UCC

Coláiste na hOllscoile Corcaigh, Éire
University College Cork, Ireland

Validation and Development of Sequence-Based Tools to Analyse the Human Gut Virome.

A Thesis Presented to the National University of Ireland

for the Degree of Doctor of Philosophy by

Thomas D.S. Sutton, BSc

School of Microbiology

University College Cork

Research Supervisors: **Prof. Colin Hill and R. Paul Ross**

Head of School: **Prof. Paul O'Toole**

December 2019

Contents

Declaration

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

Thomas D.S. Sutton

Date: _________________

Dedicated to Heather,

Thank you for making this possible and for putting up with me.

I promise to go and get a real job now.

Publications

Leading author

Sutton, T.D.S., and Hill, C. (2019). Gut bacteriophage: Current understanding and challenges. Frontiers in Endocrinology 10, 784.

Clooney, A.G.*, **Sutton, T.D.S.***, Shkoporov, A.N., Holohan, R.K., Daly, K.M., O'regan, O. et al. (2019). Whole-Virome Analysis Sheds Light on Viral Dark Matter in Inflammatory Bowel Disease. Cell Host & Microbe 26, 764-778.e765.

Sutton, T.D.S., Clooney, A.G., and Hill, C. (2019). Giant oversights in the human gut virome. Gut, gutjnl-2019-319067.

Sutton, T.D.S.*, Clooney, A.G.*, Ryan, F.J.*, Ross, R.P., and Hill, C. (2019). Choice of assembly software has a critical impact on virome characterisation. Microbiome 7, 12.

*Contributed equally

Contributing author (not featured as chapters, see Appendix 2)

Shkoporov, A.N., Clooney, A.G., **Sutton, T.D.S.**, Ryan, F.J., Daly, K.M., Nolan, J.A. et al. (2019). The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. Cell Host & Microbe 26, 527-541.e525.

Guerin, E., Shkoporov, A., Stockdale, S.R., Clooney, A.G., Ryan, F.J., **Sutton, T.D.S.** et al. (2018). Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. Cell Host & Microbe 24, 653-664.e656.

Fitzgerald, C.B., Shkoporov, A.N., **Sutton, T.D.S.**, Chaplin, A.V., Velayudhan, V., Ross, R.P. et al. (2018). Comparative analysis of Faecalibacterium prausnitzii genomes shows a high level of genome plasticity and warrants separation into new species-level taxa. BMC genomics 19, 931.

Shkoporov, A.N., Ryan, F.J., Draper, L.A., Forde, A., Stockdale, S.R., Daly, K.M. McDonnell, S.A., **Sutton T.D.S,** et al. (2018). Reproducible protocols for metagenomic analysis of human faecal phageomes. Microbiome 6, 68.

Abstract

The gut microbiome is a complex community of microorganisms that interacts closely with the human host and is believed to play an important role in the maintenance of human health. The viral component of this community is referred to as the human gut virome and is dominated by bacteriophage. Bacteriophage are central to microbial ecosystems by facilitating nutrient turnover, horizontal gene transfer and driving bacterial diversity. In this way the gut virome is believed to closely interact with the human host by shaping the composition and function of the gut microbiome. However, the gut virome also represents one of the biggest gaps in our understanding of the microbiome as it is dominated by unknown bacteriophage targeting unknown bacterial hosts and with uncharacterised downstream functions.

These challenges mean that virome research relies heavily on sequence-based approaches and metagenomics to identify compositional patterns and targets for future characterisation. A typical virome study involves physical and chemical separation of individual virions from the cellular components of the microbiome and the contents of the faecal, luminal or mucosal sample from which it came. A viral metagenome is then generated by extracting virome DNA and/or RNA for sequencing on a given platform. These sequencing reads are then quality filtered and assembled to reconstruct the viral genomes in the original sample. The abundance of these assemblies is then estimated by aligning the sequencing reads and performing statistical analysis. However, each step in a virome analysis pipeline has the potential to distort the final viral community and given the unknown nature of the virome, this distortion is difficult to identify and characterise. As a result, conclusions are often drawn from virome studies without fully appreciating the impact of the analysis methods on the findings.

This thesis examines the major steps in sequence-based virome analysis pipelines, highlighting how choices made at each step of an analysis protocol can impact the final conclusions drawn from a study. In doing so, we have changed our perspective of the human gut virome and challenged previous assumptions. Chapter One discusses the current understanding of the virome field, giving particular attention to how the analysis methods and challenges affect our view of the virome. In Chapter

Two, we focus on the assembly step of virome analysis pipelines. This step is of particular importance to virome studies, as an assembler's ability to recover viral sequences can ultimately determine the amount of sequence information used in a that study. We compared all short-read assembly programs used in virome studies to date, across mock communities, simulated and real datasets. We found that not all assemblers are equal, and choice of assembler can drastically affect the conclusions that can be drawn from a virome study. These findings call the comparability of different virome studies into question and would suggest that previous virome studies would benefit from reanalysis using improved assembly methods and re-examination of the conclusions drawn.

As discussed, the human gut virome is dominated by "viral dark matter"; those sequences which do not share homology to reference databases. However, the majority of what is currently known about the virome in human health and disease is based on the minor fraction of viral sequences collated in these databases. This presents a serious gap in our understanding and was the primary focus of Chapter Three. We reanalysed a keystone inflammatory bowel disease (IBD) dataset, which had formed the foundation of much of what we knew about the virome in IBD. We developed a new approach to analysing the virome beyond the identifiable minority and by doing so, changed our understanding of the virome in IBD significantly.

In the final chapter, we directed our attention to possibly the most important aspect of a sequence-based study, the sequencing approach itself. This step bridges the gap between the biological information in a virome and the digital information that is analysed. As with all steps in a virome analysis pipeline, this has serious implications for the final conclusions of the study. We described the use of long-read sequencing in the human gut virome and the benefits and challenges which are associated with this technology. We also found the ability of amplified short-read sequencing libraries to represent the gut virome was limited, but that alternative library preparation methods and long-read sequencing platforms may be able to address these limitations. These findings imply that much of what we know about that human gut virome may be linked to sequencing performance, rather than the biology of the community itself.

These three major aspects of virome analysis pipelines highlight the importance of considering the impact of the analysis approach when interpreting the results of virome data and complex biological systems in general.

Chapter 1

Literature review

Gut Bacteriophage: Current Understanding and Challenges

This chapter has been published as the following:

Sutton, T.D.S., and Hill, C. (2019). Gut bacteriophage: Current understanding and challenges. Frontiers in Endocrinology 10, 784

https://www.frontiersin.org/articles/10.3389/fendo.2019.00784/full

Manuscript written and conceived by Thomas Sutton, supervised and revised by Colin Hill

Abstract

The gut microbiome is widely accepted to have a significant impact on human health yet, despite years of research on this complex ecosystem, the contributions of different forces driving microbial population structure remain to be fully elucidated. The viral component of the human gut microbiome is dominated by bacteriophage, which are known to play crucial roles in shaping microbial composition, driving bacterial diversity and facilitating horizontal gene transfer. Bacteriophage are also one of the most poorly understood components of the human gut microbiome, with the vast majority of viral sequences sharing little to no homology to reference databases. If we are to understand the dynamics of bacteriophage populations, their interaction with the human microbiome and ultimately their influence on human health, we will depend heavily on sequence based approaches and *in silico* tools. This is complicated by the fact that, as with any research field in its infancy, methods of analyses vary and this can impede our ability to compare the outputs of different studies.

Here we discuss the major findings to date regarding the human virome and reflect on our current understanding of how gut bacteriophage shape the microbiome. We consider whether or not the virome field is built on shaky foundations and if so, how can we provide a solid basis for future experimentation. The virome is a challenging yet crucial piece of the human microbiome puzzle. In order to develop our understanding, we will discuss the need to underpin future studies with robust research methods and suggest some solutions to existing challenges.

Introduction

The human gastrointestinal tract (GIT) is a complex environment containing billions of microorganisms (Sender et al., 2016). Changes in oxygen concentration, pH, nutrient availability, water availability and bile salts shape the relative abundance of microorganisms from all domains of life (fungi, protists, bacteria and archaea) (Duncan et al., 2009;Espey, 2013;Ridlon et al., 2014;Vandeputte et al., 2016). Of these microorganisms, bacteria are by far the most characterized, making up the vast majority of the DNA sequences and biomass (Qin et al., 2010;Yatsunenko et al., 2012). This bacterial community also plays a central role in normal physiology of the mammalian gut by facilitating metabolic functions, protecting against pathogens and modulating the immune system (Sonnenburg et al., 2005;Sokol et al., 2008;Belkaid and Hand, 2014). Similarly, alterations in the composition and abundance of this bacterial community are closely associated with diseases such as irritable bowel syndrome (IBS), inflammatory bowel disease (IBD), colorectal cancer (CRC), *Clostridium difficile* infection (CDI), obesity and neurological disorders (Chey et al., 2015;Sharon et al., 2016;Halfvarson et al., 2017;Liu et al., 2017;Wirbel et al., 2019). However, the forces that shape the composition of these bacterial communities remain poorly understood, and this has slowed the development of microbiome based therapeutics and biomarkers.

Bacteriophage (phage) are viruses that infect prokaryotic hosts and play crucial roles in shaping the composition and diversity of bacterial communities in many environments, facilitating horizontal gene transfer and nutrient turnover through continuous cycles of predation and coevolution (Suttle, 2007;Breitbart, 2011;von Wintersdorff et al., 2016). To date, the majority of viral metagenome (virome) research has been focused on environmental communities such as those in the ocean (Hurwitz and Sullivan, 2013;Hurwitz et al., 2015). In this environment, the virome is central to the movement of dissolved organic matter across trophic levels of the ocean food chain and between the surface and the depths of the water column (Suttle, 2007;Lauro et al., 2009). A growing body of evidence also suggests the virome can shape the functional capacity of host communities encoding functions such as photosynthetic genes in the photic zones of the ocean (Sullivan et al., 2006) and bacterial virulence factors in pathogenic bacteria (Muniesa et al., 2012).

Phage make up the vast majority of the viral component of the gut microbiome(Gregory et al., 2019) . They are also believed to play a key role in shaping the composition and function of the human gut microbiome in both health and disease (Norman et al., 2015;Manrique et al., 2016;Zuo et al., 2019). However, despite being highly abundant in the gut (>10^{10} g^{-1}) (Hoyles et al., 2014;Shkoporov et al., 2018b) and having considerable impacts on microbial ecosystems, they remain one of the least understood members of the gut microbiome. Early sequencing studies of the human gut virome estimated that it was dominated by novel sequences, with only 41% sharing homology to databases(Breitbart et al., 2003). However as sequencing platforms and library preparation methods improved and yielded a more detailed view of the virome, this unknown majority or "viral dark matter" was found to make up an even greater proportion of the virome, lowering the identifiable fraction to as little as (1-14%) (Aggarwala et al., 2017).

Since phage were first identified by Frederick Twort in 1915 (Twort, 1915), culture-based methods such as plaque assays have been used to screen and quantify phage titres from many environments. Today, these methods still play a central role in identifying phage which target specific bacteria and have contributed to our understanding the mechanics of phage host interactions and replication cycles. However, as the vast majority of phage-host pairs in the gut are unknown, these methods are not suited to large-scale characterization of a complex ecosystem such as the human gut. Additionally, many of bacteria in the human gut are not routinely cultivated, despite recent advances(Forster et al., 2019). As a result, virome studies lean heavily on sequencing based metagenomic approaches to investigate gut phage communities and to try to understand their role in shaping the gut microbiome (Aggarwala et al., 2017). This involves sequencing the total viral DNA from a community following physical separation from the bacterial component, using assembly software to recreate the viral genomes within that community and characterizing abundance and function of those genomes. However, many sequence-based virome studies exclude viral dark matter from analysis, working largely with a small fraction of known phage sequences (usually 1-14% of the dataset). This can have profound implications for the conclusions drawn from these studies, as changes in the known fraction may not reflect changes in the virome as a whole. As a result, database-independent analysis methods are increasingly being used which include both known and unknown fractions of the virome (Shkoporov et al., 2018b). However, high levels of inter-individuality make biological signals across virome studies difficult to detect (Clooney et al., 2019;Gregory et al., 2019;Moreno-Gallego et al., 2019). Furthermore, virome studies are particularly susceptible to methodological bias due to difficulties in benchmarking *de novo* bioinformatic tools and the dominance of unknown sequences in virome datasets (Roux et al., 2016;Hesse et al., 2017;Sutton et al., 2019a).

We will discuss phage of the human gut virome and their interactions with the microbiome. We will also highlight how little we know about its role in human health. Finally, we will discuss critical areas of virome analysis methods which must be addressed and improved upon if we are to fully understand the role of phage in shaping the microbiome and human health.

How phage interact with bacterial hosts

Phage infection cycles

As obligate parasites of bacteria, phage persistence in a microbial ecosystem is dependent on the presence of a suitable sensitive host. Phage infection is typically followed by one of two replication cycles, lytic or lysogenic (Weinbauer, 2004) (Figure 1.A). In both cases a phage virion binds to the host cell surface using a phage receptor-binding protein triggering the insertion of its genome into the host. For lytic phage, subsequent translation of the phage genetic material by the host cell results in the replication of the phage genome, assembly of phage particles and lysis of the host. This results in the release of new phage virions into the environment that can infect nearby hosts. Alternatively, lysogenic infection results in the replication of the phage genome within the host cell without the immediate synthesis of phage virions. The phage genome may integrate into that of the host where it exists as a prophage, replicating together with the host genome and thus persisting in resulting daughter cells. In the case of pseudolysogeny, the phage genome persists as an episome within the host cell, separate to the host genome. In order to ensure subsequent daughter cells contain phage genomes pseudolysogenic phage can use maintenance systems such as toxin-antitoxin (Ravin, 2015;Cenens et al., 2016).

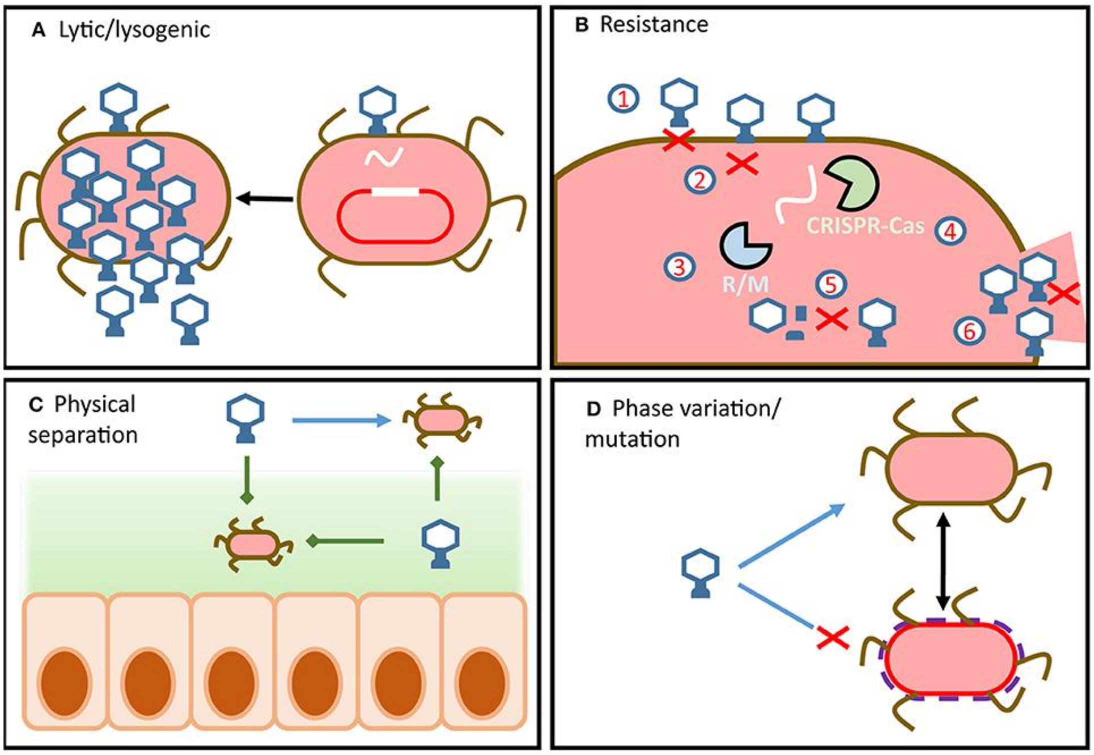


Figure 1. Overview of phage-host dynamics in the gut. (A) Phage infection can lead to virulent or temperate replication cycles. Integrated temperate phage use internal and external signals from hosts to determine if or when to enter the lytic cycle. (B) Bacteria can possess a wide array of defence mechanisms which target different steps of the phage replication cycle. Similarly, phage encode a wide array of counter-defence mechanisms which target host defences and allow the phage to remain infectious. (C) Physical separation of phage and host (e.g. in mucous or in lumen) means that dynamics change along the radial and longitudinal axes of the gut. (D) Strain-level variation can result from resistance by mutation or by phase variation.

However, cases of daughter cells lacking pseudolysogenic phage have also been reported (Cenens et al., 2016). Following an induction event, the lysogenic phage will initiate the translation of its genome and subsequent production of phage virions leading to host lysis. Additionally, phage such as M13 undergo chronic non-lethal infection cycles, where newly produced virions exit the cell without lysis (Smeal et al., 2017). However, little is known about the prevalence of these different lifecycles in the human gut.

Resistance and counter resistance

Bacterial hosts employ a wide array of phage resistance mechanisms which have been comprehensively reviewed by Labrie et al. (Labrie et al., 2010) and Rostøl et al. (Rostøl and Marraffini, 2019). To prevent phage adsorption, bacterial cells can differentially express or mutate cell surface receptors (Figure 1.B1) (Clement et al., 1983;Chung et al., 2014), S-layer proteins (Zago et al., 2017), or produce protective cell surface polysaccharides(Scholl et al., 2005). Additionally, bacterial hosts can reduce the numbers of phage particles available to infect hosts by producing outer membrane vesicles (Schwechheimer and Kuehn, 2015). These bind and sequester phage particles, reducing their numbers in the environment and thereby the risk of infection. Should the phage successfully bind to the appropriate surface receptor, the fate of the host is not yet sealed as anti-phage resistance mechanisms extend to all steps of the phage infection cycle.

Hosts can prevent phage DNA injection entirely by modifying inner-membrane proteins (Figure 1.B2) (Cumby et al., 2015), identify and degrade injected phage DNA using restriction modification systems (Figure 1.B3) (Tock and Dryden, 2005) and CRISPR-Cas systems (Figure 1.B4) (Rostøl and Marraffini, 2019), chemically block phage DNA replication (Kronheim et al., 2018), or prevent virion assembly (Figure 1.B5) (Ram et al., 2012). Should these defence mechanisms fail to prevent phage replication within the cell, the bacterial host can sacrifice itself in order to protect its sister cells (Dy et al., 2014). These are referred to as abortive infection systems and act by shutting down cellular functions to prevent phage release (Figure 1. B6). Phage defence systems are regularly encoded on mobile genetic elements that can facilitate the transfer of resistance across the bacterial community. However, due to their metabolic costs bacterial cells rarely encode more than one of these systems (van Houte et al., 2016). This gives rise to complex dynamics between hosts with the metabolic burden of resistance and susceptible hosts with fitness advantages. In addition, a cell carrying a prophage can be made resistant to other phage in what is referred to as superinfection exclusion(Hofer et al., 1995). To further complicate these relationships, antagonistic coevolution of phage-host pairs has led to the development of phage counter resistance mechanisms, which allow phage to remain infectious in the face of a resistant host population(Samson et al., 2013). Phage counter-resistance can range from glycosidases to degrade host capsules (Leiman et al., 2007) and reveal binding sites, to directed mutagenesis or hypervariable receptor binding proteins(Chatterjee and Rothenberg, 2012;Minot et al., 2013;Warwick-Dugdale et al., 2019). These systems allow phage to retain compatibility with modified host receptors and also allow for the expansion of host range. Phage can even overcome host CRISPR-Cas by mutating or deleting CRISPR target sites or expressing anti-CRISPR proteins to directly interfere with CRISPR-Cas activity (Bondy-Denomy et al., 2013). Some of the most striking phage counter-resistance mechanisms include alteration of phage DNA to evade host restriction modification mechanisms (Bair and Black, 2007) and phage encoded CRISPR-Cas systems that target and disable a range of host defence systems (Seed et al., 2013).

Ecological relevance of phage-host dynamics

These mechanisms of infection, resistance and counter resistance underpin virome-microbiome interactions. Thus, in order to investigate how the virome shapes or reflects the microbiome, we must first understand these interactions in the GIT. Understanding phage host interactions is also vital if we are to use the virome as a diagnostic or therapeutic tool in the future. To this end, numerous ecological models have been used to describe phage-host interactions in the context of a biological system. Some models focus entirely on the interplay of resistance and infectivity such as the arms race dynamics model (Scanlan, 2017). This model proposes that phage infection applies selective evolutionary pressure for mutations in the hosts, resulting in resistant host populations. These mutations in turn select for phage mutations that restore infectivity, resulting in predator prey cycles (Zhu et al., 2015). Other models take into account phage-host density and the metabolic cost of resistance such as the fluctuating-selection dynamics model (Gandon et al., 2008;Hall et al., 2011). This proposes that as phage predation selects for resistant hosts, it will also reduce the number of phage virions in the environment. This absence allows for the expansion of susceptible bacterial strains which lack the metabolic burden of resistance and therefore out-compete resistant hosts in the absence of phage. This transient resistance and infectivity of phage-host communities results in short-term fluctuations of phage and host numbers, but the long-term persistence of both (De Sordi et al., 2019).

It is important to note that strain-level fluctuations of phage-host are difficult to study in the context of the microbiome, making it difficult to verify or quantify this model. Strain-level variation within bacterial hosts cannot be detected by 16S rDNA analysis (Gandon et al., 2008) and hampers metagenomic assembly of phage (Nurk et al., 2017;Sutton et al., 2019a). Despite these challenges, recent insights have proposed that variation in capsular polysaccharides encoded by the abundant gut bacterium *Bacteroides thetaiotaomicron* play a central role in phage susceptibility (Porter et al., 2019). This mechanism supports the concept of fluctuating-selection dynamics, as phase variation of the capsular polysaccharides creates heterogeneous host phenotypes within an isogenic population. This in turn leads to transient phage resistance across the population as host phenotypes are dynamic and non-uniform (Turkington et al., 2019). Additionally, this resistance can occur without the need for horizontally transferred resistance or mutation (Figure 1.D).

Phage replication not only requires the host to be present and susceptible but it must also be metabolically active. It has therefore been proposed that the metabolic state of the host is one of the primary barriers to phage infection (De Sordi et al., 2019). In this way bacterial hosts can be transiently resistant to phage infection due to a lack of nutrient availability and a dormant growth phase, without incurring the metabolic cost of encoding resistance (Figure 1. D) (Denou et al., 2007). Additionally, this mechanism would support the proliferation of both phage and host in the GIT, independent of resistance-counter resistance dynamics. Nutrient availability varies significantly along the GIT, which implies that phage-host dynamics in the proximal or distal colon may be significantly different to those in fecal samples (Figure 1. C) (Maura et al., 2012;Galtier et al., 2017). As the majority of virome studies draw conclusions from fecal samples, our current understanding of phage host dynamics in the GIT is limited.

The kill-the-winner ecological model is an extension of lotka-volterra dynamics applied to phage-host interactions. This model describes rapid changes of diversity and abundance of both phage and their hosts. As the most abundant bacteria are killed by their phages, other bacterial taxa will take over the ecological niche and be subsequently killed by their phages. In this way high levels of phage-host diversity and abundance are maintained (Mirzaei and Maurice, 2017). In ecosystems where lotka-volterra or kill-the-winner dynamics can be applied, phage exhibit an exclusively predatory relationship on hosts and the microbial biomass is significantly below the carrying capacity of the ecosystem (Thingstad, 2000;Avrani et al., 2012). However, in the human gut ecosystem microbial biomass approaches the carrying capacity of the ecosystem and the virus to microbe ratio (VMR) is low (Shkoporov and Hill, 2019). Despite reports of kill-the-winner dynamics in infants(Breitbart et al., 2008;Lim et al., 2015), this suggests that these models cannot not fully explain phage-host interactions in the healthy adult gut (Mirzaei and Maurice, 2017). Additionally, these dynamics overlook lysogeny and conditions that govern the switch between lytic and lysogenic replication cycles. Furthermore, it has been proposed that the gut virome is dominated by temperate phage (Reyes et al., 2010) and that compositional changes in temperate phage communities are associated with disease states (Norman et al., 2015). Consequently the mechanisms that determine the switch between lytic and lysogenic replication cycles are also central to understanding virome-microbiome dynamics. Recent reports have described phage which can hijack bacterial quorum sensing machinery to determine the density and metabolic activity of the bacterial population from within the host cell (Silpe and Bassler, 2019). This in turn, could dictate whether persisting within in the host genome or excising and entering the lytic cycle would favor phage proliferation (Figure 1. A). In addition, single cell analysis of phage-host interactions between the temperate P22 phage and *Salmonella typhimurium* suggested that phage were directly involved in creating transiently resistant host subpopulations (Cenens et al., 2016). This allowed for both lysogenic and lytic replication without impeding host proliferation. Despite these intriguing insights, little is known about the dynamics of temperate to lytic switching in the mammalian GIT, highlighting a crucial target for future virome research.

Ecological models that consider the switch between lytic and lysogenic replication such as the piggyback-the-winner model appear to support experimental evidence of phage-host dynamics within the mammalian GIT (Knowles et al., 2016;Silveira and Rohwer, 2016). This model focuses on a lytic to lysogenic switch that is host density dependent. Traditionally phage were believed to enter the lysogenic cycle in cases of high VMR (i.e. increased phage abundance, decreased host abundance) as a means to persist in the environment until host density can support lytic replication cycles. However, experimental evidence from coral reef ecosystems suggested that phage also entered the lysogenic cycle in high host density situations (Knowles et al., 2016). Subsequently, phage can “piggyback” on host success in the ecosystem at that particular time. This model has also been proposed for phage-host interactions on mucosal surfaces in the GIT. It has been proposed that at the mucosal surface, high bacterial colonization and high VMR gives rise to piggyback the winner dynamics, whereas deeper mucosal layers may give rise to kill-the-winner dynamics due to the lower levels of bacterial colonization and low VMR. This model is also supported by reports that rapidly evolving Ig-like domains expressed on phage capsids interact with mammalian host mucus glycans. This in turn results in subdiffusive motion of phage within mucus and allows them to persist in mucosal layers of the gut (Barr et al., 2013;Barr et al., 2015).

While these models assist in our understanding of how the virome interacts with the microbiome and the gut environment, it is also important to consider their limitations. Phage-host dynamics can be expected to change both radially and longitudinally within the GIT (Zhao et al., 2019) to reflect physical separation of phages and hosts and metabolic changes in the host populations (Figure 1A.). They will also be heavily influenced by dietary components and the composition of the faecal matrix itself (Vandeputte et al., 2016). To this end, sampling method and sample composition must be considered when drawing conclusions from virome data (Figure 2A.). In the absence of studies of how these dynamics change along the human GIT, phage-host interactions must be interpreted as a snapshot of one particular point in space and time.

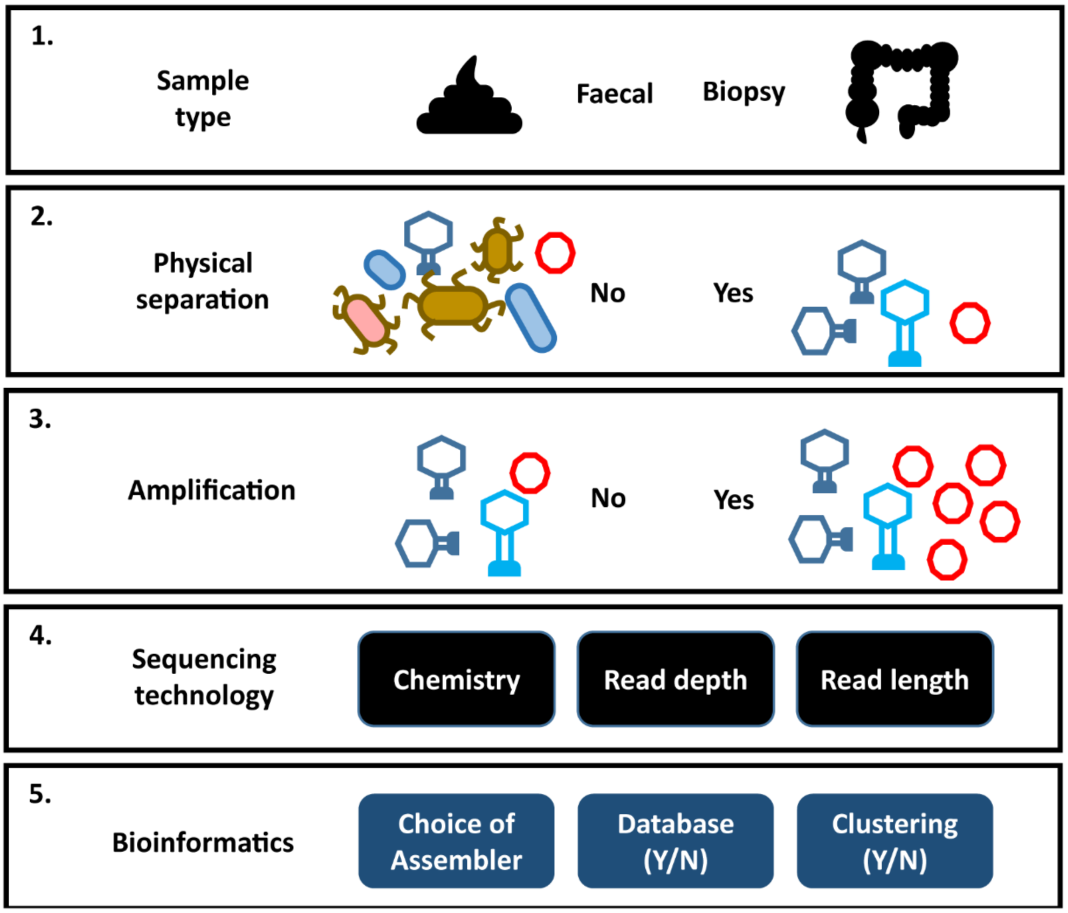


Figure 2. Impact of analysis choices on virome composition.

Each step of a virome analysis protocol presents different options, each of which may affect the final outcome. (1) Sample type. (2) Physical separation of VLPs. (3) Amplification of virome DNA can preferentially amplify certain viral taxa (see Figure 3). (4) Sequencing chemistry, depth of sequencing and read length. (5) Assembly programs vary significantly in their ability to assemble virome data (see Figure 3). Reporting on the composition of viral sequences with homology to reference databases excludes the unknown majority of the virome. Clustering viral sequences by gene composition offers a promising alternative to database dependent methods by addressing high levels of sequence divergence in viral genomes.

Composition of the Gut virome

The gut virome consists of two elements, the temperate phage located within bacterial genomes and the free virions or virus-like particles (VLP). The VLP fraction is obtained by applying several physical and enzymatic steps that remove dietary debris and prokaryotic and eukaryotic genetic material (Figure 2.B) (Shkoporov et al., 2018b). As VLPs represent only a small fraction of the mass of the microbiome, virome studies that do not carry out this viral enrichment are limited almost entirely to the prophage sequences within bacterial cells (Waller et al., 2014;Ma et al., 2018). However, these prophage sequences are in turn under-represented in studies that focus on the VLP fraction. In order to understand the virome as a whole, both elements need to be analysed in tandem. The majority virome studies focus on one of these elements in isolation and as a result the apparent structure and composition of a virome is heavily dependent on the physical preparation of virome samples in the laboratory, the sequencing strategy, and the bioinformatic methodology employed (Figure 2.). This is further complicated by the limited representation of human gut viruses in databases and a reliance on phage taxonomic classification systems which do not necessarily represent viral biology. Consequently, our understanding of the taxonomic composition of the phage populations in the human gut are predictably varied and contradictory.

Taxonomic composition

Sequence-based analysis suggests the identifiable fraction of the gut virome is dominated by small single-stranded DNA (ssDNA) phage of the *Microviridae* family and double-stranded DNA (dsDNA) phage of the order *Caudovirales* (Breitbart et al., 2003;Minot et al., 2013;Manrique et al., 2016;McCann et al., 2018;Shkoporov et al., 2018b). However, bias in extraction methods has also been reported to skew the abundance of *Microviridae,* calling into question their dominance in the virome (discussed later) (Kim and Bae, 2011) (Figure 2C.). The *Caudovirales* are classified into three families according to their distinctive virion morphology, consisting of a head and a tail structure making them easily distinguishable in microscopy studies. *Siphoviridae* exhibit long (sometimes up to 1µm in length), non-contractile tails, *Podoviridae* exhibit short non-contractile tails, while *Myoviridae* exhibit a more rigid contractile tail composed of distinctive sheath proteins. These phage families have linear genomes, can encode a relatively large gene repertoire (in extreme cases containing over 600 genes) and can exhibit both temperate and lytic replication cycles. Early sequencing studies of the virome reported changes in *Caudovirales* community composition being associated with disease (Pérez-Brocal et al., 2013;Wagner et al., 2013;Norman et al., 2015) and high *Caudovirales* abundance in EM images (Lepage et al., 2008;Hoyles et al., 2014). However, reported abundance in EM images could also be influenced by the fact that they are more easily identified than other viral taxa. Possibly as a result of these observations, *Caudovirales* composition and abundance has become a regular focus of virome studies, where it is used as a proxy for virome composition as a whole. However, as known *Caudovirales* only represent a small minority of VLP sequences in the human gut, caution should be taken when interpreting these results.

Due to the current structure-based classification system of *Caudovirales*, phage that exhibit significant functional similarities can be considered members of different families due to tail morphology. For example, phages P22 and lambda are classified as *Podoviridae* and *Siphoviridae* respectively, despite both undergoing the same replication cycle, sharing significant similarity in gene sequences and exhibiting identical genome organization (i.e. gene order and layout and regulation of transcription). Additionally this classification system is not grounded in sequence or protein sharing, yet sequence and protein homology are the primary methods of identifying *Caudovirales* in viromes (Reyes et al., 2015;Fernandes et al., 2019;Moreno-Gallego et al., 2019;Zuo et al., 2019). This classification system is therefore limited in its ability to reflect the biological role or interactions of sequences classified as *Caudovirales*. A number of recent studies have also highlighted anomalies in *Caudovirales* taxonomy (Hulo et al., 2015;Bolduc et al., 2017;Barylski et al., 2018) and have proposed novel sequence-based methods to restructure the viral order (Bolduc et al., 2017). However, these are not without their own challenges, as shared genes and gene cassettes have been found to blur the boundaries between various dsDNA viruses (Iranzo et al., 2016;Jang et al., 2019), which will be discussed in detail later.

Case study of the most dominant gut phage, crAss

A prime example of the limitations of focusing on identifiable *Caudovirales* alone when carrying out virome analysis is that of crAssphage, one of the most abundant and successful biological entities of the mammalian gut. First reported in 2014 using novel assembly methods (Dutilh et al., 2012;Dutilh et al., 2014), crAss was found to be six times more abundant in publically available gut metagenome samples than all other phage together. It made up to 90% of VLP sequencing reads and 30% of whole metagenomes in certain individuals, yet did not share homology with known reference databases (Dutilh et al., 2014). As a result, crAss would have gone entirely unnoticed using database-dependent analysis alone. CrAss has since been found to be globally distributed, with strains reflecting geographic distribution of human populations (Edwards et al., 2019). Intriguingly, the presence of distant relatives to crAss in primates may suggest that crAss has coevolved with humans for millions of years from ancestors shared with primates, to modern *Homo sapiens* (Edwards et al., 2019). These results highlight the importance of the unknown majority of the virome in human health and why it is critical to analyse the virome as a whole. Given the proposed ubiquity and dominance of crAss in the human gut it is also possible that the abundance of *Caudovirales* reported by EM studies may in fact reflect members of the extended crass-like phage family, which at the time were unknown.

The characterization of crAss and studying its role in the gut microbiome has been hampered by a lack of database representation, unknown host range and until recently (Shkoporov et al., 2018a), unsuccessful attempts to form plaques on agar overlays. However, its progression from an unknown abundant phage sequence to a characterized dominant member of the gut virome provides a useful framework in building our understanding of viral dark matter. Its initial discovery was built upon by using sensitive protein homology searches to identify an extended family of crAss-like phage from human gut virome samples (Guerin et al., 2018;Yutin et al., 2018). Although sequence similarity across family members was low, protein family-based clustering identified conserved capsid proteins and predicted crAss to encode a short tail similar to that of the *Podoviridae* family of *Caudovirales* (Yutin et al., 2018). Subsequently, crAss has also been identified in patients with diarrhoea and in Malawian infants (Guerin et al., 2018). Additionally it was found to be shared across healthy individuals and was capable of stable engraftment in FMT treatments (Draper et al., 2018).However, the host range and mechanism behind crAss ubiquity and abundance remained unknown.

Through the use of enrichment-based techniques the host for one member of the crAss family was confirmed to be *Bacteroides intestinalis*, a finding previously suggested by co-abundance profiling (Dutilh et al., 2014;Shkoporov et al., 2018a). This provided new opportunities to study the phage-host interactions of the most abundant phage family in the human gut and its role in the microbiome. The ability to culture crAss *in vitro* also led to the description of one of the most intriguing aspects of this phage, its ability to coexist with its host at high abundance. This has a profound impact on our understanding of how the virome shapes the microbiome, as it appears that in crAss-rich individuals the dominant phage populations do not restrict the growth and proliferation of their host. Furthermore, it is believed that crAss persistence and coexistence with its host is not due to crAss undergoing temperate replication, as it does not possess any of the genes typically associated with lysogeny (Shkoporov et al., 2018a). Additionally, crAss prophage sequences have never been observed in the numerous *Bacteroidales* genomes available in sequence databases. It is possible that crAss replicates in a pseudolysogenic manner, existing as an episome within the host cell (Cenens et al., 2016), that it facilitates a chronic infection without killing the host cell (Smeal et al., 2017), or that it can exist in an extracellular carrier state (Siringan et al., 2014). The high proportion of resistant host cells in culture (1-2%) could also suggest that a host cell carrying crAss in a pseudolysogenic state could confer resistance from infection as in superinfection exclusion. However, experimental evidence suggests crAss replicates in a lytic manner and its abundance in the gut is maintained by transiently resistant hosts. As phase variant CPS and transient resistance have been found to be central in phage-host dynamics in other *Bacteroides* species (Porter et al., 2019), it is possible that these mechanisms are also central to the unusual phage-host dynamics observed in crAss both *in vitro* and *in vivo*. This theory is supported by the reversion over time of some resistant clones to sensitive states (Shkoporov et al., 2018a). Given the abundance of *Bacteroides* and crAss in the human gut, phenotypic heterogeneity across a host populations may be the central mechanism to support stable interactions between some lytic phage and their hosts in the microbiome.

Another possible explanation for sustained crAss-host proliferation is the recently proposed “Royal Family” ecological model (Breitbart et al., 2018). This suggests continuous kill-the-winner dynamics occur at a strain-level rather than at a species or genus-level. In this way, the abundance of both phage and host would appear to be stable as fluctuations would occur below the level of detection. Subsequently, detailed analysis of strain-level variation between crAss and *B. intestinalis* could provide insights into the importance of this model in the GIT. A similar study carried out by De Sordi et al. (De Sordi et al., 2017) supports the “royal family model” and gave intriguing insights into its mechanics. This study described a point mutation in the tail fibre gene of P10 which is associated with strain-level host expansion from *E. coli* LF82 to *E. coli* MG1655. Interestingly, this strain-level host switching was only observed when phage and both hosts coevolved within the murine model and was not observed when the two strains were cultured separately. Subsequent experiments revealed that switching required the presence of an intermediate host *E. coli* MEc1. These observations suggest that crAss-host stability in the gut could be caused by a wide array of strain-level and phenotypic variation. However, as strain-level variation is difficult to observe *in vivo* and *in vitro*, novel analytical approaches may needed to reveal crAss phage-host dynamics in the GIT.

The proliferation of both phage and host as is seen in crAss-*B. intestinalis* dynamics could also suggest that the presence of phage could confer an ecological advantage to the host. This phenomenon has been regularly reported in the ocean, where phage infecting cyanobacteria were found to carry auxiliary metabolic genes encoding photosynthetic genes (Sullivan et al., 2006;Hurwitz et al., 2013). Similarly, phage-mediated transformation of the host has also been well established in disease such as the lysogenic phage encoding shiga toxin (Muniesa et al., 2012) and cholera toxin (Waldor and Mekalanos, 1996). It is therefore highly likely that in the dense and diverse ecosystem of the gut, extensive horizontal transfer of genes between hosts is facilitated by phage infection. In this way the virome may play a crucial role in shaping the functional capacity of the microbiome. One report of the presence of significant numbers of antibiotic resistance genes in gut virome sequences (Modi et al., 2013) was later shown to have probably resulted from bacterial contamination and confirmed that examples of phage-encoded antibiotic resistance genes were rare (Enault et al., 2017). This could be due to the efficiency of phage replication and the fitness cost of carrying antibiotic resistance genes. Without the selective pressure of antibiotics, viruses that pay the metabolic cost of carrying antibiotic resistance genes could be outcompeted (Enault et al., 2017). Additionally, the selective evolutionary pressure of remaining infectious in the face of a constantly adapting host may outweigh the need to preserve the host from an antibiotic. The follow up study also highlighted the importance of using stringent alignment criteria and validating results when classifying sequences or proteins. Due to the extensive unknown fraction of the virome and the difficulty in benchmarking classification criteria, lenient cut-offs can often lead to false conclusions of virome composition and function (Roux et al., 2013;Enault et al., 2017;Sutton et al., 2019b).

It is tempting to propose that the piggyback-the-winner hypothesis could be extended by considering the possible fitness advantages of carrying a prophage over and above superinfection exclusion. Such advantages could include carrying a virulence factor or providing access to a novel nutrient source (Brüssow et al., 2004;Obeng et al., 2016). Prophage-encoded fitness advantages have been observed in a number of pathogenic bacteria. These include prophage-encoded toxins (Waldor and Mekalanos, 1996;Muniesa et al., 2012), alteration of O antigens in *Salmonella* and *Shigella* (Wright, 1971;Verma et al., 1991) and phage-encoded glycosyl transferase operons which drive *Salmonella* LPS diversity (Davies et al., 2013). Furthermore the horizontal transfer of virulence factors through temperate phage was found to be increased in gut inflammation (Diard et al., 2017). Additionally, recent observations in *Staphylococcus aureus* prophage (Chen et al., 2018) have described the packaging of chromosomal host DNA in phage capsids through a mechanism deemed lateral transduction. This mechanism suggests that phage-mediated horizontal gene transfer occurs at much higher rates than previously thought and that it plays a role in disease. However, the extent to which lateral transduction mediates gene transfer in the gut microbiome remains unknown. Examining phage-encoded auxiliary metabolic genes and how they shape the functional capacity of the gut microbiome is hindered at a large-scale metagenomic level by the complexity of faecal samples themselves. Dietary components and the sheer abundance and diversity of bacterial cells in faeces make it difficult to completely remove bacterial sequences from virome samples. Use of density gradients such as CsCl are reasonably effective at generating viral particles devoid of cellular contamination, but will introduce bias in favour of particular viral capsid types (Castro-Mejía et al., 2015) and are not feasible for large-scale projects due to their associated manual workload. As a result, background contamination exists in the vast majority of virome samples (Roux et al., 2013;Enault et al., 2017;Shkoporov et al., 2018b) making it difficult to determine if a gene is of viral or bacterial origin. This may be further complicated by potential gene transfer agents (GTAs) (Lang and Beatty, 2007) in gut VLP samples. GTAs are defective phage virions that exclusively carry fragments of bacterial chromosomal DNA and are used by bacteria as a means of HGT. As a result sequences within GTAs are difficult to differentiate from background contamination. However the presence and prevalence of GTAs in human gut microbiome remains to be seen. Overall, despite the evidence of phage shaping the functional capacity of host communities, it is challenging to determine the extent to which phage transfer genes relevant to human health within the microbiome.

