W1), pp. W256–W259. doi: 10.1093/nar/gkz239.

Li, D. *et al.* (2016) 'MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices', *Methods*. Academic

Press Inc., 102, pp. 3–11. doi: 10.1016/j.ymeth.2016.02.020.

Li, F. *et al.* (2019) 'Comparative metagenomic and metatranscriptomic analyses reveal the breed effect on the rumen microbiome and its associations with feed efficiency in beef cattle 06 Biological Sciences 0604 Genetics 06 Biological Sciences 0605 Microbiology', *Microbiome*. BioMed Central Ltd., 7(1), p. 6. doi: 10.1186/s40168-019-0618-5.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics (Oxford, England)*. 2009/06/08. Oxford University Press, 25(16), pp. 2078–2079. doi: 10.1093/bioinformatics/btp352.

Li, H. (2016) 'Minimap and miniasm: Fast mapping and de novo assembly for noisy long sequences', *Bioinformatics*. Oxford Academic, 32(14), pp. 2103–2110. doi: 10.1093/bioinformatics/btw152.

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

Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows–Wheeler transform', *Bioinformatics*. Oxford Academic, 25(14), pp. 1754–1760. doi: 10.1093/BIOINFORMATICS/BTP324.

Li, X. *et al.* (2022) 'Metagenomic evidence for co-occurrence of antibiotic, biocide and metal resistance genes in pigs', *Environment International*. Environ Int, 158. doi: 10.1016/j.envint.2021.106899.

Liang, G. *et al.* (2020) 'The stepwise assembly of the neonatal virome is modulated by breastfeeding', *Nature 2020 581:7809*. Nature Publishing Group, 581(7809), pp. 470–474. doi: 10.1038/s41586-020-2192-1.

Lim, E. S. *et al.* (2015) 'Early life dynamics of the human gut virome and bacterial microbiome in infants', *Nature Medicine*. Nature Publishing Group, 21(10), pp. 1228–

1234. doi: 10.1038/nm.3950.

Lindell, D. *et al.* (2004) ‘Transfer of photosynthesis genes to and from Prochlorococcus viruses’, *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 101(30), pp. 11013–11018. doi: 10.1073/PNAS.0401526101.

Liu, F. *et al.* (2016) ‘Zonula occludens toxins and their prophages in Campylobacter species’, *Gut Pathogens*. BioMed Central Ltd., 8(1), p. 43. doi: 10.1186/s13099-016-0125-1.

Liu, L. *et al.* (2020) ‘High-quality bacterial genomes of a partial-nitritation/anammox system by an iterative hybrid assembly method’, *Microbiome*. BioMed Central Ltd, 8(1), pp. 1–17. doi: 10.1186/S40168-020-00937-3/FIGURES/5.

Liu, L. *et al.* (2021) ‘Charting the complexity of the activated sludge microbiome through a hybrid sequencing strategy’, *Microbiome*. BioMed Central Ltd, 9(1), pp. 1–15. doi: 10.1186/S40168-021-01155-1/FIGURES/5.

Liu, M. *et al.* (2002) ‘Reverse transcriptase-mediated tropism switching in Bordetella bacteriophage’, *Science*, 295(5562), pp. 2091–2094. doi: 10.1126/science.1067467.

Loman, N. J. *et al.* (2012) ‘Performance comparison of benchtop high-throughput sequencing platforms’, *Nature Biotechnology 2012 30:5*. Nature Publishing Group, 30(5), pp. 434–439. doi: 10.1038/nbt.2198.

López-Bueno, A. *et al.* (2009) ‘High diversity of the viral community from an Antarctic lake’, *Science (New York, N.Y.)*. Science, 326(5954), pp. 858–861. doi: 10.1126/SCIENCE.1179287.

Low, S. J. *et al.* (2019) ‘Evaluation of a concatenated protein phylogeny for classification of tailed double-stranded DNA viruses belonging to the order Caudovirales’, *Nature microbiology*. Nat Microbiol, 4(8), pp. 1306–1315. doi:

10.1038/S41564-019-0448-Z.

Lu, C. *et al.* (2021) ‘Prokaryotic virus host predictor: a Gaussian model for host prediction of prokaryotic viruses in metagenomics’, *BMC Biology*, 19(1), p. 5. doi: 10.1186/s12915-020-00938-6.

Luong, T., Salabarria, A. C. and Roach, D. R. (2020) ‘Phage Therapy in the Resistance Era: Where Do We Stand and Where Are We Going?’, *Clinical therapeutics*. Clin Ther, 42(9), pp. 1659–1680. doi: 10.1016/J.CLINTHERA.2020.07.014.

Ma, Y. *et al.* (2018) ‘A human gut phage catalog correlates the gut phageome with type 2 diabetes’, *Microbiome*. doi: 10.1186/s40168-018-0410-y.

Mann, N. H. *et al.* (2003) ‘Bacterial photosynthesis genes in a virus’, *Nature 2003 424:6950*. Nature Publishing Group, 424(6950), pp. 741–741. doi: 10.1038/424741a.

Margulies, M. *et al.* (2005) ‘Genome sequencing in microfabricated high-density picolitre reactors’, *Nature 2005 437:7057*. Nature Publishing Group, 437(7057), pp. 376–380. doi: 10.1038/nature03959.

Marine, R. *et al.* (2014) ‘Caught in the middle with multiple displacement amplification: The myth of pooling for avoiding multiple displacement amplification bias in a metagenome’, *Microbiome*. BioMed Central Ltd., 2(1), pp. 1–8. doi: 10.1186/2049-2618-2-3/TABLES/3.

Martinez-Hernandez, F. *et al.* (2019) ‘Single-cell genomics uncover Pelagibacter as the putative host of the extremely abundant uncultured 37-F6 viral population in the ocean’, *The ISME journal*. 2018/09/18. Nature Publishing Group UK, 13(1), pp. 232–236. doi: 10.1038/s41396-018-0278-7.

McMurdie, P. J. and Holmes, S. (2013) ‘Phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data’, *PLoS ONE*. Edited by

M. Watson. Public Library of Science, 8(4), p. e61217. doi: 10.1371/journal.pone.0061217.

Meier-Kolthoff, J. P. and Göker, M. (2017) 'VICTOR: genome-based phylogeny and classification of prokaryotic viruses', *Bioinformatics*. Oxford Academic, 33(21), pp. 3396–3404. doi: 10.1093/BIOINFORMATICS/BTX440.

Van Metre, D. C. (2017) 'Pathogenesis and Treatment of Bovine Foot Rot', *The Veterinary clinics of North America. Food animal practice*. Vet Clin North Am Food Anim Pract, 33(2), pp. 183–194. doi: 10.1016/J.CVFA.2017.02.003.

Michniewski, S. *et al.* (2019) 'Riding the wave of genomics to investigate aquatic coliphage diversity and activity', *Environmental microbiology*. 2019/04/04. John Wiley & Sons, Inc., 21(6), pp. 2112–2128. doi: 10.1111/1462-2920.14590.

Michniewski, S. (2020) 'Phages infecting marine Vibrios: prevalence, diversity and role in the dissemination of antibiotic resistance genes'. Available at: http://webcat.warwick.ac.uk/record=b3599901 (Accessed: 25 October 2022).

Michniewski, S. *et al.* (2021) 'A new family of "megaphages" abundant in the marine environment', *ISME Communications 2021 1:1*. Nature Publishing Group, 1(1), pp. 1–4. doi: 10.1038/s43705-021-00064-6.

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

Millard, A. D. *et al.* (2009) 'Comparative genomics of marine cyanomyoviruses reveals the widespread occurrence of Synechococcus host genes localized to a hyperplastic region: Implications for mechanisms of cyanophage evolution', *Environmental Microbiology*. John Wiley & Sons, Ltd, 11(9), pp. 2370–2387. doi: 10.1111/j.1462-2920.2009.01966.x.

Miller, E. S. *et al.* (2003) ‘Bacteriophage T4 genome’, *Microbiology and molecular biology reviews : MMBR*. Microbiol Mol Biol Rev, 67(1), pp. 86–156. doi: 10.1128/MMBR.67.1.86-156.2003.

Minot, S. *et al.* (2011) ‘The human gut virome: Inter-individual variation and dynamic response to diet’, *Genome Research*. Genome Res, 21(10), pp. 1616–1625. doi: 10.1101/gr.122705.111.

Minot, S. *et al.* (2013) ‘Rapid evolution of the human gut virome’, *Proceedings of the National Academy of Sciences of the United States of America*. 2013/07/08. National Academy of Sciences, 110(30), pp. 12450–12455. doi: 10.1073/pnas.1300833110.

Mizuno, C. M. *et al.* (2013) ‘Reconstruction of Novel Cyanobacterial Siphovirus Genomes from Mediterranean Metagenomic Fosmids’, *Applied and Environmental Microbiology*. American Society for Microbiology (ASM), 79(2), p. 688. doi: 10.1128/AEM.02742-12.

Mizuno, C. M., Ghai, R. and Rodriguez-Valera, F. (2014) ‘Evidence for metaviromic islands in marine phages’, *Frontiers in Microbiology*. Frontiers Research Foundation, 5(FEB). doi: 10.3389/fmicb.2014.00027.

Modi, S. R. *et al.* (2013) ‘Antibiotic Treatment Expands the Resistance Reservoir and Ecological Network of the Phage Metagenome’, *Nature*. NIH Public Access, 499(7457), p. 219. doi: 10.1038/NATURE12212.

Mohanraj, U. *et al.* (2019) ‘A Toxicity Screening Approach to Identify Bacteriophage-Encoded Anti-Microbial Proteins’, *Viruses*. Multidisciplinary Digital Publishing Institute (MDPI), 11(11). doi: 10.3390/V11111057.

Monaco, C. L. *et al.* (2016) ‘Altered Virome and Bacterial Microbiome in Human Immunodeficiency Virus-Associated Acquired Immunodeficiency Syndrome’, *Cell Host and Microbe*. doi: 10.1016/j.chom.2016.02.011.

Monier, A. *et al.* (2017) ‘Host-derived viral transporter protein for nitrogen uptake in infected marine phytoplankton’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 114(36), pp. E7489–E7498. doi: 10.1073/pnas.1708097114.

Moon, K. *et al.* (2020) ‘Freshwater viral metagenome reveals novel and functional phage-borne antibiotic resistance genes’, *Microbiome*. NLM (Medline), 8(1), p. 75. doi: 10.1186/s40168-020-00863-4.

Moraru, C., Varsani, A. and Kropinski, A. M. (2020) ‘VIRIDIC—A Novel Tool to Calculate the Intergenomic Similarities of Prokaryote-Infecting Viruses’, *Viruses 2020, Vol. 12, Page 1268*. Multidisciplinary Digital Publishing Institute, 12(11), p. 1268. doi: 10.3390/V12111268.

Moreno-Gallego, J. L. *et al.* (2019) ‘Virome diversity correlates with intestinal microbiome diversity in adult monozygotic twins’, *Cell Host and Microbe*. Cell Press, 25(2), pp. 261-272.e5. doi: 10.1016/j.chom.2019.01.019.

Morse, M. L. (1954) ‘TRANSDUCTION OF CERTAIN LOCI IN ESCHERICHIA-COLI K-12’, in *Genetics*. 428 EAST PRESTON ST, BALTIMORE, MD 21202, pp. 984–985.

Morse, M. L., Lederberg, E. M. and Lederberg, J. (1956a) ‘Transduction in Escherichia Coli K-12’, *Genetics*. Oxford University Press, 41(1), p. 142. doi: 10.1093/genetics/41.1.142.

Morse, M. L., Lederberg, E. M. and Lederberg, J. (1956b) ‘Transductional Heterogenotes in Escherichia Coli’, *Genetics*. Oxford University Press, 41(5), p. 758. doi: 10.1093/genetics/41.5.758.

Munita, J. M. and Arias, C. A. (2016) ‘Mechanisms of Antibiotic Resistance’, *Microbiology spectrum*. NIH Public Access, 4(2), pp. 464–472. doi: 10.1128/MICROBIOLSPEC.VMBF-0016-2015.

Nakamura, T. *et al.* (2018) ‘Parallelization of MAFFT for large-scale multiple sequence alignments’, *Bioinformatics*, 34(14), pp. 2490–2492. doi: 10.1093/bioinformatics/bty121.

Nayfach, S. *et al.* (2020) ‘CheckV: assessing the quality of metagenome-assembled viral genomes’, *bioRxiv*. Cold Spring Harbor Laboratory, p. 2020.05.06.081778. doi: 10.1101/2020.05.06.081778.

Nayfach, S. *et al.* (2021) ‘Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome’, *Nature Microbiology 2021 6:7*. Nature Publishing Group, 6(7), pp. 960–970. doi: 10.1038/s41564-021-00928-6.

Neri, U. *et al.* (2022) ‘A five-fold expansion of the global RNA virome reveals multiple new clades of RNA bacteriophages’, *bioRxiv*. Cold Spring Harbor Laboratory, p. 2022.02.15.480533. doi: 10.1101/2022.02.15.480533.