Controversy surrounding the core virome

The widespread geographical distribution and stability of crass-like phage supports the concept of a core human virome, which was initially proposed by Manrique et al. (Manrique et al., 2016). This was in response to a growing body of evidence that a core microbiome played an important role in human health. The study proposed a core of 23 viral sequences, one of which being the original crAss genome, which were shared across more than 50% of samples from an independent cohort of 62 healthy individuals. However, these findings were in stark contrast to the well-established belief that the human gut virome is highly individual-specific at a sequence level (Shkoporov et al., 2018b;Clooney et al., 2019;Moreno-Gallego et al., 2019;Shkoporov et al., 2019). This disparity is largely due to the criteria used to define the presence of a viral sequence in a sample. If a single sequencing read from an individual could align to a particular viral assembly, the assembly was deemed to be present. This lenient criteria does not account for the modular nature and extensive gene sharing that occurs across dsDNA viral genomes (Minot et al., 2012;Iranzo et al., 2016;Bolduc et al., 2017). Thus, it would not be possible to differentiate the true presence of a viral sequence in a sample from the presence of a shared gene between two unrelated phage. This in turn, would lead to an inflated number of viral sequences being shared across individuals. However, the concept of a core virome has received support from a recent study with adult monozygotic twins, in which 18 contigs were found to be present in all individuals (n=42) (Moreno-Gallego et al., 2019). Here, more stringent read recruitment criteria were applied to differentiate shared genes from the true presence of a viral sequence in a sample. Interestingly more than half of the viral assemblies identified across all individuals were homologous to crAss. It should be noted however, that these assemblies may also represent fragments of the same phage genome or family.

In contrast to these findings and to the proposed global distribution of crAss phage in human populations, the compilation of a large-scale gut virome database called into question the existence of a core human gut virome at a sequence level (Gregory et al., 2019). By examining VLP metagenomes from 572 individuals, this study proposed that a core human gut virome does not exist. Recently, Shkoporov et al., too made observations which support these findings by examining the virome of ten individuals across a 12-month period (Shkoporov et al., 2019). Here, a personal persistent virome (PPV) was observed that was composed of viral sequences detected in at least six of the 12 monthly time points. In accordance with previous longitudinal studies (Reyes et al., 2010;Minot et al., 2013), some viral sequences were present at nearly all time points within an individual. However, the virome was also highly individual-specific and viral sequences were not shared across the PPV of all individuals which supports the findings of Gregory et al. (Gregory et al., 2019). This high level of inter-individuality in the gut virome hampers our understanding of the virome in disease as it is difficult to detect common viral signals within or between cohorts. While it is likely that the individuality of the virome is driven by infection and resistance dynamics, the level of taxonomic resolution at which the virome is studied is also a contributory factor (Clooney et al., 2019). Sequence-based virome studies are carried out at the level of metagenomic assembly due to the absence of universal marker genes, limited database representation and established taxonomic organization. This represents species or strain-level resolution and is in contrast to the majority of bacterial metagenomics studies, which tend to be analysed at higher taxonomic ranks such as genus or family. It is possible to find patterns across virome cohorts using a minor subset of known viral sequences and by excluding unknown sequences (Norman et al., 2015;Monaco et al., 2016;Zhao et al., 2017;Zuo et al., 2019). However, it is not known if these subsets represent the dynamics of the virome as a whole. To this end, a number of clustering programs have been developed that group viral sequences based on shared protein families (Minot et al., 2012;Bolduc et al., 2017) such as vContact2 (Bolduc et al., 2017). This is a similar approach to that which was used to establish the extended crAss family despite low levels of nucleotide similarity between family members (Guerin et al., 2018;Yutin et al., 2018). In this way, protein-based clustering of the virome can reveal compositional patterns across individuals that were not visible at a nucleotide level. Furthermore, this approach allows for both the known and unknown components of the virome to be included in analysis giving new perspectives to the virome in health and disease.

Applying this protein family clustering approach (Jang et al., 2019) to the same longitudinal cohort can give new insights into the existence and composition of a core gut virome across individuals. This phylogenetic core was composed primarily of crAss and *Microviridae* and was not identifiable at a nucleotide level. Intriguingly, this core was not composed of temperate phage which is in contrast to observations in previous reports (Reyes et al., 2010). Furthermore, temperate phage were found to make up a minor subset of the core virome both within and between individuals. These findings suggest that mechanisms other than lysogenic replication are responsible for long term stability of the virome within healthy individuals. Moreover, they are in accordance with the global distribution and persistence of crAss in the human gut and ecological models such as the "royal family model". Upon clustering the stable fraction of individual viromes (PPV) the largest and most interconnected viral cluster as associated with known *Caudovirales* sequences. This is in accordance with previous observations (Minot et al., 2012;Bolduc et al., 2017) and reflects extensive shared genetic content across this order. It is also likely that this extensive gene sharing influenced previous database-dependent reports of temperate phage and *Caudovirales* dominance in the human gut virome. Furthermore, it highlights the importance of considering shared genes and gene cassettes when setting alignment criteria.

Gut virome in disease

Given the extensive evidence that phage can shape the composition and function of bacterial communities, the virome of the human gut has been studied in a number of diseases. However, as with the concept of the core virome, findings have been somewhat contradictory and any potential role for the virome in shaping the microbiome in disease remains elusive. Studies have reported gut phage populations were not significantly altered in diseases such as colorectal cancer and HIV-associated AIDS (Monaco et al., 2016;Hannigan et al., 2018), despite established associations between the gut microbiome and these diseases (Wirbel et al., 2019) (Dinh et al., 2014). This contradicts the established view that the virome and microbiome are closely linked but is more likely to reflect limitations of different analysis methods. These limitations include lenient alignment criteria to reference databases and the exclusion of viral dark matter from analysis. Furthermore, reports of changes in phage populations associated with diseases are limited to changes in the composition of known *Caudovirales* (Norman et al., 2015;Zhao et al., 2017;Ma et al., 2018;Fernandes et al., 2019). Given the limitations of *Caudovirales* taxonomy and the challenges presented by extensive gene sharing across the order, these findings provide little insight into any role of the virome in disease.

An intervention study by Gogokhia et al. (Gogokhia et al., 2019) sought to target cancer associated bacteria adherent invasive *Escherichia coli* and *Fusobacterium nucleatum* with lytic phage in a germ-free mouse model. However, a direct interaction between the mammalian immune system and phage virions resulted in an exacerbated colitic reaction. The authors also proposed that phage DNA plays a central role in phage interaction with the mammalian immune system as empty phage capsids did not induce an immune response. Similarly an *in vitro* study using *Staphylococcus aureus* and *Pseudomonas aeruginosa* phage observed a production of both pro and anti-inflammatory cytokines from peripheral blood mononuclear cells following endocytosis of purified phage virions (Van Belleghem et al., 2017). These observations are supported by the proposed ability of phage virions to cross the mammalian epithelial barrier *in vitro* via peptide sequences expressed on the capsid surface (Ivanenkov and Menon, 2000;Nguyen et al., 2017). In this way it is possible that phage communities in the human gut shape the gut microbiome indirectly through interactions with the mammalian immune system. This concept of phage translocation and interaction with mammalian immune system has also been discussed in a number of reviews and perspective pieces as follows (Górski et al., 2006;Górski et al., 2017;Łusiak-Szelachowska et al., 2017;Górski et al., 2018). Through the induction of a pro or anti-inflammatory response, phage could facilitate conditions that would favor a particular host or replication cycle. It is also possible that proposed phage–immune system interactions are driven by bacterial populations to facilitate infection or persistence in the human body. This was first demonstrated by lysogenic Pf phage which triggered maladaptive viral pattern recognition receptors and facilitated the chronic infection of *Pseudomonas aeruginosa* in murine and human cells (Sweere et al., 2019). This was also the first reported case of a directly pathogenic effect of phage in bacterial infection and demonstrated that phage do not need to directly encode virulence factors to impact the virulence of their host.

Faecal microbiota transplantation

FMT (faecal microbiota transplantation) is an emerging and experimental therapy that aims to restore healthy gut function through infusion of a faecal slurry from a healthy individual to the colon, cecum or duodenum of the recipient. It has been shown to be very effective in the treatment of recurrent *Clostridium difficile* infection (CDI) with donor bacteria colonizing recipients for up to a year (Jalanka et al., 2016) and reported success rates of 80–90% (Cammarota et al., 2014). The first evidence that the virome had potential as a tool to shape the microbiome and may play a role in the efficacy of FMT treatment was reported by Ott et al (Ott et al., 2017). In this study, patients with relapsing CDI received faecal filtrates from healthy donors that resulted in CDI symptoms being eliminated for up to 6 months. Furthermore recipient phage populations were substantially altered, resembling those of the donor for a minimum of six weeks. Surprisingly, *Lactococcus* phage were reported to dominate both donor and recipient virome, despite *Lactococcus* spp. representing only a minor fraction of the gut microbiome. This could reflect a dominance of lactococcal phage in the donor and recipient, implying lactococcal phage play an important role in homeostasis in CDI. However, phage sequence databases are dominated by those which are industrially relevant and cultivable, which includes lactococcal phage (Moineau and Lévesque, 2005). As a result, unknown sequences are statistically more likely to align to lactococcal phage and other cultivable or industrially relevant sequences when lenient alignment criteria are used. Thus, the dominance of lactococcal phage in the virome of FFT recipients is likely to be yet another artefact of database-dependent analysis methods. It should also be noted that, in accordance with the majority of database dependent virome studies, lactococcal phage are members of the order *Caudovirales*. This supports the concept that *Caudovirales* dominance reported at an order and family level in the gut virome could be driven by gene modules shared with *Caudovirales* in reference databases. However, the resemblance of the donor virome to that of the recipient suggests that, regardless of classification limitations, the virome plays a significant role in the maintenance of homeostasis in the gut. A subsequent study by Draper et al. (Draper et al., 2018) also reported the stable engraftment of donor phage populations in recipients for up to 12 months. In accordance with observations at a bacterial level (Jalanka et al., 2016;Wilson et al., 2019) successful phage engraftment was also dependent on specific donor recipient pairings. It should be noted that the role of other elements present in the filtrate (i.e. chromosomal DNA, plasmids, bacterial cellular components and signalling molecules) remains unknown and could also play a role in the restoration of healthy gut function in CDI. With conditions such as ulcerative colitis (UC) that are characterized by more subtle microbiome changes than CDI, successful FMT treatment (i.e. remission and mucosal healing) was not associated with changes in the phage population (Conceição-Neto et al., 2018). This is in contrast to reports of alterations in bacterial diversity following FMT in UC (Vaughn et al., 2016). Additionally, changes in the diversity of phage populations between healthy and UC cohorts were not observed (Conceição-Neto et al., 2018). These observations were in contrast with previous IBD virome studies, which reported differences in phage alpha diversity between healthy and UC cohorts (Norman et al., 2015;Fernandes et al., 2019).

Inflammatory bowel disease

Inflammatory bowel disease is a prevalent chronic disorder of the gastrointestinal tract with both genetic and environmental risk factors (Ng et al., 2017). The composition of gut bacteria and their interaction with the host immune system are believed to be central to its pathology, yet the aetiology of the disease remains poorly understood. Given the evidence that the virome can interact directly with the host immune system and shape the composition and function of the microbiome, both faecal and mucosal phage communities have been studied in IBD. Current understanding of phage populations in IBD has focused on VLPs, proposing that disease-specific patterns of

Caudovirales are linked to Crohn's disease (CD) and ulcerative colitis (UC). Furthermore, these changes in VLP composition have been reported not to reflect alterations of the bacterial community. However, the details of these compositional changes vary between studies (Lepage et al., 2008;Pérez-Brocal et al., 2013;Wagner et al., 2013;Norman et al., 2015;Pérez-Brocal et al., 2015;Zuo et al., 2019) .

Early IBD virome studies using the Roche 454 sequencing platform were limited by sequencing depth. These studies observed lower diversity and greater variation in the faecal VLP communities of patients with CD relative to healthy samples (Pérez-Brocal et al., 2013). The same group performed another sequence based analysis of the faecal and mucosal VLP community of CD and observed greater viral load and diversity in the faeces than mucosa of all individuals (Pérez-Brocal et al., 2015). Additionally, virome alpha diversity was reported to be significantly lower in disease. However, in contrast to their previous study, it was also reported that both healthy and CD cohorts were dominated by *Microviridae* rather than *Caudovirales*. Another study that analysed the first mucosal virome in paediatric CD(Wagner et al., 2013) proposed a dominance of *Caudovirales* phage overall and detected a single viral sequence in colonic mucosal samples from patients with CD. While this may suggest an extreme dominance of this virus in the mucosa of paediatric CD it more likely reflects insufficient sequencing depth resulting from low biomass samples and should be treated with caution.

As sequencing technology progressed, researchers were granted more detailed insights into the virome in IBD. However these insights contradicted previous findings and highlight the impact of sequencing platform on results. This could also suggest that our understanding of the virome in disease will continue to change as sequencing technology progresses. Illumina-based studies reported disease-specific increases in *Caudovirales* alpha diversity in the VLP viromes of UC and CD compared to healthy controls in adults (Norman et al., 2015)and children (Fernandes et al., 2019). Intriguingly, these alterations were also reported not to reflect changes in the bacterial community (bacteriome). Conversely, decreased alpha diversity was observed in *Caudovirales* families from the mucosal VLP virome of UC relative to healthy controls (Zuo et al., 2019). This supports the idea that viral communities and ecological models differ at different spatial locations within the gut. Additionally, it suggests that phage of the order *Caudovirales* play a central role shaping the microbiome in IBD. However, as with other studies of the virome in disease, these findings are limited to minor fractions of the dataset and offer little by way of insight into the role of phage in IBD.

To expand our understanding of the virome in IBD beyond the limitations of databases, we (Clooney et al.) applied the whole-virome analysis (Clooney et al., 2019) protocol discussed earlier to the keystone IBD virome dataset published by Norman et al. (Norman et al., 2015;Clooney et al., 2019). In this way, it was possible to gain the first insights into the composition viral dark matter in this disease. Contrary to the original findings of the study, alterations in the whole virome mirrored those of the bacteriome, and differences in overall virome alpha diversity were not seen at a sequence level. In accordance with current understanding the gut virome, high levels of inter-individuality were observed, and were likely to conceal any patterns in virome composition across individuals. Subsequently, we followed the protocol established in Shkoporov et al. (Shkoporov et al., 2019) to cluster viral sequences according to gene content. This revealed a core of primarily lytic phage in healthy individuals and supported observations of Shkoporov et al. (Shkoporov et al., 2019). However, this healthy core was also found to be absent in patients with IBD where it appeared to be replaced by a community of temperate phage. The majority (six) of the eight viral clusters which made up the healthy core virome in this analysis did not share homology to known viral sequences further highlighting the biological signals that may be missed when relying on database-dependent methods. Interestingly, one of these healthy core virome clusters was identified as crAss. This supports previous observations of its ubiquity in healthy human populations, low rates of detection in unhealthy individuals and its role in the core healthy human virome.

One possibility is that in the inflamed gut, environmental stresses from the human immune response such as antibodies and reactive oxygen species, leads to increased induction of the prophage present in the bacterial microbiome. The physiology of bacterial cells and the composition of the bacterial community influences whether integrated prophage enter the lytic or lysogenic replication cycle (Figure 1.A) (Casjens and Hendrix, 2015;Silpe and Bassler, 2019). It is therefore likely that temperate virome would react to environmental stress and resulting changes in the host community. This increased lytic replication and subsequent death of bacterial hosts would correspond with the observed reduction in bacterial alpha diversity associated with IBD. It would also correspond with the observed increase in in free phage virions (Lepage et al., 2008) and temperate VLPs in disease (Clooney et al., 2019). Additionally, the resulting increase of bacterial cell wall components and debris available to interact with the human immune system could perpetuate an inflammatory response. In accordance with *in vitro* reports, it is also possible that increased cell wall permeability associated with inflammation allows for increased phage translocation (Nguyen et al., 2017) and interaction with the host immune system (Górski et al., 2012;Van Belleghem et al., 2017;Gogokhia et al., 2019). The observations of (Clooney et al., 2019) provide a novel theoretical, mechanistic rationale for the interaction of the whole virome and the microbiome in disease, beyond taxonomic assignment and compositional patterns of the known minority. Additionally, they provide the first comprehensive evidence to support the mechanisms that had been previously proposed by Norman et al (Norman et al., 2015).

The extent to which the switch from temperate to lytic replication cycles and the composition of the core virome shape the gut microbiome and influence human health and disease remains speculative. However, the analysis approach outlined by(Clooney et al., 2019) paves the way to a better understanding of how the interplay between the microbiome and the virome reflects or influences human health and disease. By allowing the detection of biological signals across the entire virome it is now possible to identify viral signals associated with disease which had been previously undetected. The WVA approach is also supported by the methods used to characterize the crAss-like family (Guerin et al., 2018;Shkoporov et al., 2018a;Yutin et al., 2018). CrAss has progressed from an unknown viral sequencing anomaly to providing insights into the composition and function of the healthy human gut virome. In turn, the methods used in this progression provide a framework to characterize unknown but biologically relevant sequences identified by WVA.

It should be noted, that although many virome studies tend to report the dominance and composition of known *Caudovirales*, this provides little insight into the biological impact of these phages or how they shape the gut ecosystem. *Caudovirales* dominate reference databases, exhibit extensive gene sharing across families and orders and feature temperate phage genera (King et al., 2011;Iranzo et al., 2016;Jang et al., 2019). It is therefore likely that database-dependent methods of virome analysis classify unknown dsDNA virus sequences as known *Caudovirales* due to shared genes or gene cassettes and lenient detection criteria. As the representation of gut virome sequences in databases improves with efforts such as those carried out by Gregory et al.(Gregory et al., 2019) and the characterization of crAssphage (Guerin et al., 2018;Shkoporov et al., 2018a), database-dependent analysis methods may be able to better reflect the composition and dynamics of the virome. However, as extensive gene sharing remains a central part of dsDNA viral genomes it is essential that stringent alignment criteria are used to differentiate shared functional modules from the presence of a particular virus (Roux et al., 2016;Roux et al., 2017), regardless of the database used. Additionally, it is crucial that these alignment methods and their findings are validated to avoid misleading conclusions as to virome composition or function (Roux et al., 2013;Enault et al., 2017;Ott et al., 2017;Sutton et al., 2019b).

Addressing the current challenges of virome research

As has been discussed, studying phage of the human gut microbiome is made challenging by the composition of the sample (usually a faecal specimen) (Figure 2.A). In order to enrich for the VLP fraction of the virome, extensive mechanical, chemical and enzymatic processing is required to remove cellular DNA and dietary components (Figure 2.B) (Castro-Mejía et al., 2015;Shkoporov et al., 2018b). Unfortunately, this results in particularly low DNA yields that can complicate the generation of sequencing libraries. This challenge is more pronounced in mucosal virome studies where DNA yields are lower again (Hannigan et al., 2015;Hannigan et al., 2018). As a result, all but one (Manrique et al., 2016) virome study to date have depended on multiple displacement amplification (MDA) of viral DNA to reach sufficient quantities to sequence. As with all metagenomics, it is crucial to find the balance between sequencing chemistry, depth of sequencing and read length. These factors have profound impacts on the final virome sequences available for downstream analysis. This was highlighted by the differences in virome alpha diversity reported by 454 pyrosequencing, when compared to deeper sequencing on the Illumina platform as previously discussed. Short read platforms such as the Illumina HiSeq are a means to perform deep sequencing of the virome with low error rates and relatively low input DNA requirements. However, these libraries can also lead to fragmented assemblies and poor recovery of viral genomes (Sutton et al., 2019a). Long read sequencing offers a promising solution to this assembly challenge, as it is possible to sequence entire viral genomes on a single read (Supplementary table 1.). This overcomes hypervariable sequences and repeat regions in viral genomes which hamper assembly (Warwick-Dugdale et al., 2019). Currently long read sequencing platforms also require very large quantities of un-fragmented DNA, which can be challenging acquire from virome samples. As a result, the initial DNA yield and the amplification step directly influence the sequencing chemistry, read depth and read length which can be used with virome sequences (Figure 2.D). The MDA step has also been reported to introduce considerable bias into the composition of the resulting virome which must be considered when drawing conclusions from data (Figure 3.A) (Yilmaz et al., 2010;Probst et al., 2015). Studies have also reported MDA preferentially amplifies small circular ssDNA viruses, which include the family *Microviridae* (Figure 2.C) (Kim and Bae, 2011). This could call into question both the reported abundance and ubiquity of this family across individuals. Although it is difficult to quantify the extent to which preferential amplification occurs, recent meta-analysis of gut virome studies suggested *Microviridae* may be 10-fold lower in abundance than previously thought (Gregory et al., 2019). It is believed that priming biases of the random hexamers used in the MDA reaction do not prime equally across all genomes, making quantitative interpretation of virome data difficult.

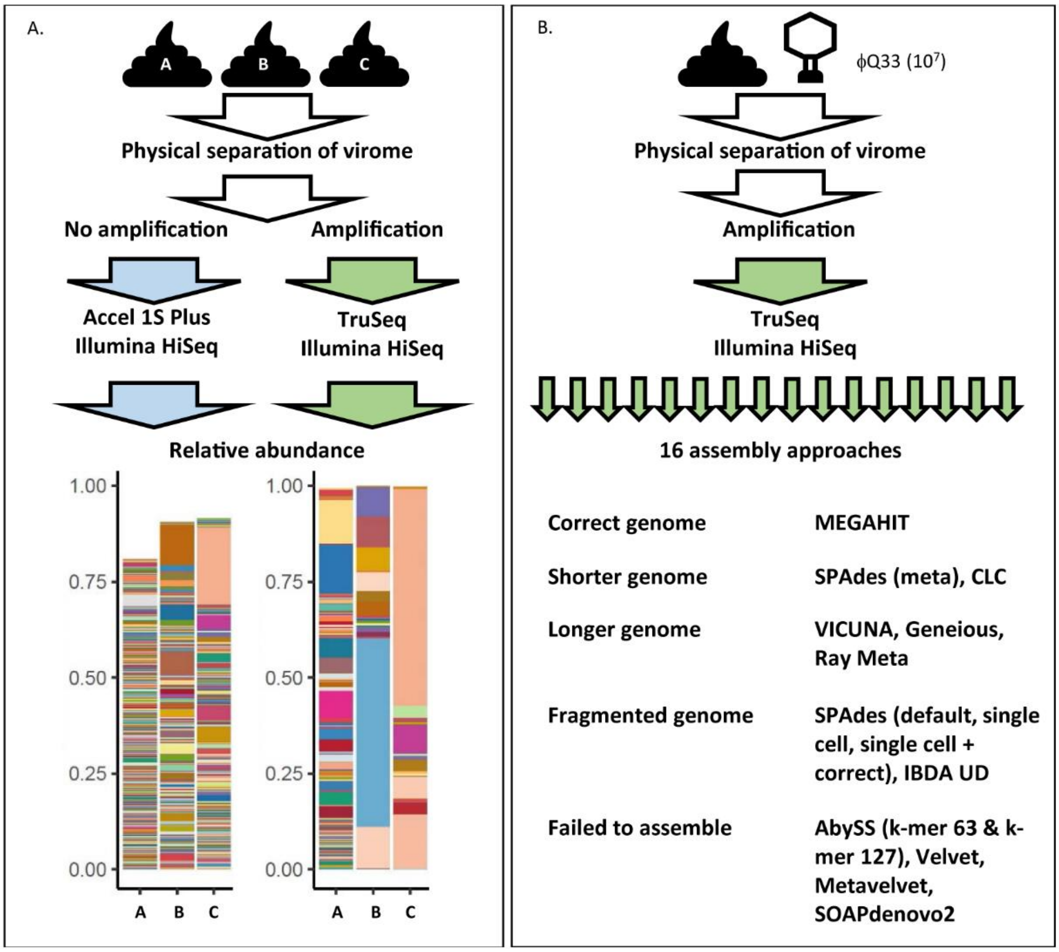
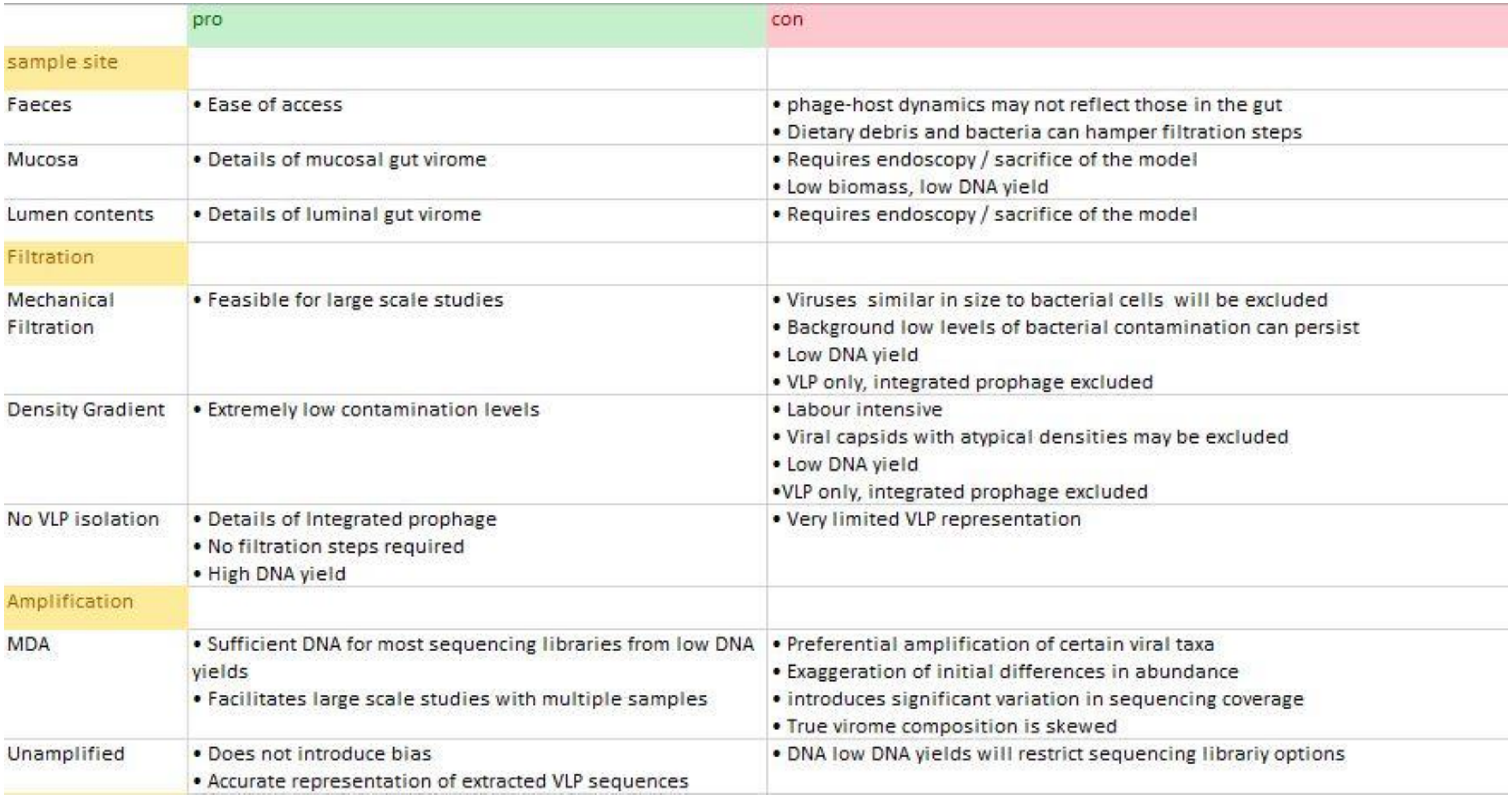


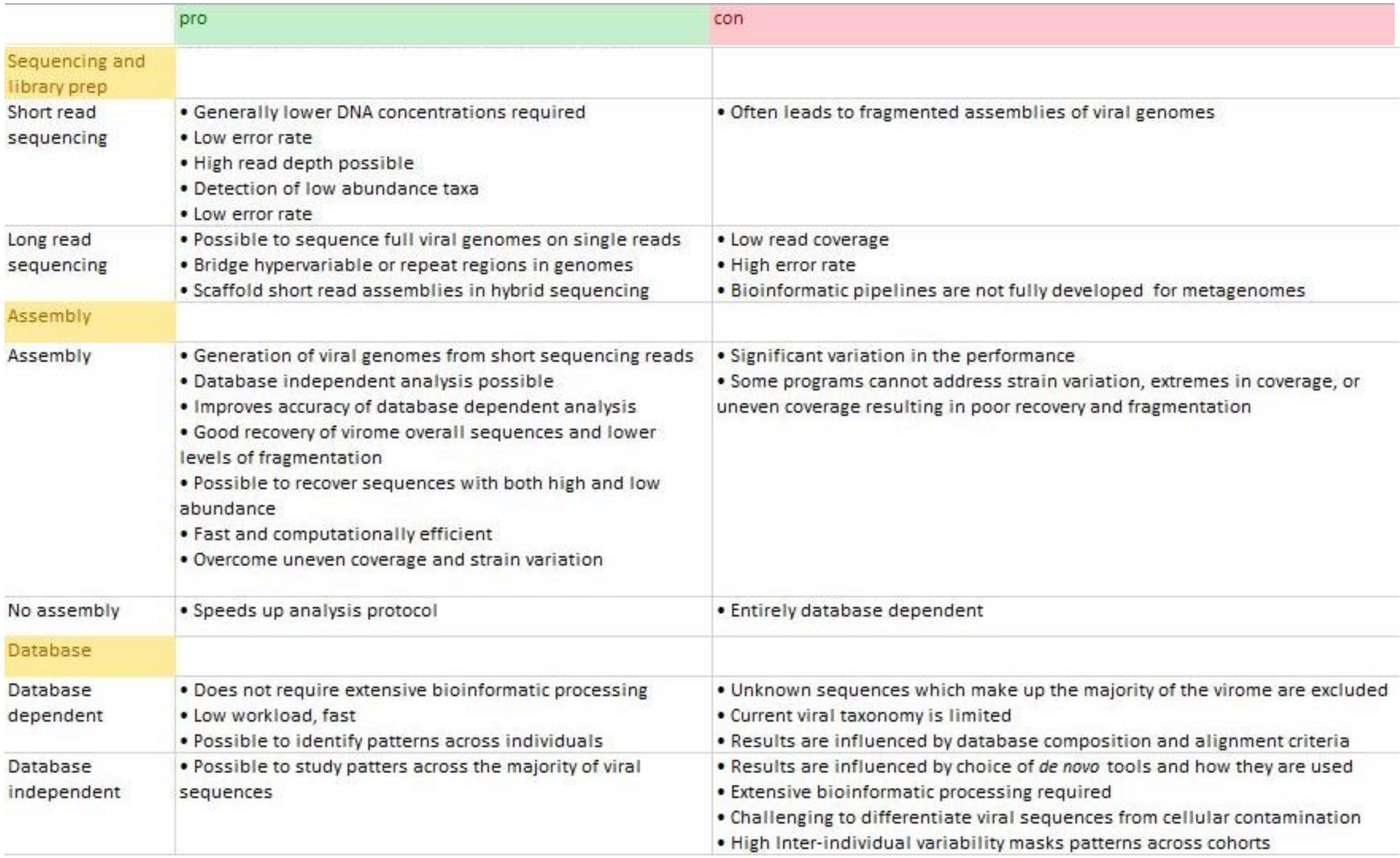
Figure 3. Examples of how virome composition is influenced by key steps in analysis.

(A) Three samples were subjected to identical filtration and DNA extraction steps. One set was amplified and prepared for sequencing using the Illumina TruSeq library kit while another set of unamplified samples were prepared using the Accel 1S Plus kit. Both sets were sequenced on the Illumina HiSeq platform. Differently treated samples differ in terms of final composition, represented in bar plots. Each colour represents the relative abundance of a unique viral contig in each sample. Abundance does not reach 100% in the unamplified sample as the higher level of richness also hampered assembly (adapted from (Shkoporov et al., 2019)). (B) Impact of assembly software on final virome composition. A faecal samples was spiked with ϕQ33 10^7 PFU ml^{-1}, extracted and sequenced. These sequences were assembled using 16 assembly programs. Only one assembler identified the genome in a single contig of the correct length. Five assemblers completely failed to assemble the genome and a further five generated fragmented assemblies. (Sutton et al., 2019a).



	pro	con
sample site		
Faeces	• Ease of access	• phage-host dynamics may not reflect those in the gut • Dietary debris and bacteria can hamper filtration steps
Mucosa	• Details of mucosal gut virome	• Requires endoscopy / sacrifice of the model • Low biomass, low DNA yield
Lumen contents	• Details of luminal gut virome	• Requires endoscopy / sacrifice of the model
Filtration		
Mechanical Filtration	• Feasible for large scale studies	• Viruses similar in size to bacterial cells will be excluded • Background low levels of bacterial contamination can persist • Low DNA yield • VLP only, integrated prophage excluded
Density Gradient	• Extremely low contamination levels	• Labour intensive • Viral capsids with atypical densities may be excluded • Low DNA yield •VLP only, integrated prophage excluded
No VLP isolation	• Details of Integrated prophage • No filtration steps required • High DNA yield	• Very limited VLP representation
Amplification		
MDA	• Sufficient DNA for most sequencing libraries from low DNA yields • Facilitates large scale studies with multiple samples	• Preferential amplification of certain viral taxa • Exaggeration of initial differences in abundance • introduces significant variation in sequencing coverage • True virome composition is skewed
Unamplified	• Does not introduce bias • Accurate representation of extracted VLP sequences	• DNA low DNA yields will restrict sequencing librariy options

Table S1. Pros and Cons of key steps in virome analysis pipelines



	pro	con
Sequencing and library prep		
Short read sequencing	• Generally lower DNA concentrations required • Low error rate • High read depth possible • Detection of low abundance taxa • Low error rate	• Often leads to fragmented assemblies of viral genomes
Long read sequencing	• Possible to sequence full viral genomes on single reads • Bridge hypervariable or repeat regions in genomes • Scaffold short read assemblies in hybrid sequencing	• Low read coverage • High error rate • Bioinformatic pipelines are not fully developed for metagenomes
Assembly		
Assembly	• Generation of viral genomes from short sequencing reads • Database independent analysis possible • Improves accuracy of database dependent analysis • Good recovery of virome overall sequences and lower levels of fragmentation • Possible to recover sequences with both high and low abundance • Fast and computationally efficient • Overcome uneven coverage and strain variation	• Significant variation in the performance • Some programs cannot address strain variation, extremes in coverage, or uneven coverage resulting in poor recovery and fragmentation
No assembly	• Speeds up analysis protocol	• Entirely database dependent
Database		
Database dependent	• Does not require extensive bioinformatic processing • Low workload, fast • Possible to identify patterns across individuals	• Unknown sequences which make up the majority of the virome are excluded • Current viral taxonomy is limited • Results are influenced by database composition and alignment criteria
Database independent	• Possible to study patters across the majority of viral sequences	• Results are influenced by choice of *de novo* tools and how they are used • Extensive bioinformatic processing required • Challenging to differentiate viral sequences from cellular contamination • High Inter-individual variability masks patterns across cohorts

Table S1. Pros and Cons of key steps in virome analysis pipelines

MDA bias can also have a significant impact on qualitative analysis of the gut virome. As the concentration of the DNA template also impacts the products of an MDA reaction, initial log-fold differences in the abundance of viral sequences are exaggerated by MDA (Zhang et al., 2006;Woyke et al., 2009). This results in extremes of both high and low read coverage and uneven representation of the initial metagenome. As high-abundance sequences sequester sequencing resources, low-abundance sequences can be insufficiently covered (García-López et al., 2015). These coverage extremes have profound impacts on a number of steps in bioinformatic pipelines, but in particular metagenomic assembly (García-López et al., 2015;Sutton et al., 2019a). A recent comparison of all assemblers used in virome studies to date observed that both high and low-coverage sequences resulted in fragmented assemblies and recovered only small proportions of viral genomes (Sutton et al., 2019a). Furthermore, samples were spiked with an abundant (10^7 plaque-forming units ml^{-1}) exogenous lactococcal phage Q33. These samples underwent identical extraction amplification and sequencing and resulting viromes were then assembled using all assembly methods which had been reported in virome studies at the time (16). Ten of the assemblers either failed to recover or significantly fragmented the Q33 genome and only one assembler recovered the genome at the correct length (Figure 3.B) (Sutton et al., 2019a). These results suggest that numerous and potentially biologically relevant viral sequences may be not only be skewed in abundance but also excluded by current virome analysis protocols. This also means that we see the virome through the lens of the extraction protocol before any decision has been made to use database-dependent or independent methods (Figure 2.).

As with many microbiome studies, conclusions from virome studies are primarily drawn from relative rather than absolute abundances of sequences. As has been discussed, these abundances are often skewed by MDA bias. This ambiguity highlights the need for quantitative analysis protocols in virome studies as was recently described (Shkoporov et al., 2018b;Clooney et al., 2019;Shkoporov et al., 2019). These studies reported significant differences in the overall viral load between individuals. This total viral load was also correlated negatively with viral alpha diversity within the sample and the presence of abundant non-temperate phage such as *Microviridae* and crAss. These results suggest that high viral load is associated with a low number of abundant phage, which consequently mask underlying temperate phage diversity. This also suggests the maintenance of high-abundance non-temperate phage may be closely linked to the health status of the gut. Subsequently, this may act as a useful biomarker for disease and give insight into the phage-host dynamics related to microbiome stability and disease status.

Future Prospects and Conclusions

Sequence-based analysis of the bacteriophage in the human gut has revolutionized our view of the gut virome and its relationship with the microbiome. However, this new insight has also revealed how little we know about this relationship. Our current understanding is founded predominantly on extending our knowledge generated in reductionist phage-host studies *in vitro*, or by large scale metagenomic studies of the VLP fraction. While these *in vitro* studies give detailed insights into the mechanics of individual phage-host interactions, the prevalence of these interactions in the gut ecosystem remains speculative. Additionally, numerous studies have also reported that these interactions can change dramatically in the gut (De Sordi et al., 2017;Gogokhia et al., 2019). Large-scale metagenomic studies suggest that the virome and the microbiome are closely linked, but these studies tend to give broad overviews of subsets of the virome and lack details on the host-phage interactions. This leads to significant gaps in our understanding of how phage-host dynamics *in vitro* differ from phage-host dynamics occurring in the gut.

Application of similar analysis methods across studies (i.e. MDA, alignment to reference databases, reports on *Caudovirales* alpha diversity) allows for comparison across samples and studies. However, many virome studies present inconclusive or contradictory results that hinder the progression of the field. Arguably the overreliance on these analysis methods is largely to blame for the gaps in our understanding of the virome. This is also supported by recent observations that methodology had greater impact on the conclusions drawn from virome studies than health or disease status (Gregory et al., 2019). As conclusions are drawn from minor fractions of data and as detection criteria do not take into account phage biology and evolutionary history, we must pose the question: is the gut virome field built on unstable foundations? With this in mind, due caution must be used when interpreting the findings of virome data. As virome data is particularly sensitive to methodological bias, conclusions must be considered in the context of the analysis methods used (Figure 2, Supplementary table 1.). These limitations highlight the need for radically new approaches to studying the virome if we are to understand its role in shaping the microbiome in health and disease.

Developments in sequencing library kits such as the Swift Biosciences Accel 1S Plus kit or extraction protocols like the linker amplified displacement LADs (Roux et al., 2016), offer potential solutions to creating unbiased sequencing libraries from low-input DNA yields. Through the removal of MDA bias and spiking known concentrations of exogenous phage, (Shkoporov et al., 2018b) it may be possible to gain new insights into the true composition and absolute abundance of the virome. As has been discussed at length, the sensitivity of virome data to methodological bias highlights the critical need for extensive optimization and validation of all steps of virome analysis protocols, from wet-lab extraction protocols (Roux et al., 2016;Roux et al., 2017;Shkoporov et al., 2018b) to bioinformatic pipelines (García-López et al., 2015;Hesse et al., 2017;Vollmers et al., 2017;Sutton et al., 2019a). However, standardization and consistency must not be at the cost of developing new methods. Furthermore, when characterizing viral sequences it is crucial to use stringent detection criteria to minimize the impact of spurious alignments and the influence of gene sharing across dsDNA viruses.

Significant progress has been made in increasing the representation of gut phage in reference databases and there is a growing consensus that viral taxonomy will soon move towards sequence-based taxonomy (Paez-Espino et al., 2016;Simmonds et al., 2017;Aiewsakun et al., 2018;Eloe-Fadrosh, 2019;Gregory et al., 2019). However, proposed protocols to add metagenomic sequences to current taxonomic systems have not yet been accepted (Simmonds and Aiewsakun, 2018). Given the dominance of unknown sequences in virome data it is therefore crucial to accept the current limitations of phage taxonomy. Rather than force the virome to fit current taxonomic systems, we propose that future studies should allow the virome to reveal its own targets for downstream characterization. Subsequently, we have outlined a method to analyse the virome in its entirety with our WVA protocol. Furthermore we have described a framework to characterize unknown but biologically relevant viral sequences that may be identified using WVA. In this way it may be possible to address the gaps in our understanding of phage-host dynamics in the human gut, and see existing datasets in a new light.

To what extent the phage of the human gut shape the microbiome will dictate whether it will be possible to use phage as a therapeutic tool in the future. There are significant gaps in our understanding of phage-host interactions which need to be addressed before we can reach any conclusions on the usefulness of phage as a biotherapeutic. By increasing our understanding of phage-host interactions in the gut, it may be possible to pave the way for therapeutic applications of phage in the human body. However, the limited insights we have been granted to date of phage-host interactions have also highlighted some significant hurdles facing phage therapy. Early evidence that the virome could play a role in the success of FMT (Ott et al., 2017;Draper et al., 2018) suggests there may be a future in using the virome to shape the microbiome in disease. To date, the majority of phage intervention studies are based on single phage-host pairs (or cocktails containing limited numbers of phage) *in vitro* which have been shown to be significantly different to phage-host dynamics *in vivo* (De Sordi et al., 2017). For example, phage have been found to switch hosts and interact directly with mammalian immune cells *in vivo*, which has serious implications for the future of phage therapy (Górski et al., 2012;Gogokhia et al., 2019). Additionally the phage-mediated transfer of host virulence factors (Waldor and Mekalanos, 1996;Friman et al., 2011;Muniesa et al., 2012;Scanlan et al., 2015) as well as direct pathogenesis of phage capsids (Sweere et al., 2019) suggests phage could be a potential risk in therapeutic settings. These challenges are confounded with regulatory issues (Brüssow, 2019) and additional gaps in our understanding of the pharmacodynamics of phage in mammalian tissue. Having the potential to directly interact with the immune system (Ivanenkov and Menon, 2000;Barr et al., 2013;Barr et al., 2015;Nguyen et al., 2017) given their larger size relative to other biological therapeutic agents makes phage a more complex therapeutic agent than any that have preceded them. However, in light of the increasing incidence of bacterial pathogens which are resistant to antibiotics, and given the promising results of some existing phage therapy trials (Wright et al., 2009;Chan et al., 2018;LaVergne et al., 2018;Garrett, 2019), overcoming these challenges is critically important. Similarly the gaps in our understanding of how phage shape bacterial communities will need to be addressed if phages are to have a role in avoiding future global health issues.

Acknowledgments

The authors thank Dr. Andrey Shkoporov, Dr. Lorraine Draper, Dr. Hugh Harris and Dr. Adam Clooney for their critical remarks during the preparation of this manuscript. The authors also thank Dr. Andrey Shkoporov for providing data featured in Figure 3.

References

Aggarwala, V., Liang, G., and Bushman, F.D. (2017). Viral communities of the human gut: metagenomic analysis of composition and dynamics. *Mobile DNA* 8**,** 12.

Aiewsakun, P., Adriaenssens, E.M., Lavigne, R., Kropinski, A.M., and Simmonds, P. (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* 99**,** 1331-1343. doi:10.1099/jgv.0.001110.

Avrani, S., Schwartz, D.A., and Lindell, D. (2012). Virus-host swinging party in the oceans: Incorporating biological complexity into paradigms of antagonistic coexistence. *Mobile genetic elements* 2**,** 88-95.

Bair, C.L., and Black, L.W. (2007). A type IV modification dependent restriction nuclease that targets glucosylated hydroxymethyl cytosine modified DNAs. *Journal of molecular biology* 366**,** 768-778.

Barr, J.J., Auro, R., Furlan, M., Whiteson, K.L., Erb, M.L., Pogliano, J. et al. (2013). Bacteriophage adhering to mucus provide a non–host-derived immunity. *Proceedings of the National Academy of Sciences* 110**,** 10771-10776.

Barr, J.J., Auro, R., Sam-Soon, N., Kassegne, S., Peters, G., Bonilla, N. et al. (2015). Subdiffusive motion of bacteriophage in mucosal surfaces increases the frequency of bacterial encounters. *Proceedings of the National Academy of Sciences* 112**,** 13675-13680.

Barylski, J., Enault, F., Dutilh, B.E., Schuller, M.B.P., Edwards, R.A., Gillis, A. et al. (2018). Analysis of Spounaviruses as a Case Study for the Overdue Reclassification of Tailed Bacteriophages. BioRxiv [Preprint].

Available at: https://www.biorxiv.org/content/10.1101/220434v2.full (Accessed August 09, 2019) *bioRxiv*. doi:10.1101/220434.