Nguyen, L.-T. *et al.* (2015) ‘IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies’, *Molecular biology and evolution*. 2014/11/03. Oxford University Press, 32(1), pp. 268–274. doi: 10.1093/molbev/msu300.

Nicholls, S. M. *et al.* (2019) ‘Ultra-deep, long-read nanopore sequencing of mock microbial community standards’, *GigaScience*. Oxford Academic, 8(5), pp. 1–9. doi: 10.1093/gigascience/giz043.

Nishimura, Y. *et al.* (2017) ‘ViPTree: the viral proteomic tree server’, *Bioinformatics (Oxford, England)*. Bioinformatics, 33(15), pp. 2379–2380. doi: 10.1093/BIOINFORMATICS/BTX157.

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

Nurk, S. *et al.* (2017) ‘MetaSPAdes: A new versatile metagenomic assembler’,

Genome Research. Cold Spring Harbor Laboratory Press, 27(5), pp. 824–834. doi: 10.1101/GR.213959.116/-/DC1.

O'Brien, A. D. *et al.* (1984) 'Shiga-like toxin-converting phages from Escherichia coli strains that cause hemorrhagic colitis or infantile diarrhea', *Science*. doi: 10.1126/science.6387911.

O'Leary, N. A. *et al.* (2016) 'Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation', *Nucleic Acids Research*. Oxford University Press, 44(Database issue), p. D733. doi: 10.1093/NAR/GKV1189.

O'Neill, J. (2015) 'ANTIMICROBIALS IN AGRICULTURE AND THE ENVIRONMENT: REDUCING UNNECESSARY USE AND WASTE'.

Ofir, G. and Sorek, R. (2018) 'Contemporary Phage Biology: From Classic Models to New Insights', *Cell*. Cell Press, 172(6), pp. 1260–1270. doi: 10.1016/J.CELL.2017.10.045.

Oksanen, J. *et al.* (2020) 'vegan: Community Ecology Package'.

Oliver, S. P., Murinda, S. E. and Jayarao, B. M. (2011) 'Impact of antibiotic use in adult dairy cows on antimicrobial resistance of veterinary and human pathogens: a comprehensive review', *Foodborne pathogens and disease*. Foodborne Pathog Dis, 8(3), pp. 337–355. doi: 10.1089/FPD.2010.0730.

Olson, N. D. *et al.* (2019) 'Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes', *Briefings in bioinformatics*. Oxford University Press, 20(4), pp. 1140–1150. doi: 10.1093/bib/bbx098.

Ondov, B. D. *et al.* (2016) 'Mash: Fast genome and metagenome distance estimation using MinHash', *Genome Biology*. BioMed Central Ltd., 17(1), p. 132. doi:

10.1186/s13059-016-0997-x.

Orlova, E. V. (2012) ‘Bacteriophages and Their Structural Organisation’, *Bacteriophages*. IntechOpen. doi: 10.5772/34642.

Oude Munnink, B. B. *et al.* (2014) ‘Unexplained diarrhoea in HIV-1 infected individuals’, *BMC infectious diseases*. BioMed Central, 14, p. 22. doi: 10.1186/1471-2334-14-22.

Overholt, W. A. *et al.* (2020) ‘Inclusion of Oxford Nanopore long reads improves all microbial and viral metagenome-assembled genomes from a complex aquifer system’, *Environmental Microbiology*. John Wiley & Sons, Ltd, 22(9), pp. 4000–4013. doi: 10.1111/1462-2920.15186.

Paez-Espino, D. *et al.* (2016) ‘Uncovering Earth’s virome’, *Nature*. Nature Publishing Group, 536(7617), pp. 425–430. doi: 10.1038/nature19094.

Paez-Espino, D. *et al.* (2017) ‘IMG/VR: A database of cultured and uncultured DNA viruses and retroviruses’, *Nucleic Acids Research*. doi: 10.1093/nar/gkw1030.

Pal, C. *et al.* (2014) ‘BacMet: antibacterial biocide and metal resistance genes database’, *Nucleic acids research*. Nucleic Acids Res, 42(Database issue). doi: 10.1093/NAR/GKT1252.

Pal, C. *et al.* (2015) ‘Co-occurrence of resistance genes to antibiotics, biocides and metals reveals novel insights into their co-selection potential’, *BMC Genomics*. BioMed Central Ltd., 16(1), pp. 1–14. doi: 10.1186/S12864-015-2153-5/FIGURES/9.

Panis, G., Méjean, V. and Ansaldi, M. (2007) ‘Control and Regulation of KplE1 Prophage Site-specific Recombination: A NEW RECOMBINATION MODULE ANALYZED*’, *Journal of Biological Chemistry*, 282(30), pp. 21798–21809. doi: https://doi.org/10.1074/jbc.M701827200.

Park, J. and Kim, E. B. (2019) ‘Differences in microbiome and virome between cattle

and horses in the same farm', *Asian-Australasian Journal of Animal Sciences*. Asian-Australasian Association of Animal Production Societies (AAAP) and Korean Society of Animal Science and Technology (KSAST), 33(6), pp. 1042–1055. doi: 10.5713/AJAS.19.0267.

Park, K. S. *et al.* (2018) 'PNGM-1, a novel subclass B3 metallo-β-lactamase from a deep-sea sediment metagenome', *Journal of Global Antimicrobial Resistance*. Elsevier Ltd, 14, pp. 302–305. doi: 10.1016/j.jgar.2018.05.021.

Parmar, K. *et al.* (2018) 'An Insight into Phage Diversity at Environmental Habitats using Comparative Metagenomics Approach', *Current Microbiology*. Springer New York LLC, 75(2), pp. 132–141. doi: 10.1007/S00284-017-1357-0/FIGURES/4.

Payne, A. *et al.* (2019) 'BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files', *Bioinformatics*. Oxford University Press, 35(13), pp. 2193–2198. doi: 10.1093/bioinformatics/bty841.

Perez Sepulveda, B. *et al.* (2016) 'Marine phage genomics: the tip of the iceberg', *FEMS Microbiology Letters*. Edited by K. Hantke, 363(15), p. fnw158. doi: 10.1093/femsle/fnw158.

Peterson, C. B. and Mitloehner, F. M. (2021) 'Sustainability of the Dairy Industry: Emissions and Mitigation Opportunities', *Frontiers in Animal Science*. Frontiers, 0, p. 51. doi: 10.3389/FANIM.2021.760310.

Pfeifer, E., Bonnin, R. A. and Rocha, E. P. C. (2022) 'Phage-Plasmids Spread Antibiotic Resistance Genes through Infection and Lysogenic Conversion', *mBio*. American Society for Microbiology. doi: 10.1128/MBIO.01851-22/SUPPL_FILE/MBIO.01851-22-S0005.XLSX.

Polson, S. W., Wilhelm, S. W. and Wommack, K. E. (2011) 'Unraveling the viral tapestry (from inside the capsid out)', *The ISME journal*. ISME J, 5(2), pp. 165–168.

doi: 10.1038/ISMEJ.2010.81.

Prapaspongsa, T. *et al.* (2010) ‘LCA of comprehensive pig manure management incorporating integrated technology systems’, *Journal of Cleaner Production*, 18(14), pp. 1413–1422. doi: https://doi.org/10.1016/j.jclepro.2010.05.015.

Prosser, N. S. *et al.* (2020) ‘Serogroups of Dichelobacter nodosus, the cause of footrot in sheep, are randomly distributed across England’, *Scientific Reports 2020 10:1*. Nature Publishing Group, 10(1), pp. 1–13. doi: 10.1038/s41598-020-73750-5.

Pruss, G. J. and Calendar, R. (1978) ‘Maturation of bacteriophage P2 DNA’, *Virology*, 86(2), pp. 454–467. doi: https://doi.org/10.1016/0042-6822(78)90085-5.

Puxty, R. J. *et al.* (2016) ‘Viruses inhibit CO 2 fixation in the most abundant phototrophs on Earth’, *Current Biology*. Cell Press, 26(12), pp. 1585–1589.

Rakonjac, J. *et al.* (2011) ‘Filamentous bacteriophage: biology, phage display and nanotechnology applications’, *Current issues in molecular biology*. Curr Issues Mol Biol, 13(2), pp. 51–76. doi: 10.21775/cimb.013.051.

Ren, J. *et al.* (2017) ‘VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data’, *Microbiome*. Microbiome, 5(1), p. 69. doi: 10.1186/s40168-017-0283-5.

Ren, J. *et al.* (2018) ‘Identifying viruses from metagenomic data by deep learning’.

Ren, J. *et al.* (2020) ‘Identifying viruses from metagenomic data using deep learning’, *Quantitative Biology*. Higher Education Press, 8(1), pp. 64–77. doi: 10.1007/s40484-019-0187-4.

Reyes, A. *et al.* (2010) ‘Viruses in the faecal microbiota of monozygotic twins and their mothers’, *Nature*. NIH Public Access, 466(7304), pp. 334–338. doi: 10.1038/nature09199.

Reyes, A. *et al.* (2015) ‘Gut DNA viromes of Malawian twins discordant for severe

acute malnutrition', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.1514285112.

Rezaei Javan, R. *et al.* (2019) 'Prophages and satellite prophages are widespread in Streptococcus and may play a role in pneumococcal pathogenesis', *Nature Communications*. Nature Publishing Group, 10(1), pp. 1–14. doi: 10.1038/s41467-019-12825-y.

Rihtman, B. *et al.* (2016) 'Assessing Illumina technology for the high-throughput sequencing of bacteriophage genomes.', *PeerJ*. PeerJ, 4, p. e2055. doi: 10.7717/peerj.2055.

Rihtman, B. *et al.* (2019) 'Cyanophage MazG is a pyrophosphohydrolase but unable to hydrolyse magic spot nucleotides', *Environmental Microbiology Reports*. Wiley-Blackwell, 11(3), pp. 448–455. doi: 10.1111/1758-2229.12741.

Rodriguez-Brito, B. *et al.* (2010) 'Viral and microbial community dynamics in four aquatic environments', *The ISME Journal 2010 4:6*. Nature Publishing Group, 4(6), pp. 739–751. doi: 10.1038/ismej.2010.1.

Rohwer, F. and Edwards, R. (2002) 'The phage proteomic tree: A genome-based taxonomy for phage', *Journal of Bacteriology*. doi: 10.1128/JB.184.16.4529-4535.2002.

Rolain, J. M. *et al.* (2011) 'Bacteriophages as vehicles of the resistome in cystic fibrosis', *The Journal of antimicrobial chemotherapy*. J Antimicrob Chemother, 66(11), pp. 2444–2447. doi: 10.1093/JAC/DKR318.

Romero, P. *et al.* (2009) 'Comparative genomic analysis of ten Streptococcus pneumoniae temperate bacteriophages', *Journal of Bacteriology*. American Society for Microbiology (ASM), 191(15), pp. 4854–4862. doi: 10.1128/JB.01272-08.

Ross, E. M. *et al.* (2013) 'Metagenomics of rumen bacteriophage from thirteen

lactating dairy cattle.', *BMC microbiology*. BioMed Central, 13(1), pp. 1–11. doi: 10.1186/1471-2180-13-242/TABLES/3.

Roux, S. *et al.* (2012) 'Assessing the Diversity and Specificity of Two Freshwater Viral Communities through Metagenomics', *PLOS ONE*. Public Library of Science, 7(3), p. e33641. doi: 10.1371/JOURNAL.PONE.0033641.

Roux, S. *et al.* (2014) 'Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics', *eLife*. eLife Sciences Publications Ltd, 2014(3). doi: 10.7554/ELIFE.03125.001.

Roux, S. *et al.* (2015) 'VirSorter: mining viral signal from microbial genomic data', *PeerJ*, 3, p. e985. doi: 10.7717/peerj.985.

Roux, S., Brum, J. R., *et al.* (2016) 'Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses', *Nature*. Nature Publishing Group, 537(7622), pp. 689–693. doi: 10.1038/nature19366.

Roux, S., Solonenko, N. E., *et al.* (2016) 'Towards quantitative viromics for both double-stranded and single-stranded DNA viruses', *PeerJ*. PeerJ Inc., 4, pp. e2777–e2777. doi: 10.7717/peerj.2777.

Roux, S. *et al.* (2017) 'Benchmarking viromics: An in silico evaluation of metagenome-enabled estimates of viral community composition and diversity', *PeerJ*, 2017(9), p. e3817. doi: 10.7717/peerj.3817.

Roux, S. *et al.* (2021) 'IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses', *Nucleic Acids Research*. Oxford Academic, 49(D1), pp. D764–D775. doi: 10.1093/NAR/GKAA946.

Roux, S. *et al.* (2022) 'iPHoP: an integrated machine-learning framework to maximize host prediction for metagenome-assembled virus genomes', *bioRxiv*. Cold Spring Harbor Laboratory, p. 2022.07.28.501908. doi: 10.1101/2022.07.28.501908.

Ruan, J. and Li, H. (2020) ‘Fast and accurate long-read assembly with wtdbg2’, *Nature Methods*. Nature Publishing Group, 17(2), pp. 155–158. doi: 10.1038/s41592-019-0669-3.