Belkaid, Y., and Hand, T.W. (2014). Role of the microbiota in immunity and inflammation. *Cell* 157**,** 121-141.

Bolduc, B., Jang, H.B., Doulcier, G., You, Z.-Q., Roux, S., and Sullivan, M.B. (2017). vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. *PeerJ* 5**,** e3243.

Bondy-Denomy, J., Pawluk, A., Maxwell, K.L., and Davidson, A.R. (2013). Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. *Nature* 493**,** 429.

Breitbart, M. (2011). Marine viruses: truth or dare.

Breitbart, M., Bonnain, C., Malki, K., and Sawaya, N.A. (2018). Phage puppet masters of the marine microbial realm. *Nature Microbiology* 3**,** 754-766. doi:10.1038/s41564-018-0166-y.

Breitbart, M., Haynes, M., Kelley, S., Angly, F., Edwards, R.A., Felts, B. et al. (2008). Viral diversity and dynamics in an infant gut. *Research in Microbiology* 159**,** 367-373. doi:https://doi.org/10.1016/j.resmic.2008.04.006.

Breitbart, M., Hewson, I., Felts, B., Mahaffy, J.M., Nulton, J., Salamon, P. et al. (2003). Metagenomic analyses of an uncultured viral community from human feces. *Journal of bacteriology* 185**,** 6220-6223.

Brüssow, H. (2019). Hurdles for Phage Therapy to Become a Reality—An Editorial Comment. *Viruses* 11. doi:10.3390/v11060557.

Brüssow, H., Canchaya, C., and Hardt, W.-D. (2004). Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol. Mol. Biol. Rev.* 68**,** 560-602.

Cammarota, G., Ianiro, G., and Gasbarrini, A. (2014). Fecal microbiota transplantation for the treatment of Clostridium difficile infection: a systematic review. *Journal of clinical gastroenterology* 48**,** 693-702.

Casjens, S.R., and Hendrix, R.W. (2015). Bacteriophage lambda: Early pioneer and still relevant. *Virology* 479-480**,** 310-330. doi:https://doi.org/10.1016/j.virol.2015.02.010.

Castro-Mejía, J.L., Muhammed, M.K., Kot, W., Neve, H., Franz, C.M., Hansen, L.H. et al. (2015). Optimizing protocols for extraction of bacteriophages prior to metagenomic analyses of phage communities in the human gut. *Microbiome* 3**,** 64.

Cenens, W., Makumi, A., Govers, S.K., Lavigne, R., and Aertsen, A. (2016). Viral Transmission Dynamics at Single-Cell Resolution Reveal Transiently Immune Subpopulations Caused by a Carrier State Association. *PLOS Genetics* 11**,** e1005770. doi:10.1371/journal.pgen.1005770.

Chan, B.K., Turner, P.E., Kim, S., Mojibian, H.R., Elefteriades, J.A., and Narayan, D. (2018). Phage treatment of an aortic graft infected with Pseudomonas aeruginosa. *Evolution, medicine, and public health* 2018**,** 60-66.

Chatterjee, S., and Rothenberg, E. (2012). Interaction of bacteriophage l with its E. coli receptor, LamB. *Viruses* 4**,** 3162-3178.

Chen, J., Quiles-Puchalt, N., Chiang, Y.N., Bacigalupe, R., Fillol-Salom, A., Chee, M.S.J. et al. (2018). Genome hypermobility by lateral transduction. *Science* 362**,** 207-212. doi:10.1126/science.aat5867.

Chey, W.D., Kurlander, J., and Eswaran, S. (2015). Irritable bowel syndrome: a clinical review. *Jama* 313**,** 949-958.

Chung, I.-Y., Jang, H.-J., Bae, H.-W., and Cho, Y.-H. (2014). A phage protein that inhibits the bacterial ATPase required for type IV pilus assembly. *Proceedings of the National Academy of Sciences* 111**,** 11503-11508.

Clement, J.-M., Lepouce, E., Marchal, C., and Hofnung, M. (1983). Genetic study of a membrane protein: DNA sequence alterations due to 17 lamB point mutations affecting adsorption of phage lambda. *The EMBO journal* 2**,** 77-80.

Clooney, A.G., Sutton, T.D.S., Shkoporov, A.N., Holohan, R.K., Daly, K.M., O'regan, O. et al. (2019). Whole-Virome Analysis Sheds Light on Viral Dark Matter in Inflammatory Bowel Disease. *Cell Host & Microbe* 26**,** 764-778.e765. doi:https://doi.org/10.1016/j.chom.2019.10.009.

Conceição-Neto, N., Deboutte, W., Dierckx, T., Machiels, K., Wang, J., Yinda, K.C. et al. (2018). Low eukaryotic viral richness is associated with faecal microbiota transplantation success in patients with UC. *Gut* 67**,** 1558-1559. doi:10.1136/gutjnl-2017-315281.

Cumby, N., Reimer, K., Mengin-Lecreulx, D., Davidson, A.R., and Maxwell, K.L. (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* 96**,** 437-447. doi:10.1111/mmi.12918.

Davies, M.R., Broadbent, S.E., Harris, S.R., Thomson, N.R., and Van Der Woude, M.W. (2013). Horizontally Acquired Glycosyltransferase Operons Drive Salmonellae Lipopolysaccharide Diversity. *PLOS Genetics* 9**,** e1003568. doi:10.1371/journal.pgen.1003568.

De Sordi, L., Khanna, V., and Debarbieux, L. (2017). The gut microbiota facilitates drifts in the genetic diversity and infectivity of bacterial viruses. *Cell host & microbe* 22**,** 801-808. e803.

De Sordi, L., Lourenço, M., and Debarbieux, L. (2019). "I will survive": A tale of bacteriophage-bacteria coevolution in the gut. *Gut Microbes* 10**,** 92-99. doi:10.1080/19490976.2018.1474322.

Denou, E., Berger, B., Barretto, C., Panoff, J.-M., Arigoni, F., and Brüssow, H. (2007). Gene expression of commensal Lactobacillus johnsonii strain NCC533 during in vitro growth and in the murine gut. *Journal of bacteriology* 189**,** 8109-8119.

Diard, M., Bakkeren, E., Cornuault, J.K., Moor, K., Hausmann, A., Sellin, M.E. et al. (2017). Inflammation boosts bacteriophage transfer between Salmonella spp. *Science* 355**,** 1211-1215.

Dinh, D.M., Volpe, G.E., Duffalo, C., Bhalchandra, S., Tai, A.K., Kane, A.V. et al. (2014). Intestinal microbiota, microbial translocation, and systemic inflammation in chronic HIV infection. *The Journal of infectious diseases* 211**,** 19-27.

Draper, L.A., Ryan, F.J., Smith, M.K., Jalanka, J., Mattila, E., Arkkila, P.A. et al. (2018). Long-term colonisation with donor bacteriophages following successful faecal microbial transplantation. *Microbiome* 6**,** 220. doi:10.1186/s40168-018-0598-x.

Duncan, S.H., Louis, P., Thomson, J.M., and Flint, H.J. (2009). The role of pH in determining the species composition of the human colonic microbiota. *Environmental microbiology* 11**,** 2112-2122.

Dutilh, B.E., Cassman, N., Mcnair, K., Sanchez, S.E., Silva, G.G., Boling, L. et al. (2014). A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nature communications* 5**,** ncomms5498.

Dutilh, B.E., Schmieder, R., Nulton, J., Felts, B., Salamon, P., Edwards, R.A. et al. (2012). Reference-independent comparative metagenomics using cross-assembly: crAss. *Bioinformatics* 28**,** 3225-3231.

Dy, R.L., Przybilski, R., Semeijn, K., Salmond, G.P., and Fineran, P.C. (2014). A widespread bacteriophage abortive infection system functions through a Type IV toxin–antitoxin mechanism. *Nucleic acids research* 42**,** 4590-4605.

Edwards, R.A., Vega, A.A., Norman, H.M., Ohaeri, M., Levi, K., Dinsdale, E.A. et al. (2019). Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. *Nature Microbiology*. doi:10.1038/s41564-019-0494-6.

Eloe-Fadrosh, E.A. (2019). Towards a genome-based virus taxonomy. *Nature Microbiology* 4**,** 1249-1250. doi:10.1038/s41564-019-0511-9.

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

Espey, M.G. (2013). Role of oxygen gradients in shaping redox relationships between the human intestine and its microbiota. *Free Radical Biology and Medicine* 55**,** 130-140.

Fernandes, M.A., Verstraete, S.G., Phan, T.G., Deng, X., Stekol, E., Lamere, B. et al. (2019). Enteric virome and bacterial microbiota in children with ulcerative colitis and Crohn disease. *Journal of pediatric gastroenterology and nutrition* 68**,** 30-36.

Forster, S.C., Kumar, N., Anonye, B.O., Almeida, A., Viciani, E., Stares, M.D. et al. (2019). A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nature biotechnology* 37**,** 186.

Friman, V.-P., Hiltunen, T., Jalasvuori, M., Lindstedt, C., Laanto, E., Örmälä, A.-M. et al. (2011). High Temperature and Bacteriophages Can Indirectly Select for Bacterial Pathogenicity in Environmental Reservoirs. *PLOS ONE* 6**,** e17651. doi:10.1371/journal.pone.0017651.

Galtier, M., Sordi, L.D., Sivignon, A., De Vallée, A., Maura, D., Neut, C. et al. (2017). Bacteriophages targeting adherent invasive Escherichia coli strains as a promising new treatment for Crohn’s disease. *Journal of Crohn's and Colitis* 11**,** 840-847.

Gandon, S., Buckling, A., Decaestecker, E., and Day, T. (2008). Host–parasite coevolution and patterns of adaptation across time and space. *Journal of evolutionary biology* 21**,** 1861-1866.

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

Garrett, L. (2019). Seven circles of antimicrobial hell. *The Lancet* 393**,** 865-867.

Gogokhia, L., Buhrke, K., Bell, R., Hoffman, B., Brown, D.G., Hanke-Gogokhia, C. et al. (2019). Expansion of Bacteriophages Is Linked to

Aggravated Intestinal Inflammation and Colitis. *Cell Host & Microbe* 25**,** 285-299.e288. doi:https://doi.org/10.1016/j.chom.2019.01.008.

Górski, A., Dąbrowska, K., Międzybrodzki, R., Weber-Dąbrowska, B., Łusiak-Szelachowska, M., Jończyk-Matysiak, E. et al. (2017). Phages and immunomodulation. *Future Microbiology* 12**,** 905-914. doi:10.2217/fmb-2017-0049.

Górski, A., Jończyk-Matysiak, E., Międzybrodzki, R., Weber-Dąbrowska, B., Łusiak-Szelachowska, M., Bagińska, N. et al. (2018). Phage Therapy: Beyond Antibacterial Action. *Frontiers in Medicine* 5. doi:10.3389/fmed.2018.00146.

Górski, A., Międzybrodzki, R., Borysowski, J., Dąbrowska, K., Wierzbicki, P., Ohams, M. et al. (2012). "Chapter 2 - Phage as a Modulator of Immune Responses: Practical Implications for Phage Therapy," in *Advances in Virus Research*, eds. M. Łobocka & W. Szybalski. Academic Press), 41-71.

Górski, A., Ważna, E., Dąbrowska, B.-W., Dąbrowska, K., Świtała-Jeleń, K., and Międzybrodzki, R. (2006). Bacteriophage translocation. *Pathogens and Disease* 46**,** 313-319. doi:10.1111/j.1574-695X.2006.00044.x.

Gregory, A.C., Zablocki, O., Howell, A., Bolduc, B., and Sullivan, M.B. (2019). The human gut virome database. BioRxiv [Preprint]. Available at: https://www.biorxiv.org/content/10.1101/655910v1.full (Accessed August 09, 2019). *BioRxiv*.

Guerin, E., Shkoporov, A., Stockdale, S.R., Clooney, A.G., Ryan, F.J., Sutton, T.D.S. et al. (2018). Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. *Cell Host & Microbe* 24**,** 653-664.e656. doi:https://doi.org/10.1016/j.chom.2018.10.002.

Halfvarson, J., Brislawn, C.J., Lamendella, R., Vázquez-Baeza, Y., Walters, W.A., Bramer, L.M. et al. (2017). Dynamics of the human gut microbiome in inflammatory bowel disease. *Nature microbiology* 2**,** 17004.

Hall, A.R., Scanlan, P.D., Morgan, A.D., and Buckling, A. (2011). Host-parasite coevolutionary arms races give way to fluctuating selection. *Ecol Lett* 14**,** 635-642. doi:10.1111/j.1461-0248.2011.01624.x.

Hannigan, G.D., Duhaime, M.B., Ruffin, M.T., Koumpouras, C.C., and Schloss, P.D. (2018). Diagnostic Potential and Interactive Dynamics of the Colorectal Cancer Virome. *mBio* 9**,** e02248-02218. doi:10.1128/mBio.02248-18.

Hannigan, G.D., Meisel, J.S., Tyldsley, A.S., Zheng, Q., Hodkinson, B.P., Sanmiguel, A.J. et al. (2015). The human skin double-stranded DNA virome: topographical and temporal diversity, genetic enrichment, and dynamic associations with the host microbiome. *MBio* 6**,** e01578-01515.

Hesse, U., Van Heusden, P., Kirby, B.M., Olonade, I., Van Zyl, L.J., and Trindade, M. (2017). Virome Assembly and Annotation: A Surprise in the Namib Desert. *Frontiers in Microbiology* 8. doi:10.3389/fmicb.2017.00013.

Hofer, B., Ruge, M., and Dreiseikelmann, B. (1995). The superinfection exclusion gene (sieA) of bacteriophage P22: identification and overexpression of the gene and localization of the gene product. *Journal of bacteriology* 177**,** 3080-3086.

Hoyles, L., Mccartney, A.L., Neve, H., Gibson, G.R., Sanderson, J.D., Heller, K.J. et al. (2014). Characterization of virus-like particles associated with the human faecal and caecal microbiota. *Research in microbiology* 165**,** 803-812.

Hulo, C., Masson, P., Le Mercier, P., and Toussaint, A. (2015). A structured annotation frame for the transposable phages: A new proposed family “Saltoviridae” within the Caudovirales. *Virology* 477**,** 155-163. doi:https://doi.org/10.1016/j.virol.2014.10.009.

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. *The ISME journal* 9**,** 472.

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

Hurwitz, B.L., and Sullivan, M.B. (2013). The Pacific Ocean Virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PloS one* 8**,** e57355.

Iranzo, J., Krupovic, M., and Koonin, E.V. (2016). The Double-Stranded DNA Virosphere as a Modular Hierarchical Network of Gene Sharing. *mBio* 7**,** e00978-00916. doi:10.1128/mBio.00978-16.

Ivanenkov, V.V., and Menon, A.G. (2000). Peptide-Mediated Transcytosis of Phage Display Vectors in MDCK Cells. *Biochemical and Biophysical Research Communications* 276**,** 251-257. doi:https://doi.org/10.1006/bbrc.2000.3358.

Jalanka, J., Mattila, E., Jouhten, H., Hartman, J., De Vos, W.M., Arkkila, P. et al. (2016). Long-term effects on luminal and mucosal microbiota and commonly acquired taxa in faecal microbiota transplantation for recurrent Clostridium difficile infection. *BMC Medicine* 14**,** 155. doi:10.1186/s12916-016-0698-z.

Jang, H.B., Bolduc, B., Zablocki, O., Kuhn, J.H., Roux, S., Adriaenssens, E.M. et al. (2019). Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nature biotechnology* 37**,** 632.

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***,** AEM. 00289-00211.

King, A.M., Lefkowitz, E., Adams, M.J., and Carstens, E.B. (2011). *Virus taxonomy: ninth report of the International Committee on Taxonomy of Viruses.* Elsevier.

Knowles, B., Silveira, C.B., Bailey, B.A., Barott, K., Cantu, V.A., Cobián-Güemes, A.G. et al. (2016). Lytic to temperate switching of viral communities. *Nature* 531**,** 466. doi:10.1038/nature17193.

Kronheim, S., Daniel-Ivad, M., Duan, Z., Hwang, S., Wong, A.I., Mantel, I. et al. (2018). A chemical defence against phage infection. *Nature* 564**,** 283-286. doi:10.1038/s41586-018-0767-x.

Labrie, S.J., Samson, J.E., and Moineau, S. (2010). Bacteriophage resistance mechanisms. *Nature Reviews Microbiology* 8**,** 317.

Lang, A.S., and Beatty, J.T. (2007). Importance of widespread gene transfer agent genes in α-proteobacteria. *Trends in Microbiology* 15**,** 54-62. doi:https://doi.org/10.1016/j.tim.2006.12.001.

Lauro, F.M., Mcdougald, D., Thomas, T., Williams, T.J., Egan, S., Rice, S. et al. (2009). The genomic basis of trophic strategy in marine bacteria. *Proceedings of the National Academy of Sciences* 106**,** 15527-15533.

Lavergne, S., Hamilton, T., Biswas, B., Kumaraswamy, M., Schooley, R.T., and Wooten, D. (2018). Phage Therapy for a Multidrug-Resistant Acinetobacter baumannii Craniectomy Site Infection. *Open Forum Infectious Diseases* 5. doi:10.1093/ofid/ofy064.

Leiman, P.G., Battisti, A.J., Bowman, V.D., Stummeyer, K., Mühlenhoff, M., Gerardy-Schahn, R. et al. (2007). The structures of bacteriophages K1E and K1-5 explain processive degradation of polysaccharide capsules and evolution of new host specificities. *Journal of molecular biology* 371**,** 836-849.

Lepage, P., Colombet, J., Marteau, P., Sime-Ngando, T., Doré, J., and Leclerc, M. (2008). Dysbiosis in inflammatory bowel disease: a role for bacteriophages? *Gut* 57**,** 424-425.

Lim, E.S., Zhou, Y., Zhao, G., Bauer, I.K., Droit, L., Ndao, I.M. et al. (2015). Early life dynamics of the human gut virome and bacterial microbiome in infants. *Nature medicine* 21**,** 1228.

Liu, R., Hong, J., Xu, X., Feng, Q., Zhang, D., Gu, Y. et al. (2017). Gut microbiome and serum metabolome alterations in obesity and after weight-loss intervention. *Nature medicine* 23**,** 859.

Łusiak-Szelachowska, M., Weber-Dąbrowska, B., Jończyk-Matysiak, E., Wojciechowska, R., and Górski, A. (2017). Bacteriophages in the gastrointestinal tract and their implications. *Gut Pathogens* 9**,** 44. doi:10.1186/s13099-017-0196-7.

Ma, Y., You, X., Mai, G., Tokuyasu, T., and Liu, C. (2018). A human gut phage catalog correlates the gut phageome with type 2 diabetes. *Microbiome* 6**,** 24-24. doi:10.1186/s40168-018-0410-y.

Manrique, P., Bolduc, B., Walk, S.T., Van Der Oost, J., De Vos, W.M., and Young, M.J. (2016). Healthy human gut phageome. *Proceedings of the National Academy of Sciences* 113**,** 10400-10405.

Maura, D., Galtier, M., Le Bouguénec, C., and Debarbieux, L. (2012). Virulent bacteriophages can target O104: H4 enteroaggregative Escherichia coli in the mouse intestine. *Antimicrobial agents and chemotherapy* 56**,** 6235-6242.

Mccann, A., Ryan, F.J., Stockdale, S.R., Dalmasso, M., Blake, T., Ryan, C.A. et al. (2018). Viromes of one year old infants reveal the impact of birth mode on microbiome diversity. *PeerJ* 6**,** e4694.

Minot, S., Bryson, A., Chehoud, C., Wu, G.D., Lewis, J.D., and Bushman, F.D. (2013). Rapid evolution of the human gut virome. *Proceedings of the National Academy of Sciences* 110**,** 12450-12455.

Minot, S., Wu, G.D., Lewis, J.D., and Bushman, F.D. (2012). Conservation of Gene Cassettes among Diverse Viruses of the Human Gut. *PLOS ONE* 7**,** e42342. doi:10.1371/journal.pone.0042342.

Mirzaei, M.K., and Maurice, C.F. (2017). Ménage à trois in the human gut: interactions between host, bacteria and phages. *Nature Reviews Microbiology* 15**,** 397. doi:10.1038/nrmicro.2017.30.

Modi, S.R., Lee, H.H., Spina, C.S., and Collins, J.J. (2013). Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature* 499**,** 219.

Moineau, S., and Lévesque, C. (2005). Control of bacteriophages in industrial fermentations. *Bacteriophages: Biology and applications***,** 285-296.

Monaco, Cynthia l., Gootenberg, David b., Zhao, G., Handley, Scott a., Ghebremichael, Musie s., Lim, Efrem s. et al. (2016). Altered Virome and Bacterial Microbiome in Human Immunodeficiency Virus-Associated Acquired Immunodeficiency Syndrome. *Cell Host & Microbe* 19**,** 311-322. doi:https://doi.org/10.1016/j.chom.2016.02.011.

Moreno-Gallego, J.L., Chou, S.-P., Di Rienzi, S.C., Goodrich, J.K., Spector, T.D., Bell, J.T. et al. (2019). Virome diversity correlates with intestinal microbiome diversity in adult monozygotic twins. *Cell host & microbe* 25**,** 261-272. e265.

Muniesa, M., Hammerl, J.A., Hertwig, S., Appel, B., and Brüssow, H. (2012). Shiga toxin-producing Escherichia coli O104: H4: a new challenge for microbiology. *Appl. Environ. Microbiol.* 78**,** 4065-4073.

Ng, S.C., Shi, H.Y., Hamidi, N., Underwood, F.E., Tang, W., Benchimol, E.I. et al. (2017). Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. *The Lancet* 390**,** 2769-2778.

Nguyen, S., Baker, K., Padman, B.S., Patwa, R., Dunstan, R.A., Weston, T.A. et al. (2017). Bacteriophage Transcytosis Provides a Mechanism To Cross Epithelial Cell Layers. *mBio* 8**,** e01874-01817. doi:10.1128/mBio.01874-17.

Norman, J.M., Handley, S.A., Baldridge, M.T., Droit, L., Liu, C.Y., Keller, B.C. et al. (2015). Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* 160**,** 447-460.

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome research***,** gr. 213959.213116.

Obeng, N., Pratama, A.A., and Van Elsas, J.D. (2016). The significance of mutualistic phages for bacterial ecology and evolution. *Trends in microbiology* 24**,** 440-449.

Ott, S.J., Waetzig, G.H., Rehman, A., Moltzau-Anderson, J., Bharti, R., Grasis, J.A. et al. (2017). Efficacy of Sterile Fecal Filtrate Transfer for Treating Patients With Clostridium difficile Infection. *Gastroenterology* 152**,** 799-811.e797. doi:https://doi.org/10.1053/j.gastro.2016.11.010.

Paez-Espino, D., Eloe-Fadrosh, E.A., Pavlopoulos, G.A., Thomas, A.D., Huntemann, M., Mikhailova, N. et al. (2016). Uncovering Earth's virome. *Nature* 536**,** 425.

Pérez-Brocal, V., García-López, R., Nos, P., Beltrán, B., Moret, I., and Moya, A. (2015). Metagenomic analysis of Crohn's disease patients identifies changes in the virome and microbiome related to disease status and therapy, and detects potential interactions and biomarkers. *Inflammatory bowel diseases* 21**,** 2515-2532.

Pérez-Brocal, V., García-López, R., Vázquez-Castellanos, J.F., Nos, P., Beltrán, B., Latorre, A. et al. (2013). Study of the viral and microbial communities associated with Crohn's disease: a metagenomic approach. *Clinical and translational gastroenterology* 4**,** e36.

Porter, N.T., Hryckowian, A.J., Merrill, B.D., Gardner, J.O., Singh, S., Sonnenburg, J.L. et al. (2019). Multiple phase-variable mechanisms, including capsular polysaccharides, modify bacteriophage susceptibility in Bacteroides thetaiotaomicron. BioRxiv [Preprint]. Available at: https://www.biorxiv.org/content/10.1101/521070v1.full (Accessed August 09, 2019). *bioRxiv*. doi:10.1101/521070.

Probst, A.J., Weinmaier, T., Desantis, T.Z., Santo Domingo, J.W., and Ashbolt, N. (2015). New Perspectives on Microbial Community Distortion after Whole-Genome Amplification. *PLOS ONE* 10**,** e0124158. doi:10.1371/journal.pone.0124158.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C. et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *nature* 464**,** 59.

Ram, G., Chen, J., Kumar, K., Ross, H.F., Ubeda, C., Damle, P.K. et al. (2012). Staphylococcal pathogenicity island interference with helper phage reproduction is a paradigm of molecular parasitism. *Proceedings of the National Academy of Sciences* 109**,** 16300-16305.

Ravin, N.V. (2015). "Replication and maintenance of linear phage-plasmid N15," in *Plasmids: Biology and Impact in Biotechnology and Discovery*. American Society of Microbiology), 71-82.

Reyes, A., Blanton, L.V., Cao, S., Zhao, G., Manary, M., Trehan, I. 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* 112**,** 11941-11946. doi:10.1073/pnas.1514285112.

Reyes, A., Haynes, M., Hanson, N., Angly, F.E., Heath, A.C., Rohwer, F. et al. (2010). Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* 466**,** 334.

Ridlon, J.M., Kang, D.J., Hylemon, P.B., and Bajaj, J.S. (2014). Bile acids and the gut microbiome. *Current opinion in gastroenterology* 30**,** 332.

Rostøl, J.T., and Marraffini, L. (2019). (Ph) ighting phages: how bacteria resist their parasites. *Cell host & microbe* 25**,** 184-194.

Roux, S., Emerson, J.B., Eloe-Fadrosh, E.A., and Sullivan, M.B. (2017). Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ* 5**,** e3817.

Roux, S., Krupovic, M., Debroas, D., Forterre, P., and Enault, F. (2013). Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. *Open biology* 3**,** 130160.

Roux, S., Solonenko, N.E., Dang, V.T., Poulos, B.T., Schwenck, S.M., Goldsmith, D.B. et al. (2016). Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. *PeerJ* 4**,** e2777.

Samson, J.E., Magadán, A.H., Sabri, M., and Moineau, S. (2013). Revenge of the phages: defeating bacterial defences. *Nature Reviews Microbiology* 11**,** 675. doi:10.1038/nrmicro3096.

Scanlan, P.D. (2017). Bacteria–Bacteriophage Coevolution in the Human Gut: Implications for Microbial Diversity and Functionality. *Trends in*

Microbiology 25**,** 614-623. doi:https://doi.org/10.1016/j.tim.2017.02.012.

Scanlan, P.D., Buckling, A., and Hall, A.R. (2015). Experimental evolution and bacterial resistance: (co)evolutionary costs and trade-offs as opportunities in phage therapy research. *Bacteriophage* 5**,** e1050153. doi:10.1080/21597081.2015.1050153.

Scholl, D., Adhya, S., and Merril, C. (2005). Escherichia coli K1's capsule is a barrier to bacteriophage T7. *Appl. Environ. Microbiol.* 71**,** 4872-4874.

Schwechheimer, C., and Kuehn, M.J. (2015). Outer-membrane vesicles from Gram-negative bacteria: biogenesis and functions. *Nature reviews microbiology* 13**,** 605.

Seed, K.D., Lazinski, D.W., Calderwood, S.B., and Camilli, A. (2013). A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature* 494**,** 489. doi:10.1038/nature11927.

Sender, R., Fuchs, S., and Milo, R. (2016). Revised estimates for the number of human and bacteria cells in the body. *PLoS biology* 14**,** e1002533.

Sharon, G., Sampson, T.R., Geschwind, D.H., and Mazmanian, S.K. (2016). The central nervous system and the gut microbiome. *Cell* 167**,** 915-932.

Shkoporov, A.N., Clooney, A.G., Sutton, T.D.S., Ryan, F.J., Daly, K.M., Nolan, J.A. et al. (2019). The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. *Cell Host & Microbe* 26**,** 527-541.e525. doi:https://doi.org/10.1016/j.chom.2019.09.009.

Shkoporov, A.N., and Hill, C. (2019). Bacteriophages of the Human Gut: The “Known Unknown” of the Microbiome. *Cell Host & Microbe* 25**,** 195-209. doi:https://doi.org/10.1016/j.chom.2019.01.017.

Shkoporov, A.N., Khokhlova, E.V., Fitzgerald, C.B., Stockdale, S.R., Draper, L.A., Ross, R.P. et al. (2018a). ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. *Nature Communications* 9**,** 4781. doi:10.1038/s41467-018-07225-7.

Shkoporov, A.N., Ryan, F.J., Draper, L.A., Forde, A., Stockdale, S.R., Daly, K.M. et al. (2018b). Reproducible protocols for metagenomic analysis of human faecal phageomes. *Microbiome* 6**,** 68.

Silpe, J.E., and Bassler, B.L. (2019). A host-produced quorum-sensing autoinducer controls a phage lysis-lysogeny decision. *Cell* 176**,** 268-280. e213.

Silveira, C.B., and Rohwer, F.L. (2016). Piggyback-the-Winner in host-associated microbial communities. *NPJ biofilms and microbiomes* 2**,** 16010.

Simmonds, P., Adams, M.J., Benkő, M., Breitbart, M., Brister, J.R., Carstens, E.B. et al. (2017). Virus taxonomy in the age of metagenomics. *Nature Reviews Microbiology* 15**,** 161. doi:10.1038/nrmicro.2016.177.

Simmonds, P., and Aiewsakun, P. (2018). Virus classification – where do you draw the line? *Archives of Virology* 163**,** 2037-2046. doi:10.1007/s00705-018-3938-z.

Siringan, P., Connerton, P.L., Cummings, N.J., and Connerton, I.F. (2014). Alternative bacteriophage life cycles: the carrier state of Campylobacter jejuni. *Open biology* 4**,** 130200.

Smeal, S.W., Schmitt, M.A., Pereira, R.R., Prasad, A., and Fisk, J.D. (2017). Simulation of the M13 life cycle I: Assembly of a genetically-structured deterministic chemical kinetic simulation. *Virology* 500**,** 259-274. doi:https://doi.org/10.1016/j.virol.2016.08.017.

Sokol, H., Pigneur, B., Watterlot, L., Lakhdari, O., Bermúdez-Humarán, L.G., Gratadoux, J.-J. et al. (2008). Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proceedings of the National Academy of Sciences* 105**,** 16731-16736.

Sonnenburg, J.L., Xu, J., Leip, D.D., Chen, C.-H., Westover, B.P., Weatherford, J. et al. (2005). Glycan foraging in vivo by an intestine-adapted bacterial symbiont. *Science* 307**,** 1955-1959.

Sullivan, M.B., Lindell, D., Lee, J.A., Thompson, L.R., Bielawski, J.P., and Chisholm, S.W. (2006). Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS biology* 4**,** e234.

Suttle, C.A. (2007). Marine viruses—major players in the global ecosystem. *Nature Reviews Microbiology* 5**,** 801.

Sutton, T.D., Clooney, A.G., Ryan, F.J., Ross, R.P., and Hill, C. (2019a). Choice of assembly software has a critical impact on virome characterisation. *Microbiome* 7**,** 12.

Sutton, T.D.S., Clooney, A.G., and Hill, C. (2019b). Giant oversights in the human gut virome. *Gut***,** gutjnl-2019-319067. doi:10.1136/gutjnl-2019-319067.

Sweere, J.M., Van Belleghem, J.D., Ishak, H., Bach, M.S., Popescu, M., Sunkari, V. et al. (2019). Bacteriophage trigger antiviral immunity and prevent clearance of bacterial infection. *Science* 363**,** eaat9691. doi:10.1126/science.aat9691.

Thingstad, T.F. (2000). Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnology and Oceanography* 45**,** 1320-1328.

Tock, M.R., and Dryden, D.T. (2005). The biology of restriction and anti-restriction. *Current opinion in microbiology* 8**,** 466-472.

Turkington, C.J.R., Morozov, A., Clokie, M.R.J., and Bayliss, C.D. (2019). Phage-Resistant Phase-Variant Sub-populations Mediate Herd Immunity Against Bacteriophage Invasion of Bacterial Meta-Populations. *Frontiers in Microbiology* 10. doi:10.3389/fmicb.2019.01473.

Twort, F.W. (1915). An Investigation on the Nature of Ultra-Microscopic Viruses. *The Lancet* 186**,** 1241-1243. doi:https://doi.org/10.1016/S0140-6736(01)20383-3.

Van Belleghem, J.D., Clement, F., Merabishvili, M., Lavigne, R., and Vaneechoutte, M. (2017). Pro- and anti-inflammatory responses of peripheral blood mononuclear cells induced by Staphylococcus aureus and Pseudomonas aeruginosa phages. *Scientific reports* 7**,** 8004-8004. doi:10.1038/s41598-017-08336-9.

Van Houte, S., Buckling, A., and Westra, E.R. (2016). Evolutionary ecology of prokaryotic immune mechanisms. *Microbiol. Mol. Biol. Rev.* 80**,** 745-763.

Vandeputte, D., Falony, G., Vieira-Silva, S., Tito, R.Y., Joossens, M., and Raes, J. (2016). Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates. *Gut* 65**,** 57-62.

Vaughn, B.P., Vatanen, T., Allegretti, J.R., Bai, A., Xavier, R.J., Korzenik, J. et al. (2016). Increased intestinal microbial diversity following fecal microbiota transplant for active Crohn's disease. *Inflammatory bowel diseases* 22**,** 2182-2190.

Verma, N., Brandt, J., Verma, D., and Lindberg, A. (1991). Molecular characterization of the O-acetyl transferase gene of converting bacteriophage SF6 that adds group antigen 6 to Shigella flexneri. *Molecular microbiology* 5**,** 71-75.

Vollmers, J., Wiegand, S., and Kaster, A.-K. (2017). Comparing and evaluating metagenome assembly tools from a microbiologist’s perspective-Not only size matters! *PloS one* 12**,** e0169662.

Von Wintersdorff, C.J., Penders, J., Van Niekerk, J.M., Mills, N.D., Majumder, S., Van Alphen, L.B. et al. (2016). Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. *Frontiers in microbiology* 7**,** 173.

Wagner, J., Maksimovic, J., Farries, G., Sim, W.H., Bishop, R.F., Cameron, D.J. et al. (2013). Bacteriophages in gut samples from pediatric Crohn’s disease patients: metagenomic analysis using 454 pyrosequencing. *Inflammatory bowel diseases* 19**,** 1598-1608.

Waldor, M.K., and Mekalanos, J.J. (1996). Lysogenic Conversion by a Filamentous Phage Encoding Cholera Toxin. *Science* 272**,** 1910-1914. doi:10.1126/science.272.5270.1910.

Waller, A.S., Yamada, T., Kristensen, D.M., Kultima, J.R., Sunagawa, S., Koonin, E.V. et al. (2014). Classification and quantification of bacteriophage taxa in human gut metagenomes. *The ISME journal* 8**,** 1391.

Warwick-Dugdale, J., Solonenko, N., Moore, K., Chittick, L., Gregory, A.C., Allen, M.J. et al. (2019). Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. *PeerJ* 7**,** e6800.

Weinbauer, M.G. (2004). Ecology of prokaryotic viruses. *FEMS microbiology reviews* 28**,** 127-181.

Wilson, B.C., Vatanen, T., Cutfield, W.S., and O'sullivan, J.M. (2019). The Super-Donor Phenomenon in Fecal Microbiota Transplantation. *Frontiers in Cellular and Infection Microbiology* 9. doi:10.3389/fcimb.2019.00002.

Wirbel, J., Pyl, P.T., Kartal, E., Zych, K., Kashani, A., Milanese, A. et al. (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nature medicine* 25**,** 679.

Woyke, T., Xie, G., Copeland, A., González, J.M., Han, C., Kiss, H. et al. (2009). Assembling the Marine Metagenome, One Cell at a Time. *PLOS ONE* 4**,** e5299. doi:10.1371/journal.pone.0005299.

Wright, A. (1971). Mechanism of conversion of the Salmonella O antigen by bacteriophage ε34. *Journal of bacteriology* 105**,** 927-936.

Wright, A., Hawkins, C.H., Änggård, E.E., and Harper, D.R. (2009). A controlled clinical trial of a therapeutic bacteriophage preparation in chronic otitis due to antibiotic-resistant Pseudomonas aeruginosa; a preliminary report of efficacy. *Clinical Otolaryngology* 34**,** 349-357. doi:10.1111/j.1749-4486.2009.01973.x.

Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M. et al. (2012). Human gut microbiome viewed across age and geography. *nature* 486**,** 222.

Yilmaz, S., Allgaier, M., and Hugenholtz, P. (2010). Multiple displacement amplification compromises quantitative analysis of metagenomes. *Nature Methods* 7**,** 943. doi:10.1038/nmeth1210-943.

Yutin, N., Makarova, K.S., Gussow, A.B., Krupovic, M., Segall, A., Edwards, R.A. et al. (2018). Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nature microbiology* 3**,** 38.

Zago, M., Orrù, L., Rossetti, L., Lamontanara, A., Fornasari, M.E., Bonvini, B. et al. (2017). Survey on the phage resistance mechanisms displayed by a dairy Lactobacillus helveticus strain. *Food Microbiology* 66**,** 110-116. doi:https://doi.org/10.1016/j.fm.2017.04.014.

Zhang, K., Martiny, A.C., Reppas, N.B., Barry, K.W., Malek, J., Chisholm, S.W. et al. (2006). Sequencing genomes from single cells by polymerase cloning. *Nature Biotechnology* 24**,** 680-686. doi:10.1038/nbt1214.

Zhao, G., Droit, L., Gilbert, M.H., Schiro, F.R., Didier, P.J., Si, X. et al. (2019). Virome biogeography in the lower gastrointestinal tract of rhesus macaques with chronic diarrhea. *Virology* 527**,** 77-88. doi:https://doi.org/10.1016/j.virol.2018.10.001.

Zhao, G., Vatanen, T., Droit, L., Park, A., Kostic, A.D., Poon, T.W. et al. (2017). Intestinal virome changes precede autoimmunity in type I diabetes-susceptible children. *Proceedings of the National Academy of Sciences* 114**,** E6166-E6175. doi:10.1073/pnas.1706359114.

Zhu, A., Sunagawa, S., Mende, D.R., and Bork, P. (2015). Inter-individual differences in the gene content of human gut bacterial species. *Genome biology* 16**,** 82.

Zuo, T., Lu, X.-J., Zhang, Y., Cheung, C.P., Lam, S., Zhang, F. et al. (2019). Gut mucosal virome alterations in ulcerative colitis. *Gut***,** gutjnl-2018-318131.

Chapter 2

Choice of assembly software has a critical impact on virome characterisation.

This chapter has been published as the following:

Sutton, T.D.S.*, Clooney, A.G.*, Ryan, F.J.*, Ross, R.P., and Hill, C. (2019).

Choice of assembly software has a critical impact on virome characterisation. Microbiome 7,12.

https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-019-0626-5

* Contributed Equally

Manuscript and editorial revisions written by Thomas Sutton and Adam Clooney. Study design and constructive comments provided by Feargal Ryan. Bioinformatic analysis carried out by Thomas Sutton (Figures 1-3) and Adam Clooney (Figure 4.) Constructive comments and supervision provided by Colin Hill and Paul Ross.

Abstract

The viral component of microbial communities play a vital role in driving bacterial diversity, facilitating nutrient turnover and shaping community composition. Despite their importance, the vast majority of viral sequences are poorly annotated and share little or no homology to reference databases. As a result, investigation of the viral metagenome (virome) relies heavily on *de novo* assembly of short sequencing reads to recover compositional and functional information. Metagenomic assembly is particularly challenging for virome data, often resulting in fragmented assemblies and poor recovery of viral community members. Despite the essential role of assembly in virome analysis and difficulties posed by these data, current assembly comparisons have been limited to subsections of virome studies or bacterial datasets. This study presents the most comprehensive virome assembly comparison to date, featuring 16 metagenomic assembly approaches which have featured in human virome studies. Assemblers were assessed using four independent virome datasets, namely; simulated reads, two mock communities, viromes spiked with a known phage and human gut viromes. Assembly performance varied significantly across all test datasets, with SPAdes (meta) performing consistently well. Performance of MIRA and VICUNA varied, highlighting the importance of using a range of datasets when comparing assembly programs. It was also found that while some assemblers addressed the challenges of virome data better than others, all assemblers had limitations. Low read coverage and genomic repeats resulted in assemblies with poor genome recovery, high degrees of fragmentation and low accuracy contigs across all assemblers. These limitations must be considered when setting thresholds for downstream analysis and when drawing conclusions from virome data.

Introduction

The rapid evolution of metagenomics and high throughput sequencing technologies has revolutionised the study of microbial communities, giving new insights into the role and identity of the uncultivated microbes which account for the majority of metagenomic sequences (Solden et al., 2016). However, the majority of microbial sequencing efforts have focused on the characterisation of prokaryotic microbes. Viral metagenomes (viromes) are dominated by novel sequences, often with up to 90% of sequences sharing little to no homology to reference databases (Aggarwala et al., 2017). Bacteriophage, the most abundant members of viral communities, play a key role in the shaping the composition of microbial communities and facilitate horizontal gene transfer (Paul, 2008). Viromes have been shown to play a role in global geochemical cycles (Breitbart, 2011) and have been studied in varied ecosystems including the ocean (Hurwitz and Sullivan, 2013). Viromes of the human body are of particular interest, where they have been linked to disease status (Norman et al., 2015), maintaining human health (Manrique et al., 2016) and shaping the gut microbiome in early life (Lim et al., 2015;McCann et al., 2018). Due to the predominance of uncharacterised viral sequences "viral dark matter"; (Roux et al., 2015), and the lack of a universal marker gene, virome studies rely on database independent analysis methods and depend heavily on *de novo* assembly to resolve viral genomes from metagenomic sequencing reads.

Metagenomic assemblers typically use de Bruijn graph (DBG) approaches to address the complexity and size of metagenomic datasets in an accurate and efficient manner. Microbial metagenomes pose significant challenges to DBG assembly when compared to single genome assemblies often complicating the DBG and leading to fragmentation and/or misassembly (Olson et al., 2017). These challenges include uneven sequencing coverage of organisms within the metagenome, the presence of conserved regions across different species, repeat regions within genomes and the introduction of false *k*-mers by both closely related genomes at differing abundances and sequencing errors at high read coverage. This hampers the use of coverage statistics to resolve repeat regions between and within genomes (Olson et al., 2017).

A wide array of metagenomic assembly programs have been employed, each addressing aspects of metagenomic challenges to varying degrees. However, many of these programs have been designed and optimised for bacterial metagenomes, which share many assembly challenges of viromes but to a lesser degree. Virome data is characterised by high proportions of repeat regions within viral genomes (Minot et al., 2012), hypervariable genomic regions associated with host interaction (Warwick-Dugdale et al., 2018) and high mutation rates which lead to increased metagenomic complexity and strain variation (Roux et al., 2017). Low DNA yields also limit read coverage and often require a multiple displacement amplification (MDA) step which has been shown to preferentially amplify small single stranded DNA viruses (Kim and Bae, 2011). Extremes in read coverage caused by MDA bias and dominant viral taxa such as crAssphage, which can make up large proportions of human gut viromes (Dutilh et al., 2014), sequester sequencing resources and result in insufficient coverage of low abundance viruses. These challenges result in fragmented virome assemblies (García-López et al., 2015), limiting their use in downstream analysis. Despite benchmarks of bacterial metagenomes having highlighted failings and benefits of particular assembly programs, many poorly performing assemblers have featured in virome studies (Foulongne et al., 2012;Hannigan et al., 2015;Guo et al., 2017).

Accurate comparison of metagenomic assemblers is complicated by the unknown composition of metagenomic datasets and the limited applicability of general assembly statistics such as N50 (Deng et al., 2015;Vollmers et al., 2017). To address this, the accuracy and efficacy of metagenomic assembly programs is often evaluated using simulated datasets and mock communities of known composition. Although these simulated datasets are undergoing constant improvements (Sczyrba et al., 2017;Fritz et al., 2018), they have focused primarily on bacterial metagenomes and remain limited in their ability to accurately replicate the challenges of true metagenomes. While some virome-specific assembly benchmarks have been performed, many have been limited to a small number of assemblers, 454 data or subsections of virome studies and have exclusively used simulated data (Aguirre de Cárcer et al., 2014;Smits et al., 2014;Vázquez-Castellanos et al., 2014;García-López et al., 2015;Hesse et al., 2017;Roux et al., 2017).

Here we expand upon previous studies and present a detailed investigation of assembly software for virome analysis which compares all those previously used in human virome studies to date, as well as other popular or more recently published assemblers (Table 1). We compare assembly efficacy and accuracy using simulated viromes, mock viral communities and human gut viromes spiked with a known exogenous bacteriophage. Furthermore we confirm these findings using human virome data from published datasets and assess computational parameters such as runtime and RAM usage. We also investigate in detail the impact of sequencing coverage and genomic repeats on assembly performance and highlight important considerations for future virome studies. Together these data comprise most comprehensive virome assembly benchmark to date.