Rückert, C. (2016) ‘Sulfate reduction in microorganisms — recent advances and biotechnological applications’, *Current Opinion in Microbiology*. Elsevier Ltd, pp. 140–146. doi: 10.1016/j.mib.2016.07.007.

Ruegg, P. L. and Petersson-Wolfe, C. S. (2018) ‘Mastitis in Dairy Cows’, *Veterinary Clinics of North America - Food Animal Practice*, pp. ix–x. doi: 10.1016/j.cvfa.2018.08.001.

Ruohan, W. *et al.* (2022) ‘DeepHost: phage host prediction with convolutional neural network’, *Briefings in Bioinformatics*. Oxford Academic, 23(1), pp. 1–10. doi: 10.1093/BIB/BBAB385.

Russel, M. and Model, P. (2006) ‘Filamentous bacteriophages’, in Abedon, S. T. (ed.) *The Bacteriophages*. 2nd edn. Oxford University Press, pp. 140–160.

Sakaguchi, M. (2011) ‘Practical Aspects of the Fertility of Dairy Cattle’, *Journal of Reproduction and Development*. The Society for Reproduction and Development, 57(1), pp. 17–33. doi: 10.1262/JRD.10-197E.

Salmond, G. P. C. and Fineran, P. C. (2012) ‘A century of the phage: past, present and future’, in *Bacteriophages*. InTechOpen, pp. 777–786. doi: 10.1038/nrmicro3564.

Sandars, D. L. *et al.* (2003) ‘Environmental benefits of livestock manure management practices and technology by life cycle assessment’, *Biosystems Engineering*, 84(3), pp. 267–281. doi: https://doi.org/10.1016/S1537-5110(02)00278-7.

Sanger, F. *et al.* (1977) ‘Nucleotide sequence of bacteriophage φx174 DNA’, *Nature*,

265(5596), pp. 687–695. doi: 10.1038/265687a0.

Santoro, F. *et al.* (2022) ‘Streptococcus pyogenes ϕ1207.3 is a temperate bacteriophage carrying the macrolide efflux gene pair mef(A)-msr(D) and capable to lysogenise different Streptococci’, *bioRxiv*. Cold Spring Harbor Laboratory, p. 2022.10.13.512196. doi: 10.1101/2022.10.13.512196.

Sato, M. P. *et al.* (2019) ‘Comparison of the sequencing bias of currently available library preparation kits for Illumina sequencing of bacterial genomes and metagenomes’, *DNA research : an international journal for rapid publication of reports on genes and genomes*. DNA Res, 26(5), pp. 391–398. doi: 10.1093/DNARES/DSZ017.

Sayers, E. W. *et al.* (2020) ‘GenBank’, *Nucleic Acids Research*. Oxford University Press, 48(D1), p. D84. doi: 10.1093/NAR/GKZ956.

Sazinas, P. *et al.* (2018) ‘Comparative genomics of bacteriophage of the genus Seuratvirus’, *Genome Biology and Evolution*. Oxford University Press, 10(1), pp. 72–76. doi: 10.1093/gbe/evx275.

Sazinas, P. *et al.* (2019) ‘Metagenomics of the viral community in three cattle slurry samples’, *Microbiology Resource Announcements*. Am Soc Microbiol, 8(7), pp. e01442-18. doi: 10.1128/mra.01442-18.

Schmidt, E., Kelly, S. M. and van der Walle, C. F. (2007) ‘Tight junction modulation and biochemical characterisation of the zonula occludens toxin C-and N-termini’, *FEBS Letters*. No longer published by Elsevier, 581(16), pp. 2974–2980. doi: 10.1016/j.febslet.2007.05.051.

Schmieger, H. (1982) ‘Packaging signals for phage P22 on the chromosome of Salmonella typhimurium’, *Molecular and General Genetics MGG 1982 187:3*. Springer, 187(3), pp. 516–518. doi: 10.1007/BF00332637.

Seemann, T. (2014) ‘Prokka: Rapid prokaryotic genome annotation’, *Bioinformatics*, 30(14), pp. 2068–2069. doi: 10.1093/bioinformatics/btu153.

Seemann, T. (no date a) ‘Abricate’. Github.

Seemann, T. (no date b) *snippy: Rapid haploid variant calling and core genome alignment*, *2015*. Available at: https://github.com/tseemann/snippy (Accessed: 29 May 2020).

Seiffert, S. N. *et al.* (2013) ‘Extended-spectrum cephalosporin-resistant gram-negative organisms in livestock: An emerging problem for human health?’, *Drug Resistance Updates*. Churchill Livingstone, 16(1–2), pp. 22–45. doi: 10.1016/J.DRUP.2012.12.001.

Shan, T. *et al.* (2011) ‘The Fecal Virome of Pigs on a High-Density Farm’, *Journal of Virology*. American Society for Microbiology (ASM), 85(22), p. 11697. doi: 10.1128/JVI.05217-11.

Shannon, P. *et al.* (2003) ‘Cytoscape: A software environment for integrated models of biomolecular interaction networks’, *Genome Research*. Cold Spring Harbor Laboratory Press, 13(11), pp. 2498–2504. doi: 10.1101/gr.1239303.

Sharon, I. *et al.* (2011) ‘Comparative metagenomics of microbial traits within oceanic viral communities’, *ISME Journal*. Nature Publishing Group, 5(7), pp. 1178–1190. doi: 10.1038/ismej.2011.2.

Shaw, L. M. *et al.* (2019) ‘DirtyGenes: testing for significant changes in gene or bacterial population compositions from a small number of samples’, *Scientific Reports*. Nature Publishing Group, 9(1), pp. 1–10. doi: 10.1038/s41598-019-38873-4.

Sherrill-Mix, S. (2018) ‘taxonomizr: Functions to Work with NCBI Accessions and Taxonomy’, *R package version 0.5.1*. Available at: https://cran.r-

project.org/web/packages/taxonomizr/ (Accessed: 29 May 2020).

Shkoporov, A. N. *et al.* (2018) 'ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis', *Nature Communications*. Nature Publishing Group, 9(1), pp. 1–8. doi: 10.1038/s41467-018-07225-7.

Shkoporov, A. N. *et al.* (2019) 'The human gut virome is highly diverse, stable, and individual specific', *Cell Host and Microbe*. Elsevier Inc., 26(4), pp. 527-541.e5. doi: 10.1016/j.chom.2019.09.009.

Shousha, A. *et al.* (2015) 'Bacteriophages Isolated from Chicken Meat and the Horizontal Transfer of Antimicrobial Resistance Genes', *Applied and environmental microbiology*. Appl Environ Microbiol, 81(14), pp. 4600–4606. doi: 10.1128/AEM.00872-15.