Methods

Each assembler with the exception of Geneious and CLC was run as per manual with default parameters (unless stated) using a Lenovo x3650 M5 server with an intel Xeon processor E5-2690 v3 and 512Gb RAM running Ubuntu 14.04.5. Geneious assembly approach mirrored that used in (Manrique et al., 2016) by generating consensus sequences from the assemblies of both MIRA and Vicuna. CLC and Geneious were run on a 64-bit windows 10 computer with an i5-4690 CPU and 16 GB of RAM.

Data sources

Sequencing reads from mock communities A and B featured in (Roux et al., 2016), Simulated Virome dataset featured in (Hesse et al., 2017), reads used to compare the impact of sequencing depth on time and RAM usage featured in (Manrique et al., 2016) and human viromes spiked with 10^7 PFU of Lactococcal phage Q33 (Mahony et al., 2013) and originated from (Shkoporov et al., 2018) .

Read Pre-processing

Raw read quality was assessed with FastQC v0.11.5 and sequencing adapters were removed with cutadapt v1.9.1 (Martin, 2011) for the mock, Spiked and healthy gut virome data sets. Trimming and filtering was carried out with Trimmomatic v0.36 (Bolger et al., 2014) using parameters specific to each dataset. A sliding window size of 4 with a minimum Phred score of 30 and a minimum length of 60bp was used with reads from both mock communities. The leading 15bp and trailing 60bp were removed from “Healthy human gut phageome” reads and a sliding window of 4bp with a minimum phred score of 20 was applied. The leading 10bp and trailing 100bp were removed from the Q33 spiked virome reads and a sliding window size of 4bp with a minimum Phred score of 30. Filtered reads were through a minimum length filter of 60bp.

Analysis methods

Quality filtered reads from the Q33 spiked dataset consisted of 3 individual viromes which were pooled and subsequently assembled. Contigs were aligned to the published Q33 using Blastn with an e-value cut-off of 1e^{-20}. Top hit alignments to the Q33 genome with a minimum alignment length of 800 bases and which shared 95% identity were included in further analysis using QUAST (v. 4.4) (Gurevich et al., 2013) with "--unique mapping" flag. Further comparison and visualisation of Q33 assemblies was carried out using Mauve (v. 20150226, build 10) (Darling et al., 2010).

Alignment and comparison of assemblies from mock and simulated data sets to reference genomes was carried using MetaQUAST (v. 4.4) (Mikheenko et al., 2015) with "--unique mapping" flag and default parameters (minimum contig length of 500bp, minimum alignment length of 65bp, minimum identity threshold of 95%). Correlations were carried out using Spearman method and plots were generated using the package ggplot2 (v 3.0.0) package in R (v.3.4.3). These correlations were validated using a linear model in R base library. For data which was not normally distributed, log transformation was carried out.

Reads from the "healthy human gut phageome" were analysed to compare the overall assembler efficiency and the impact of sequencing depth. Reads were randomly subset in pairs (both the forward and reverse read of a pair were retained) to different depths using an in-house python script. Samples were subset in increments of 300,000 reads to their respective maximum depth (2.7, 3.5, 3 and 3.3 million reads). GNU time was utilised to measure the maximum RAM and length of time for each assembly to reach completion. All assemblers were run using 5 threads where possible with the exception of CLC, Geneious, Ray Meta, Velvet and Vicuna. Ray Meta and Velvet were run with 10, 1 thread(s) respectively. Ray Meta failed to run with 5 while Velvet ran with 1 core despite 5 being allocated. Vicuna was also allocated 5 threads however used upwards of 20. MetaVelvet was run, but after 7 days had failed to reach completion and was therefore removed from the subsequent analysis of these metrics. Contig statistics and filtering (contigs greater than 1kb retained) were performed using the assembly-stats script (v1.0.1) from the Pathogen Informatics group at the Wellcome Sanger Institute (https://github.com/sanger-pathogens/assembly-stats).

Software	Link	Version used	Reference
ABySS	http://www.bcgsc.ca/downloads/abyss/	v2.0.2	(Simpson et al., 2009)
CLC	https://www.qiagenbioinformatics.com/products/clc-assembly-cell/	v5.0.5	https://www.qiagenbioinformatics.com/
Geneious	https://www.geneious.com/features/assembly-mapping/	v11.0.3	(Kearse et al., 2012)
IDBA UD	https://i.cs.hku.hk/~alse/hkubrg/projects/idba_ud	v1.1.1	(Peng et al., 2012)
MEGAHIT	https://github.com/voutcn/megahit	v1.1.1-2	(Li et al., 2016)
MetaVelvet	https://metavelvet.dna.bio.keio.ac.jp/	v1.2.02	(Namiki et al., 2012)
MIRA	http://www.chevreux.org/mira_downloads.html	v4.0.2	(García-López et al., 2015)
Ray Meta	http://denovoassembler.sourceforge.net/	v2.3.0	(Boisvert et al., 2012)
SOAPdenovo2	http://soap.genomics.org.cn/soapdenovo.html	v2.04	(Luo et al., 2012)
SPAdes	http://cab.spbu.ru/software/spades/	v3.10.0	(Bankevich et al., 2012)
SPAdes meta	http://cab.spbu.ru/software/spades/ (variation of SPAdes applied with flag)	v3.10.0	(Nurk et al., 2017)
Velvet	https://www.ebi.ac.uk/~zerbino/velvet/	v1.2.10	(Zerbino and Birney, 2008)
VICUNA	https://github.com/broadinstitute/mvicuna	v1.3	(Vázquez-Castellanos et al., 2014)

Table 1: A list of assemblers used in this study

Results

Simulated virome dataset

Normalised genome abundance of 572 members of a published simulated community (Figure. 1A) (Hesse et al., 2017) and the degree of fragmentation, was assessed by aligning the resulting contigs from each assembler to the reference genomes (Figure. 1B). MetaVelvet was not included in this analysis as it failed to reach completion after seven days. Approximately half of the genomes in the community featured an average recovered genome fraction less than 75% and exhibited higher degrees of fragmentation (>10 contigs per genome on average) across all assemblers. For 87 of the 572 genomes there was an average recovered genome fraction of less than 20% across all assemblers (the low recovered genome fraction of VICUNA was excluded as an outlier). Of these genomes, 84 were present at low abundance (lowest 40% of all abundances normalised to genome length). The remaining three genomes were present at higher normalised abundances (50 – 80th percentile) but featured the some of the highest proportions of genomic repeats (70th-90th percentile).

Normalised genome abundance within the community had a strong positive correlation with recovered genome fraction across all assemblers (Supplementary Table 1, Additional file 5) and was verified using a linear model (Supplementary Table 2, Additional file 5), with the exception of SOAPdenovo2, which was negative. Normalised abundance also correlated negatively with the degree of fragmentation (number of contigs) across all assemblers except Velvet which was positively correlated and Geneious which was not significant (Supplementary Table1, Additional file 5). None of the genomes in the lower 30th percentile of normalised abundance featured an average recovered genome fraction greater than 75%, further exemplifying the impact of low sequencing coverage.

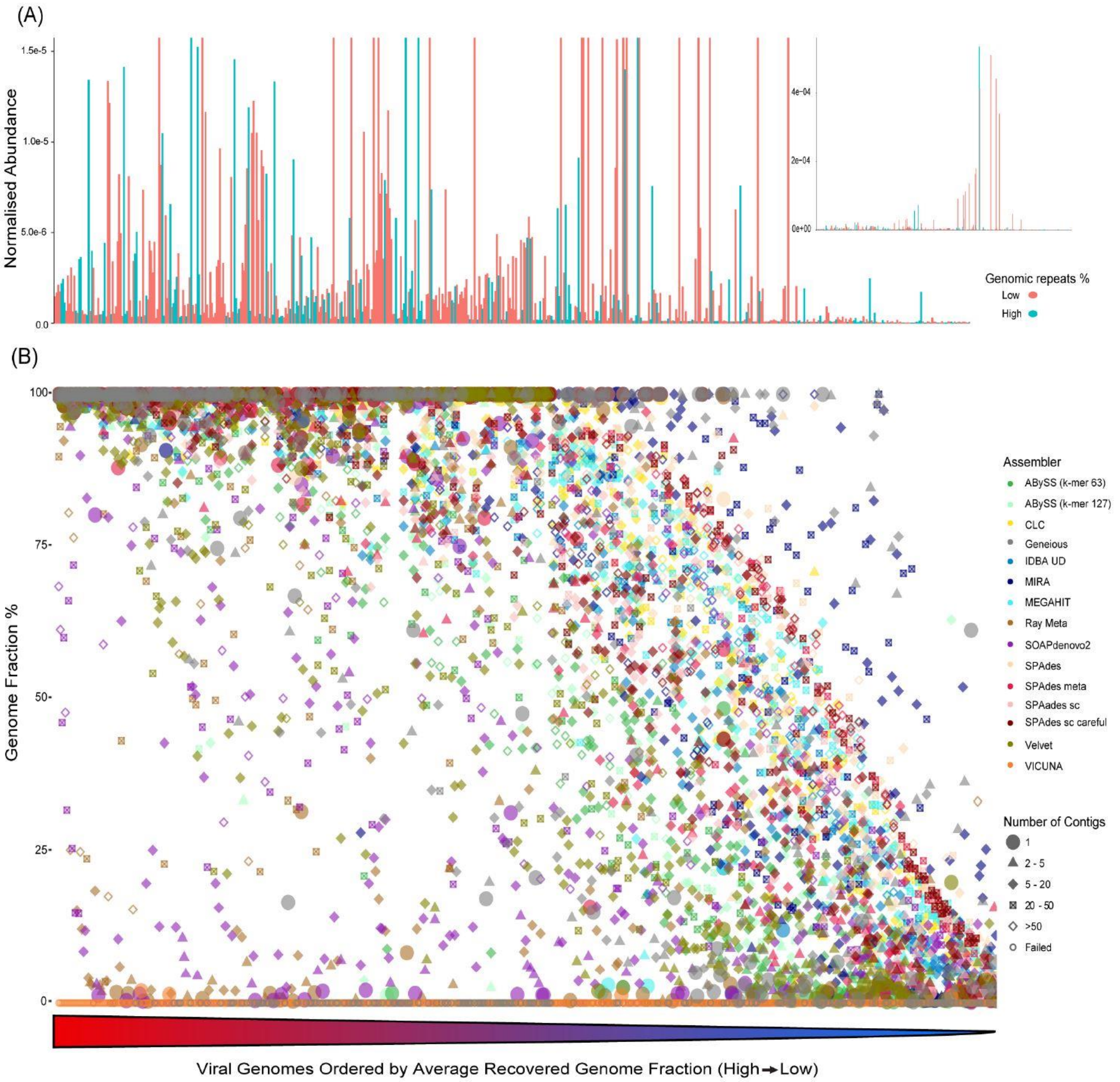


Figure 1: Relationship between percentage of each genome recovered (genome fraction), the number of contigs required for each combination of genome and assembler and the abundance and proportion of repeats for each genome. (A and B) Genomes are ordered by their average genome fraction across all assemblers from high to low along the x-axis. (A main) Relative abundance, normalized by genome length is plotted along y-axis with upper limit of 0.75% and colour of bars determined by proportion of repeat regions in each genome. Blue bars represent genomes with a high proportion of genomic repeats (4th quartile of all genomes), red represents all other genomes below this quartile. (A insert) Expanded view of (A) without an upper limit of y value. (B) Percentage genome recovered is plotted along the y axis. Points are coloured by assembler with shape of the point is denoting number of contigs generated by each assembler for each genome.

However high abundance did not consistently improve genome recovery and of the 172 genomes in the top 30% of normalised abundance, 20 featured an average genome fraction below 50%. The distance of the log transformed (due to extremes in values) normalised abundances from the mean was negatively correlated with recovered genome fraction across all assemblers (correlation coefficient: -0.42, p-value < $2.2e^{-16}$). Of 171 genomes in the 40th – 60th percentile of normalised abundance 29 featured an average genome fraction below 50%. This indicates factors other than abundance may hamper genome recovery. MIRA and Geneious both recovered a greater fraction of low abundance genomes with fewer contigs than other assemblers. However, MIRA assemblies of 13 of the most abundant genomes in the community (highest 10%) exhibited the highest degree of fragmentation in the study, generating between 401 and 2983 contigs per genome.

The proportion of inverted repeats, palindromic repeats, tandem repeats and a total proportion of genomic repeats was calculated for each genome. The total percentage of repeat regions predicted in each genome was positively correlated with the degree of fragmentation observed in each assembly across all assemblers with the exception of Ray Meta (Supplementary Table 3, Additional file 5), and negatively correlated with recovered genome fraction across all assemblers except ABySS (*k*-mer 63/127), Geneious, and SOAPdenovo2. When this relationship between repeat regions and the recovered genome fraction was assessed using a linear model, correlations were significant for CLC, MIRA, Ray Meta, Velvet, and all parameters of SPAdes (Supplementary Table 2, Additional file 5). Both the proportion of repeat regions in a genome and the relative abundance of that genome contribute to the variation in recovered genome fraction, though each explain a separate aspect of this variation. No interaction was found between these two metrics.

VICUNA, Ray Meta, SOAPdenovo2, Geneious, ABySS (both *k*-mer sizes) and Velvet recovered under 50% of the total genome fraction (all genomes in the community). VICUNA produced just four contigs in total with high levels of mismatches (174 per 100kb on average) which could possibly be linked to the format of the artificial reads as this was not observed in real sequencing data. The five assemblers which recovered the highest genome fraction overall were SPAdes (default), MEGAHIT, SPAdes (single cell), SPAdes (single cell + careful) and CLC. All assemblers achieving a minimum average genome fraction of 50% were subjected to a ranking system (Supplementary Table 4, Additional file 5). To compare both recovery and fragmentation assemblers were ordered from best to worst based on genome recovery and number of aligned contigs. The average rank resulted in Spades (default) performing best, recovering 72.2% overall genome sequences with 8230 contigs. The remaining top five assemblers of this combined rank were SPAdes (meta) 68.2% with 7419 contigs, SPAdes (single cell) 68.9% with 9506 contigs, CLC 68.6% with 9152 contigs and MEGAHIT 69.6% with 10083 contigs. The number of assemblies which recovered greater than 90% of the target genome in one single contig was compared (Figure. 2). SPAdes (default) performed best, recovering 210, SPAdes (meta), SPAdes (single cell + careful), CLC, and SPAdes (single cell) each recovered 179, 168, 162 and 160 genomes respectively.

The accuracy of assemblies was assessed by calculating the average count of indels, mismatches, and misassemblies per 100kb across all genomes. These counts were normalised to the number of genomes each assembler recovered with a minimum genome fraction of 50%. These were ranked according to their performance in all three metrics (Supplementary Table 4, Additional file 5), with assemblies from Velvet having the lowest overall counts followed by ABySS, IDBA UD, MEGAHIT and Ray Meta. With the exception of Ray Meta and SOAPdenovo2, the number of mismatches per 100kb was negatively correlated with both genome abundance and recovered genome fraction across all assemblers (Supplementary Table 1, Additional file 5).

The rate of false positive (no alignment to reference genomes) and false negative (recovered genome fraction of 0%) contigs assembled allowed for the determination of sensitivity. A number of assemblers had a sensitivity greater than 97%, however each returned greater than 7,000 contigs, inferring a high degree of fragmentation (Table 2). MIRA assembled (partial or complete) 559 of the genomes with a false positive count of just four. However, this was achieved from more than 27,000 contigs. ABySS (both *k*-mer sizes), Geneious, Ray Meta and Velvet returned very few false positives but failed to detect many of the genomes present. SPAdes (meta) performed best with 558 of the 572 genomes detected and only five false positives resulting from 7419 contigs.

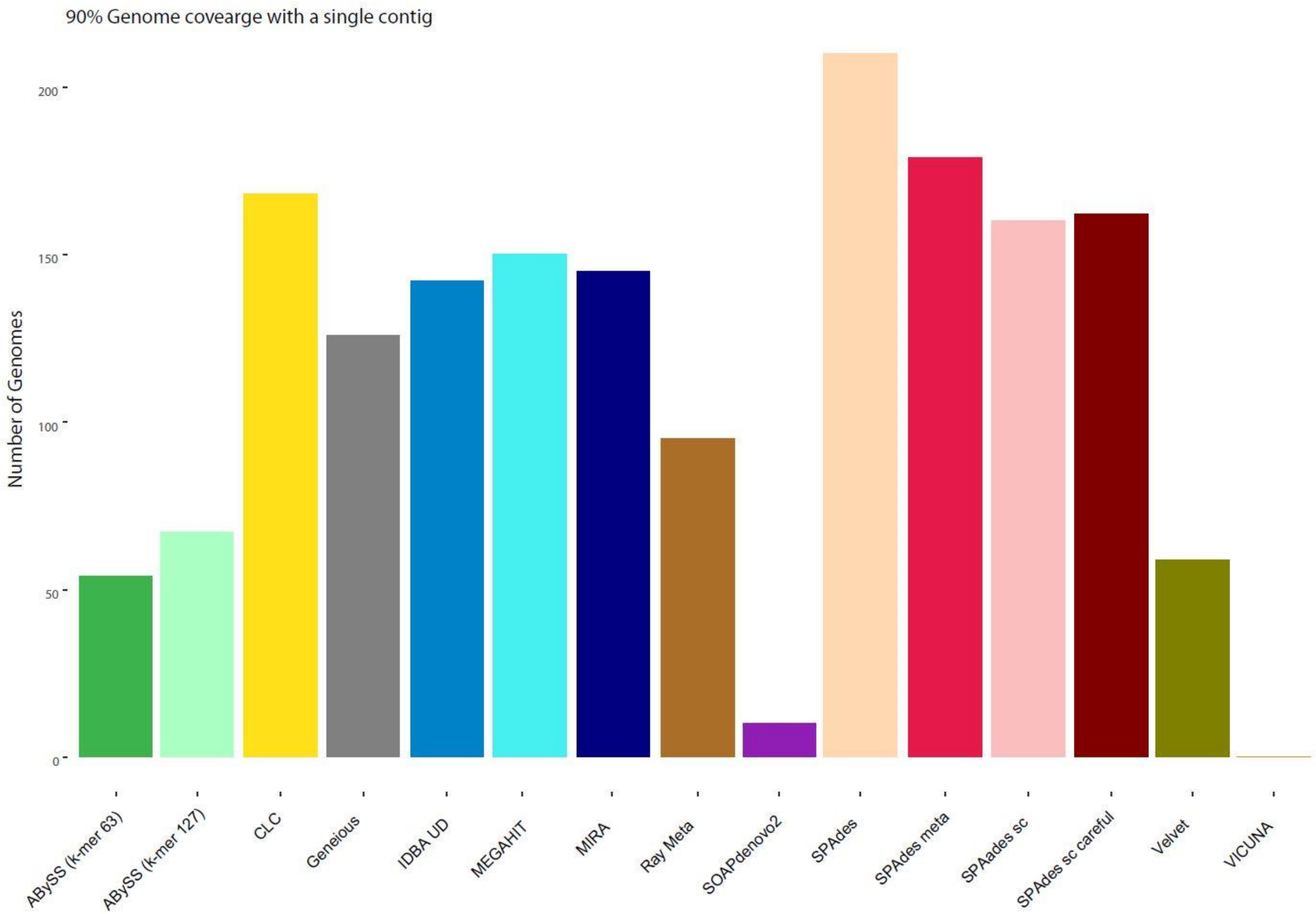


Figure 2. Number of contigs each assembler recovered to a minimum genome fraction of 90% in a single contig

	False Positives	False Negative	True Positives	No. of contigs returned*	Sensitivity
ABSS (*k*-mer 63)	0	111	461	7957	80.59
ABySS (*k*-mer 127)	1	123	449	7732	78.50
CLC	34	5	567	9152	99.13
Geneious	9	190	382	958	66.78
IDBA UD	25	9	563	8999	98.43
MEGAHIT	21	8	564	10083	98.60
MetaVelvet	N/A	N/A	N/A	N/A	N/A
MIRA	4	13	559	27600	97.73
Ray Meta	0	213	359	4224	62.76
SOAPdenovo2	536	116	456	11548	79.72
SPAdes	29	3	569	8230	99.48
SPAdes meta	5	14	558	7419	97.55
SPAdes sc	38	7	565	9506	98.78
SPAdes sc careful	40	6	566	9724	98.95
Velvet	1	65	507	6343	88.64
VICUNA	0	558	14	4	2.45
				*572 in community	

Table 2: The number of false positive, false negative contigs generated by each assembler for the Simulated community, together with the sensitivity rates

Mock community dataset

Two mock viral communities were used to investigate the impact of high and low abundance ssDNA viruses on assembly performance. Mock A (Table 3a) contained 12 viral genomes, 10 of which were at equal abundance (9.82% of the total community) and two ssDNA genomes (NC_001330 and NC_001422) at low abundance (0.92%). Analysis of this community showed that although some assemblers, namely CLC, Geneious, SPAdes (single cell) and VICUNA, detected all 12 genomes, this was at the expense of a large number of false positives (1143, 53, 1513 and 4969 respectively). Velvet and MetaVelvet generated no false positives, but failed to assemble three genomes, while ABySS (for both *k*-mers) generated a large number of false positives and failed to assemble four and six genomes, respectively. IDBA UD and Ray Meta outperformed the other assemblers with an equal number of contigs to genomes (12), followed by MEGAHIT, SPAdes (default) and SPAdes (meta) with 13, 14 and 14. Mock B (Table 3b) also contained 12 genomes but with a higher abundance of ssDNA genomes NC_001330 and NC_001422 (32.47%). VICUNA assemblies of Mock B improved upon those from Mock A as no false positives were generated, while the false positive rate in the MIRA assembler increased to 94 from none in Mock A. IDBA UD performed best followed by SPAdes (default), Ray Meta, MEGAHIT and SPAdes (meta) based on sensitivity and number of contigs, while ABySS (both *k*-mer sizes) and SOAPdenovo2 had the lowest sensitivity. Despite being a relatively simple community consisting of 12 members, not all assemblers were able to recover all members (Supplementary Table 5-6, Additional file 5). A greater number of assemblers (six) failed to assemble all members of Mock B than Mock A (four). ABySS(*k*-mer 63), ABySS(*k*-mer 127), Velvet and MetaVelvet failed to assemble 6, 4, 3 and 3 genomes respectively, in Mock A and 6, 4 ,1 and 1 genomes, respectively in Mock B. In addition, MIRA and SOAPdenovo2 failed to assemble 1 and 2 genomes respectively in Mock B.

A)

	False Positives	False Negative	True Positive	No. of contigs returned*	Sensitivity
ABySS (*k*-mer 63)	52	4	8	61	66.67
ABySS (*k*-mer 127)	50	6	6	56	50.00
CLC	1143	0	12	1299	100.00
Geneious	53	0	12	65	100.00
IDBA UD	0	0	12	12	100.00
MEGAHIT	0	0	12	13	100.00
MetaVelvet	0	3	9	26	75.00
MIRA	0	0	12	89	100.00
Ray Meta	0	0	12	12	100.00
SOAPdenovo2	2	0	12	23	100.00
SPAdes	0	0	12	14	100.00
SPAdes meta	0	0	12	14	100.00
SPAdes sc	1513	0	12	1527	100.00
SPAdes sc careful	0	0	12	15	100.00
Velvet	0	3	9	26	75.00
VICUNA	4969	0	12	5385	100.00
				*12 in community	

B)

	False Positive	False Negative	True Positives	No. of contigs returned*	Sensitivity
ABySS (*k*-mer 63)	60	4	8	69	66.67
ABySS (*k*-mer 127)	132	6	6	139	50.00
CLC	450	0	12	505	100.00
Geneious	14	0	12	30	100.00
IDBA UD	0	0	12	12	100.00
MEGAHIT	0	0	12	14	100.00
MetaVelvet	0	1	11	24	91.67
MIRA	94	1	11	157	91.67
Ray Meta	0	0	12	13	100.00
SOAPdenovo2	2	2	10	27	83.33
SPAdes	0	0	12	13	100.00
SPAdes meta	0	0	12	14	100.00
SPAdes sc	593	0	12	607	100.00
SPAdes sc careful	0	0	12	14	100.00
Velvet	0	1	11	24	91.67
VICUNA	0	0	12	15	100.00
				*12 in community	

Table 3: The number of false positive, false negative contigs generated by each assembler for a) Mock community A (overleaf) and b) Mock community B) along with the sensitivity rates for each

All but three VICUNA assemblies in Mock A exhibited a high level of fragmentation, generating 34.7 ± 35 (mean ± standard deviation) contigs per genome. Fragmentation was also seen in MIRA assemblies to a lesser degree with 7.4 ± 10 (mean ± standard deviation) contigs per genome on average. There was a high rate of fragmentation in CLC with one community member generating 144 contigs for genome KF302035. Average recovered genome fraction of 85.4 ± 6.4 % was skewed by ABySS (*k*-mer 63), ABySS (*k*-mer 127), Velvet, MetaVelvet, SOAPdenovo2, and VICUNA which recovered on average 49.5%, 66.6%, 73.8%, 73.8%, 29.7% and 76.6%, respectively. All other assemblers recovered over 99% of each genome in the community (Supplementary Figure 1).

Closer inspection of the two ssDNA genomes present at lower relative abundance highlighted significant differences in the average number of indels across all assemblies of the NC_001330 and NC_001422 genomes versus other members of the community (p-value = 0.037). These genomes exhibited an average of 41.7 ± 18.5 and 9.4 ± 20.4 indels per 100kb, while all other genomes featured an average of 7.8 ± 18.9 indels per 100kb. The low abundant ssDNA genomes NC_001330 and NC_001422 also featured the highest average mismatches per 100kb at 148.7 ± 3 and 302.5 ± 10.7, respectively (Supplementary Figure 1).

The degree of fragmentation observed by VICUNA and MIRA in Mock B was lower than in Mock A with a mean of 1.3 ± 0.89 and 5.3 ± 7.7 contigs per genome, respectively. CLC fragmented genome KF302035 in Mock B (44 contigs), but to a lesser degree than Mock A (144 contigs). MEGAHIT, which recovered at least 98% of all genomes in Mock A, also recovered over 98% of all genomes in Mock B except for the ssDNA genome NC_001422, of which 56.5% was recovered in two contigs. The majority of assemblies exhibited 147.9 ± 0 and 297 ± 1 mismatches per 100kb for NC_001330 and NC_001422 (high abundance ssDNA), respectively, identical values to those measured in Mock A. Velvet and MetaVelvet were exceptions with 184.2 and 860.2 for genome NC_001422 and NC_001330. A similar pattern of high values across a narrow range was also observed with the number of indels, with 49.3 to 32.9 present in all assemblies NC_001330. Genome NC_001422 featured 18.57 indels across all SPAdes assemblies (all parameters) and 860.2 across both Velvet and

Metavelvet assemblies. All other assemblers which successfully recovered this genome did not feature any indels (Supplementary Figure 1).

Q33

Five assemblers failed to generate contigs which met alignment thresholds and were subsequently excluded from further analysis - namely ABySS (*k*-mer 63), ABySS (*k*-mer 127), SOAPdenovo2, Velvet and MetaVelvet. All remaining assemblers recovered over 90% of the spiked Q33 genome with the exception of MIRA (8.5%). Six assemblers recovered over 99% of the Q33 genome in a single contig - SPAdes (meta) 99.74%, MEGAHIT (99.6%), VICUNA (99.6%), Ray meta (99.6%), CLC (99.5%) and Geneious (99.1) (Figure. 3). However, only MEGAHIT assembled the Q33 genome with a contig equal in length to the genome itself. SPAdes (meta) and CLC generated assemblies shorter than the reference genome by 86 and 141 bases. VICUNA (723), Geneious (1765), and Ray Meta (9884) each generated assemblies longer than the reference genome. SPAdes (default) SPAdes (single cell), IDBA UD and SPAdes (single cell + careful) each assembled Q33 in 2, 3, 4, 5 and 5 contigs, respectively. Ray Meta and VICUNA assemblies had the lowest number of mismatches and indels per 100kb, however Ray Meta exhibited the highest rate of misassemblies (2 relocations, 1 inversion). All assemblers featured a minimum of one local misassembly with the exception of SPAdes (meta) did not feature any. The six best assemblies of the Q33 genome and the genome itself are syntenic (although occasionally on the reverse strand) and the start and end point were not conserved (Figure .3).

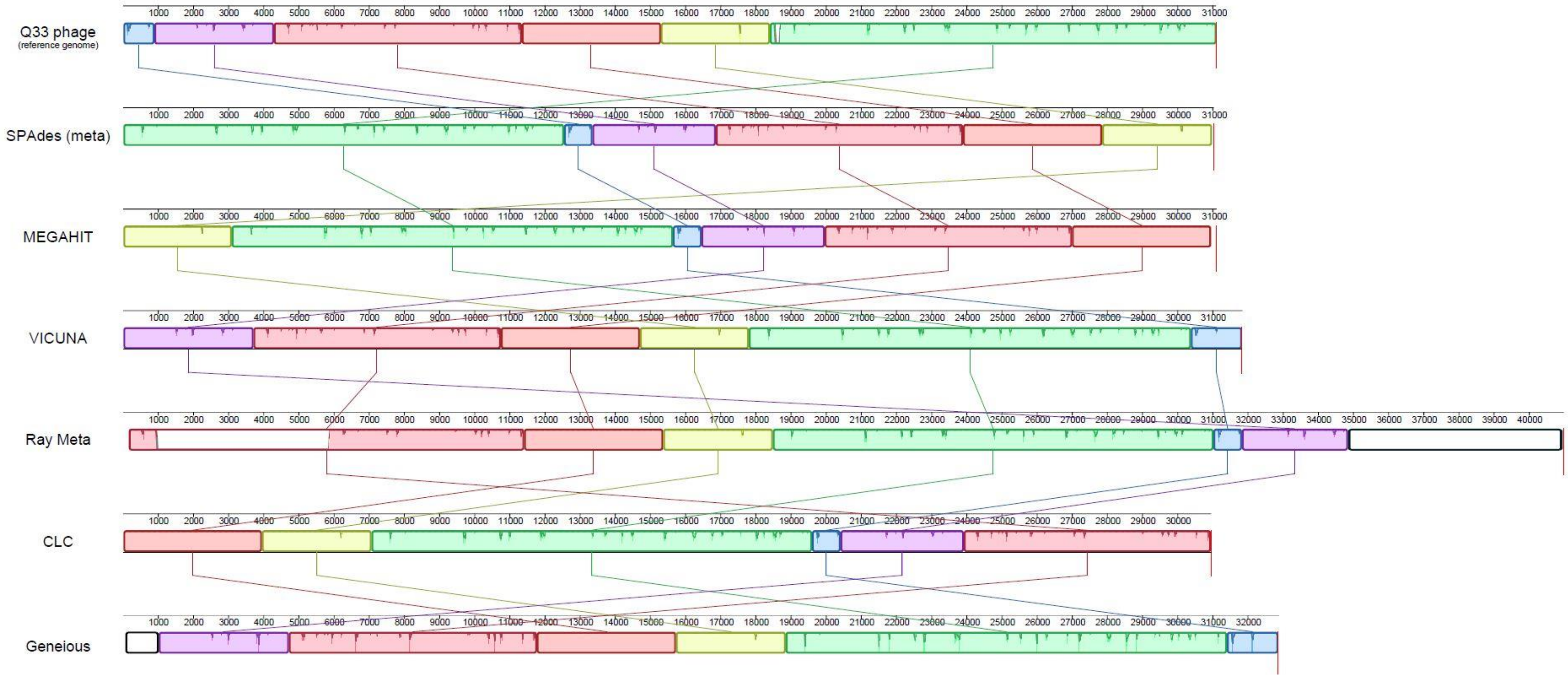


Figure 2. Mauve output of the Q33 reference genome (top) along with of the six assemblers which recovered > 99% of the genome with a single contig. Assembly regions outside of locally collinear blocks which do not share homology to the reference genome are highlighted by a black outline. Reverse complement of assemblies in the opposite orientation to the reference were plotted for visualisation purposes (VICUNA, CLC, Geneious).

Read depth analysis (Time and RAM)

Assemblers were compared for practicality by measuring the time to reach completion and maximum RAM usage via four published healthy human gut viromes (Manrique et al., 2016) and various sequencing depths . It must be noted that all assembly tasks were allocated five threads, however specifying the number of threads did not change the number of threads used by certain programs. MetaVelvet was not included in this analysis as it failed to reach completion after running for seven days. CLC and Geneious were performed on a desktop computer and therefore excluded from time and RAM analysis. Run time is dependent upon the number of reads and this is largely linear in scale with more reads leading to an increased assembly time (Figure. 4a). MIRA and Vicuna (Figure. 4a insert) were the slowest with MIRA taking over 15 times longer than the other software to assemble 3.5 million reads. SOAPdenovo2 had the shortest completion time followed by IBDA UD and Velvet. Most assemblers were consistent across samples (observed via error bars) with the exception of MIRA and Ray Meta. MIRA, Vicuna and Velvet (Figure. 4b insert) had the highest max RAM usage while the lowest was Ray Meta, IDBA UD and SPAdes (meta) (Figure. 4b). The majority of assemblers observed a linear scale pattern similar to that of run time.

Read depth analysis N50 and Longest contig length

For both the N50 (Figure. 4c) and the longest contig length (Figure. 4d), there was a large amount of variation between samples for the majority of assemblers. The longest contig length showed a large increase at the final sequencing depth. Particular assemblers, namely SPAdes (default), SPAdes (meta), MEGAHIT and ABySS (*k*-mer 127), produced longer contigs as the sequence depth was increased.

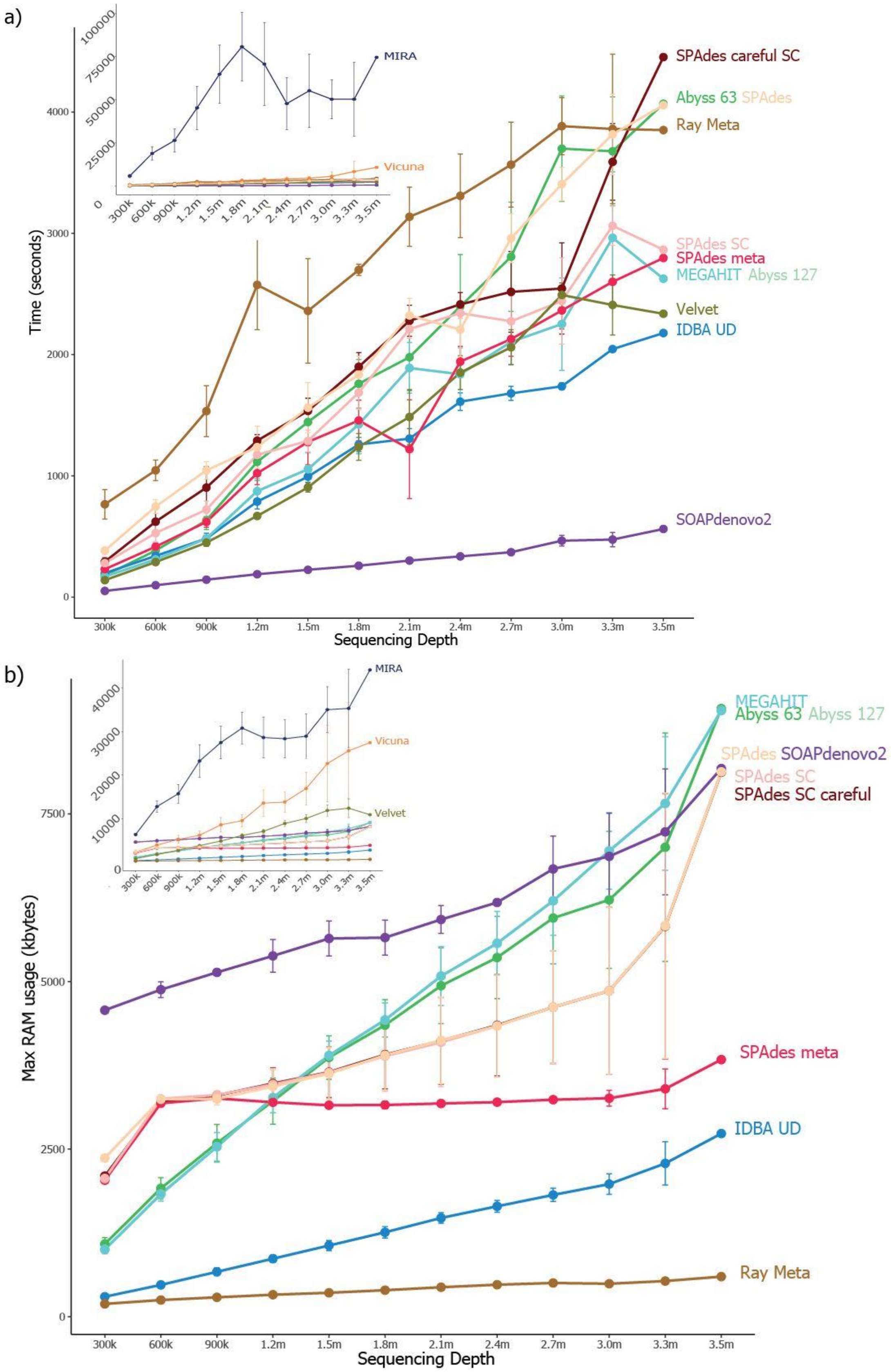


Figure 4. **A)** Time, measured in seconds, for each assembly to reach completion successfully for each read subset. **B)** The maximum RAM, measured in MB, used for each assembly for each read subset.

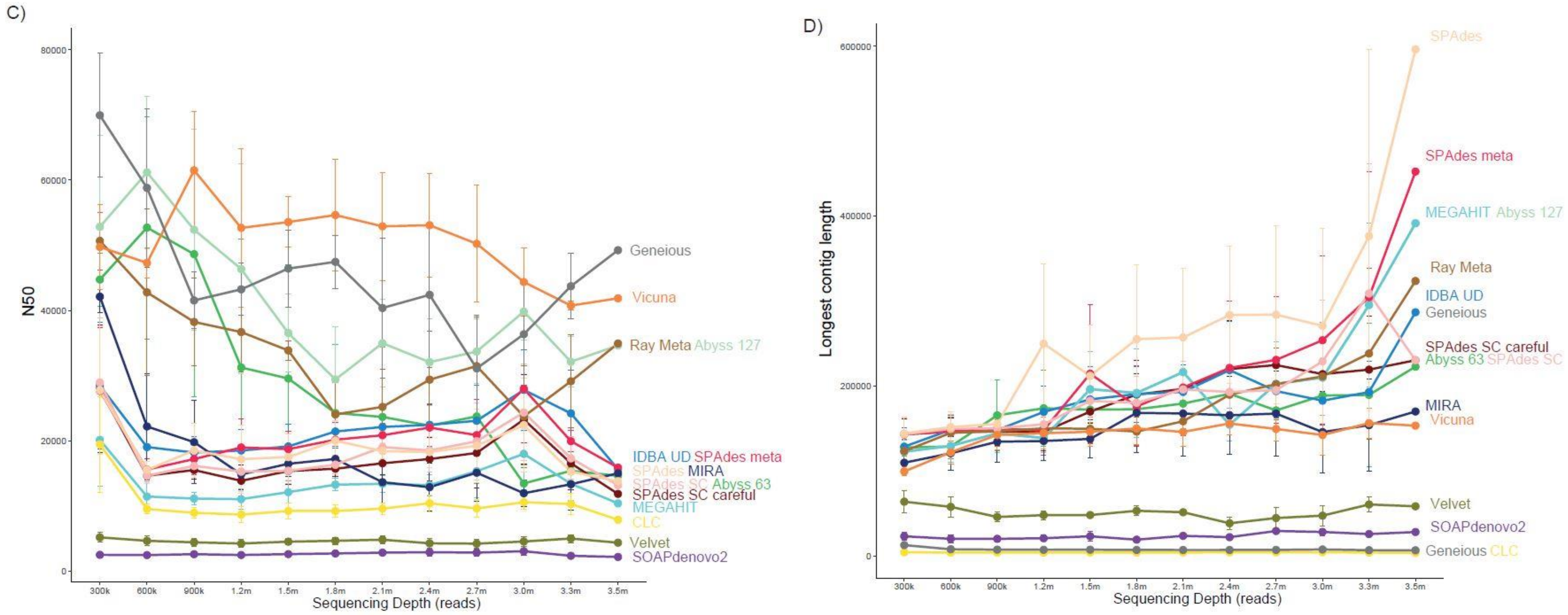


Figure 4. C) **Mean N50 length and D) mean contig length for four samples for each assembly across the read subsets after filtering contigs less than 1000 bases. Points represent the mean time for the four samples while error bars are the standard error**

Discussion

Many bacterial metagenomic assembly comparisons have highlighted that the choice of assembler has a significant impact on downstream analysis and the accuracy of the reconstructed metagenome (Mavromatis et al., 2007;Lindgreen et al., 2016;Greenwald et al., 2017;Vollmers et al., 2017). We have found this also to be true for viral metagenomes, where accurate and complete assembly are of particular importance given the lack of viral representation in reference databases. Virome studies depend heavily on the assembly step and possess many features which are challenging to successful assembly. In this study we compared the performance of those assemblers used to date in human viral metagenomics studies using datasets of known and unknown composition and varying complexity. These included a Q33 spiked virome, mock virome communities, a simulated virome and the "Healthy human gut phageome" (Manrique et al., 2016). Each dataset provided unique attributes allowing for comparison of assembly performance on a number of levels. The combination of artificial and real viromes used in this study allows for the comparison of various aspects of assembly performance across a range of datasets rather than depending on simulated viromes alone, as is commonly carried out in assembly comparisons (Mavromatis et al., 2007;Fritz et al., 2018) .

The Simulated dataset featured 572 viral genomes at various relative abundances as published by Vázquez-Castellanos and colleagues (Vázquez-Castellanos et al., 2014). Fragmented assemblies of individual genomes within microbial communities hamper downstream analysis and limit the conclusions which can be drawn from metagenomic data such as taxonomic and functional profiles (Florea et al., 2011). Consequently, the percentage genome recovery and degree of fragmentation was assessed across each assembler, with SPAdes (default, meta and single cell) each performing well. VICUNA performed very poorly, recovering only four contigs with high numbers of mismatches and misassemblies, despite having performed well with other datasets and being designed to address challenges of heterogeneous viral populations (Yang et al., 2012). This failure may reflect the computational challenges relating to the format of the simulated reads, as benchmarks carried out within the VICUNA study itself only include actual sequencing reads (Yang et al., 2012). However, similar poor performance has been previously observed in virome assembly comparison using VICUNA and 454 reads (Vázquez-Castellanos et al., 2014). For those assemblers which could recover greater than 90% of the reference genome in a single contig, SPAdes (default) outperformed SPAdes (meta). This may be explained by a lack of strain variants in the dataset and the fact that SPAdes (meta) was optimised to combine strain variants of each species to form consensus sequences.

A subset of genomes were poorly recovered (<20% genome fraction) by nearly all assemblers. This observation indicates that there are challenging aspects of viral genomes and metagenomes which cannot be overcome with current assembly strategies. The strong positive correlations between the relative abundance and genome fraction suggest that a low abundance threshold applies to virome assembly, below which assemblies will consist of small fractions of the viral genome, and in most cases be highly fragmented. This detrimental impact of low coverage has been well established in previous assembly comparison studies (García-López et al., 2015;Roux et al., 2017;Fritz et al., 2018). Highly abundant genomes also caused similar recovery and fragmentation issues across all assemblers, which is of particular importance due to the prevalence of extremely high abundance genomes in viral data (crAssphage, certain ssDNA viruses). As both abundance extremes are common in virome data, their impact must be considered when designing virome studies (i.e. sequencing depth). As relative abundance alone did not fully explain the variation in genome fraction recovered, the role of genomic repeats (a well-established assembly challenge (Acuña-Amador et al., 2018) was also investigated. However, genomic repeats could explain the variation in genome fraction recovered to a lesser degree than relative abundance, suggesting other factors contribute to poor genome recovery.

Compositional differences between final assemblies and viromes themselves must be taken into account when drawing conclusions about virome composition and setting parameters for downstream analysis. Challenges such as genomic content and strain variation are not currently addressed in human virome assembly strategies and impact the reconstruction of certain members of a virome. Hybrid sequencing, which uses both long and short reads to resolve genomic regions associated with poor assembly (Warwick-Dugdale et al., 2018) is a promising new technology which could address current virome assembly challenges. Library preparation methods which may reduce the bias introduced by MDA steps include using Swift Biosciences 1S Plus kit (Roux et al., 2016) and/or increasing overall sequencing depth or read length to improve recovery of lowly abundant viral genomes will be key. Furthermore, utilizing an assembler which can robustly deal with ultra-high coverage genomes (> 1000x coverage) is an important but often not appreciated aspect of virome assembly analysis. While promising, these potential solutions highlight a requirement for ongoing optimisation and extermination of virome analysis protocols.

Performance of some assemblers in this study was hampered by high coverage sequences (primarily overlap consensus assemblers). VICUNA assemblies exhibited the highest degree of fragmentation of all assemblers with Mock A, despite having resolved both high abundance ssDNA genomes of Mock B to a single contig. MIRA also exhibited a high degree of fragmentation with high abundance genomes in both simulated and mock datasets. However, MIRA was least affected by low abundance reads, recovering a greater genome fraction of low abundance genomes than other assemblers with fewer contigs. Performance of assemblers hampered by high coverage sequences in viromes may potentially be improved by sub-setting reads similar to the assembly approach used by SLICEMBLER (Mirebrahim et al., 2015).

Multi-assembler approaches such as the use of Geneious to generate consensus sequences from separate assemblers have been developed (Koren et al., 2014;Schürch et al., 2014;Deng et al., 2015) but have not been included in human virome studies using short reads. MIRA assemblies of the Q33 genome and some low abundance genomes in the Simulated dataset were improved using Geneious, resolving greater genome fractions with fewer contigs (despite Geneious recovering a lower genome fraction of the Simulated dataset overall). It is possible that using these approaches could address issues facing each assembler, i.e. combine the assemblies of SPAdes (meta) which performs well across all 4 datasets but struggles to recover low abundant genomes, with MIRA assemblies which are less affected by low abundance but has difficulty resolving genomes of higher abundance. Comparison of multi-assembler approaches and combinations of various assemblers was not within the scope of this study, but should not be ruled out as a potential method of improving virome assembly in cases where composition could be assessed and obvious assembly challenges were known to be present.

Across all analysis methods in this study, SPAdes (meta) performed consistently well and would be our recommendation. It performed best in the

Simulated data based on false positives, true positives and false negatives, best assembled the Q33 genome (recovery, fragmentation, misassemblies and genome size) and performed well with both mock communities in recovering all members accurately in one or two contigs. SPAdes (meta) RAM usage was low and did not increase to the same degree as other assemblers with increasing sequencing depth. This recommendation is in agreement with previous comparisons (Vollmers et al., 2017) which also suggested using SPAdes (meta) due to its ability to accurately assemble members of bacterial metagenomes. SPAdes (meta) is less able to accurately reconstruct micro-diversity as it generates a consensus assembly of "strain–contigs" in a metagenome, which means it is better equipped to address the high mutation rates observed in virome data (Nurk et al., 2017). This recommendation is also concurrent with a previous study (Roux et al., 2017) which found IDBA UD, MEGAHIT and SPAdes (meta) to perform equally well when assessed using 14 simulated viromes. However, we found that SPAdes (meta) outperformed IDBA UD and MEGAHIT in the Q33 spiked dataset, RAM usage in relation to increasing sequencing depth, and in its ability to recover members of the Simulated virome in a single contig. This recommendation contradicts two previous assembly comparisons which found CLC (Hesse et al., 2017) and Velvet (White, Wang et al. 2017) to be best suited to virome data. However, SPAdes (meta) was not included in either study. Though SPAdes (meta) was out performed by MIRA in the assembly of low abundance genomes in the Simulated dataset, MIRA has limited application to large datasets. MEGAHIT also performed well across all datasets performing well in relation to recovery, fragmentation and accuracy, but encountered some recovery issues in mock datasets and minor accuracy issues with the Q33 genome.

The higher levels of accuracy (low mismatch indel and misassembly counts) of assemblers which performed poorly in other metrics namely (velvet and ABySS (*k*-mer 63), highlights the trade-off between accuracy and contiguity observed in previous assembly studies (Gritsenko et al., 2012;Lin and Liao, 2013). However, both IDBA-UD and MEGAHIT performed well for accuracy, genome recovery and fragmentation. These assemblers may be worth considering if strain level detail is of particular importance. The mock A and B datasets were used to assess the impact of amplification bias on assembly performance. All ssDNA assemblies featured an equal minimum number of mismatches across both Mock A and B. This may be caused by challenges in the genomes themselves hampering accurate assembly, but is more likely to reflect strain variation between genome sequence featured in the original publication and the genome of the phage used in the community itself.

The Q33 spiked virome consisted of pooled reads from three healthy human faecal samples, each of which having been spiked with 10^7 PFU ml^{-1} of lactococcal phage Q33 prior to virome extraction. This allowed for assembly comparison of one abundant member of a challenging viral community. Despite the high relative abundance of the Q33 genome, only 6 assemblers could recover over 90% of the genome in a single contig, of these SPAdes (meta) and MEGAHIT reconstructed the Q33 genome accurately without the introduction foreign or chimeric DNA. It was also noted that the genome synteny was conserved across these six assemblies. This may reflect circularization of the linear Q33 genome during DNA extraction as the presence of cos sites has been previously predicted (Mahony et al., 2013).

The longest contigs of each assembler were only detected at the highest sequencing depths and varied across assemblers, which may indicate that high coverage is necessary to recover the largest viral genomes in a community. However, it is also possible that these long contigs may reflect misassemblies and duplication events caused by read errors at high sequencing depths which must be considered when analysing high coverage data. At almost all sequencing depths Geneious, Vicuna, Ray Meta and ABySS (k-mer 127) exhibited the highest N50 values, despite performing poorly in other metrics. This further highlights the limitation of using N50 alone as a metric of metagenomic assembly (Vollmers et al., 2017).

A further important consideration when performing any metagenomic assembly is practicality; size of dataset, computational resources, bioinformatic resources, and how much hands-on time is required from the end user. Both CLC and Geneious are available as a GUI (albeit requiring a licence fee) which widens their audience to researchers with limited command-line experience (CLC can also be run using the windows command line). However, this limits their practicality for large scale virome studies as they are limited to the computational power of desktop computers and are not suited to the assembly of large numbers of samples. Despite the limitations of computational power, CLC performed well in all datasets in terms of genome recovery and fragmentation. Of the freely available open source assemblers, MIRA and VICUNA are the least efficient in terms of RAM usage and assembly time, reflecting limitations of the overlap consensus approach to assembly. This limits their applicability to large virome datasets, and further increases the time required to carry out the Geneious assembly approach which requires the output of both assemblers. Despite the long runtime, VICUNA did not adhere to the number of cores specified, instead using all available cores. All other assemblers had a similar time requirements (with the exception of SOAPdenovo2 which performed poorly across all datasets). Of the assemblers which consistently performed well in terms of accuracy, genome fraction recovered and fragmentation, SPAdes (meta) was most efficient in terms of RAM usage, which did not increase to the same degree as other assemblers with increasing sequencing depth. MIRA stood out in terms of impracticality by generating by far the largest intermediate files of any assembler, requiring several gigabytes of storage space for intermediate files.

The combination of results from four datasets facilitates accurate comparison of assemblers as the limitations of each individual dataset vary. Application of Phi29 MDA to amplify virome DNA to sufficient quantities for sequencing can introduce significant bias and skew the original composition of the virome, making quantitative viromics difficult (Kim and Bae, 2011;Roux et al., 2016). As a result, it is likely that true diversity of viral metagenomes is not being accurately captured using current virome extraction methods. However, as these procedures move away from steps known to introduce bias, greater diversity will be observed. In this sense, the level of complexity of the Q33 dataset, which pooled three independent human viromes, provides a useful testbed for metagenomic assemblers in future virome studies as extraction methods improve. Additionally, Q33 was not present in the viromes prior to spiking, assemblers were not challenged by the presence of native strain variations of Q33 genome. In this study, assemblers were compared without individual optimisation to the specific dataset. Feasibility dictates that, this “straight out of the box” approach to assembly is used by almost all metagenomic assembly comparisons. Additionally, as the true composition of metagenomes is unknown, any impact of parameter optimisation must be estimated from general assembly statistics such as N50 and longest contig which have been highlighted to be of limited usefulness (Aguirre de Cárcer et al., 2014;Vollmers et al., 2017). Any parameter optimisation performed in the study (i.e. ABySS *k*-mer lengths, SPAdes careful vs. single cell) reflected parameters used in published virome studies and were not analysed in greater depth. While it is possible that parameter optimisation could improve individual assemblers we believe that the differences in assembly algorithms are the primary drivers of assembly performance.

This study describes the impact of a crucial analysis step on virome characterisation and highlights the need for a standardised analysis protocol across future virome studies. Such a protocol would allow for comparison across studies and facilitate accurate meta and cross analyses. This will be crucial should virome sequencing be utilised in diagnostic and clinical settings. However, it must be noted that any workflow will be somewhat limited biased to the detection of particular viral taxa. Consequently, studies (e.g. identifying novel viruses) may benefit from implementing multiple assembly approaches due to the large number of factors, both technical (read length, quality, paired-end information etc.) and biological (genetic complexity, evenness etc.) which impact virome assembly.

Conclusions

Of all assembly programs used in human virome studies, SPAdes (meta) addressed the challenges of virome data most effectively. However, all assemblers have limitations and are hampered by aspects of virome data. Low read coverage and high genomic repeats lead to assemblies with low recovered genome fraction and a higher degree of fragmentation, with the assemblies themselves being less accurate. This pattern was seen across all assemblers used in this study.

As assembler choice has significant implications for virome composition and the conclusions which can be drawn from a dataset, assemblers which performed poorly in this study (i.e. low genome recovery or accuracy and high degree of fragmentation) highlight a potential untapped resource in the sequence data of previously conducted virome studies. It is highly likely that many viral sequences were poorly assembled and reanalysis using a more effective assembler may yield new insights. Researchers conducting meta-analysis of virome sequencing studies should take particular care when evaluating viral assemblies from different assembly programs. Design of future virome studies should carefully consider the impact of sequencing depth, as extremes in read coverage will prevent the assembly and detection of viral genomes at both high and low abundance.

Abbreviations/Glossary

The following terms; Genome fraction, N50, number of contigs, misassemblies, local misassemblies, are defined by QUAST (Mikheenko et al., 2015)

Genome fraction "is the total number of aligned bases in the reference, divided by the genome size. A base in the reference genome is counted as aligned if there is at least one contig with at least one alignment to this base. Contigs from repeat regions may map to multiple places, and thus may be counted multiple times in this quantity."

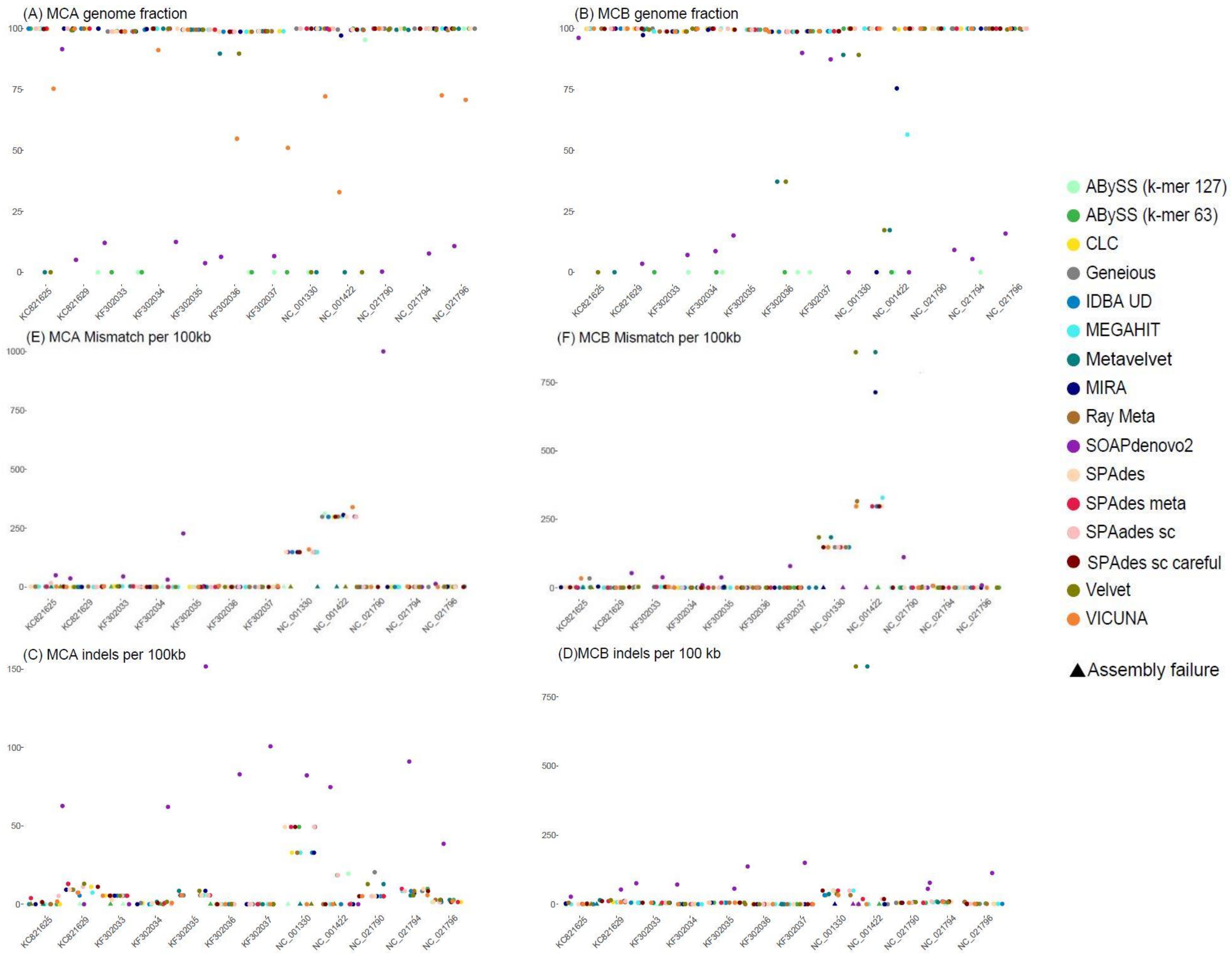
N50 "is the contig length such that using longer or equal length contigs produces half (50%) of the bases of the assembly. Usually there is no value that produces exactly 50%, so the technical definition is the maximum length x such that using contigs of length at least x accounts for at least 50% of the total assembly length."

Number of contigs "is the total number of contigs in the assembly."

Misassemblies “is the number of positions in the assembled contigs where the left flanking sequence aligns over 1 kbp away from the right flanking sequence on the reference (relocation) or they overlap on more than 1 kbp (relocation) or flanking sequences align on different strands (inversion) or different chromosomes (translocation).”

Local misassemblies “A local misassembly has two or more distinct alignments covering the breakpoint, the gap between left and right flanking sequences is less than 1 kbp and the left and right flanking sequences both are on the same strand of the same chromosome of the reference genome.”

Figure S1. Analysis of recovered genome fraction and indel/mismatch counts for mock communities A and B. Triangles represent N/A values for mismatches and indels caused by assembly failures.



Data availability

The Sequencing reads which support this study are available from the following links.

Mock communities A and B are available at: http://datacommons.cyverse.org/browse/iplant/home/shared/iVirus/DNA_Viromes_library_comparison .

Simulated virome reads are available at: https://figshare.com/articles/Simulated_virome_datasest_for_assembly_and_annotation_tests/5151163

Reads used to compare the impact of sequencing depth on time and RAM usage are available from the NCBI SRA; http://www.ncbi.nlm.nih.gov/sra under the accession numbers SAMN04415496 to SAMN04415499

Reads from human viromes spiked with 10^7 PFU of Lactococcal phage Q33 phage are available at http://www.ncbi.nlm.nih.gov/sra under the accession numbers SRX3240741, SRX3240716, SRX3240715

Supplementary material

Supplementary tables 1-6 are available from at the following link

https://static-content.springer.com/esm/art%3A10.1186%2Fs40168-019-0626-5/MediaObjects/40168_2019_626_MOESM5_ESM.xlsx

Table S1. Spearman correlation values from the relationships of indel, mismatch and misassembly counts, recovered genome fraction, abundance and total proportion of genomic repeats within the simulated virome. *GF–recovered genome fraction.

Table S2. Linear modelling correlation values comparing recovered genome fraction, total proportion of genomic repeats and abundance for the Simulated virome.

Table S3. Spearman correlation values from the relationships of inverted, tandem, palindromic and total repeats, abundance and the number of contigs generated by each assembler for the Simulated virome.

Table S4. (A) Ranking table comparing recovered genome fraction and contig numbers for assemblers which recovered at least 50% of the total genome fraction. (B) Ranking table of indel, mismatch and misassembly counts per 100 kb, normalised to the number of genomes recovered to at least 50%.

Table S5. Number of aligned and unaligned contigs generated by each assembler for mock community A.

Table S6. Number of aligned and unaligned contigs generated by each assembler for mock community B.

MetaQUAST outputs for each dataset are available at the following links

Simulated virome

https://static-content.springer.com/esm/art%3A10.1186%2Fs40168-019-0626-5/MediaObjects/40168_2019_626_MOESM1_ESM.html

Mock virome A.
https://static-content.springer.com/esm/art%3A10.1186%2Fs40168-019-0626-5/MediaObjects/40168_2019_626_MOESM2_ESM.html

Mock virome B.

https://static-content.springer.com/esm/art%3A10.1186%2Fs40168-019-0626-5/MediaObjects/40168_2019_626_MOESM3_ESM.html

Q33-spiked virome

https://static-content.springer.com/esm/art%3A10.1186%2Fs40168-019-0626-5/MediaObjects/40168_2019_626_MOESM4_ESM.html

Acknowledgements

We thank the authors of all datasets used in this study for the availability of their data.

References

Acuña-Amador, L., Primot, A., Cadieu, E., Roulet, A., and Barloy-Hubler, F. (2018). Genomic repeats, misassembly and reannotation: a case study with long-read resequencing of Porphyromonas gingivalis reference strains. *BMC genomics* 19**,** 54.

Aggarwala, V., Liang, G., and Bushman, F.D. (2017). Viral communities of the human gut: metagenomic analysis of composition and dynamics. *Mobile DNA* 8**,** 12.

Aguirre De Cárcer, D., Angly, F.E., and Alcamí, A. (2014). Evaluation of viral genome assembly and diversity estimation in deep metagenomes. *BMC Genomics* 15**,** 989. doi:10.1186/1471-2164-15-989.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S. et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology* 19**,** 455-477.

Boisvert, S., Raymond, F., Godzaridis, É., Laviolette, F., and Corbeil, J. (2012). Ray Meta: scalable de novo metagenome assembly and profiling. *Genome biology* 13**,** R122.

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

Breitbart, M. (2011). Marine viruses: truth or dare.

Darling, A.E., Mau, B., and Perna, N.T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PloS one* 5**,** e11147.

Deng, X., Naccache, S.N., Ng, T., Federman, S., Li, L., Chiu, C.Y. et al. (2015). An ensemble strategy that significantly improves de novo assembly of microbial genomes from metagenomic next-generation sequencing data. *Nucleic acids research* 43**,** e46-e46.

Dutilh, B.E., Cassman, N., Mcnair, K., Sanchez, S.E., Silva, G.G., Boling, L. et al. (2014). A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nature communications* 5**,** ncomms5498.

Florea, L., Souvorov, A., Kalbfleisch, T.S., and Salzberg, S.L. (2011). Genome assembly has a major impact on gene content: a comparison of annotation in two Bos taurus assemblies. *PLoS One* 6**,** e21400.

Foulongne, V., Sauvage, V., Hebert, C., Dereure, O., Cheval, J., Gouilh, M.A. et al. (2012). Human skin microbiota: high diversity of DNA viruses identified on the human skin by high throughput sequencing. *PloS one* 7**,** e38499.

Fritz, A., Hofmann, P., Majda, S., Dahms, E., Droege, J., Fiedler, J. et al. (2018). CAMISIM: Simulating metagenomes and microbial communities. *bioRxiv***,** 300970.

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

Greenwald, W.W., Klitgord, N., Seguritan, V., Yooseph, S., Venter, J.C., Garner, C. et al. (2017). Utilization of defined microbial communities enables effective evaluation of meta-genomic assemblies. *BMC genomics* 18**,** 296.

Gritsenko, A.A., Nijkamp, J.F., Reinders, M.J., and Ridder, D.D. (2012). GRASS: a generic algorithm for scaffolding next-generation sequencing assemblies. *Bioinformatics* 28**,** 1429-1437.

Guo, L., Hua, X., Zhang, W., Yang, S., Shen, Q., Hu, H. et al. (2017). Viral metagenomics analysis of feces from coronary heart disease patients reveals the genetic diversity of the Microviridae. *Virologica Sinica* 32**,** 130-138.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29**,** 1072-1075.

Hannigan, G.D., Meisel, J.S., Tyldsley, A.S., Zheng, Q., Hodkinson, B.P., Sanmiguel, A.J. et al. (2015). The human skin double-stranded DNA virome: topographical and temporal diversity, genetic enrichment, and dynamic associations with the host microbiome. *MBio* 6**,** e01578-01515.

Hesse, U., Van Heusden, P., Kirby, B.M., Olonade, I., Van Zyl, L.J., and Trindade, M. (2017). Virome Assembly and Annotation: A Surprise in the Namib Desert. *Frontiers in Microbiology* 8. doi:10.3389/fmicb.2017.00013.

Hurwitz, B.L., and Sullivan, M.B. (2013). The Pacific Ocean Virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PloS one* 8**,** e57355.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S. et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28**,** 1647-1649.

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***,** AEM. 00289-00211.

Koren, S., Treangen, T.J., Hill, C.M., Pop, M., and Phillippy, A.M. (2014). Automated ensemble assembly and validation of microbial genomes. *BMC bioinformatics* 15**,** 126.

Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K. et al. (2016). MEGAHIT v1. 0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* 102**,** 3-11.

Lim, E.S., Zhou, Y., Zhao, G., Bauer, I.K., Droit, L., Ndao, I.M. et al. (2015). Early life dynamics of the human gut virome and bacterial microbiome in infants. *Nature medicine* 21**,** 1228.

Lin, S.-H., and Liao, Y.-C. (2013). CISA: contig integrator for sequence assembly of bacterial genomes. *PloS one* 8**,** e60843.

Lindgreen, S., Adair, K.L., and Gardner, P.P. (2016). An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific reports* 6**,** 19233.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J. et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 1**,** 18.

Mahony, J., Martel, B., Tremblay, D.M., Neve, H., Heller, K.J., Moineau, S. et al. (2013). Identification of a new P335 subgroup through molecular analysis of lactococcal phages Q33 and BM13. *Applied and environmental microbiology* 79**,** 4401-4409.

Manrique, P., Bolduc, B., Walk, S.T., Van Der Oost, J., De Vos, W.M., and Young, M.J. (2016). Healthy human gut phageome. *Proceedings of the National Academy of Sciences* 113**,** 10400-10405.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* 17**,** pp. 10-12.

Mavromatis, K., Ivanova, N., Barry, K., Shapiro, H., Goltsman, E., Mchardy, A.C. et al. (2007). Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature methods* 4**,** 495.

Mccann, A., Ryan, F.J., Stockdale, S.R., Dalmasso, M., Blake, T., Ryan, C.A. et al. (2018). Viromes of one year old infants reveal the impact of birth mode on microbiome diversity. *PeerJ* 6**,** e4694.

Mikheenko, A., Saveliev, V., and Gurevich, A. (2015). MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* 32**,** 1088-1090.

Minot, S., Grunberg, S., Wu, G.D., Lewis, J.D., and Bushman, F.D. (2012). Hypervariable loci in the human gut virome. *Proceedings of the National Academy of Sciences* 109**,** 3962-3966.

Mirebrahim, H., Close, T.J., and Lonardi, S. (2015). De novo meta-assembly of ultra-deep sequencing data. *Bioinformatics* 31**,** i9-i16.

Namiki, T., Hachiya, T., Tanaka, H., and Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic acids research* 40**,** e155-e155.

Norman, J.M., Handley, S.A., Baldridge, M.T., Droit, L., Liu, C.Y., Keller, B.C. et al. (2015). Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* 160**,** 447-460.

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome research***,** gr. 213959.213116.

Olson, N.D., Treangen, T.J., Hill, C.M., Cepeda-Espinoza, V., Ghurye, J., Koren, S. et al. (2017). Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Briefings in bioinformatics*.

Paul, J.H. (2008). Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas? *The ISME journal* 2**,** 579.

Peng, Y., Leung, H.C., Yiu, S.-M., and Chin, F.Y. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28**,** 1420-1428.

Roux, S., Emerson, J.B., Eloe-Fadrosh, E.A., and Sullivan, M.B. (2017). Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ* 5**,** e3817.

Roux, S., Hallam, S.J., Woyke, T., and Sullivan, M.B. (2015). Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *Elife* 4**,** e08490.

Roux, S., Solonenko, N.E., Dang, V.T., Poulos, B.T., Schwenck, S.M., Goldsmith, D.B. et al. (2016). Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. *PeerJ* 4**,** e2777.

Schürch, A.C., Schipper, D., Bijl, M.A., Dau, J., Beckmen, K.B., Schapendonk, C.M. et al. (2014). Metagenomic survey for viruses in Western Arctic caribou, Alaska, through iterative assembly of taxonomic units. *PLoS One* 9**,** e105227.

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J. et al. (2017). Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nature methods* 14**,** 1063.

Shkoporov, A.N., Ryan, F.J., Draper, L.A., Forde, A., Stockdale, S.R., Daly, K.M. et al. (2018). Reproducible protocols for metagenomic analysis of human faecal phageomes. *Microbiome* 6**,** 68.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J., and Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome research***,** gr. 089532.089108.

Smits, S.L., Bodewes, R., Ruiz-Gonzalez, A., Baumgärtner, W., Koopmans, M.P., Osterhaus, A.D. et al. (2014). Assembly of viral genomes from metagenomes. *Frontiers in microbiology* 5**,** 714.

Solden, L., Lloyd, K., and Wrighton, K. (2016). The bright side of microbial dark matter: lessons learned from the uncultivated majority. *Current opinion in microbiology* 31**,** 217-226.

Vázquez-Castellanos, J.F., García-López, R., Pérez-Brocal, V., Pignatelli, M., and Moya, A. (2014). Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut. *BMC genomics* 15**,** 37.

Vollmers, J., Wiegand, S., and Kaster, A.-K. (2017). Comparing and evaluating metagenome assembly tools from a microbiologist's perspective-Not only size matters! *PloS one* 12**,** e0169662.

Warwick-Dugdale, J., Solonenko, N., Moore, K., Chittick, L., Gregory, A.C., Allen, M.J. et al. (2018). Long-read metagenomics reveals cryptic and abundant marine viruses. *bioRxiv*. doi:10.1101/345041.

Yang, X., Charlebois, P., Gnerre, S., Coole, M.G., Lennon, N.J., Levin, J.Z. et al. (2012). De novo assembly of highly diverse viral populations. *BMC genomics* 13**,** 475.

Zerbino, D., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research***,** gr. 074492.074107.

Chapter 3

Whole-virome analysis sheds light on viral dark matter in inflammatory bowel disease

This chapter has been published as the following:

Clooney, A.G.*, Sutton, T.D.S.*, Shkoporov, A.N., Holohan, R.K., Daly, K.M., O'regan, O., Ryan F.J., Draper L.A., Plevy S.E., Ross R.P, Hill C.(2019). Whole-Virome Analysis Sheds Light on Viral Dark Matter in Inflammatory Bowel Disease. *Cell Host & Microbe* 26**,** 764-778.e765. doi:https://doi.org/10.1016/j.chom.2019.10.009

* Contributed Equally

Manuscript written by Thomas Sutton and Adam Clooney. Editorial revisions carried out by Thomas Sutton. Study design by Adam Clooney, Thomas Sutton, Andrey Shkoporov, Lorraine Draper, Colin Hill, Scot Plevy and Paul Ross. Bioinformatic analysis carried out by Adam Clooney (Figures 1-6) and Thomas Sutton (Figures 1-4 and 7). Bioinformatic support provided by Andrey Shkoporov, Ross Holohan and Feargal Ryan. DNA extraction and library preparation carried out by Karen Daly, and Orla O' Regan. Supervision provided by Colin Hill and Paul Ross.

Abstract

The human gut virome is thought to significantly impact the microbiome and human health. However, most virome analyses have been performed on a limited fraction of known viruses. Using whole-virome analysis on a published keystone inflammatory bowel disease (IBD) cohort and an in-house ulcerative colitis data set, we shed light on the composition of the human gut virome in IBD beyond this identifiable minority. We observe IBD-specific changes to the virome and increased numbers of temperate phage sequences in individuals with Crohn's disease. Unlike prior database-dependent methods, no changes in viral richness were observed. Among IBD subjects, the changes in virome composition reflected alterations in bacterial composition. Furthermore, incorporating both bacteriome and virome composition offered greater classification power between health and disease. This approach to analysing whole-virome across cohorts highlights significant IBD signals, which may be crucial for developing future biomarkers and therapeutics.

Introduction

The virome is likely to be one of the major forces shaping the human gut microbiome, but is perhaps its least understood component. The virome is dominated by bacteriophage (phage) which play vital roles in many microbial communities by driving diversity, aiding nutrient turnover (Weitz et al., 2015) and facilitating horizontal gene transfer (Canchaya et al., 2003). Understanding the role of bacteriophages in microbial community structures will be essential if we are to understand or control the alterations in human gut microbiome composition and diversity associated with many diseases, including Inflammatory Bowel Disease (IBD) (Gevers et al., 2014;Halfvarson et al., 2017), obesity (Le Chatelier et al., 2013) and diabetes (Forslund et al., 2015).

Many gut bacteria (and potential phage hosts) remain difficult to culture (Forster et al., 2019), which means that analysing the virome depends heavily on metagenomic sequencing and bioinformatic approaches. However, a lack of universal marker genes on phage (similar to 16S rRNA in bacteria) and a subsequent lack of taxonomic information due to poorly populated databases (Krishnamurthy and Wang, 2017) means that database-independent methods are required and that virome analysis must be carried out at the level of metagenomic assembly or individual viral genomes.

Early metagenomic studies highlighted the novelty and diversity of the human gut virome, but were able to classify only a minor fraction (2%) of sequenced DNA (Minot et al., 2011). Improvements in high throughput sequencing technologies have allowed the virome to be analysed in unprecedented detail with studies sequencing up to 50 million reads per sample (Zuo et al., 2019). It has been confirmed that the virome is incredibly diverse, that the majority do not align to known sequences in databases (i.e. viral dark matter)(Roux et al., 2015b), and that composition is highly unique to individuals (Reyes et al., 2010;Moreno-Gallego et al., 2019;Shkoporov et al., 2019).

Inflammatory Bowel Disease, including Crohn's disease (CD) and ulcerative colitis (UC), is a chronic disorder of the intestinal tract resulting in periods of flare and remission. Although the aetiology of IBD remains unclear, it appears to be multifactorial and has been repeatedly associated with alterations in the human gut microbiome. These include decreased bacterial diversity and a reduced abundance of certain *Firmicutes* and *Bacteroides*. There is an emerging body of data providing evidence that the gut virome is altered in IBD (Norman et al., 2015;Fernandes et al., 2019;Zuo et al., 2019) with increased overall virome diversity, and an increased relative abundance of the family *Caudovirales*. Yet nearly all findings have been drawn from compositional changes of the identifiable fraction of the virome, which can be as little as 15% of the data (Norman et al., 2015). This limits the overall understanding of the virome and hampers the identification of potential disease biomarkers.

A database independent analysis method is essential if we are to fully characterise changes in the gut virome in health and disease. This approach begins with metagenomic assembly of short reads to resolve viral genomes and subsequent alignment of reads to these assemblies to determine their relative abundance. Spurious alignments to repeat and conserved regions are removed from further analysis by using a breadth of coverage filter (Roux et al., 2017). However, at this level of resolution the virome exhibits enormous diversity and interpersonal variation (Reyes et al., 2010), obscuring any patterns in the virome across individuals and cohorts. As part of this study, we reanalysed a previously published keystone data set (Norman et al., 2015) consisting of subjects with CD, UC and healthy controls. We overcame strain-level resolution using protein homology and MCL (Markov Cluster Algorithm) to group viral sequences into putative higher taxonomic ranks. In this way it was possible to describe compositional changes across the entire virome in health and disease, beyond the known minority. We propose that a core virome in healthy individuals shifts towards a community that is less stable and dominated by temperate phage and IBD. We show that virome alterations mimic those of the bacteriome and that when used together, they offer an improved method for classifying IBD patients from healthy subjects. We also validated our results using a longitudinal cohort of patients with UC in both active and inactive states of disease. This analysis approach supports future virome studies by providing insight into changes in composition across the entire data set. By comparing the whole-virome composition of other published datasets, it may also reveal further disease-specific alterations that had been previously obscured. For details of the analysis methods used in this study, see (STAR Methods).

Results

Data sets

We reanalysed a previously published dataset of healthy and IBD gut viromes (Norman et al., 2015). The data set comprised of 165 virome samples from 130 subjects, which consisted of 61 healthy controls, 27 subjects with CD and 42 with UC. Of these, six samples were known to be collected during CD flare, eight in CD remission, 13 in UC flare and 20 in UC remission. To expand upon these findings, we explored a second data set of longitudinal samples for 40 subjects with UC, focusing on the impact of disease on gut virome composition. The cohort was part of the PURSUIT-M phase 3 clinical trial (STAR Methods). This data set was generated in-house and consisted of samples from periods of flare (82) and remission (31). For this data set, disease activity was determined by Mayo score. For all subjects, initial samples were taken during a period of flare (week 0). Two further time points were taken for each subject at weeks 6 and 30. For both data sets, 16S rRNA gene sequencing data was also obtained and performed on 149 (data set 1) and 109 (data set 2) samples.

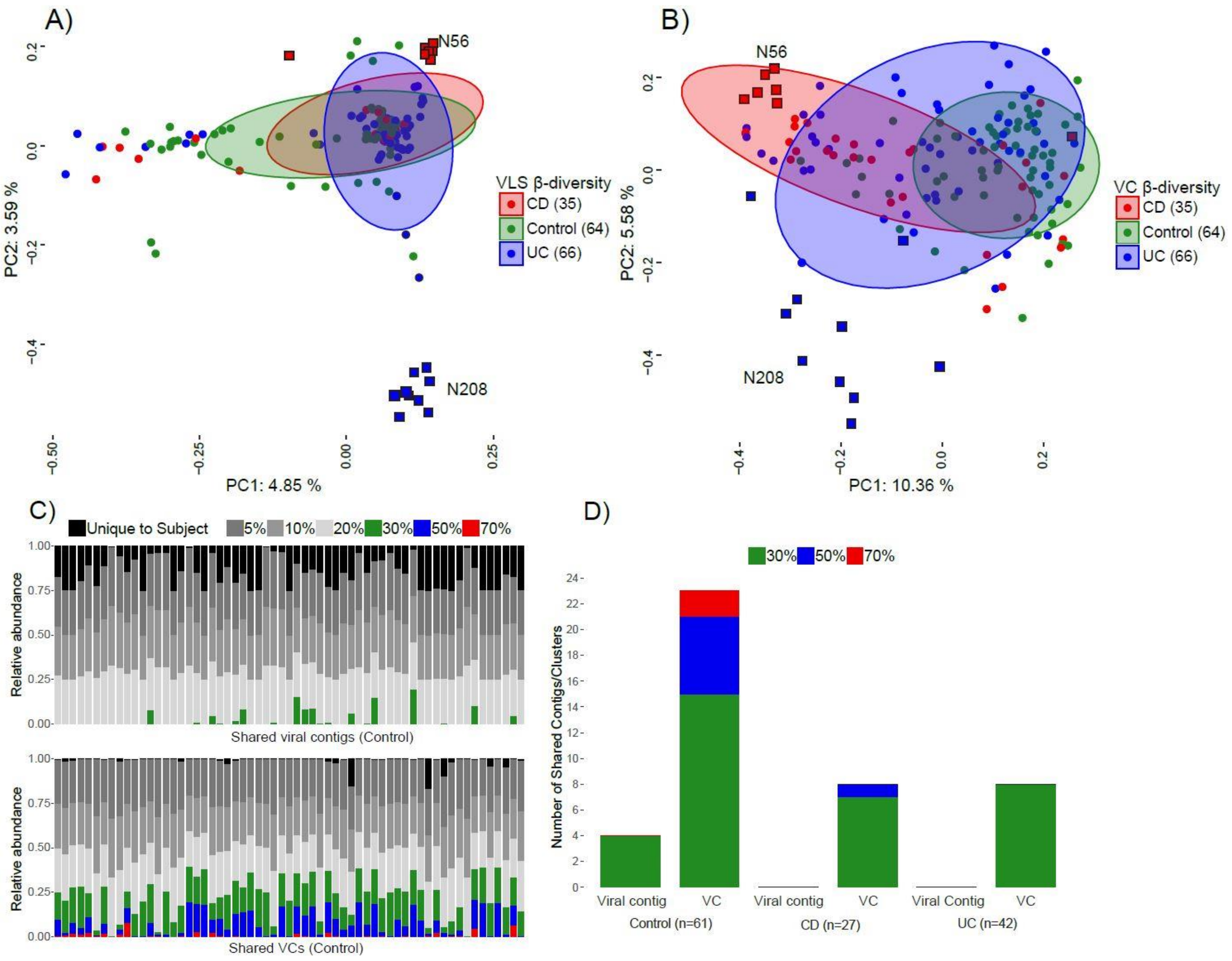
Clustering is required to overcome virome individuality and allow cohort comparisons

Initially, virome analysis was performed on the Norman data set by aligning quality filtered reads to the final set of virus-like sequences (VLS, see STAR Methods), made non-redundant at 90% identity over 90% of length. This resulted in a mean of 80.38%

(± 29.29%) quality filtered reads per sample being used in the final analysis. As VLS represent groups of highly related viral genomes (whole or partial), analysis was carried out at a strain or species level. This was reflected in the extremely high levels of individuality amongst subjects. It was also observed that the individuals themselves were the primary drivers of separation and longitudinal samples grouped together (Fig.1A). This individual specificity masked compositional differences of the virome and each of the cohorts (control, CD and UC) showed little divergence. PC-axes 1 and 2 described very little of the variation (4.85% and 3.59%), suggesting that disease-specific changes in virome composition were not visible at the level of VLS (patterns of β-diversity were reproducible across various metrics, data not shown).

Lower taxonomic resolution (i.e. a higher taxonomic rank) was required to overcome this high level of interpersonal variation and identify compositional changes in the virome associated with disease. This was achieved by clustering VLS based on protein-coding gene sharing networks (Bin Jang et al., 2019) (see STAR Methods). The VLS clustered into 472 Viral Clusters (VCs) of >2 members with 2,382 singletons remaining, henceforward referred to as a VC with 1 member. The resulting VCs formed a new count table and a VC-based analysis of β-diversity was carried out (Figure. 1B). Samples largely grouped per condition with noticeable increases in the eigenvalues to 10.36% and 5.58% variation explained for PC-axes 1 and 2, respectively, meaning the biological signals that drove separation between samples were considerably stronger. However, it should be noted that samples with true deviation from the main cohort (such as subjects N208 and N56) remained distinctive, suggesting that the clustering process retains true compositional differences. To further determine if clustering VLS could overcome the masking effect of inter-individual variation, the relative abundances of shared and unique VLS and VCs were plotted for control subjects (Figure. 1C).

Figure. 1: A comparison of shared viral features across the Norman et al. data set samples pre and post clustering of VLS and VCs. PCoA of Spearman distances using A) VLS (pre-clustering) and B) viral clusters (post-clustering) count tables. Points from subjects N208 and N56 are represented with a square and have a black border. C) The relative abundance of VLS (top) and VCs (bottom) for control subjects at varying thresholds of shared viral features across subjects. D) The number of VLS/VCs shared across 30%, 50% and 70% of subjects in each cohort.

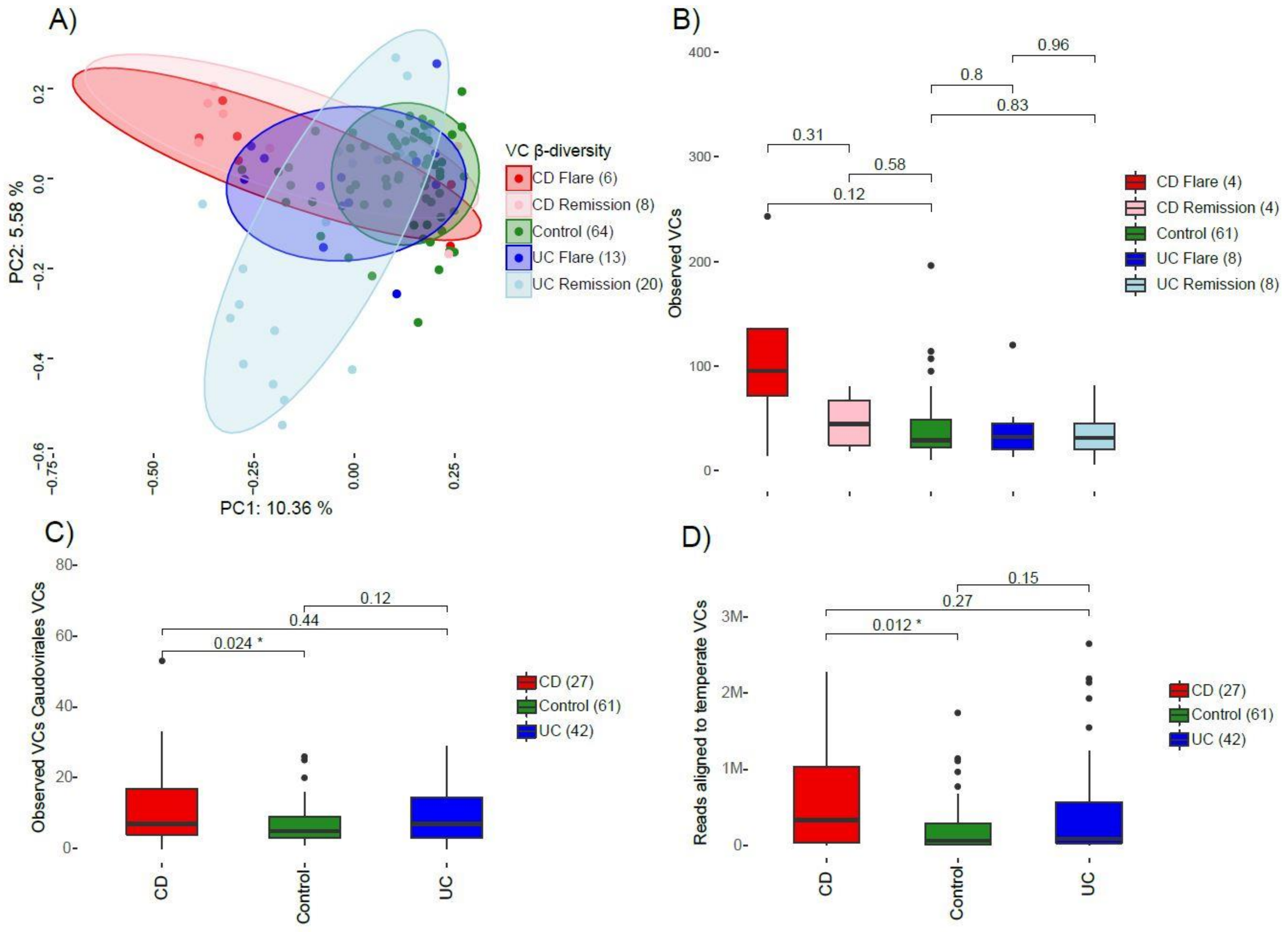


At a VLS level, inter-individual variation was represented by a high proportion of sequences unique to a given subject (relative abundance 14% ± 8%). Furthermore, sequences that were shared across 30% individuals made up a minor proportion of the virome (relative abundance 1.7% ± 4%) and no VLS was shared across 50% or more of individuals. In contrast, inter-individuality was far less evident at a VC level and the proportion of VCs unique to an individual was lower (relative abundance 1.3% ± 3%). Shared VCs also made up a substantially larger proportion of the virome with a relative abundance of 15% ± 6% per subject shared across 30% of the cohort, 7.1% ± 6.6% across 50% and 0.7% ± 1.4% across 70%. Furthermore, a total of eight VCs were shared across 30% of CD and UC cohorts (Figure. 1D). Analysis was continued at a VC level as these shared features made it possible to compare viromes across and between cohorts.