Smith, K. A. and Williams, A. G. (2016) 'Production and management of cattle manure in the UK and implications for land application practice', *Soil Use and Management*. Edited by F. Nicholson. Blackwell Publishing Ltd, 32, pp. 73–82. doi: 10.1111/sum.12247.

Smith, R. *et al.* (2015) 'Draft genome sequences of 14 Escherichia coli phages isolated from cattle slurry', *Genome Announcements*. American Society for Microbiology, 3(6), pp. e01364-15. doi: 10.1128/genomeA.01364-15.

Song, Y. *et al.* (2021) 'Characterization of a Novel Group of Listeria Phages That Target Serotype 4b Listeria monocytogenes', *Viruses*. Viruses, 13(4). doi: 10.3390/V13040671.

Spruit, C. M. *et al.* (2020) 'Discovery of Three Toxic Proteins of Klebsiella Phage fHe-Kpn01', *Viruses*. Multidisciplinary Digital Publishing Institute (MDPI), 12(5). doi: 10.3390/V12050544.

St-Pierre, B. and Wright, A. D. G. (2017) ‘Implications from distinct sulfate-reducing bacteria populations between cattle manure and digestate in the elucidation of H2S production during anaerobic digestion of animal slurry’, *Applied Microbiology and Biotechnology*. Springer Verlag, 101(13), pp. 5543–5556. doi: 10.1007/s00253-017-8261-1.

Steen, A. D. *et al.* (2019) ‘High proportions of bacteria and archaea across most biomes remain uncultured’, *The ISME Journal 2019 13:12*. Nature Publishing Group, 13(12), pp. 3126–3130. doi: 10.1038/s41396-019-0484-y.

Su, L. K. *et al.* (2010) ‘Lysogenic infection of a Shiga toxin 2-converting bacteriophage changes host gene expression, enhances host acid resistance and motility’, *Molecular Biology*, 44(1), pp. 54–66. doi: 10.1134/S0026893310010085.

Subirats, J. *et al.* (2016) ‘Metagenomic analysis reveals that bacteriophages are reservoirs of antibiotic resistance genes’, *International journal of antimicrobial agents*. Int J Antimicrob Agents, 48(2), pp. 163–167. doi: 10.1016/J.IJANTIMICAG.2016.04.028.

Sullivan, M. B. *et al.* (2010) ‘Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments’, *Environmental microbiology*. Blackwell Publishing Ltd, 12(11), pp. 3035–3056. doi: 10.1111/j.1462-2920.2010.02280.x.

Suttle, C. A. (2007) ‘Marine viruses--major players in the global ecosystem.’, *Nature reviews. Microbiology*, 5(10), pp. 801–12. doi: 10.1038/nrmicro1750.

Sutton, T. D. S. *et al.* (2019) ‘Choice of assembly software has a critical impact on virome characterisation’, *Microbiome*. BioMed Central Ltd., 7(1), pp. 1–15. doi: 10.1186/S40168-019-0626-5/FIGURES/4.

Sutton, T. D. S. and Hill, C. (2019) ‘Gut bacteriophage: Current understanding and

challenges', *Frontiers in endocrinology*. Frontiers Media S.A., 10, p. 784. doi: 10.3389/fendo.2019.00784.

Swanson, M. M. *et al.* (2009) 'Viruses in soils: morphological diversity and abundance in the rhizosphere', *Annals of Applied Biology*. John Wiley & Sons, Ltd, 155(1), pp. 51–60. doi: 10.1111/J.1744-7348.2009.00319.X.

Tatusov, R. L. *et al.* (2000) 'The COG database: a tool for genome-scale analysis of protein functions and evolution', *Nucleic acids research*. Oxford University Press, 28(1), pp. 33–36. doi: 10.1093/nar/28.1.33.

Team, R. C. (2018) 'R: A language and environment for statistical computing'. Vienna: R Foundation for Statistical Computing. Available at: https://www.r-project.org/.

Temperton, B. and Giovannoni, S. J. (2012) 'Metagenomics: Microbial diversity through a scratched lens', *Current Opinion in Microbiology*. Elsevier Current Trends, pp. 605–612. doi: 10.1016/j.mib.2012.07.001.

Terzian, P. *et al.* (2021) 'PHROG: families of prokaryotic virus proteins clustered using remote homology', *NAR Genomics and Bioinformatics*. Oxford Academic, 3(3). doi: 10.1093/NARGAB/LQAB067.

Thierauf, A., Perez, G. and Maloy, A. S. (2009) 'Generalized transduction', *Methods in molecular biology (Clifton, N.J.)*. Methods Mol Biol, 501, pp. 267–286. doi: 10.1007/978-1-60327-164-6_23.

Thomassen, M. A. *et al.* (2008) 'Life cycle assessment of conventional and organic milk production in the Netherlands', *Agricultural Systems*, 96(1), pp. 95–107. doi: https://doi.org/10.1016/j.agsy.2007.06.001.

Thrash, J. C. (2021) 'Towards culturing the microbe of your choice', *Environmental Microbiology Reports*. John Wiley & Sons, Ltd, 13(1), pp. 36–41. doi: 10.1111/1758-

2229.12898.

Tsao, Y.-F. *et al.* (2018) ‘Phage Morons Play an Important Role in Pseudomonas aeruginosa Phenotypes’, *Journal of bacteriology*. American Society for Microbiology, 200(22), pp. e00189-18. doi: 10.1128/JB.00189-18.

Twort, F. W. (1915) ‘AN INVESTIGATION ON THE NATURE OF ULTRA-MICROSCOPIC VIRUSES.’, *The Lancet*. Elsevier, 186(4814), pp. 1241–1243. doi: 10.1016/S0140-6736(01)20383-3.

Tye, B.-K., Huberman, J. A. and Botstein, D. (1974) ‘Non-random circular permutation of phage P22 DNA’, *Journal of Molecular Biology*, 85(4), pp. 501–527. doi: https://doi.org/10.1016/0022-2836(74)90312-X.

UK Government (2013) ‘Nitrate Vulnerable Zones (NVZs)’, *European Commission Nitrates Directive*, p. 1996. Available at: https://www.gov.uk/government/collections/nitrate-vulnerable-zones (Accessed: 19 June 2020).

UK Government (2016) *Handling of manure and slurry to reduce antibiotic resistance - GOV.UK*. Available at: https://www.gov.uk/guidance/handling-of-manure-and-slurry-to-reduce-antibiotic-resistance (Accessed: 23 February 2022).