Analysis of viral clusters reveals IBD specific alterations in the gut virome

In the Norman et al. data set, β-diversity PCoA (Spearman distances) yielded the greatest degree of separation between the viromes of CD patients (relapse/remission) and healthy controls (PERMANOVA, p = 0.0023/0.0032, respectively) followed by UC relapse/remission (PERMANOVA, p = 0.002/0.0023) (Figure. 2A). Variations observed between disease states of each condition, were not significant, which may be due to small sample sizes (PERMANOVA, p > 0.05). PCoA without the division of relapse/remission status showed CD and UC β-diversity significantly differed from healthy controls (PERMANOVA, p = 0.0002 and 0.0002; Figure. S1A). The healthy cohort also had the greatest similarity across subjects, having the lowest pairwise distances between points (Figure. S1B), which supports the previous observations of shared VCs across individuals (Figure. 1D). This core virome (defined as presence across >50% of subjects) in the healthy cohort was composed of two VCs (vc2 and vc7) shared across >70% of subjects and six (vc1, vc10, vc23, vc25, vc32, vc39) across >50%. vc1 was classified as temperate *Siphoviridae* with CRISPR hits to various *Firmicutes* and *Parascardovia* (phylum - *Actinobacteria*) and vc10 was classified as a crAss-like phage (Guerin et al., 2018). However, the majority of these VCs were unclassified (i.e. did not cluster with known viral genomes). This highlights the important biological signals which are often overlooked by database-dependent analysis methods.

Figure. 2: Virome composition comparison of the Norman et al. IBD cohorts to controls. A) PCoA using Spearman distances, B) α-diversity (observed VCs, sample sizes are reduced relative to A due to the retention of only 1 sample per subject). A Wilcoxon test was employed for tests of significance. C) α-diversity of observed VLPs for any VCs classified as *Caudovirales* (Wilcoxon test). D) Read alignment for samples in each cohort to VCs classified as temperate, (Wilcoxon test).



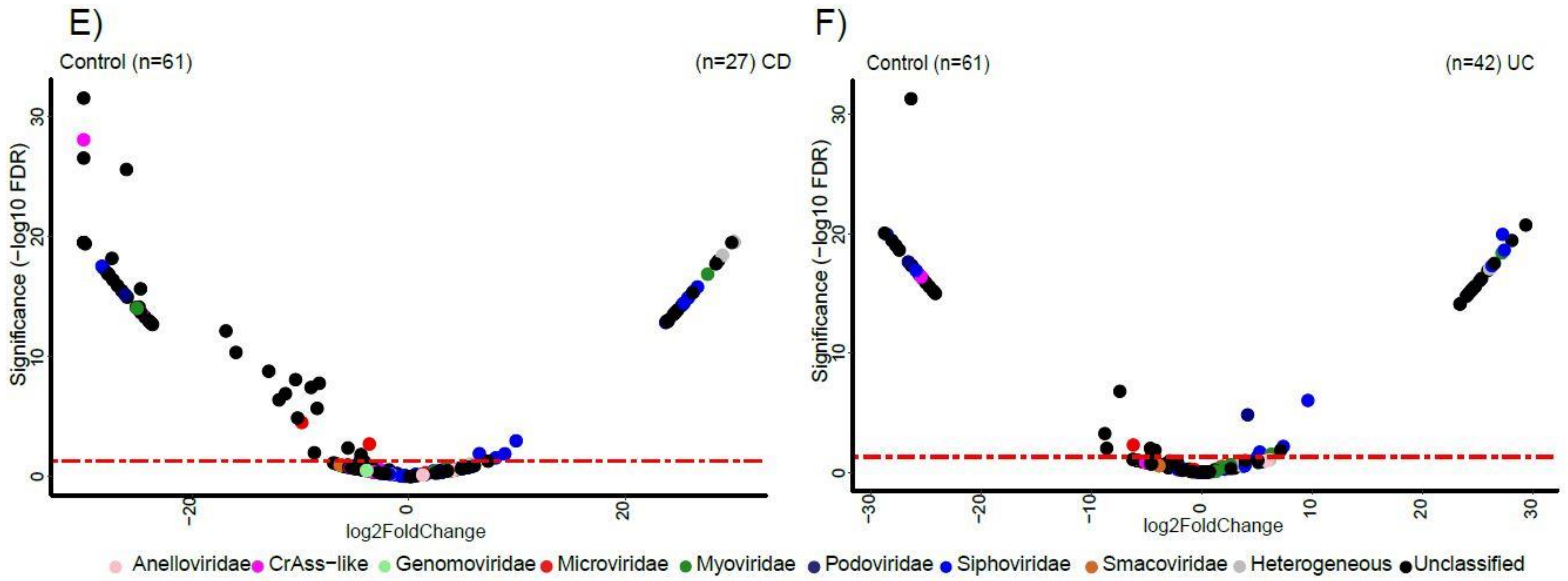
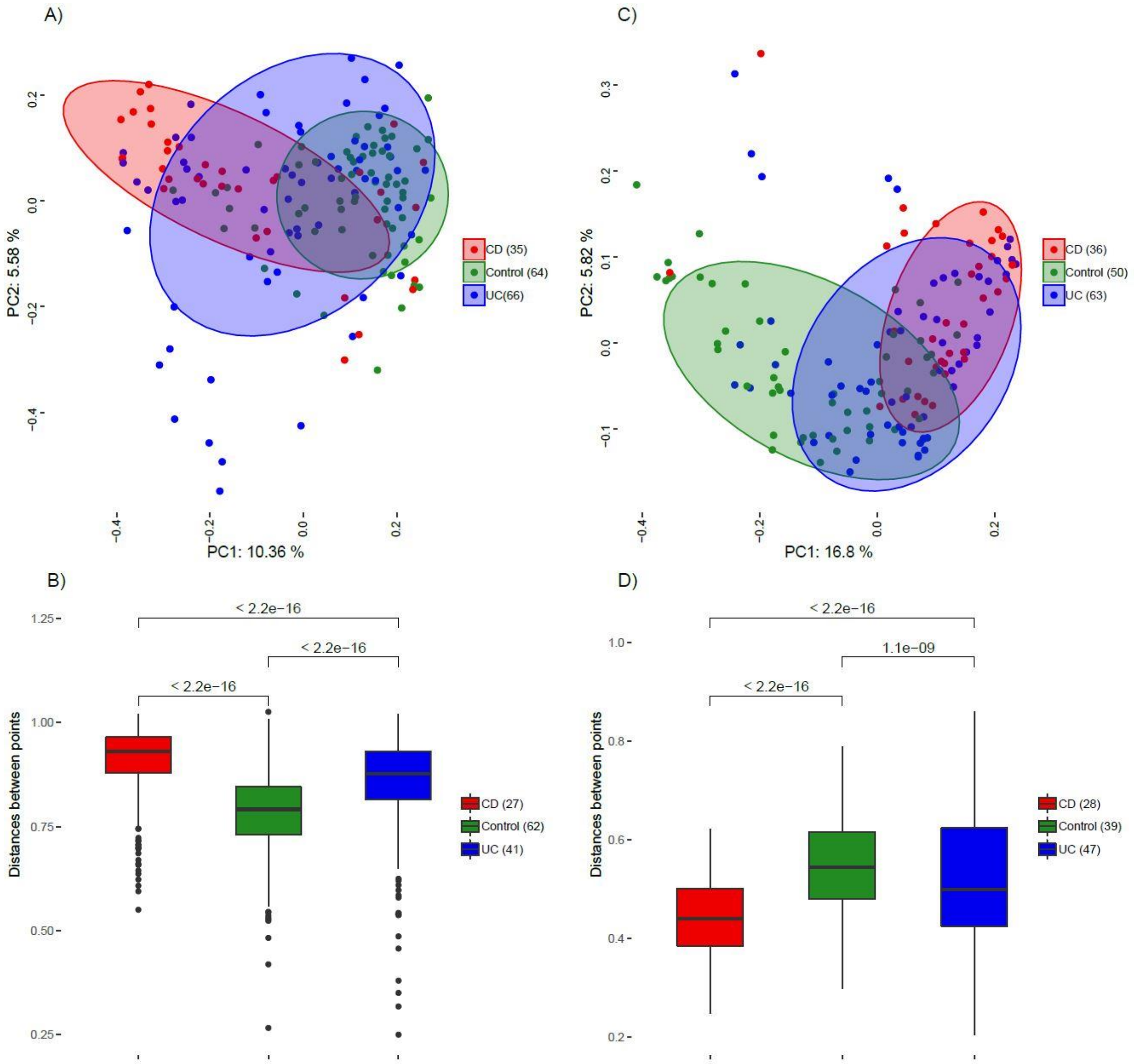


Figure. 2: E) volcano plots of differential abundance results between controls and CD (DESeq2), and F) control and UC (DESeq2). All points above the red dotted line are significant.23,

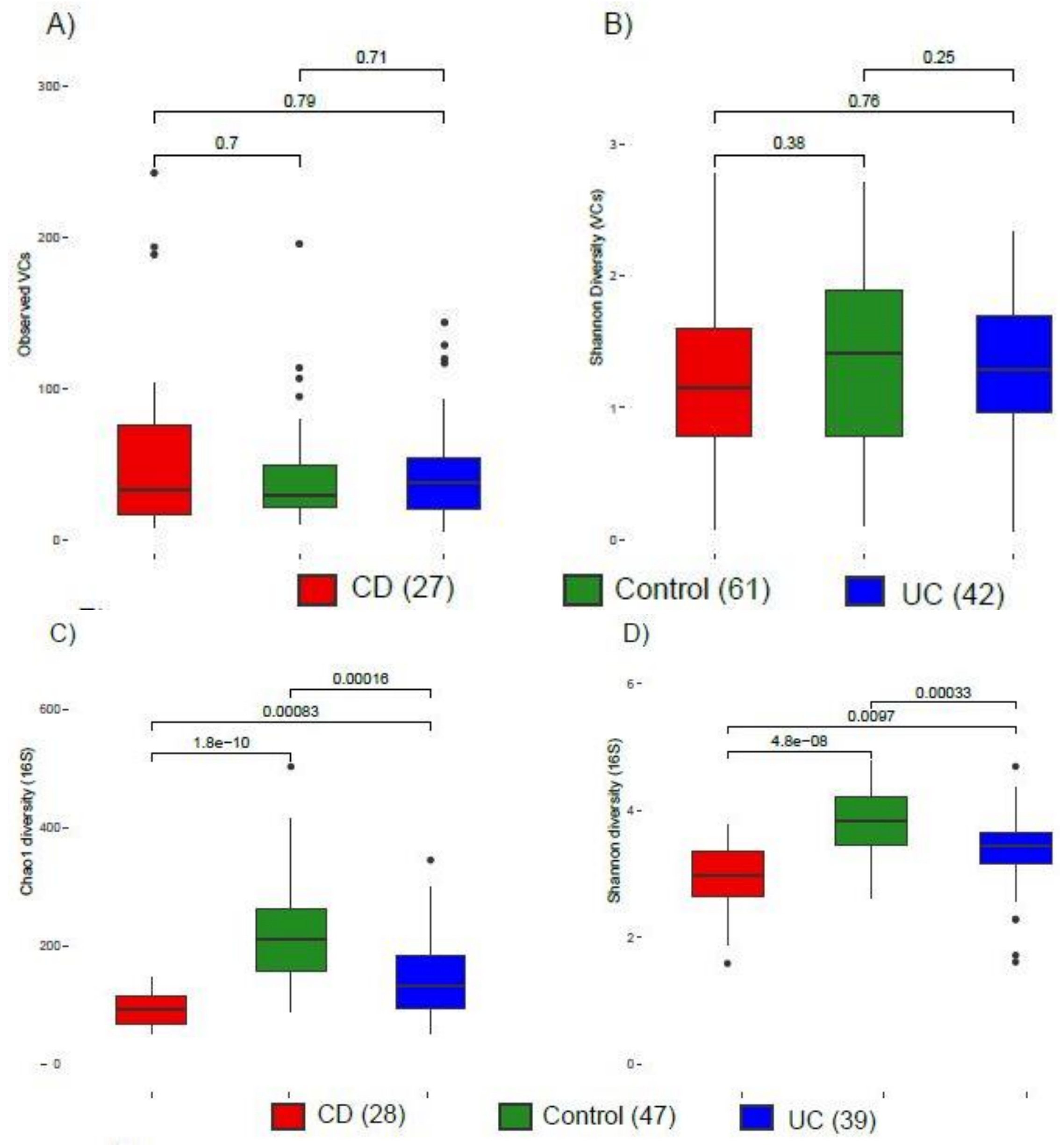


S. Figure. 1 (related to Figure2, Figure3): A) VC PCoA using Spearman distances comparing the 3 cohorts CD, UC and controls. B) Distances between points in each cohort for the VC spearman PCoA (Wilcoxon test). C) 16S PCoA using unweighted UniFrac distances comparing the 3 cohorts. D) Distances between points in each cohort for the 16S unweighted UniFrac PCoA (Wilcoxon test).

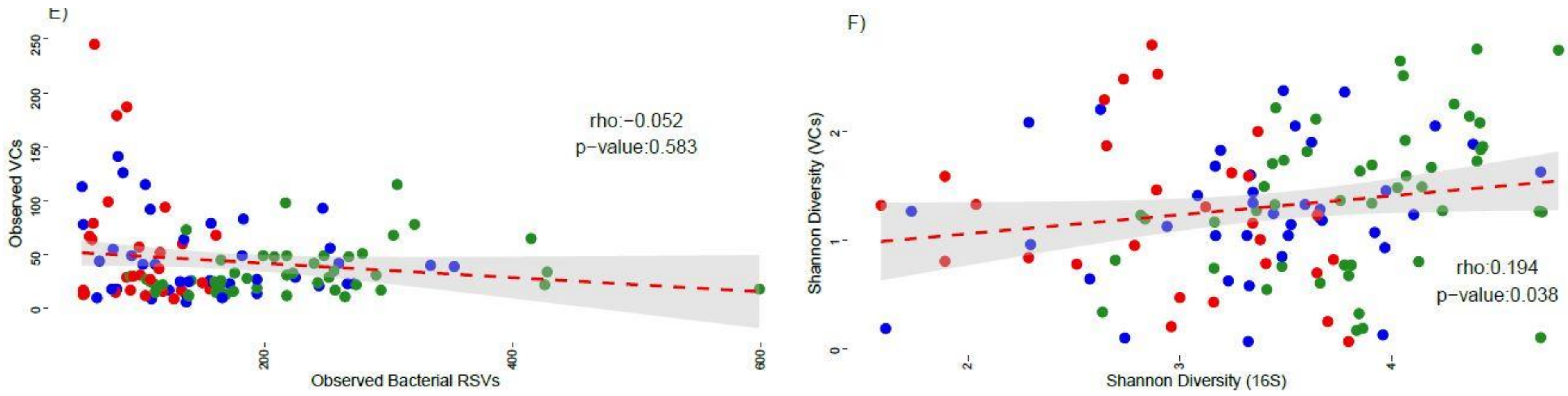
Core VC's were not found across UC subjects and just one VC (vc32 unclassified, CRISPR hits to *Bacteroides dorei*) was found across > 50% of CD subjects.

Significant differences had been observed in the richness of both *Caudovirales* and the virome overall between health and disease in the original analysis (Norman et al., 2015). Contrary to those previous findings, there were no significant differences in virome richness across the cohorts or disease states when VC count tables were used (Wilcoxon test, CD flare vs remission, p = 0.31, UC flare vs remission p = 0.96, CD vs healthy p = 0.12, UC vs healthy p = 0.83, Figure. 2B). The Shannon diversity metric also did not yield a significant difference (Wilcoxon test, CD vs. healthy p = 0.38, UC vs healthy p = 0.25, UC vs. CD p = 0.76, Figure. S2A-2B). When only VCs classified as the order *Caudovirales* were considered (Figure. 2C), a significant increase in richness was observed in CD versus healthy only (Wilcoxon test, p = 0.024). This suggests that changes in the composition of identifiable fraction of the virome may not reflect the virome as a whole. Furthermore, although *Anelloviridae* were detected in our reanalysis of this dataset, significant differences in abundance were not observed across control CD or UC cohorts, contradictory to previous findings (Wilcoxon test, p > 0.05).

Differential abundance analysis identified 37 VCs which were increased in CD relative to controls and 34 increased in UC relative to controls. Importantly, of these VCs increased in disease, over 80% appeared to be temperate (30 of the 37 VCs increased in CD and 28 of the 34 VCs increased in UC). Furthermore, temperate VCs made up just 32% and 24% of VCs increased in controls relative CD and UC, respectively. Further investigation of temperate VC abundance in each cohort indicated that temperate VCs recruited significantly more reads from CD subjects than healthy controls (Wilcoxon test, p = 0.012, Figure. 2D). The temperate/virulent switch was also reflected in the taxonomic classification of VCs which were most differentially abundant. VCs that were increased in healthy cohorts were classified as the predominantly lytic *Microviridae* (two) and crAss (one) (Figure. 2E, Figure. 2F). Similarly, VCs increased in disease were classified as *Siphoviridae* (nine in CD, eight in UC) and *Myoviridae* (one in CD, two in UC), which feature a number of known temperate species. These findings also support the increased *Caudovirales* richness observed in CD.



S. Figure. 2 (related to Figure2, Figure3): α-diversity of patients with IBD versus healthy controls. A) Observed VCs, B) Shannon diversity of VCs, C) Chao1 diversity of 16S counts, D) Shannon diversity of 16S counts (Wilcoxon test),



S. Figure. 2 (related to Figure2, Figure3): E) Spearman correlations between observed VC counts and observed bacterial species counts and F) Shannon diversity of VCs and 16S counts.

Furthermore, of the 17 *Siphoviridae* VCs increased in IBD relative to healthy (nine in CD, eight in UC), 15 were classified as temperate and had CRISPR hits to *Firmicutes.* These observations correspond with the reduced *Firmicutes* abundance observed in IBD (Frank et al., 2007) and further support evidence that increased temperate phage abundance is linked to disease. Induction of *Firmicutes* prophage in the IBD virome would explain the observed reduction in host abundance and increased temperate *Firmicutes* phage virions. Many of the most differentially abundant clusters were taxonomically unassigned and represent viral dark matter (49 VCs increased in control vs. CD, 25 VCs increased in control vs. CD, Tables S1, S2).

The bacteriome also differs between patients with IBD and controls. Bacterial β-diversity assessed through 16S rRNA gene fragment sequencing showed CD (relapse/remission) samples grouping furthest from controls (PERMANOVA, p = 0.0065/0.0332) followed by UC (relapse/remission) (PERMANOVA, p = 0.018/0.001) (Figure. 3A), which was reflected in the virome composition. Interestingly, and in contrast to the virome, the largest degree of variation amongst samples was observed in the control cohort. Furthermore, the CD cohort exhibited the smallest distances between points (Figure. S1C/D).

Decreased α-diversity was observed in the IBD cohorts versus healthy controls with the largest differences observed in CD flare (Wilcoxon test, p = 0.012) and remission (Wilcoxon test, p = 0.018) along with UC Flare (Wilcoxon test, p = 0.051) (Figure. 3B). Due to the small sample sizes this analysis was also repeated without the division of disease status and using various metrics (Figure. S2C-D). For both Chao1 diversity (Wilcoxon test, CD vs healthy p = 1.8e−10, UC vs healthy p = 1.6e-4) and Shannon diversity (Wilcoxon test, CD vs healthy p = 4.8e-10, UC vs healthy p = 3.3e-4), the healthy cohort was significantly higher than both IBD cohorts, while UC was also significantly increased when compared to CD (Wilcoxon test, Chao1 CD vs UC, p = 8.3e-4, Shannon CD vs UC, p = 9.7e-3).

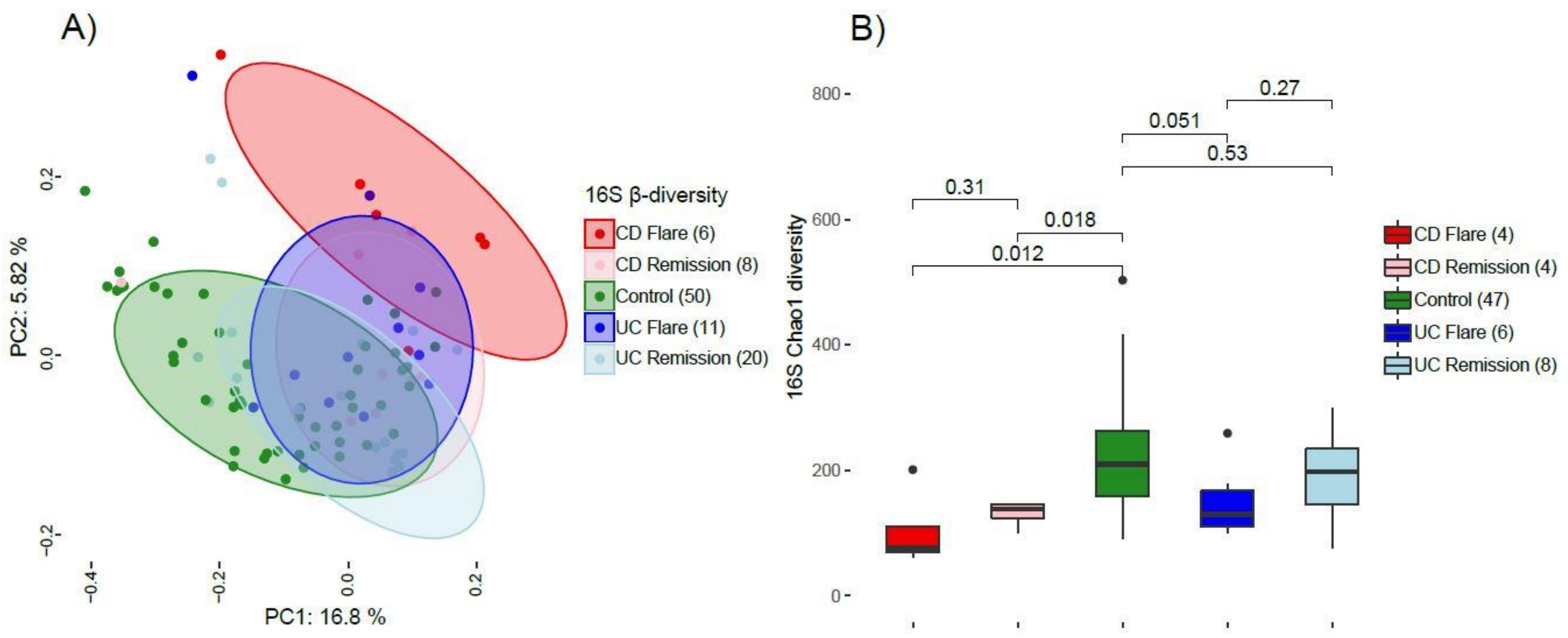


Figure. 3: Bacterial compositional comparison of the Norman et al. IBD cohorts and controls. A) PCoA using unweighted UniFrac distances, B) α-diversity (Chao1 diversity, sample sizes are reduced relative to A due to the retention of only 1 sample per subject)

A large number of taxa were found to be differentially abundant between control and CD (Figure. 3C) and control versus UC (Figure. 3D). A total of 113 RSVs (Ribosomal Sequence Variants) were decreased in CD versus controls while only 17 were increased. Similarly, 69 were increased in control vs UC and only 21 significantly increased in UC (Tables S3, S4). Many of the taxa increased in controls versus both IBD cohorts (such as the genus *Faecalibacterium*) were in accordance with previous reports (Gevers et al., 2014;Machiels et al., 2014;García-López et al., 2015;Pascal et al., 2017;Lopez-Siles et al., 2018). The most differentially abundant RSVs increased in IBD (such as *Fusobacterium* and *Veillonella* in CD, or *Clostridium sensu stricto* and Lachnospiraceae family in UC) were also in accordance with the literature (Willing et al., 2010;Joossens et al., 2011;Strauss et al., 2011;Gevers et al., 2014;Pascal et al., 2017).

Correlations between PCoA and abundance counts reveal key drivers of gut microbiome composition

The drivers of significant shifts in β-diversity were assessed through correlations between principal coordinates and the relative abundances of VCs (for the virome) and RSVs (for the bacteriome). There were 25 VCs significantly correlated to PC-axes 1 and/or 2 (Spearman, $p < 0.05$, Figure 4A, Table S5). Dependent upon the correlation coefficient, the associations could further be broken down into four quadrants and largely supported differential abundance data. In quadrant 1 (top left), towards subjects with IBD, there were 18 significant correlations, which comprised of *Siphoviridae* and *Myoviridae* VCs, as well as some heterogeneous and unclassified VCs. In quadrant 3 (bottom left), one *Myoviridae* and 1 unclassified VC were significantly correlated towards subjects with IBD. VCs classed as *Microviridae*, crAss-like phages and two unclassified VCs were significantly correlated towards the healthy controls (quadrant 4, bottom right). This provides further evidence that a shift from a lytic core of crAss-like phage and *Microviridae* to one of primarily temperate phage may be associated with IBD.

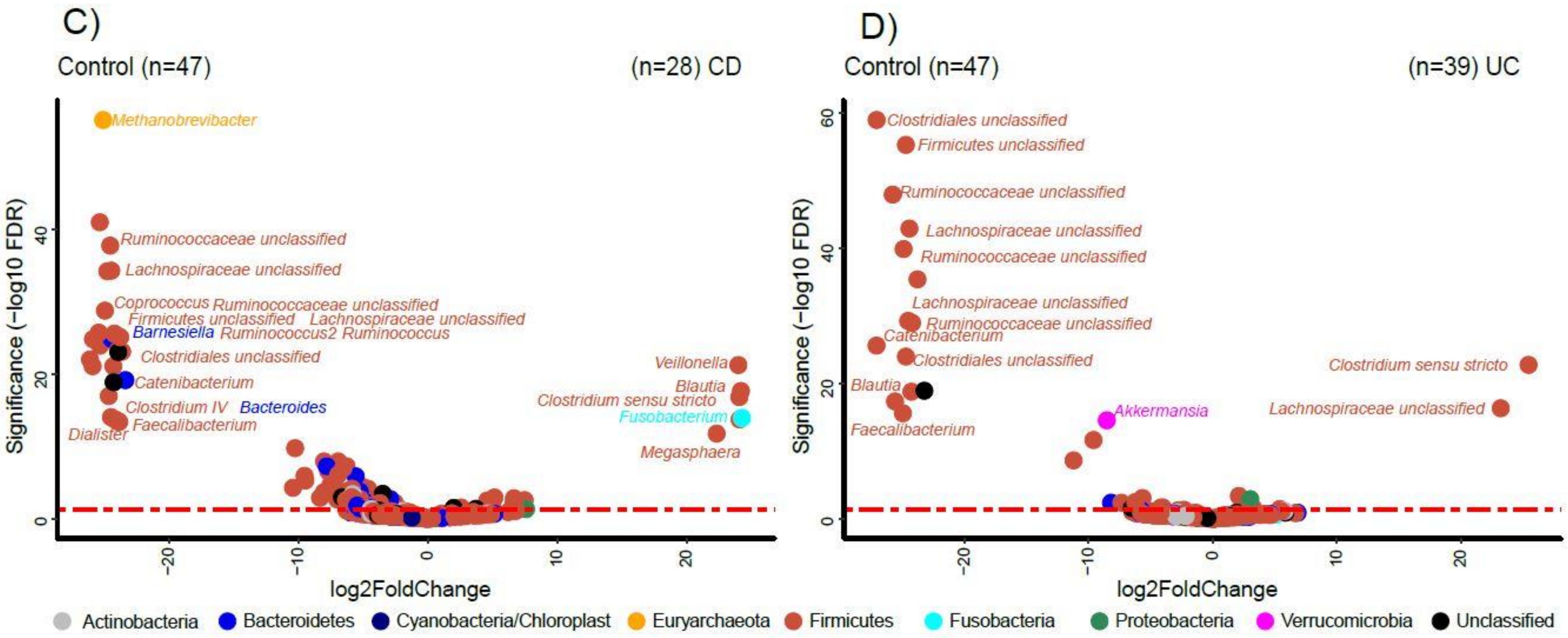


Figure. 3: Bacterial compositional comparison of the Norman et al. IBD cohorts and controls.C) differential abundance results between controls and CD, and D) control and UC. All points above the red dotted line are significant.

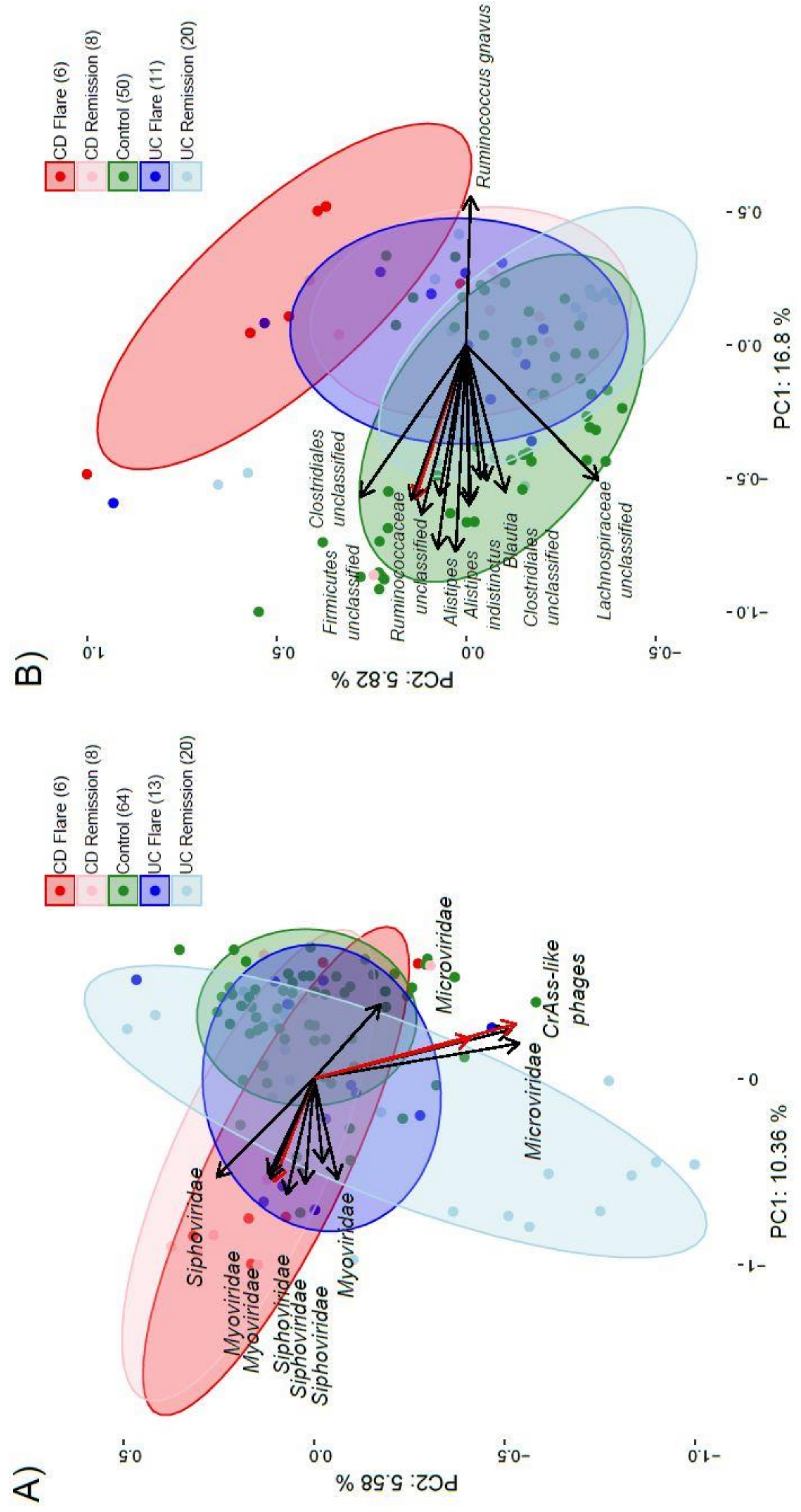

A)
PC2: 5.58 %
PC1: 10.36 %
0.5
0.0
−0.5
−1.0
-1
0
CD Flare (6)
CD Remission (8)
Control (64)
UC Flare (13)
UC Remission (20)
Siphoviridae
Myoviridae
Myoviridae
Siphoviridae
Siphoviridae
Siphoviridae
Myoviridae
Microviridae
CrAss-like phages
Microviridae
B)
PC2: 5.82 %
PC1: 16.8 %
1.0
0.5
0.0
−0.5
−1.0
−0.5
0.0
0.5
CD Flare (6)
CD Remission (8)
Control (50)
UC Flare (11)
UC Remission (20)
Ruminococcus gnavus
Clostridiales unclassified
Firmicutes unclassified
Ruminococcaceae unclassified
Alistipes
Alistipes indistinctus
Blautia
Clostridiales unclassified
Lachnospiraceae unclassified

Figure. 4: (overleaf) Drivers of PCoA separation (Norman et al. data set) for A) the virome (spearman distances) and B) 16S unweighted UniFrac. VC and RSV abundances were correlated, using Spearman correlations, with PC-axes 1 and 2. Only significant spearman correlations with a rho of greater than 0.35 or -0.35 were graphed for the virome or ± .5 for the 16S (or a maximum of the top 6 for each quadrant). Red arrows indicate unclassified VCs/RSVs. The length of the arrow represents the degree of correlation to the PC-axes.

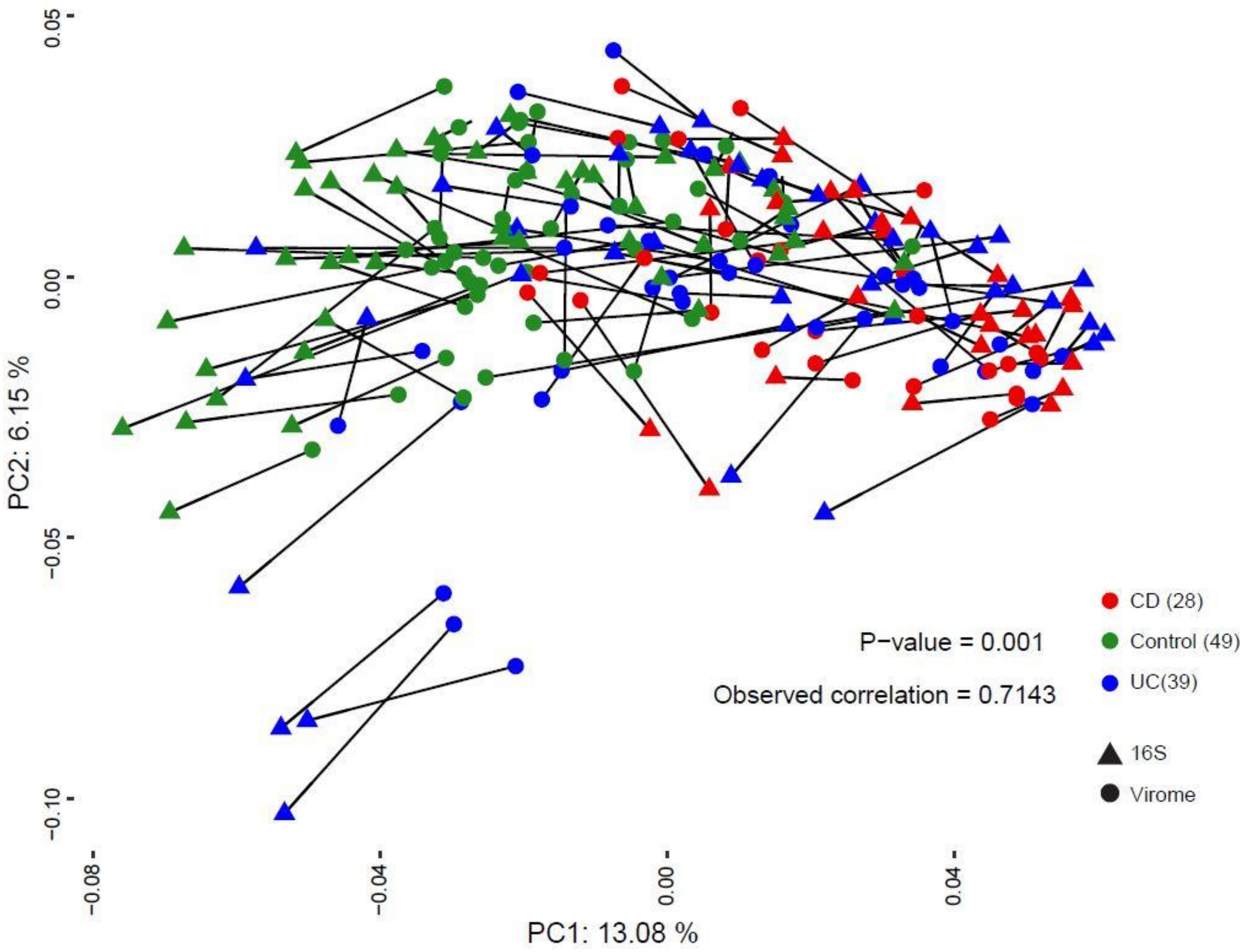
There were 76 bacterial RSVs significantly correlated towards controls (quadrant 1) (Figure. 4B, Table S5). The highest correlation coefficient corresponded to RSVs assigned to the phylum *Firmicutes*, family *Ruminococcaceae* or genus *Alistipes*. Quadrant 3 correlations, also towards controls, contained 46 RSVs including *Alistipes indistinctus* and the order *Clostridiales*. For Quadrant 4, towards IBD subjects, four RSVs were significantly correlated including *Ruminococcus gnavus* and *Flavonifractor plautii*.

Procrustes analysis revealed significant associations between bacterial and viral community composition (procuste.randtest, correlation coefificient of 0.714 p = 0.001, Fig.S3). However, overall VC α-diversity did not significantly correlate with observed bacterial richness (Spearman, p = 0.58, Figure. S2E), although there was significant weak correlation with bacterial Shannon diversity (Spearman, p = 0.038, ρ = 0.194) (Figure S2F). It is possible that this reflects an underlying biological signal which is being masked by temporal variation in phage-host dynamics across the various VCs and RSVs.

Alterations in virome composition are less distinct between UC activity states

Differences in disease states (flare and remission) were investigated using a second cohort of 40 subjects with UC, sampled longitudinally resulting in 113 virome and 109 16S rRNA amplicon samples (bacteriome). β-diversity analysis of virome composition using VCs (Figure. 5A) did not show significant separation between flare and remission (PERMANOVA, p = 0.17), despite one uncharacterised VC correlating to the PC coordinates (vc40). In bacteriome analysis, the shift between flare and remission in β-diversity was significant (PERMANOVA, p = 0.022) and 14 RSVs correlated to PC-axes 1 or 2 (Spearman, p > 0.05, Figure. 5B, Table S6). Those which correlated towards UC remission (quadrant 1) included *Faecalibacterium prausnitzii*, *Dorea longicatena* and *Coprococcus comes*. An RSV classified as *Ruminococcus gnavus* was the only one correlated towards UC flare. This agrees with recent reports that genes involved in oxidative stress responses in *Ruminococcus gnavus* strains may confer facilitate colonisation of the inflamed gut (Hall et al., 2017).

S. Figure. 3 (Related to Figure 4) Procrustes plot of the Virome PCoA using Spearman distances and the 16S PCoA with unweighted UniFrac. Lines connect samples from the same subject. A Monte-Carlo test (procruste.randtest) was performed to test for signicance.



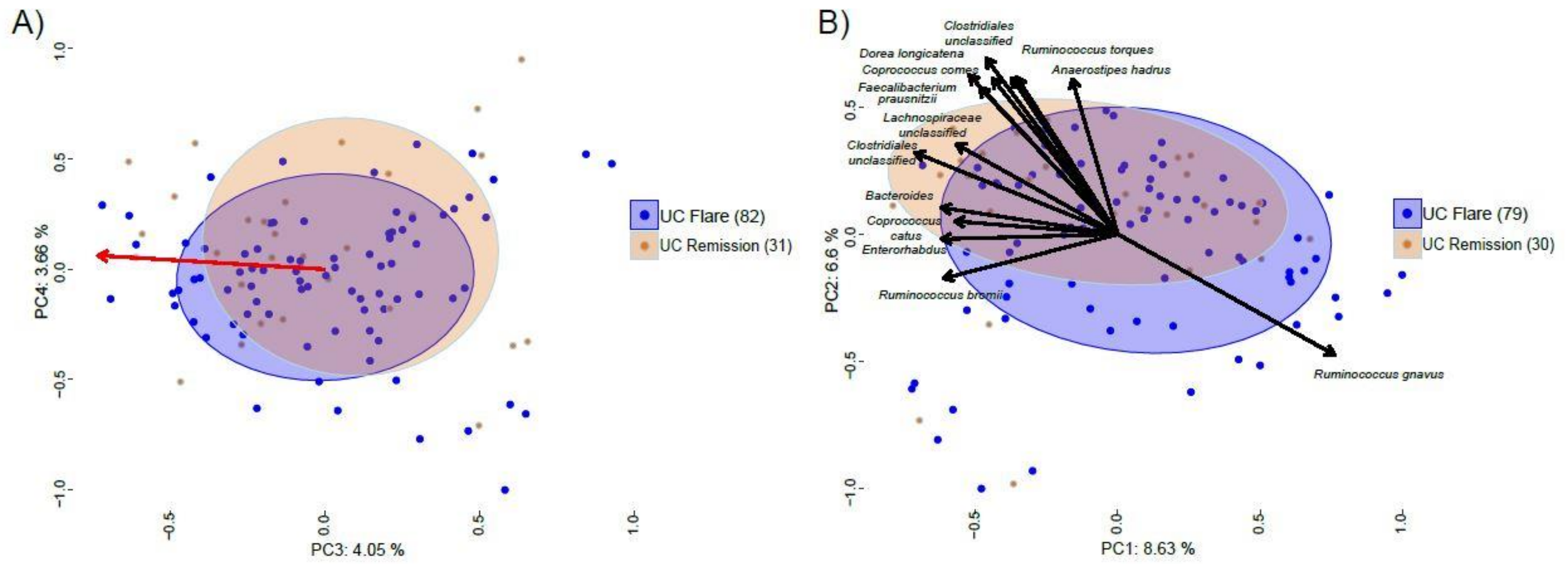


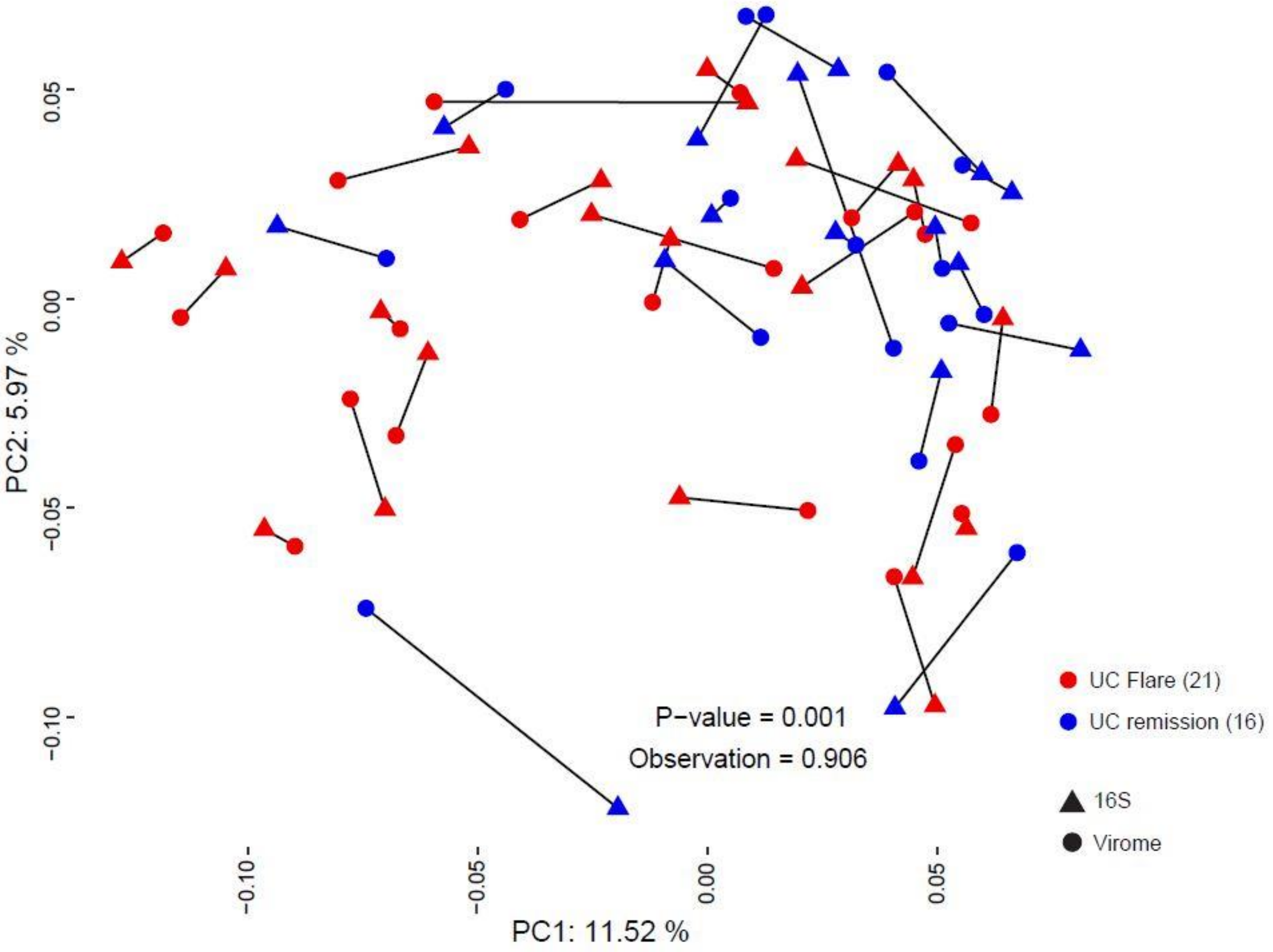
Figure. 5: Investigation of differences in virome and 16S composition between subjects from the longitudinalUC data set in UC flare and UC remission. β-diversity for A) viromes (using Spearman distances) and B) 16S (unweighted UniFrac). VCs and RSV abundance were correlated with PC-axes 1 and 2. Only significant spearman correlations with a rho of greater than ± 0.35 were graphed for the virome or ± .5 (or top 6 for each quadrant) for the 16S. Red arrows indicate unclassified VCs/RSVs. The length of the arrow represents the degree of correlation to the PC-axes.

The virome and bacteriome composition were correlated with each other in Procrustes analysis in agreement with the above results (procuste.randtest, correlation coefficient of 0.906, p = 0.001, Figure. S4).

Although the median α-diversity was higher in the virome for UC flare (Figure. 5C) and UC remission for the bacteriome (Figure. 5D), these values were not significant for either Chao1 or Shannon diversity metrics, in agreement with the results obtained above for the Norman et al. data set (Wilcoxon test, $p > 0.05$). Viral load was estimated by spiking a known concentration of lactococcal Q33 phage and was found to be negatively correlated with viral α-diversity (Spearman, $\rho = -0.415$, $p = 0.009$, Figure. S5A). Viral diversity was also investigated over time and in relation to disease status (Figure. S5B) and although there were fluctuations in the time series, there was no observable trend with disease status and a comparison did not and a comparison did not yield significant differences (Spearman, p = 0.383).

Virome stability across the UC cohort was assessed by identifying VCs present in all 3 time points of subjects, similar to the methods used in (Shkoporov et al., 2019). One key VC (vc39) was present in all three time points of 37% of individuals (13 out of 35) and was the most shared VC across the stable fraction of individuals' viromes. Of the individuals which featured vc39 in all 3 time points, 84% (11 out of 13) also featured RSVs classified as *Lachnospiraceae* in all time points supporting previous CRISPR host prediction. *Lachnospiraceae* was also one of the most stable bacterial families across the cohort, being present in all timepoints of 82% of all individuals. However, there was no significant difference between the numbers of stable VCs per individual (i.e. VCs present in all 3 time points) and remission status (Wilcoxon test, p = 0.738).

S. Figure. 4 (Related to Figure 5) Procrustes plot of the Virome PCoA using Spearman distances and the 16S PCoA with unweighted UniFrac. Lines connect samples from the same subject. A Monte-Carlo test (procruste.randtest) was performed to test for signicance.



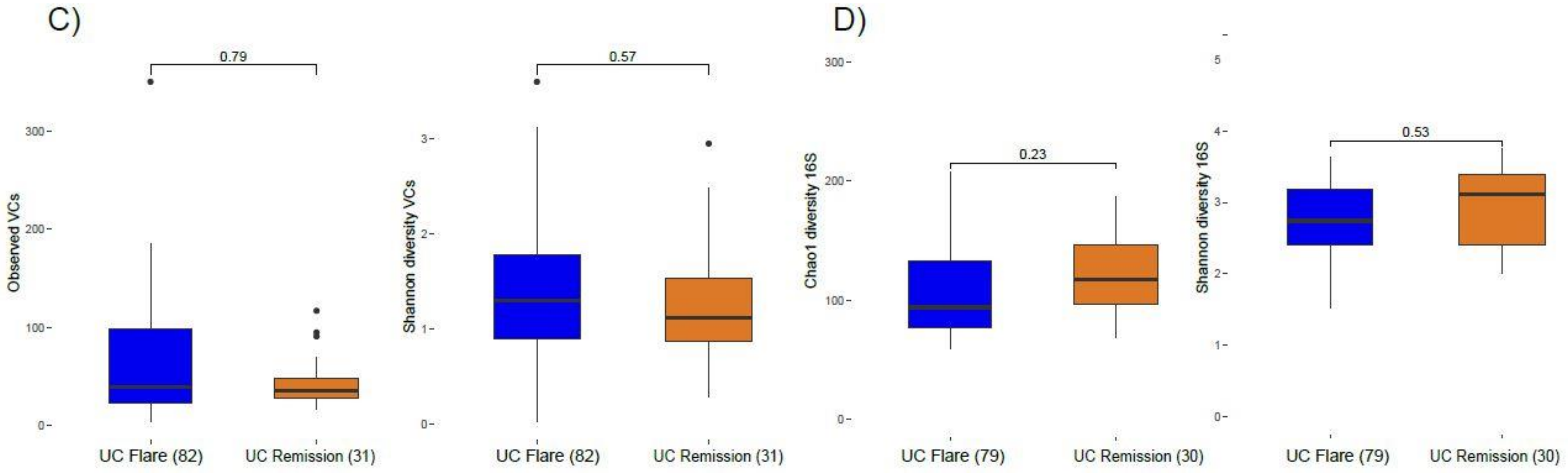
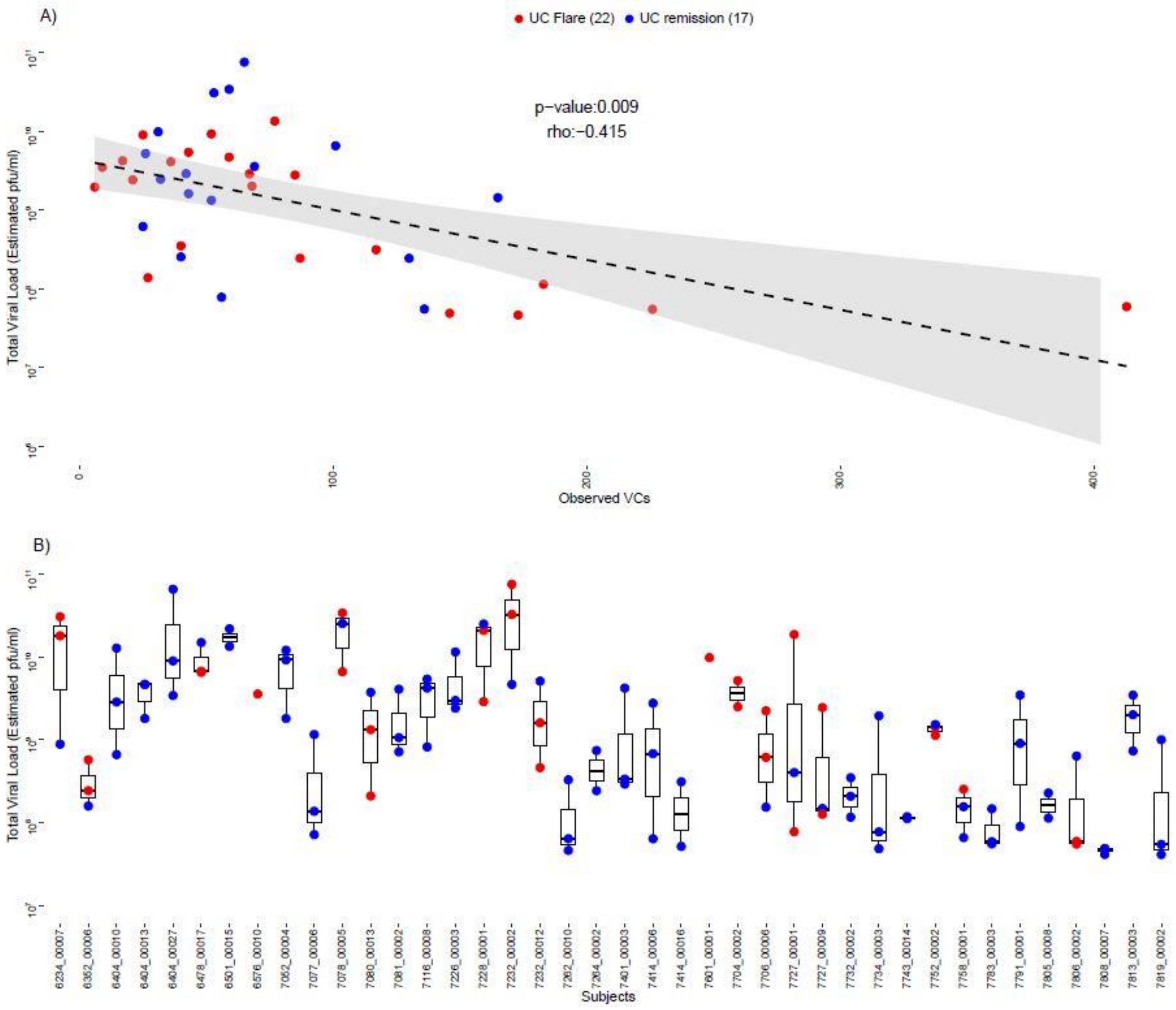


Figure. 5: Investigation of differences in virome and 16S composition between subjects from the longitudinalUC data set in UC flare and UC remission. α-diversity for C) VC (Observed VCs and Shannon) (Wilcoxon test), and D) 16S (Chao1 and Shannon diversity)(Wilcoxon test)



S. Figure. 5 (Related to Figure 5) A) Spearman correlation between estimated viral load and observed VCs. B) Viral load plotted per subject with points coloured by disease status.

Two crAss-like phages were increased in subjects in remission when compared to flare along with two *Siphoviridae*, one *Microviridae* and seven unclassified phage (Figure. 5E, Table S7). Conversely there were 39 VCs increased in flare. These included two *Anelloviridae*, one *Myoviridae*, ten *Siphoviridae* and 24 unclassified. *Bacteroides* and *Dialister* were the only RSVs increased in remission while seven RSVs were increased in flare including *Enterococcus, Prevotella* and *Streptococcus* (Figure. 5F, Table S8). These findings suggest that the changes in virome composition between flare and remission in UC reflect those of the bacteriome and are more subtle than those observed between UC and healthy cohorts.

Virome composition aids the classification between Health and Disease

The ability of the virome and bacteriome composition to differentiate between patients with IBD and healthy controls was tested through machine learning. Sample sizes were increased by combining CD and UC samples to form a composite IBD cohort. The virome alone (Figure. 6A) yielded an accuracy of 0.769 (p = 0.032) and four of the top five contributors (based on gain) (vc39, vc23, vc38 and vc45) were increased in controls versus both IBD states. All five of these clusters were unclassified, highlighting the importance of including viral dark matter in virome analysis pipelines. The top two contributing VCs had CRISPR protospacer alignments to *Lachnospiraceae* (vc39) and *Parabacteroides* (vc23) while the remaining two had alignments to *Bacteroides* (vc32 and vc38). Given the association of *Bacteroides* species with healthy mammalian gut (Delday et al., 2018), these findings further support evidence that the healthy virome closely reflects the healthy bacteriome. The bacteriome alone had a greater predictive power than the virome (accuracy: 0.824, p = 0.008) with an RSV classified as *Ruminococcaceae* contributing the largest gain followed by a *Clostridiales* and *Odoribacter splanchnicus* (Figure. 6B). The virome and the 16S data were combined and the predictive power was again measured (Figure. 6C). The accuracy increased to 0.853 (p = 0.0026) with the virome contributing five of the top 20 most important features. Of these, four had CRISPR protospacers to bacteria including the order *Clostridiales*, the family *Lachnospiraceae*, genera *Pseudoflavonifractor*, *Clostridium* and *Johnsonella* along with *Fusobacterium* and *Bacteroides* (Figure. 7).

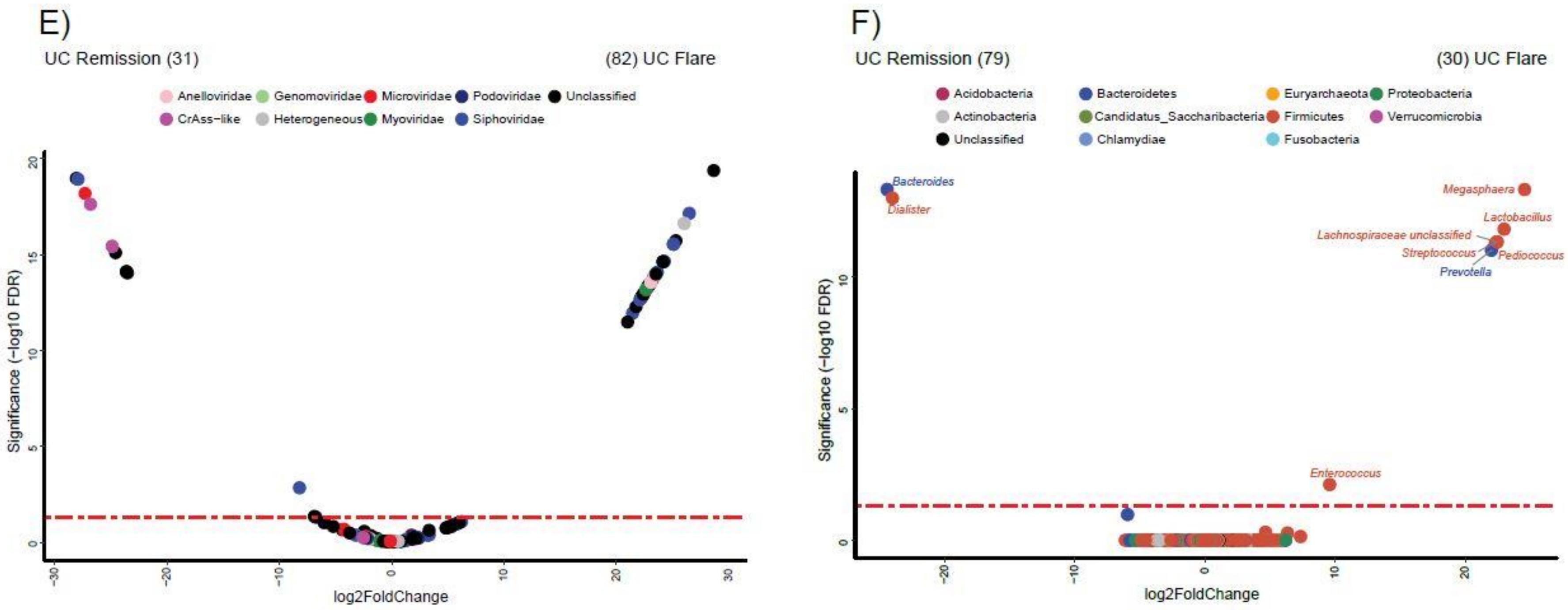


Figure. 5: Investigation of differences in virome and 16S composition between subjects from the longitudinalUC data set in UC flare and UC remission. E) differential abundance results between UC flare and remission for E) VCs (DESeq2) and F) 16S (DESeq2). All points above the red dotted line are significant.

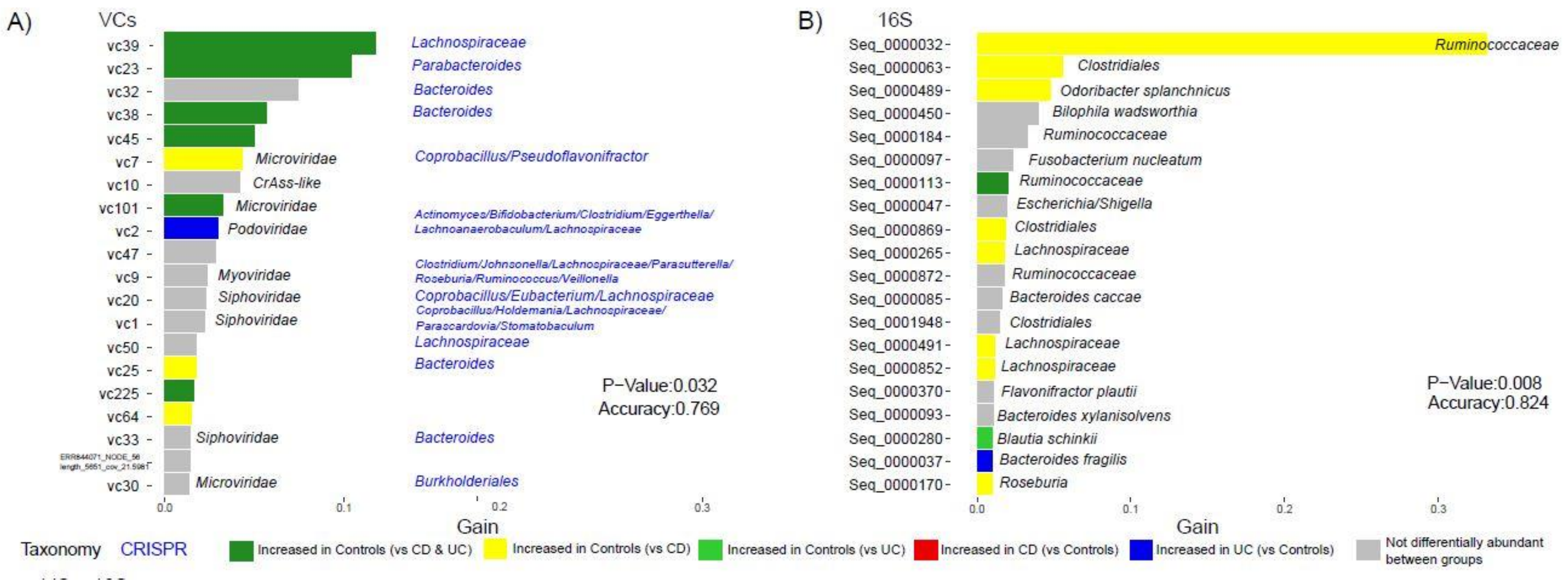


Figure. 6: The classification between healthy controls and patients with IBD, from the Norman et al. data set, using VC and 16S composition. The top 20 importance factors for each model for A) VCs, B) 16S

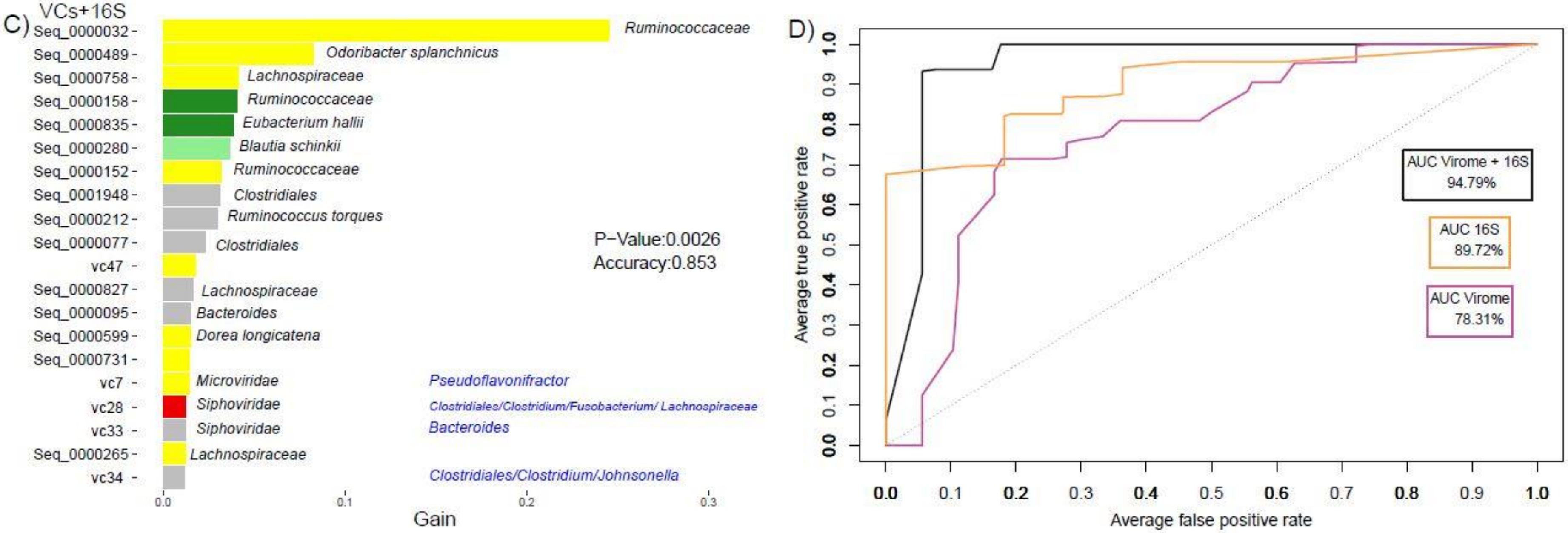
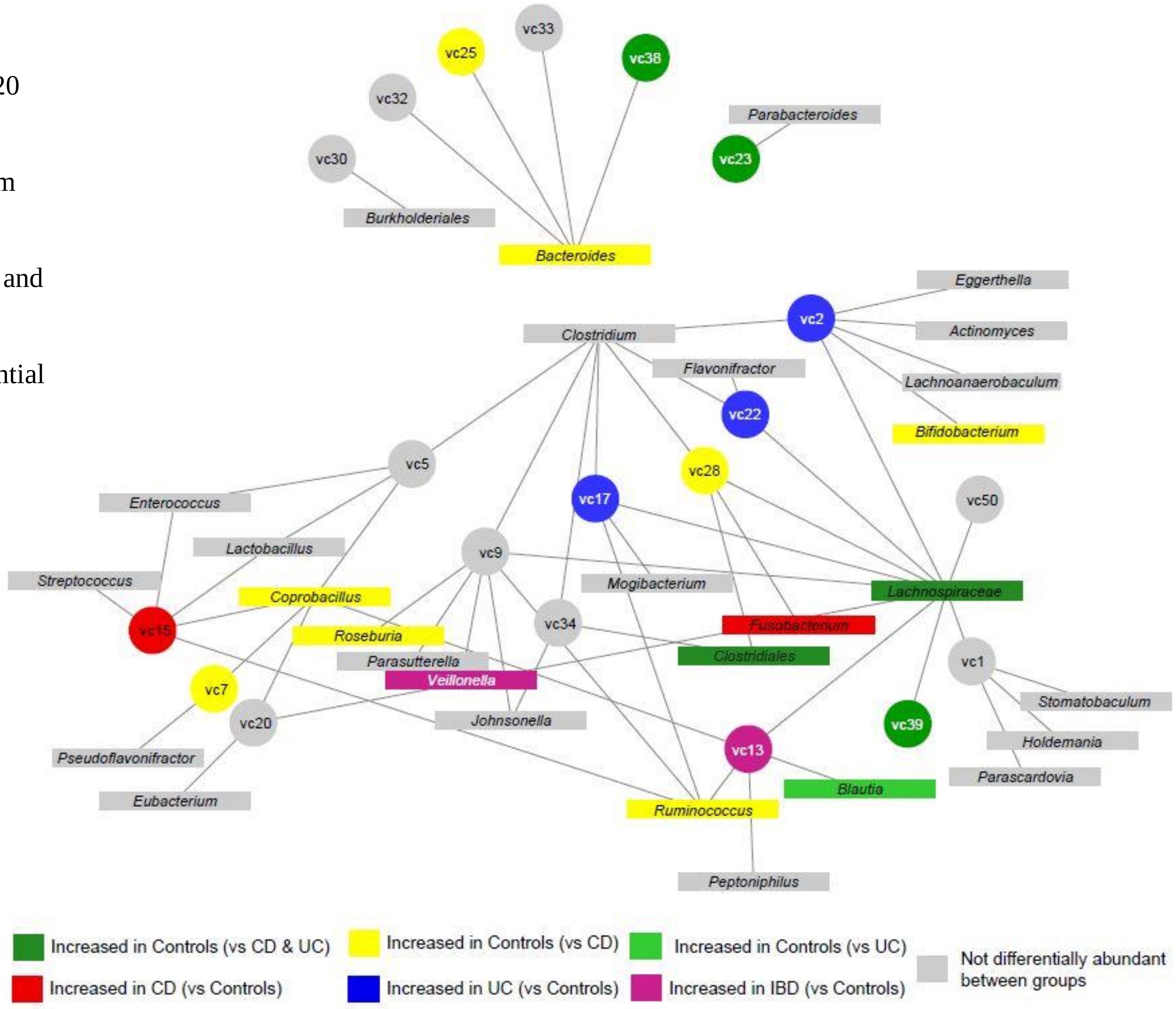


Figure. 6: The classification between healthy controls and patients with IBD, from the Norman et al. data set, using VC and 16S composition. The top 20 importance factors for each model for C) VCs and 16S combined. Colours of bars correspond to differential abundance between groups, black text are the classifications, while blue text is the bacterial annotation to CRISPR protospacers. Gain refers to the relative contribution of the feature to the model. D) The ROC curve analysis for each of the 3 models including the % accuracy.

Figure. 7: Network plot of CRISPR protospacers to the 20 most relevant VCs (key and additional important VCs from machine learning) from the Norman et al. study. Clusters and CRISPR protospacers are coloured according to differential abundance results.

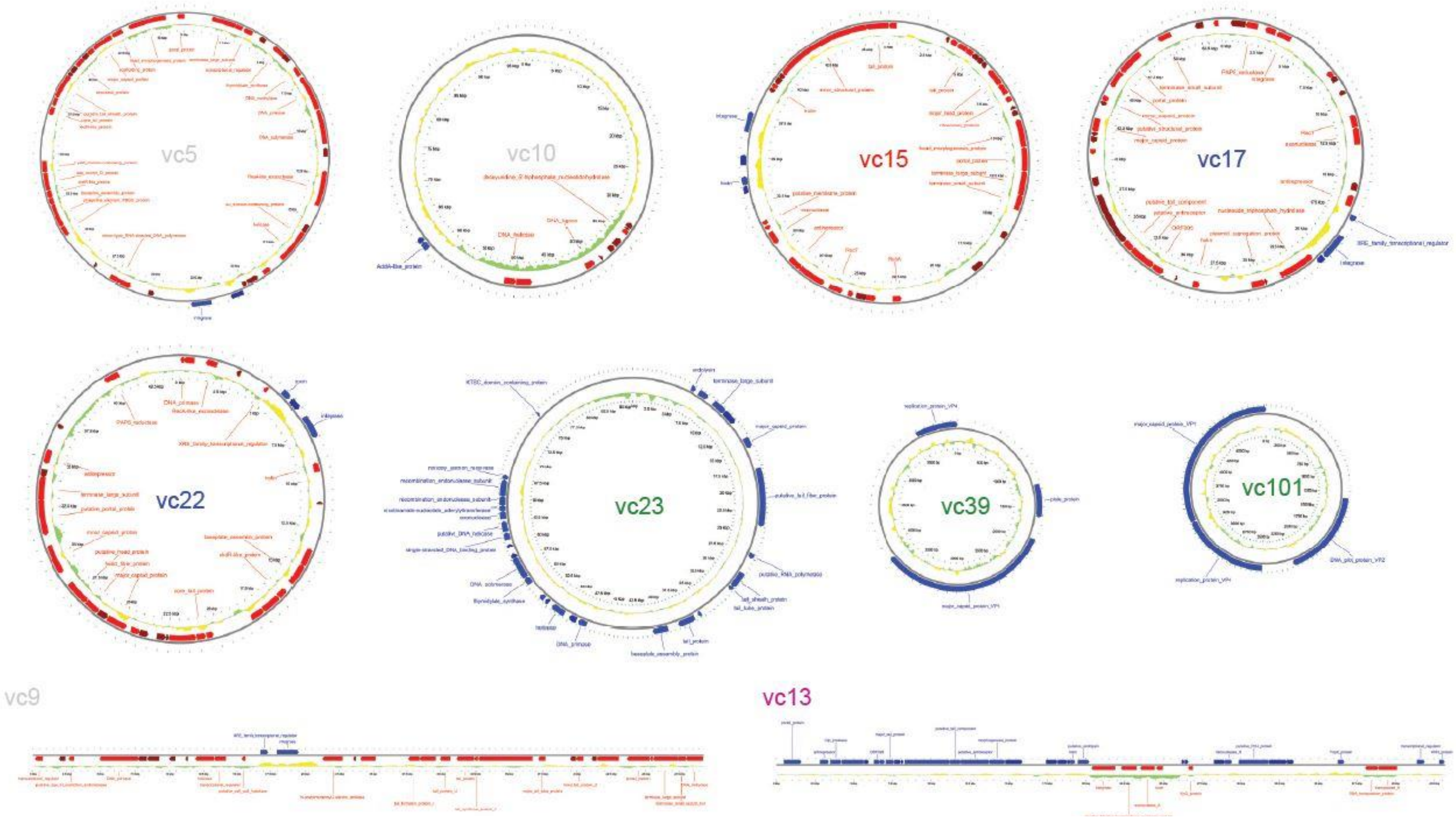


Differences between CD and healthy proved to be the main predictors of disease with 11 VCs/RSVs being decreased in CD alone and one increased when compared to controls.

ROC curve analysis was performed as a second measure of accuracy of each model (Figure. 6D). The AUC (area under the curve) of the virome alone was 78.31%, a decrease compared to bacteriome which yielded an AUC of 89.72%. However, the virome and 16S combined had the largest AUC with 94.79%, successfully predicting all 16 patients with IBD and only misclassifying five controls as UC.

Key VCs revealed by the analysis of IBD viromes

Through various approaches of virome analysis ten key VCs consistently emerged. A key VC was defined as any which was core in one cohort and largely absent from another and/or significantly correlated in the PCoA axes and differentially abundant between the cohorts. vc23, vc39 and vc10 were present in the healthy core and largely absent from the subjects with CD (7, 14 and 26%, respectively) and UC (12, 14 and 40%, respectively). These three VCs were all in the top seven importance factors in the machine learning, while vc39 and vc23 were in the top two. vc23, although unclassified, contained CRISPR protospacers to *Parabacteroides*, while vc39, also unclassified, had hits to undefined *Lachnospiraceae* (Figure. S6). vc10, a crAss-like phage, did not have any CRISPR protospacer alignments. Intriguingly, None of these three key VCs associated with the core virome in healthy individuals featured genes associated with lysogeny, which supports our previous observations and those made in recent studies (Shkoporov et al., 2019). Additionally, hosts predicted for these VCs have been found to be depleted in IBD (i.e. *Lachnospiraceae, Parabacteroides*) (Frank et al., 2007;Kverka et al., 2011) or have been shown to reduce the symptoms of inflammation in the mammalian gut (*Parabacteroides*)(Kverka et al., 2011). This supports the idea that the virome and bacteriome are closely linked and that in lytic populations, phage abundance reflects that of the host (Shkoporov et al., 2018a).



S. Figure. 6 (Related to Figure 6) Images of represetative genomes from the 10 key VCs which drove the separation of IBD and controls. Annotaion carried out using the pVOGs database.

Six of the remaining seven key VCs (vc17, vc13, vc5, vc15, vc9 and vc22) were significantly correlated to the PC-axes and were at significantly increased abundance in CD and/or UC compared to healthy controls. vc13, vc15, vc17, all classified as *Siphoviridae*, had CRISPR protospacer hits to a number of genera of the phylum *Firmicutes*, including *Blautia*, *Coprobacillus*, *Peptoniphilus*, *Ruminococcus*, *Enterococcus*, *Lactobacillus*, *Streptococcus* and *Clostridium* (Figure. S6). vc5, vc9, vc22, classified as *Myoviridae*, contained CRISPR protospacers to *Firmicutes* genera *Clostridium*, *Coprobacillus*, *Enterococcus*, *Lactobacillus*, *Johnsonella*, *Roseburia*, *Ruminococcus*, *Veillonella* and *Flavonifractor* along with the Proteobacteria *Parasutterella* (Figure. S6). All six of these key VCs featured genes associated with lysogeny which provides compelling evidence that IBD is associated with a shift from predominantly lytic virome to that dominated by temperate phage. Furthermore, these VCs also had predicted hosts which are found to be elevated in IBD (*Enterococcus*, *Ruminococcus*, *Streptococcus*, *Veilonella*, *Parasutterella*) (Ricanek et al., 2012;Gevers et al., 2014;Zhou et al., 2016;Hall et al., 2017;Pascal et al., 2017) , or are thought to play an important immunomodulatory role in the gut, by producing short-chain fatty acids (*Roseburia*, *Blautia*) (Zhang et al., 2012;Li et al., 2018;Qing et al., 2019). The remaining key VC was classified as *Microviridae* (vc101) and was increased in control and UC relative CD. This may reflect the subtle changes between CD and UC which were observed throughout the study. It did not have CRISPR protospacer alignments, nor genes associated with lysogeny.

With the exception of two VCs (vc17, vc101), strong, significant correlations were not observed between key VCs and RSVs in the Norman et al. data set (Spearman, p > 0.05). These findings contradict those of the Procrustes analysis, and further suggest that temporal fluctuations in the abundance of phage-host populations across individuals may mask underlying signals. Despite this, vc17 exhibited a strong positive correlation with an RSV classified as *Gammaproteobacteria* (Spearman, ρ = 0.51, p = 9.7e-9), and vc101 exhibited a strong positive correlation with three RSVs classified as *Firmicutes* (ρ ≈ 0.52-0.6, p < 1e-08), one as *Bacterioidetes* (ρ = 0.52, p = 2.6e-9), and one *Fusobacteria* (ρ = 0.51, p = 5.6e-9). In the longitudinal UC cohort, again the only strong, significant correlation was observed between vc101 and an RSV classified as *Bacteroidetes* (ρ = 0.61, p = 1.3e-12). It is possible that a uniform signal across individuals is masked by inverse positive and negative relationships of phage host pairs in different replication cycles. Although little is known about phage-host dynamics in the gut, it has been suggested that factors such as host population density (Silpe and Bassler, 2019), phase-variation (Turkington et al., 2019), host switching (De Sordi et al., 2017), and the biogeography of the gut itself (Maura et al., 2012;Zhao et al., 2019) significantly influence these dynamics and may therefore mask patterns in the abundance of specific phage-host pairs.

Discussion

Here we performed whole-virome analysis on two IBD virome data sets; a keystone published data set (Norman et al., 2015) and an in-house UC cohort which investigated subjects longitudinally through periods of flare and remission. We apply an analysis approach to the gut virome, which interrogates both known and unknown sequences and provides insights into viral dark matter in human health and disease. By applying shared genes network approach (Bin Jang et al., 2019) and replacing individual VLS with viral clusters, it is possible to reveal compositional patterns in the virome across individuals that had been previously masked by high inter-individual variation.

This comprehensive virome analysis revealed a core virome in healthy subjects that did not exhibit identifiable temperate features. This supports recent reports of a stable, predominantly lytic core virome observed in healthy individuals (Shkoporov et al., 2019). The healthy core virome was found to be absent in IBD and appeared to be replaced by an individual-specific shift towards induced temperate phage. This suggests that a stable core of predominantly virulent or pseudolysogenic viruses is associated with the maintenance of a healthy human gut. These observations also suggest that environmental stresses associated with the inflamed gut, such as reactive oxygen species (Rigottier-Gois, 2013), cause a reservoir of integrated prophage to enter the lytic cycle. This would correspond with the reduction in bacterial α-diversity and counts (Vandeputte et al., 2017) and an increase in the relative abundance of *Caudovirales* associated with IBD. Additionally, large scale prophage induction and cell lysis would lead to increased levels of pro-inflammatory bacterial debris available to interact with local innate immune receptors and mucosa-associated lymphoid tissue. These observations are also supported by reports of integrated prophage directly using host physiology and population density to influence the switch from lysogenic to lytic replication (Casjens and Hendrix, 2015;Silpe and Bassler, 2019). In this way the temperate virome can respond to environmental stress and changes in the host population, such as those observed in IBD. Given these observations, we theorise that a switch from lysogenic to lytic replication cycles is linked to an increase in the relative abundance of temperate VLPs in disease. Furthermore, we propose that conditions in the inflamed gut do not support the maintenance of a stable, predominantly virulent core virome.

We identified ten key VCs that were consistently associated with healthy or IBD cohorts which also provide compelling evidence that a core gut virome of predominantly virulent phages is closely associated with human health. None of the three key VCs associated with the core virome in healthy individuals featured genes associated with lysogeny. Additionally, all but one (vc101) of the seven key VCs associated with IBD did these feature these genes. Furthermore, nearly all of the predicted hosts for these key VCs play have been associated with IBD. Key VCs found in the healthy core had predicted hosts found to be significantly reduced in IBD, or are thought to attenuate the symptoms of inflammation. Similarly, many key VCs associated with disease also had predicted hosts which are significantly increased in IBD (*Enterococcus*, *Ruminococcus*, *Streptococcus*, *Veilonella*, *Parasutterella*) (Ricanek et al., 2012;Gevers et al., 2014;Zhou et al., 2016;Hall et al., 2017;Pascal et al., 2017), or are believed to interact with the mammalian immune system (*Roseburia*, *Blautia*) (Zhang et al., 2012;Li et al., 2018;Qing et al., 2019).

The reduced abundance or absence of a core virome of IBD subjects compared to healthy controls was previously described (Manrique et al., 2016) using the Norman et al. data set. A core of 23 bacteriophages was reported, contradictory to what we observed pre-clustering. We suspect that is due to the lenient criteria used by the authors to define the presence of a contig – a view shared by a recent publication (Gregory et al., 2019). In this study, we employed a breadth of coverage filter (Roux et al., 2017) of >75% on the basis that if a VLS was truly present, it would recruit reads across the whole genome, thus removing spurious read alignments to repeats and regions conserved across broadly different viral genomes.

Whole-virome analysis did not reveal differences between the viral richness of the subjects with IBD and healthy controls, a finding contradictory to previous analysis of this data set (Norman et al., 2015). This implies a virome dominated by temperate phage in disease replaces rather than adds to, the shared lytic core in healthy controls. This in turn could reflect an absence of hosts for the lytic core virome, supported by a reduction in bacterial α-diversity in disease. It is also possible that rather than being replaced, the lytic core has fallen below the detection threshold of the analysis method, overshadowed by an induced lysogenic fraction in disease. This issue could be exaggerated by multiple displacement amplification (MDA) as it known to skew the abundance of dominant sequences in metagenomes (Parras-Molto et al., 2018). It is also important to note that an absolute decline of virulent phage following the reduction of their host abundance in disease would in turn lead to a relative increase in temperate phage. These findings highlight the need for future virome studies to quantify total viral load as done previously (Shkoporov et al., 2018b) or to use MDA-free library preparation methods to definitively conclude whether a healthy core virome is replaced or supplemented by induced prophage in disease.

Replicating the work of Norman and colleagues (Norman et al., 2015), we assessed the richness of VCs classified as *Caudovirales*. In agreement with previous observations *Caudovirales* diversity was increased in the CD cohort. However, we did not observe significant changes in the whole-virome diversity between UC and healthy controls. It is possible that changes in the identifiable subsets of the virome do not reflect the virome as a whole. There was no alteration in viral load between disease status in UC, however there was an inverse relationship between viral load and diversity. This suggests that higher viral loads are a result of a dominance of a particular phage or phages rather than the detection of new members.

It has been previously reported that the human gut virome exhibits high levels of inter-individual variation (Reyes et al., 2010;Minot et al., 2011;Shkoporov et al., 2019) which is exacerbated by the need to analyse the virome at an assembly level resulting in analysis being carried out at a strain level. Unlike the bacteriome analysis, which is typically performed at higher taxonomic ranks such as family and genus, viral taxonomy does not have a similar defined structure which makes comparisons of cohorts very difficult. Strain-level resolution hampers cohort comparisons due to a lack of shared sequences across subjects in the data set thereby masking compositional patterns occurring at higher taxonomic ranks across cohorts. Our initial analysis was carried out using VLS (virus-like sequences made non-redundant at 90% identity over 90% of the length), but this level of resolution did not reveal shared signals across cohorts. We overcame this issue by clustering viral genomes based on their protein-coding gene content using vConTACT2 (Bin Jang et al., 2019). This gene network clustering approach revealed shared virome features while retaining relevant biological signals across the cohort (as seen by VCs across subjects), increased the variation explained in the β-diversity and decreased the abundance of unique viral VCs per subject across the data set. Viral clustering also enabled the detection of a core virome in healthy subjects, consisting of eight VCs across >50% of the cohort. This proved to be a key differentiator between health and disease throughout the analysis. Many of these core VCs were differentially abundant in health and disease and were primary drivers of cohort separation and machine learning predictions. In β-diversity analysis, IBD subjects shifted significantly away from healthy controls thus providing evidence of compositional differences in the gut viromes. Drivers of these separations were associated with temperate phage such as clusters of *Myoviridae* and *Siphoviridae* in IBD and clusters of *Microviridae* and crAss-like phages in healthy subjects which are predominantly non-temperate.

Alterations in the bacteriome agreed with previous observations. However, the current study provides evidence that alterations in the whole human gut virome in IBD occur in conjunction with changes in the bacteriome. Although it is not possible from a cross-sectional study to know whether the virome alters the bacteriome or *vice-versa*, the viral/bacterial data sets were shown to be complementary. Although the bacteriome was more accurate in classifying subjects with IBD from controls, the addition of the virome improved upon this classification to over 94% AUC and to over 85% accuracy. We acknowledge that this current model is not designed to be used for diagnosis but does provide evidence that alterations in the gut microbiota are present in both the bacteriome and the virome. Studies building the fundamental understanding of interactions between the virome and bacteriome in disease, as conducted here, are a crucial foundation to the future of virome-based tools to shape the microbiome.

Alterations in both the virome and bacteriome were more severe in CD patients relative to UC, which may reflect the severity of the condition relative to UC. The CD virome was furthest from healthy controls, was the least stable, and exhibited the greatest number of differential abundant VCs and RSVs relative to healthy controls. The CD cohort had the least bacterial variability across the cohort which may also be linked to having the lowest bacterial diversity. Interestingly, CD had a significantly higher diversity of *Caudovirales* and an increased number of reads aligned to temperate VCs when compared to healthy controls. This again supports the idea that shift from lysogenic to lytic replication cycles may drive a change in the bacteriome linked with CD. We did not observe any differences in *Caudovirales* diversity between UC and controls along with UC flare versus UC remission in either of our data sets contradictory to Zuo and colleagues (Zuo et al., 2019). This may be due to the analysis protocols, whole-virome versus database-dependent, or the sample type, faecal vs mucosa. It is likely that bacterial-phage dynamics at the mucosal surface in disease are significantly different to that of faecal samples as previously seen in bacterial profiles (Gevers et al., 2014). Furthermore, we did not detect any giant viruses such as *Mimivirus* or *C. ericina* virus in our UC cohorts as was suggested by Zuo and colleagues and believe their previous detection may have been a result of the analysis pipeline chosen in that study (Sutton et al., 2019a).

Virome compositional changes between UC and control cohorts were more pronounced than changes observed between flare and remission in UC. This finding, in conjunction with the overall comparison between UC, CD and healthy controls, suggests the virome is not only less perturbed between healthy and UC, but also between flare and remission. This may reflect the disease severity of UC relative to CD or suggest that these conditions interact with the host in different ways. Variation in disease location, severity and risk factors such as the potential paradoxical relationship between CD and UC with smoking (Berkowitz et al., 2018), have previously alluded to differences in disease aetiologies. It is possible that virome changes between disease states are more subtle than those between health and IBD. However, as the virome changes in CD relative to healthy were more exaggerated than those seen in UC, it is possible that disease status in CD is reflected more significantly in the virome.

There are a number of future improvements that can be undertaken to expand upon our current findings. Increased sample sizes, particularly for disease state, would increase our ability to detect any potential alterations between flare and remission. Given that diet is a key factor in shaping the microbiota (Singh et al., 2017), inclusion of food frequency questionnaires would be beneficial, as many subjects with IBD undergo significant diet alterations. Furthermore, as many subjects with IBD are on various medications, detailed medical and medication history (Maier et al., 2018) would likely give deeper insight into this data set. Unfortunately, we did not have access to extensive metadata, including household controls, which can assist in statistical analysis and allow the exploration of environmental effects. Certain analysis and figures were limited by missing information at the time of sampling, particularly information relating to disease activity status (i.e. flare or remission). It has also been shown that faecal water content (Bristol stool chart) has been associated with bacterial composition (Vandeputte et al., 2016). Future virome studies would benefit from the inclusion of water content data in samples. Strain-level variation is believed to play an important role in phage-host dynamics, particularly when explaining proliferation of both virulent phage and hosts in an environment (Breitbart et al., 2018;Shkoporov et al., 2019). Consequently, these dynamics may play an important role in the maintenance of the healthy core virome. However, despite allowing for the identification of compositional patterns across individuals, analysis at a VC level also masks these strain-level dynamics. The DNA in these samples was subjected to an MDA amplification step which has been known to skew overall abundance values (Parras-Molto et al., 2018). More modern shotgun DNA library kits remove the need for amplification and should give a more reliable indication of diversity (Roux et al., 2016). It is also possible that we have excluded some valid viral sequences in our efforts to prevent contamination and conversely, despite our best efforts we may have also erroneously included some bacterial contigs. This is ongoing work which will be improved upon when more hallmark viral genes are identified and prediction software become available.