UK Government (no date) *Use organic manures and manufactured fertilisers on farmland*. Available at: https://www.gov.uk/government/publications/nitrates-and-phosphates-plan-organic-fertiliser-and-manufactured-fertiliser-use/use-organic-manures-and-manufactured-fertilisers-on-farmland.

Unterer, M., Khan Mirzaei, M. and Deng, L. (2021) ‘Gut Phage Database: phage mining in the cave of wonders’, *Signal Transduction and Targeted Therapy 2021 6:1*. Nature Publishing Group, 6(1), pp. 1–2. doi: 10.1038/s41392-021-00615-2.

Vaser, R. *et al.* (2017) ‘Fast and accurate de novo genome assembly from long

uncorrected reads', *Genome Research*. Genome Res, 27(5), pp. 737–746. doi: 10.1101/gr.214270.116.

Villarroel, J. *et al.* (2016) 'HostPhinder: A Phage Host Prediction Tool', *Viruses*. Multidisciplinary Digital Publishing Institute (MDPI), 8(5). doi: 10.3390/V8050116.

De Vries, J. W., Groenestein, C. M. and De Boer, I. J. M. (2012) 'Environmental consequences of processing manure to produce mineral fertilizer and bio-energy', *Journal of Environmental Management*, 102, pp. 173–183. doi: https://doi.org/10.1016/j.jenvman.2012.02.032.

Wagner, P. L. *et al.* (2002) 'Bacteriophage control of Shiga toxin 1 production and release by Escherichia coli', *Molecular Microbiology*. John Wiley & Sons, Ltd, 44(4), pp. 957–970. doi: 10.1046/j.1365-2958.2002.02950.x.

Waldor, M. K. and Mekalanos, J. J. (1996) 'Lysogenic Conversion by a Filamentous Phage Encoding Cholera Toxin', *Science*. doi: 10.1126/science.272.5270.1910.

Walker, B. J. *et al.* (2014) 'Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement', *PLoS ONE*, 9(11), p. e112963. doi: 10.1371/journal.pone.0112963.

Walker, P. J. *et al.* (2021) 'Changes to virus taxonomy and to the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2021)', *Archives of Virology*. Springer, 166(9), pp. 2633–2648. doi: 10.1007/S00705-021-05156-1/TABLES/1.

Wang, Jianbin *et al.* (2005) 'Complete genome sequence of bacteriophage T5', *Virology*, 332(1), pp. 45–65. doi: https://doi.org/10.1016/j.virol.2004.10.049.

Wang, X. *et al.* (2010) 'Cryptic prophages help bacteria cope with adverse environments', *Nature communications*. Nat Commun, 1(9). doi: 10.1038/NCOMMS1146.

Wang, Yunhao *et al.* (2021) ‘Nanopore sequencing technology, bioinformatics and applications’, *Nature Biotechnology 2021 39:11*. Nature Publishing Group, 39(11), pp. 1348–1365. doi: 10.1038/s41587-021-01108-x.

Warwick-Dugdale, J. *et al.* (2019) ‘Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands’, *PeerJ*, 2019(4). doi: 10.7717/peerj.6800.

Watson, M. and Warr, A. (2019) ‘Errors in long-read assemblies can critically affect protein prediction’, *Nature Biotechnology*, 37(2), pp. 124–126. doi: 10.1038/s41587-018-0004-z.

Whist, A. C., Østerås, O. and Sølverød, L. (2007) ‘Streptococcus dysgalactiae isolates at calving and lactation performance within the same lactation’, *Journal of Dairy Science*. American Dairy Science Association, 90(2), pp. 766–778. doi: 10.3168/jds.S0022-0302(07)71561-8.

Wichmann, F. *et al.* (2014) ‘Diverse antibiotic resistance genes in dairy cow manure’, *mBio*. American Society for Microbiology, 5(2). doi: 10.1128/MBIO.01017-13/SUPPL_FILE/MBO002141797ST3.DOCX.

Wick, R. R. *et al.* (2017) ‘Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads’, *PLoS computational biology*. PLoS Comput Biol, 13(6). doi: 10.1371/JOURNAL.PCBI.1005595.

Wick, R. R. and Holt, K. E. (2021) ‘Benchmarking of long-read assemblers for prokaryote whole genome sequencing’, *F1000Research*. F1000Res, 8, p. 2138. doi: 10.12688/f1000research.21782.4.

Wick, R. R. and Holt, K. E. (2022) ‘Polypolish: Short-read polishing of long-read bacterial genome assemblies’, *PLOS Computational Biology*. Public Library of Science, 18(1), p. e1009802. doi: 10.1371/JOURNAL.PCBI.1009802.

Wigington, C. H. *et al.* (2016) ‘Re-examination of the relationship between marine virus and microbial cell abundances’, *Nature microbiology*. Nat Microbiol, 1(3). doi: 10.1038/NMICROBIOL.2015.24.

Wipf, J. R. K., Schwendener, S. and Perreten, V. (2014) ‘The novel macrolide-lincosamide-streptogramin B resistance gene erm(44) is associated with a prophage in Staphylococcus xylosus’, *Antimicrobial Agents and Chemotherapy*. American Society for Microbiology, 58(10), pp. 6133–6138. doi: 10.1128/AAC.02949-14/SUPPL_FILE/ZAC010143332SO1.PDF.

Wittmann, J. *et al.* (2020) ‘From Orphan Phage to a Proposed New Family–The Diversity of N4-Like Viruses’, *Antibiotics 2020, Vol. 9, Page 663*. Multidisciplinary Digital Publishing Institute, 9(10), p. 663. doi: 10.3390/ANTIBIOTICS9100663.

Wolf, Y. I. *et al.* (2020) ‘Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome’, *Nature Microbiology 2020 5:10*. Nature Publishing Group, 5(10), pp. 1262–1270. doi: 10.1038/s41564-020-0755-4.

Wommack, K. E. *et al.* (1992) ‘Distribution of viruses in the Chesapeake Bay’, *Applied and environmental microbiology*. Appl Environ Microbiol, 58(9), pp. 2965–2970. doi: 10.1128/AEM.58.9.2965-2970.1992.

Wood, D. E., Lu, J. and Langmead, B. (2019) ‘Improved metagenomic analysis with Kraken 2’, *Genome Biology*. BioMed Central Ltd., 20(1), pp. 1–13. doi: 10.1186/s13059-019-1891-0.

Wu, L. *et al.* (2018) ‘Diversity-generating retroelements: Natural variation, classification and evolution inferred from a large-scale genomic survey’, *Nucleic Acids Research*, 46(1), pp. 11–24. doi: 10.1093/nar/gkx1150.

Xie, H. *et al.* (2020) ‘PacBio Long Reads Improve Metagenomic Assemblies, Gene Catalogs, and Genome Binning’, *Frontiers in Genetics*. Frontiers Media S.A., 11, p.

1077. doi: 10.3389/fgene.2020.516269.

Xie, X. *et al.* (2018) 'Persistence of cellulolytic bacteria fibrobacter and treponema after short-term corn stover-based dietary intervention reveals the potential to improve rumen fibrolytic function', *Frontiers in Microbiology*. Frontiers Media S.A., 9(JUN), p. 1363. doi: 10.3389/FMICB.2018.01363/BIBTEX.

Yahara, K. *et al.* (2021) 'Long-read metagenomics using PromethION uncovers oral bacteriophages and their interaction with host bacteria', *Nature Communications 2021 12:1*. Nature Publishing Group, 12(1), pp. 1–12. doi: 10.1038/s41467-020-20199-9.

Yan, F. *et al.* (2019) 'Discovery and characterization of the evolution, variation and functions of diversity-generating retroelements using thousands of genomes and metagenomes', *BMC Genomics*. BioMed Central Ltd., 20(1), p. 595. doi: 10.1186/s12864-019-5951-3.

Yarza, P. *et al.* (2014) 'Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences', *Nature Reviews Microbiology 2014 12:9*. Nature Publishing Group, 12(9), pp. 635–645. doi: 10.1038/nrmicro3330.

Yilmaz, S., Allgaier, M. and Hugenholtz, P. (2010) 'Multiple displacement amplification compromises quantitative analysis of metagenomes', *Nature methods*. Nat Methods, 7(12), pp. 943–944. doi: 10.1038/NMETH1210-943.

Yooseph, S. *et al.* (2007) 'The Sorcerer II global ocean sampling expedition: Expanding the universe of protein families', *PLoS Biology*. Public Library of Science, 5(3), pp. 0432–0466. doi: 10.1371/journal.pbio.0050016.

York, A. (2017) 'Marine microbiology: Algal virus boosts nitrogen uptake in the ocean', *Nature Reviews Microbiology*. Nature Publishing Group, 15(10), p. 573. doi: 10.1038/nrmicro.2017.113.

Yuan, Y. and Gao, M. (2017) ‘Jumbo bacteriophages: An overview’, *Frontiers in Microbiology*. doi: 10.3389/fmicb.2017.00403.

Yukgehnaish, K. *et al.* (2022) ‘PhageLeads: Rapid Assessment of Phage Therapeutic Suitability Using an Ensemble Machine Learning Approach’, *Viruses 2022, Vol. 14, Page 342*. Multidisciplinary Digital Publishing Institute, 14(2), p. 342. doi: 10.3390/V14020342.

Yutin, N. *et al.* (2018) ‘Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut’, *Nature microbiology*. 2017/11/13, 3(1), pp. 38–46. doi: 10.1038/s41564-017-0053-y.

Zablocki, O. *et al.* (2021) ‘VirION2: A shortand long-read sequencing and informatics workflow to study the genomic diversity of viruses in nature’, *PeerJ*. PeerJ Inc., 9, p. e11088. doi: 10.7717/PEERJ.11088/SUPP-10.

Zadoks, R. N. *et al.* (2011) ‘Molecular epidemiology of mastitis pathogens of dairy cattle and comparative relevance to humans’, *Journal of Mammary Gland Biology and Neoplasia*. Springer, 16(4), pp. 357–372. doi: 10.1007/s10911-011-9236-y.

Zankari, E. *et al.* (2012) ‘Identification of acquired antimicrobial resistance genes’, *The Journal of antimicrobial chemotherapy*. 2012/07/10. Oxford University Press, 67(11), pp. 2640–2644. doi: 10.1093/jac/dks261.

Zaragoza-Solas, A. *et al.* (2022) ‘Long-Read Metagenomics Improves the Recovery of Viral Diversity from Complex Natural Marine Samples’, *mSystems*. American Society for Microbiology, 7(3). doi: 10.1128/MSYSTEMS.00192-22/SUPPL_FILE/REVIEWER-COMMENTS.PDF.

Zhang, H. *et al.* (2018) ‘dbCAN2: a meta server for automated carbohydrate-active enzyme annotation’, *Nucleic Acids Research*, 46(W1), pp. W95–W101. doi: 10.1093/nar/gky418.

Zhang, R., Wei, W. and Cai, L. (2014) ‘The fate and biogeochemical cycling of viral elements’, *Nature Reviews Microbiology*. Nature Publishing Group, 12(12), pp. 850–851. doi: 10.1038/nrmicro3384.

Zhao, Y. *et al.* (2013) ‘Abundant SAR11 viruses in the ocean’, *Nature*. Nature Publishing Group, 494(7437), pp. 357–360. doi: 10.1038/nature11921.

Zhou, B. *et al.* (2016) ‘Prevalence and dissemination of antibiotic resistance genes and coselection of heavy metals in Chinese dairy farms’, *Journal of Hazardous Materials*. Elsevier, 320, pp. 10–17. doi: 10.1016/J.JHAZMAT.2016.08.007.

Zhou, Z. *et al.* (2020) ‘The EnteroBase user’s guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity’, *Genome Research*. Cold Spring Harbor Laboratory Press, 30(1), pp. 138–152. doi: 10.1101/gr.251678.119.

Zhou, Z. *et al.* (2022) ‘METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks’, *Microbiome*. BioMed Central Ltd, 10(1), pp. 1–22. doi: 10.1186/S40168-021-01213-8/FIGURES/10.

Zinder, N. D. and Lederberg, J. (1952) ‘Genetic exchange in Salmonella’, *Journal of bacteriology*, 64(5), pp. 679–699. doi: 10.1128/jb.64.5.679-699.1952.

Zolfo, M. *et al.* (2019) ‘Detecting contamination in viromes using ViromeQC’, *Nature Biotechnology*. Nature Research, 37(12), pp. 1408–1412. doi: 10.1038/s41587-019-0334-5.

Appendices

The supplementary files below are available on FigShare at https://figshare.com/s/b7ad8c844288a4325deb

Supplementary File 1: In-house Perl script that modifies fasta headers so that previously predicted genes can be used as input for MetaPop.

Supplementary File 2: Fasta file of nucleotide sequences for putative MBLs.

Supplementary File 3: Excel workbook containing supplementary tables that are referenced within this work.

Rihtman, B., Meaden, S., Clokie, M. R. J., Koskella, B., Millard, A. D., & Rihtman B Clokie MRJ, K. B., Millard AD., Meaden S. (2016). Assessing Illumina technology for the high-throughput sequencing of bacteriophage genomes. In *PeerJ* (Vol. 4, pp. e2055): PeerJ.