This study uses a whole-virome analysis approach to give detailed insights into the function of the gut virome and its potential role in IBD. We confirm previously reported disease-specific alterations in the IBD virome but, in contrast to previous findings, did not see changes in overall viral alpha diversity. However, we did find evidence to suggest that a predominantly virulent core virome is linked to healthy gut and shift from lysogenic to lytic replication in the temperate phage population may be linked to IBD. It should be noted however, that it is not yet possible to conclude if virome composition reflects or shapes the structure of the bacteriome in the human gut. This whole-virome analysis approach identified compositional changes across the entire human gut virome associated with health and disease. These findings are a significant step towards identifying targets for further wet-lab characterisation and future virome biomarkers. This analysis approach also facilitates the comparison of whole viromes across cohorts in diseases other than IBD and highlights how we can benefit from a fuller understanding of the role of the microbiome in human health.

Acknowledgements

We thank Dr. Nilufer Seth and Dr. Prerak Desai from Janssen Biologics for their assistance and guidance throughout this study.

Author Contributions

AGC, TDSS, AS and FJR performed the bioinformatics and statistical analysis and drafted the manuscript. RH performed 16S read processing. KD and OOR extracted the DNA and performed sequencing library preparations. AC, TS, LD, SEP, PR and CH conceived the study. All authors approved and contributed to the final manuscript.

S. Table 1 (Related to Figure 2): Differential abundance results of VC counts comparing Healthy controls to CD.

https://ars.els-cdn.com/content/image/1-s2.0-S1931312819305335-mmc2.xlsx

S. Table 2 (Related to Figure 2): Differential abundance results of VC counts comparing Healthy controls to UC.

https://ars.els-cdn.com/content/image/1-s2.0-S1931312819305335-mmc3.xlsx

S. Table 3 (Related to Figure 3): Differential abundance results of 16S RVS counts comparing Healthy controls to CD.

https://ars.els-cdn.com/content/image/1-s2.0-S1931312819305335-mmc4.xlsx

S. Table 4 (Related to Figure 3): Differential abundance results of 16S RVS counts comparing Healthy controls to UC.

https://ars.els-cdn.com/content/image/1-s2.0-S1931312819305335-mmc5.xlsx

S. Table 5 (Related to Figure 4): Spearman correlations between VC/RSV counts and PC-axes 1 and 2 from Spearman/ unweighted UniFrac distances.

https://ars.els-cdn.com/content/image/1-s2.0-S1931312819305335-mmc6.xlsx

S. Table 6 (Related to Figure 5): Spearman correlations between VC/RSV counts and PC-axes 1 and 2 from Spearman/ unweighted UniFrac distances.

https://ars.els-cdn.com/content/image/1-s2.0-S1931312819305335-mmc7.xlsx

S. Table 7 (Related to Figure 5): Differential abundance results of VC counts comparing UC Flare to UC Remission.

https://ars.els-cdn.com/content/image/1-s2.0-S1931312819305335-mmc8.xlsx

S. Table 8 (Related to Figure 5): Differential abundance results of RSV counts comparing UC Flare to UC Remission.

https://ars.els-cdn.com/content/image/1-s2.0-S1931312819305335-mmc9.xlsx

Full Article including all Supplemental information is available at

https://ars.els-cdn.com/content/image/1-s2.0-S1931312819305335-mmc10.pdf

Methods

Faecal samples (Longitudinal UC data set)

Faecal samples from the PURSUIT-M (NCT00488631) phase 3 trial were provided by Janssen Biotech. This was a multicentre, placebo-controlled, double-blind, randomized-withdrawal study conducted at 251 centres between September 2007 and October 2011. The institutional review board or ethics committee at each site approved the protocol, and patients provided written informed consent. The study cohort consisted of 40 UC subjects with active disease at baseline, and varying disease states throughout the study. Disease activity was determined by Mayo score, an index based on rectal bleeding, stool frequency, physician’s global assessment and endoscopic findings. The cohort, of which 23 male, had a mean age of 39.55 (±16.60) and mean BMI of 23.74 (±4.46), while 31 were Caucasian and 9 non-Caucasian. Participants who had a partial/total colectomy or an ostomy, signs of latent or active granulomatous infection, or signs/symptoms of malignancy were excluded from this study. All subjects were taking golimumab at baseline and a full medical and medication history of all participants in the PURSUIT-M trial are found at https://clinicaltrials.gov/ct2/show/results/NCT00488631.

Extraction of faecal VLP DNA, library preparation and sequencing (Longitudinal UC data set)

Extraction of faecal VLP DNA from the longitudinal UC data set cohort samples, subsequent library preparation and sequencing were performed as described by Shkoporov et al. (Shkoporov et al., 2018b;Shkoporov et al., 2019) with the following modifications: All samples were spiked with the lactococcal Q33 (Mahony et al., 2013) into the faecal homogenate at 10^6 pfu/ml, which allowed for quantification of the total bacteriophage loads in faecal samples. Shotgun library preparation was carried out using TruSeq Nano DNA HT Library Prep Kit (Illumina) following reverse transcription of the samples with ThermoFisher Scientific SuperScript IV First Strand Synthesis System and multiple displacement amplification (MDA) with the Illustra GenomiPhi V2 kit (GE Healthcare). Libraries were normalised as per the standard manufacturer's protocol. Ready-to-load libraries were sequenced using 2 × 150 bp paired-end chemistry on an Illumina HiSeq 4000 platform (Illumina, San Diego, California) at GATC Biotech AG, Germany.

Extraction of total faecal DNA for 16S rRNA amplicon sequencing (Longitudinal UC data set)

Extraction of total faecal DNA from the longitudinal UC data set cohort samples, subsequent library preparation and sequencing were performed as described by Shkoporov et al., 2018 (Shkoporov et al., 2018b). Ready-to-load libraries were sequenced using a proprietary modified protocol using 2 × 300 bp paired-end chemistry on an Illumina HiSeq platform (Illumina, San Diego, California) at GATC Biotech AG, Germany.

Bioinformatic viral processing

Norman et al., data set

Raw sequence (2,199,754 ± 983,529 per sample) quality was assessed using FASTQC and filtered utilising Trimmomatic (Bolger et al., 2014) using the following parameters; SLIDINGWINDOW: 4:20, MINLEN: 60 HEADCROP 15; CROP 225. Human reads were removed using Kraken (v.0.10.5) (Wood and Salzberg, 2014) and version 38 of the human genome, which resulted in a mean of 1,130,518 ± 436,424 sequences per sample. SPAdes meta (Nurk et al., 2017) with default parameters, was chosen to assemble the reads into contigs per sample, based on a recent virome assembly comparison (Sutton et al., 2019b) Assemblies were subsequently pooled and retained if longer than 1kb. Redundancy was removed with 90% identity over 90% of the length (of the shorter) retaining the longest contig in each case. This was calculated by carrying out an all vs all BLASTn and parsing resulting alignments with an in-house script, as described in (Shkoporov et al., 2018b;Shkoporov et al., 2019). Briefly, all local alignments between two sequences above an e-value threshold of 1e^{-5} were summed, removing overlaps between alignments. The length of the combined local alignments was then given as a percentage of the length shorter sequence. Bacterial contamination was removed by using an extensive set of inclusion criteria to select viral sequences only. Briefly, contigs were required to fulfil one of the following criteria; 1) Categories 1-6 from VirSorter when run with default parameters and Refseqdb (--db 1) (Roux et al., 2015a) positive, 2) circular, 3) a minimum of 2 pVogs with at least 3 per 1kb (Grazziotin et al., 2017), 4) BLASTn alignment to an in-house crAssphage database (e-value threshold: 1e^{-10}) (Guerin et al., 2018), 5) greater than 3kb with no BLASTn alignments to the NT database (January '19) (e-value threshold: 1e^{-10}), 6) BLASTn alignments to viral RefSeq database (v.89) (e-value threshold: 1e^{-10}), and 7) less than 3 ribosomal proteins as predicted using the COG database (Tatusov et al., 2000). HMMscan was used to search the pVOGs hmm profile database using predicted protein sequences on VLS with and e-value filter of 1e^{-5}, retaining the top hit in each case.

Quality filtered reads were subsequently aligned to the reference set of viral sequences (n = 7,605) using bowtie2 (Langmead and Salzberg, 2012). Using SAMtools (Li et al., 2009), a count table was generated and finally a 75% breadth of coverage filter was employed to exclude any spurious bowtie2 alignments being identified as true viral hits. Any viral sequences which did not feature a recruited read coverage of at least 1 over 75% of the total sequence length were set to 0. These criteria yielded a final database of 7,582 viruses like sequences.

Longitudinal UC data set

The same processing as described above was performed for the longitudinal UC data set cohort where 2,523,262 ± 1,289,619 raw reads were quality filtered (Trimmomatic: SLIDINGWINDOW: 4:20, MINLEN: 60 HEADCROP 15; CROP 135 (fwd), 120 (rev)) and assembled yielding 8,089 VLS in the final count table. VLS classified as the spiked exogenous Q33 phage and the internal Illumina control PhiX were excluded from further analysis. Subsequent clustering lead to 484 clusters of > 2 members and 4,521 of 1 member.

Clustering and Taxonomy

Protein sequences were predicted using Prodigal (n=121,021) (Hyatt et al., 2010) and subsequently clustered using vConTACT2 (Bin Jang et al., 2019) using a pc-inflation and vc-inflation of 1.5, pcs-mode set to MCL and all other parameters set to default. This resulted in 472 viral clusters of ≥2 members and 2,382 singletons, hereby referred to as a viral cluster (VC) with 1 member. A cluster count table was generated by summing all the counts from the previous table in each cluster. Taxonomic classification was assigned to a cluster using vConTACT2 and a custom database of viral genomes formed from the concatenation of the taxonomically classified portion of the NCBI's Viral RefSeq (v.89) and the JGI's IMG-VR (downloaded 9 January 2019). The resulting clusters were classified to family-level based on the presence of reference genomes within. Clusters containing genomes from multiple families, were termed "heterogeneous", and may arise from disagreement between protein-based phylogeny and current taxonomic classification discussed further by Bolduc *et al*. CRISPR protospacers were predicted from the human microbiome project bacterial reference genomes (The Human Microbiome Jumpstart Reference Strains Consortium 2010) using PILRCR (Edgar, 2007). These were aligned to VLS using blastn (-task “blastn-short”) and formatted with blastn_formatter. (The top alignments with an e-value score <1e-5 to each VLS was retained in each case). A VC was deemed temperate if it contained VLS with alignments to pVOGs featuring annotated integrase genes or site specific recombinase genes. These pVOGs were identified through string matching “specific recombinase” or “integrase” within the functional annotations of each pVOG. This yielded 28 pVOGs in total. (VOG0221,VOG0275,VOG0286,VOG0303,VOG0375,VOG0559,VOG0944,VOG10948,VOG2142,VOG2405,VOG2773,VOG2780,VOG3344,VOG3995,VOG4609,VOG4650,VOG4942,VOG5508,VOG5717,VOG6225,VOG6237,VOG6282,VOG6466,VOG7017,VOG7518,VOG8218,VOG8244,VOG9501).

Bioinformatic 16S processing - Norman et al. data set

Read quality was assessed on the raw reads (68,146 ± 32,196) using FastQC before and after quality filtering using Trimmomatic under the following parameters;

HEADCROP:15 CROP:235 SLIDINGWINDOW:4:20 MINLEN:30. The trimmed reads of the Norman et al. 16S data set were then processed using DADA2 (Callahan et al., 2016) (v1.10.1). To do this, reads were quality filtered further (truncLen=230, maxEE=1.4, truncQ=11), before dereplication and *de novo* chimera removal (method = "consensus"). 16S reads published in this study were processed using the same method (truncLen=c(180,100), maxEE=1.4, truncQ=2) and the resulting sequence tables of both data sets merged in DADA2. Chimeras were removed *de novo* from the combined data sets (method="consensus"), followed by a round of reference based chimera removal using UCHIME (Edgar et al., 2011) (v4.2) against the ChimeraSlayer Gold database. Resulting non-chimeric RSVs were sorted by length, with all RSVs having a minimum length of 200 bp and a maximum of 260bp retained. The final count table resulted in a mean of 41,060 ± 17,131 counts per sample. Classification of retained RSVs was achieved using mothur (Schloss et al., 2009) (v1.38.0, bootstrap >=80), while SPINGO (Allard et al., 2015) (v1.3, bootstrap >= 0.8, similarity >=0.5) was used for species-level classification. The RDP v11.4 database was used in both instances.

Longitudinal UC data set

The same methods are above were employed to process the 16S raw data from the UC longitudinal data set. There were 382,602 ± 181,911 raw reads and the following Trimmomatic parameters were applied: HEADCROP:20 SLIDINGWINDOW:4:20 CROP:210 MINLEN:50, resulting in a mean of 76,619 ± 40,278 counts in the final count table per sample after being subjected to the bioinformatics pipeline.

Quantification and Statistical Analysis

All statistics and figure generation was performed in R (v.3.5.1). α and β-diversity was calculated using phyloseq (v.1.26) while differential abundance was carried out using DESeq2 (v.1.22.1). DESeq2 performs an internal normalization, in which a geometric mean is calculated for each sequence across all samples. Counts for a sequence in each sample are then divided by this mean. The median of these ratios in a sample becomes the size factor for that sample. This procedure corrects for library size and composition bias within samples. P-values are determined using the Wald test and adjusted with Benjamini-Hochberg. For further details see https://bioconductor.org/packages/release/bioc/manuals/DESeq2/man/DESeq2.pdf.

All correlations except the relative abundance of key VCs and RSVs were carried were performed with the cor.test from the stats R package (v3.6.1) using the spearman method. Correlations between the relative abundance of key VCs and RSVs were carried out using rcorr from the Hmisc R package (v.4.2-0) with Spearman method. PERMANOVA was carried out using Adonis from the vegan R package (v.2.5-3) to investigate for significance in the β-diversity and measure the degree of variation explained. Procrustes coordinates and significance was generated using procuste and procuste.randtest also from the vegan library. Machine learning was carried out in R using the XgBoost package (0.71.2). In each case the model was trained on 70% of the data and results refer to the remaining 30% of the data which was used to test the performance of the model. Parameters were optimised for each model using 5-fold cross validation employing an n.round of 1000 across 200 iterations. ROC curves and accuracy were performed using the R library ROCR (v.1.0-7). Feature importance was based on the gain values (i.e. relative contribution of the feature to the model), with increasing gain referring to increased importance for generating a prediction.

Tests for significance between groups for α-diversity, *Caudovirales* abundance, temperate VC abundance, and virome stability in relation to remission status, were performed using a Wilcoxon test. For all statistical tests significance was defined as less than 0.05 and all adjustments (where required) was using the Benjamini-hochberg method and one sample was chosen at random per subject. All figures were generated using ggplot2 (v.3.1.0).

Data Availability

The longitudinal UC data set 16S and virome reads are available on the SRA under the following accession number: PRJNA552448 (16S) PRJNA552463 (virome). Raw sequencing reads (virome and 16S) for the Norman et al., 2015 cohort were downloaded from the EMBL-EBI database using a the accession number PRJEB7772 as stated in the original publication (Norman et al., 2015).

References

Allard, G., Ryan, F.J., Jeffery, I.B., and Claesson, M.J. (2015). SPINGO: a rapid species-classifier for microbial amplicon sequences. *BMC Bioinformatics* 16**,** 324. doi:10.1186/s12859-015-0747-1.

Berkowitz, L., Schultz, B.M., Salazar, G.A., Pardo-Roa, C., Sebastian, V.P., Alvarez-Lobos, M.M. et al. (2018). Impact of Cigarette Smoking on the Gastrointestinal Tract Inflammation: Opposing Effects in Crohn's Disease and Ulcerative Colitis. *Front Immunol* 9**,** 74. doi:10.3389/fimmu.2018.00074.

Bin Jang, H., Bolduc, B., Zablocki, O., Kuhn, J.H., Roux, S., Adriaenssens, E.M. et al. (2019). Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat Biotechnol*. doi:10.1038/s41587-019-0100-8.

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

Breitbart, M., Bonnain, C., Malki, K., and Sawaya, N.A. (2018). Phage puppet masters of the marine microbial realm. *Nature Microbiology* 3**,** 754-766. doi:10.1038/s41564-018-0166-y.

Callahan, B.J., Mcmurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J., and Holmes, S.P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* 13**,** 581-583. doi:10.1038/nmeth.3869.

Canchaya, C., Fournous, G., Chibani-Chennoufi, S., Dillmann, M.-L., and Brüssow, H. (2003). Phage as agents of lateral gene transfer. *Current opinion in microbiology* 6**,** 417-424.

Casjens, S.R., and Hendrix, R.W. (2015). Bacteriophage lambda: Early pioneer and still relevant. *Virology* 479-480**,** 310-330. doi:https://doi.org/10.1016/j.virol.2015.02.010.

De Sordi, L., Khanna, V., and Debarbieux, L. (2017). The gut microbiota facilitates drifts in the genetic diversity and infectivity of bacterial viruses. *Cell host & microbe* 22**,** 801-808. e803.

Delday, M., Mulder, I., Logan, E.T., and Grant, G. (2018). Bacteroides thetaiotaomicron Ameliorates Colon Inflammation in Preclinical Models of Crohn’s Disease. *Inflammatory Bowel Diseases* 25**,** 85-96. doi:10.1093/ibd/izy281.

Edgar, R.C. (2007). PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics* 8**,** 18. doi:10.1186/1471-2105-8-18.

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27**,** 2194-2200. doi:10.1093/bioinformatics/btr381.

Fernandes, M.A., Verstraete, S.G., Phan, T.G., Deng, X., Stekol, E., Lamere, B. et al. (2019). Enteric Virome and Bacterial Microbiota in Children With Ulcerative Colitis and Crohn Disease. *J Pediatr Gastroenterol Nutr* 68**,** 30-36. doi:10.1097/MPG.0000000000002140.

Forslund, K., Hildebrand, F., Nielsen, T., Falony, G., Le Chatelier, E., Sunagawa, S. et al. (2015). Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* 528**,** 262-266. doi:10.1038/nature15766.

Forster, S.C., Kumar, N., Anonye, B.O., Almeida, A., Viciani, E., Stares, M.D. et al. (2019). A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nat Biotechnol* 37**,** 186-192. doi:10.1038/s41587-018-0009-7.

Frank, D.N., St Amand, A.L., Feldman, R.A., Boedeker, E.C., Harpaz, N., and Pace, N.R. (2007). Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proceedings of the National Academy of Sciences of the United States of America* 104**,** 13780-13785. doi:10.1073/pnas.0706625104.

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

Gevers, D., Kugathasan, S., Denson, L.A., Vazquez-Baeza, Y., Van Treuren, W., Ren, B. et al. (2014). The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* 15**,** 382-392. doi:10.1016/j.chom.2014.02.005.

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 Res* 45**,** D491-D498. doi:10.1093/nar/gkw975.

Gregory, A.C., Zablocki, O., Howell, A., Bolduc, B., and Sullivan, M.B. (2019). The human gut virome database. *bioRxiv***,** 655910. doi:10.1101/655910.

Guerin, E., Shkoporov, A., Stockdale, S.R., Clooney, A.G., Ryan, F.J., Sutton, T.D.S. et al. (2018). Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. *Cell Host Microbe* 24**,** 653-664 e656. doi:10.1016/j.chom.2018.10.002.

Halfvarson, J., Brislawn, C.J., Lamendella, R., Vazquez-Baeza, Y., Walters, W.A., Bramer, L.M. et al. (2017). Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat Microbiol* 2**,** 17004. doi:10.1038/nmicrobiol.2017.4.

Hall, A.B., Yassour, M., Sauk, J., Garner, A., Jiang, X., Arthur, T. et al. (2017). A novel Ruminococcus gnavus clade enriched in inflammatory bowel disease patients. *Genome Medicine* 9**,** 103. doi:10.1186/s13073-017-0490-5.

Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11**,** 119. doi:10.1186/1471-2105-11-119.

Joossens, M., Huys, G., Cnockaert, M., De Preter, V., Verbeke, K., Rutgeerts, P. et al. (2011). Dysbiosis of the faecal microbiota in patients with Crohn's disease and their unaffected relatives. *Gut* 60**,** 631-637. doi:10.1136/gut.2010.223263.

Krishnamurthy, S.R., and Wang, D. (2017). Origins and challenges of viral dark matter. *Virus Res* 239**,** 136-142. doi:10.1016/j.virusres.2017.02.002.

Kverka, M., Zakostelska, Z., Klimesova, K., Sokol, D., Hudcovic, T., Hrncir, T. et al. (2011). Oral administration of Parabacteroides distasonis antigens attenuates experimental murine colitis through modulation of immunity and microbiota composition. *Clinical and experimental immunology* 163**,** 250-259. doi:10.1111/j.1365-2249.2010.04286.x.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9**,** 357-359. doi:10.1038/nmeth.1923.

Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G. et al. (2013). Richness of human gut microbiome correlates with metabolic markers. *Nature* 500**,** 541-546. doi:10.1038/nature12506.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N. et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25**,** 2078-2079. doi:10.1093/bioinformatics/btp352.

Li, M., Van Esch, B.C.a.M., Wagenaar, G.T.M., Garssen, J., Folkerts, G., and Henricks, P.a.J. (2018). Pro- and anti-inflammatory effects of short chain fatty acids on immune and endothelial cells. *European Journal of Pharmacology* 831**,** 52-59. doi:https://doi.org/10.1016/j.ejphar.2018.05.003.

Lopez-Siles, M., Enrich-Capo, N., Aldeguer, X., Sabat-Mir, M., Duncan, S.H., Garcia-Gil, L.J. et al. (2018). Alterations in the Abundance and Co-occurrence of Akkermansia muciniphila and Faecalibacterium prausnitzii in the Colonic Mucosa of Inflammatory Bowel Disease Subjects. *Front Cell Infect Microbiol* 8**,** 281. doi:10.3389/fcimb.2018.00281.

Machiels, K., Joossens, M., Sabino, J., De Preter, V., Arijs, I., Eeckhaut, V. et al. (2014). A decrease of the butyrate-producing species Roseburia hominis and Faecalibacterium prausnitzii defines dysbiosis in patients with ulcerative colitis. *Gut* 63**,** 1275-1283. doi:10.1136/gutjnl-2013-304833.

Mahony, J., Martel, B., Tremblay, D.M., Neve, H., Heller, K.J., Moineau, S. et al. (2013). Identification of a new P335 subgroup through molecular analysis of lactococcal phages Q33 and BM13. *Applied and environmental microbiology* 79**,** 4401-4409. doi:10.1128/AEM.00832-13.

Maier, L., Pruteanu, M., Kuhn, M., Zeller, G., Telzerow, A., Anderson, E.E. et al. (2018). Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature* 555**,** 623-628. doi:10.1038/nature25979.

Manrique, P., Bolduc, B., Walk, S.T., Van Der Oost, J., De Vos, W.M., and Young, M.J. (2016). Healthy human gut phageome. *Proc Natl Acad Sci U S A* 113**,** 10400-10405. doi:10.1073/pnas.1601060113.

Maura, D., Galtier, M., Le Bouguénec, C., and Debarbieux, L. (2012). Virulent Bacteriophages Can Target O104:H4 Enteroaggregative &lt;span class=&quot;named-content genus-species&quot; id=&quot;named-content-1&quot;&gt;Escherichia coli&lt;/span&gt; in the Mouse Intestine. *Antimicrobial Agents and Chemotherapy* 56**,** 6235. doi:10.1128/AAC.00602-12.

Minot, S., Sinha, R., Chen, J., Li, H., Keilbaugh, S.A., Wu, G.D. et al. (2011). The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res* 21**,** 1616-1625. doi:10.1101/gr.122705.111.

Moreno-Gallego, J.L., Chou, S.P., Di Rienzi, S.C., Goodrich, J.K., Spector, T.D., Bell, J.T. et al. (2019). Virome Diversity Correlates with Intestinal Microbiome Diversity in Adult Monozygotic Twins. *Cell Host Microbe* 25**,** 261-272 e265. doi:10.1016/j.chom.2019.01.019.

Norman, J.M., Handley, S.A., Baldridge, M.T., Droit, L., Liu, C.Y., Keller, B.C. et al. (2015). Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* 160**,** 447-460. doi:10.1016/j.cell.2015.01.002.

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Res* 27**,** 824-834. doi:10.1101/gr.213959.116.

Parras-Molto, M., Rodriguez-Galet, A., Suarez-Rodriguez, P., and Lopez-Bueno, A. (2018). Evaluation of bias induced by viral enrichment and random amplification protocols in metagenomic surveys of saliva DNA viruses. *Microbiome* 6**,** 119. doi:10.1186/s40168-018-0507-3.

Pascal, V., Pozuelo, M., Borruel, N., Casellas, F., Campos, D., Santiago, A. et al. (2017). A microbial signature for Crohn's disease. *Gut* 66**,** 813-822. doi:10.1136/gutjnl-2016-313235.

Qing, Y., Xie, H., Su, C., Wang, Y., Yu, Q., Pang, Q. et al. (2019). Gut Microbiome, Short-Chain Fatty Acids, and Mucosa Injury in Young Adults with Human Immunodeficiency Virus Infection. *Digestive Diseases and Sciences* 64**,** 1830-1843. doi:10.1007/s10620-018-5428-2.

Reyes, A., Haynes, M., Hanson, N., Angly, F.E., Heath, A.C., Rohwer, F. et al. (2010). Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* 466**,** 334-338. doi:10.1038/nature09199.

Ricanek, P., Lothe, S.M., Frye, S.A., Rydning, A., Vatn, M.H., and Tønjum, T. (2012). Gut bacterial profile in patients newly diagnosed with treatment-naïve Crohn's disease. *Clinical and experimental gastroenterology* 5**,** 173-186. doi:10.2147/CEG.S33858.

Rigottier-Gois, L. (2013). Dysbiosis in inflammatory bowel diseases: the oxygen hypothesis. *ISME J* 7**,** 1256-1261. doi:10.1038/ismej.2013.80.

Roux, S., Emerson, J.B., Eloe-Fadrosh, E.A., and Sullivan, M.B. (2017). Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ* 5**,** e3817. doi:10.7717/peerj.3817.

Roux, S., Enault, F., Hurwitz, B.L., and Sullivan, M.B. (2015a). VirSorter: mining viral signal from microbial genomic data. *PeerJ* 3**,** e985. doi:10.7717/peerj.985.

Roux, S., Hallam, S.J., Woyke, T., and Sullivan, M.B. (2015b). Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *Elife* 4. doi:10.7554/eLife.08490.

Roux, S., Solonenko, N.E., Dang, V.T., Poulos, B.T., Schwenck, S.M., Goldsmith, D.B. et al. (2016). Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. *PeerJ* 4**,** e2777. doi:10.7717/peerj.2777.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B. et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75**,** 7537-7541. doi:10.1128/AEM.01541-09.

Shkoporov, A.N., Clooney, A.G., Sutton, T.D.S., Ryan, F.J., Daly, K.M., Nolan, J.A. et al. (2019). The human gut virome is highly diverse, stable and individual-specific. *bioRxiv***,** 657528. doi:10.1101/657528.

Shkoporov, A.N., Khokhlova, E.V., Fitzgerald, C.B., Stockdale, S.R., Draper, L.A., Ross, R.P. et al. (2018a). ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. *Nature Communications* 9**,** 4781. doi:10.1038/s41467-018-07225-7.

Shkoporov, A.N., Ryan, F.J., Draper, L.A., Forde, A., Stockdale, S.R., Daly, K.M. et al. (2018b). Reproducible protocols for metagenomic analysis of human faecal phageomes. *Microbiome* 6**,** 68. doi:10.1186/s40168-018-0446-z.

Silpe, J.E., and Bassler, B.L. (2019). A Host-Produced Quorum-Sensing Autoinducer Controls a Phage Lysis-Lysogeny Decision. *Cell* 176**,** 268-280.e213. doi:https://doi.org/10.1016/j.cell.2018.10.059.

Singh, R.K., Chang, H.W., Yan, D., Lee, K.M., Ucmak, D., Wong, K. et al. (2017). Influence of diet on the gut microbiome and implications for human health. *J Transl Med* 15**,** 73. doi:10.1186/s12967-017-1175-y.

Strauss, J., Kaplan, G.G., Beck, P.L., Rioux, K., Panaccione, R., Devinney, R. et al. (2011). Invasive potential of gut mucosa-derived Fusobacterium nucleatum positively correlates with IBD status of the host. *Inflamm Bowel Dis* 17**,** 1971-1978. doi:10.1002/ibd.21606.

Sutton, T.D.S., Clooney, A.G., and Hill, C. (2019a). Giant oversights in the human gut virome. *Gut*. doi:10.1136/gutjnl-2019-319067.

Sutton, T.D.S., Clooney, A.G., Ryan, F.J., Ross, R.P., and Hill, C. (2019b). Choice of assembly software has a critical impact on virome characterisation. *Microbiome* 7**,** 12. doi:10.1186/s40168-019-0626-5.

Tatusov, R.L., Galperin, M.Y., Natale, D.A., and Koonin, E.V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* 28**,** 33-36. http://www.ncbi.nlm.nih.gov/pubmed/10592175.

Turkington, C.J.R., Morozov, A., Clokie, M.R.J., and Bayliss, C.D. (2019). Phage-Resistant Phase-Variant Sub-populations Mediate Herd Immunity Against Bacteriophage Invasion of Bacterial Meta-Populations. *Frontiers in Microbiology* 10. doi:10.3389/fmicb.2019.01473.

Vandeputte, D., Falony, G., Vieira-Silva, S., Tito, R.Y., Joossens, M., and Raes, J. (2016). Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates. *Gut* 65**,** 57-62. doi:10.1136/gutjnl-2015-309618.

Vandeputte, D., Kathagen, G., D'hoe, K., Vieira-Silva, S., Valles-Colomer, M., Sabino, J. et al. (2017). Quantitative microbiome profiling links gut community variation to microbial load. *Nature* 551**,** 507-511. doi:10.1038/nature24460.

Weitz, J.S., Stock, C.A., Wilhelm, S.W., Bourouiba, L., Coleman, M.L., Buchan, A. et al. (2015). A multitrophic model to quantify the effects of marine viruses on microbial food webs and ecosystem processes. *ISME J* 9**,** 1352-1364. doi:10.1038/ismej.2014.220.

Willing, B.P., Dicksved, J., Halfvarson, J., Andersson, A.F., Lucio, M., Zheng, Z. et al. (2010). A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes. *Gastroenterology* 139**,** 1844-1854 e1841. doi:10.1053/j.gastro.2010.08.049.

Wood, D.E., and Salzberg, S.L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 15**,** R46. doi:10.1186/gb-2014-15-3-r46.

Zhang, X., Zhao, Y., Zhang, M., Pang, X., Xu, J., Kang, C. et al. (2012). Structural Changes of Gut Microbiota during Berberine-Mediated Prevention of Obesity and Insulin Resistance in High-Fat Diet-Fed Rats. *PLOS ONE* 7**,** e42529. doi:10.1371/journal.pone.0042529.

Zhao, G., Droit, L., Gilbert, M.H., Schiro, F.R., Didier, P.J., Si, X. et al. (2019). Virome biogeography in the lower gastrointestinal tract of rhesus macaques with chronic diarrhea. *Virology* 527**,** 77-88. doi:https://doi.org/10.1016/j.virol.2018.10.001.

Zhou, Y., Chen, H., He, H., Du, Y., Hu, J., Li, Y. et al. (2016). Increased Enterococcus faecalis infection is associated with clinically active Crohn disease. *Medicine* 95**,** e5019-e5019. doi:10.1097/MD.0000000000005019.

Zuo, T., Lu, X.J., Zhang, Y., Cheung, C.P., Lam, S., Zhang, F. et al. (2019). Gut mucosal virome alterations in ulcerative colitis. *Gut*. doi:10.1136/gutjnl-2018-318131.

Chapter 4

The Long and the Short of it: Impact of Sequencing Approaches on the Human Gut Virome.

Manuscript in preparation

Thomas D.S. Sutton, Joan Colom, Ana Zhu, Karen M. Daly, Muireann K. Smith, Lorraine A. Draper, Andrey N. Shkoporov, Trevor D. Lawley, R. Paul Ross, Colin Hill

Drafting of manuscript, study design and bioinformatic analysis carried out by Thomas Sutton. DNA extraction and library preparation carried out by Thomas Sutton, Joan Colom, Karen Daly, Muireann Smith and Lorraine Draper. PacBio sequencing carried out by Ana Zhu and Trevor Lawley. Nanopore sequencing carried out by Andrey Shkoporov. Constructive comments provided by Andrey Shkoporov and Lorraine Draper. Supervision provided by Colin Hill and Paul Ross.

Abstract

The human gut virome is a largely unsolved piece of the gut microbiome puzzle. To date, our understanding of this community is founded in sequence based studies using short read libraries, DNA samples amplified using multiple displacement amplification (MDA) and *de novo* bioinformatic approaches. While it is known that MDA introduces bias and hampers downstream processing steps such as assembly and diversity estimates, the impact of sequencing approaches on the virome has not been fully characterised. Recent developments in long-read sequencing platforms such as the Oxford Nanopore MinIon and Pacific Biosciences Sequel are promising solutions to some of the assembly challenges of virome data, yet to date they have not been used to our knowledge in virome studies. Here we characterise the impact of sequencing approach on virome data and describe the use of long read sequencing in the human gut virome. We report significant limitations in the ability of amplified short read libraries to represent the human gut virome, and propose the use of alternative sequencing approaches as a means to address these limitations.

Introduction

The virome is a particularly challenging microbial community to study, primarily because it heavily depends on sequence-based analysis approaches, and *de novo* assembly tools (Clooney et al., 2019;Shkoporov et al., 2019;Sutton et al., 2019a). Early virome analysis was founded on platforms such as the IonTorrent (Abeles et al., 2014) or 454 pyrosequenceing (Wagner et al., 2013)which gave the first insights into virome composition, but was limited by sequencing depth. The field then progressed to high-throughput short-read platforms such as the Illumina HiSeq or Miseq (Minot et al., 2013;Norman et al., 2015;Kang et al., 2017), and it is these platforms which laid the foundation of our current understanding of the virome. The relatively affordable cost and sequencing depth of these platforms allow for large-scale multi-cohort studies (Norman et al., 2015) and intensive deep-sequencing of individuals within these cohorts (Zuo et al., 2019). However, they are also limited in their ability to address some of the challenges associated with virome data and are known to introduce significant bias to the composition of virome samples (Kim and Bae, 2011;Roux et al., 2016). Owing to the dominance of unknown sequences (viral dark matter) in virome samples most studies depend on *de novo* assembly of these short sequencing reads to resolve viral contigs or genomes in a given sample. However, as discussed extensively (Hesse et al., 2017;Sutton et al., 2019a;Sutton and Hill, 2019) assemblers vary significantly in their ability to overcome the challenges of virome data (e.g. genomic features such as repeats and extremes in coverage) and even the best performing assembly programs are unable to recover all members of a viral community. As short read assemblers must balance the trade-off between contiguity and accuracy the resulting contigs and scaffolds are a consensus of multiple closely related strains (Nurk et al., 2017). However, in viromes where multiple strains are abundant and diverse this often leads to fragmented assemblies or failure to assemble the hypervariable regions of the genome which often relate to host interaction (Warwick-Dugdale et al., 2019). To make matters worse, these regions are often flanked by repeats and/or areas of low coverage which further hamper assembly, which often excludes them from downstream analysis (Warwick-Dugdale et al., 2019).

To our knowledge, only one virome study to date has been carried out on the Oxford Nanopore (ONT) MinIon platform and none have been carried out on the Pacific Biosciences Sequel. That study used the VirIon protocol and the Oxford Nanopore (ONT) MinIon platform to analyse a marine virome (Warwick-Dugdale et al., 2019). Dugdale et al. could scaffold hypervariable regions of abundant *Pelagibacter* phage which had proven difficult to assemble using short reads alone. These hypervariable regions were associated with host interaction and highlighted the potential benefits of long read sequencing in virome studies. Currently, long read sequencing has not yet been applied to the human virome, despite this ecosystem presenting similar assembly challenges and potentially playing an important role in shaping the composition of the gut microbiome.

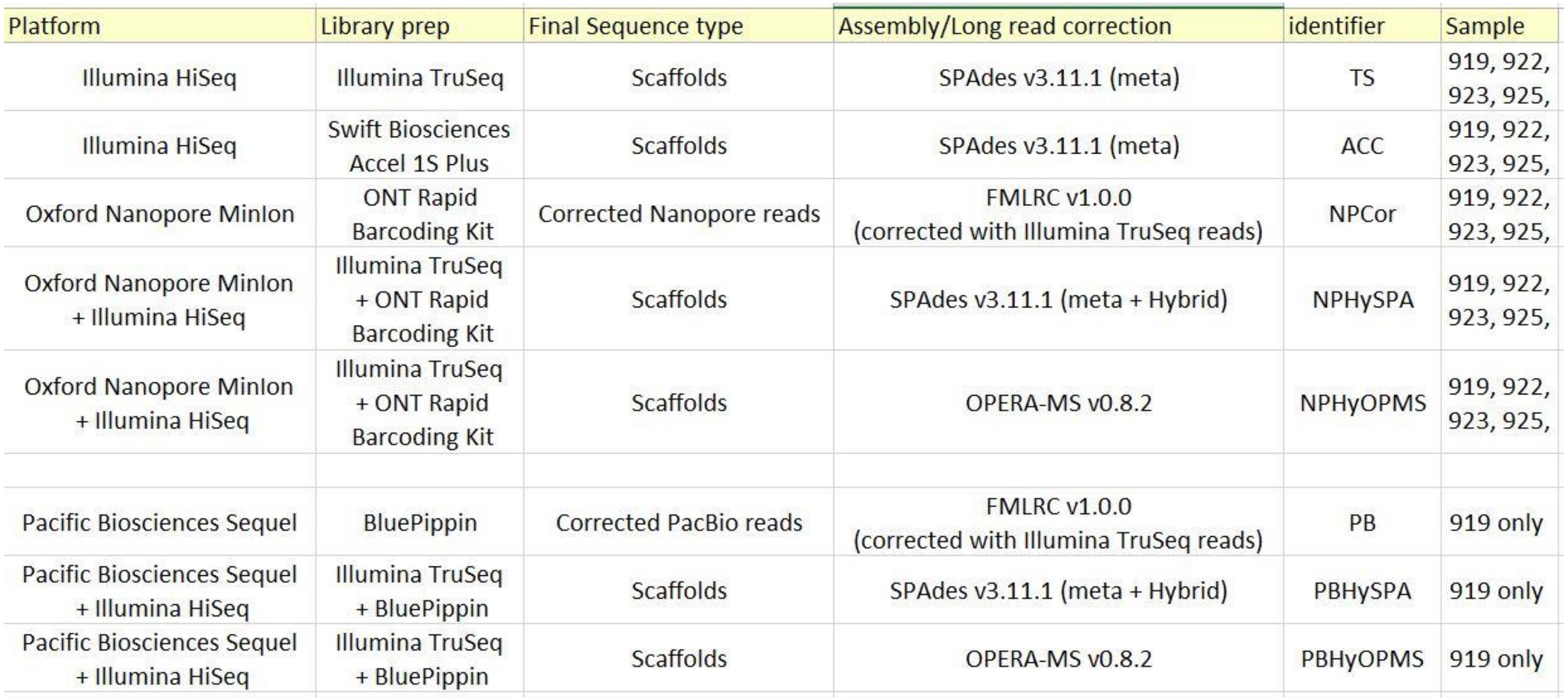
Long-read sequencing offers a number of potential benefits over traditional short read sequencing and may provide solutions to some of the assembly challenges faced by virome analysis. Given that long-read platforms can potentially sequence entire viral genomes in a single read, or make up significant fractions of even the longest phage genomes (e.g. 127.4 kb read on the PacBio Sequel platform, Table 2.) they offer an opportunity to reduce our reliance on the assembly step and its associated challenges. Alternatively they can be used to resolve regions which present challenges to short read assembly such as repeats or regions of varied coverage. However, long-read metagenomic sequencing has been primarily carried out on low diversity samples or mock communities (Tsai et al., 2016;Driscoll et al., 2017;Nicholls et al., 2019;Sevim et al., 2019;Somerville et al., 2019) and its use in complex metagenomes is relatively rare (Slaby et al., 2017;Warwick-Dugdale et al., 2019). Therefore its efficacy and practicality in resolving these complex metagenomes remains to be determined (Olson et al., 2017). Significant improvements have also been made with the accuracy and depth of long-read sequencing platforms which had both been limitations when compared to the high accuracy and depth of sequencing offered by high-throughput short-read platforms. These have been highlighted by recent benchmark studies using the ONT PromethION and GridION which generated 150-153Gbp and 14-16Gbp, respectively (Nicholls et al., 2019) and in this study that generated 10.7 Gbp using the PacBio Sequel platform.

However, long-read sequencing platforms also present a number of challenges which is likely to explain their limited used in metagenomics and viromics. These platforms currently require several micrograms of input DNA as opposed to nanograms required by some for short read libraries such as the Accel 1S Plus library. Given that DNA yields of the human gut virome tend to be low (e.g often < 500ng (Shkoporov et al., 2018b)) this poses a serious problem. As a result, viromes must be amplified using either MDA or other amplification methods which introduce significant bias to the composition of the virome (discussed below). Furthermore Nanopore and PacBio reads exhibit high indel error rates relative to short read sequences (5-10%) (Weirather et al., 2017) which cause problems with downstream ORF prediction software. Indels can shift reading frames to introduce artificial stop codons and so viral genes can seem truncated (Watson and Warr, 2019). As many viral prediction methods such as VirSorter (Roux et al., 2015a) or alignment to the pVOGS database (Grazziotin et al., 2016) depend on accurate ORF calling to identify viral sequences, these high error rates hamper an already limited ability to identify viral community members. For this reason long-reads are often used in combination with, rather than replacing, accurate short reads for metagenomic sequencing (Sevim et al., 2019;Warwick-Dugdale et al., 2019). Programs such as FMLRC (Wang et al., 2018) and Plion (Walker et al., 2014) use the accuracy of short reads to correct high error rates in long-reads. Another challenging aspect of using long read sequencing in metagenomics and in particular in viromics, is its novelty and the consequence that relatively few bioinformatic tools are available that are compatible with metagenomes. The majority of assembly and error correction tools for long reads require the user to input estimated depth of coverage (Koren et al., 2017), which in the case of metagenomes is unavailable, or require higher coverage than is feasible with viromes. Furthermore very few benchmarking studies of long-read metagenomics have been performed (Nicholls et al., 2019;Sevim et al., 2019). This means that the few tools which are compatible with long read metagenomics have not been extensively validated and their impact on the final community composition is unknown.

Here we present a pilot study in which virome samples from four individuals were sequenced using multiple sequencing platforms and library prep methods (Table 1). These included the ONT Minion, and two separate library prep methods for the illumin HiSeq platform, one that used MDA amplification (Illumina TruSeq) and one which did not (Swift Biosciences Accel 1S Plus). These sequencing approaches were supplemented by one extremely deep long-read sequencing run for one of the four viromes using the PacBio Sequel platform to yield 10.7 Gbp across 976,772 reads. To our knowledge this is the first time the PacBio platform has been used on a virome sample and is the deepest published PacBio run to be carried out on the human microbiome to date. As with the ONT MinIon runs, the PacBio run was analysed individually as corrected reads and in combination with TruSeq reads using two hybrid assembly methods. These viromes were analysed on an individual basis using two combinations of the platforms (i.e. hybrid assembly) and the platforms individually to characterize the elements of the virome which are missed by TruSeq approaches alone. This is particularly important in the virome as MDA short-read sequencing such as the TruSeq approach form the foundation of the majority of gut virome research. Furthermore the impact of sequencing platform and library prep on virome composition has yet to be fully characterised, despite recent evidence to suggest choice of virome analysis methods have significant impacts on the conclusions drawn from virome studies (Sutton et al., 2019a) and can be more pronounced than health or disease status (Gregory et al., 2019).

With the exception of the Swift Biosciences Accel 1S Plus library on the illumina HiSeq platform all viromes were amplified using MDA. While MDA is a crucial step to generating sufficient quantities of high molecular weight required for sequencing libraries, particularly those used in long-read platforms, it is also known to introduce significant bias and is to skew the composition of virome samples (Lasken and Stockwell, 2007;Kim and Bae, 2011;Sabina and Leamon, 2015;Roux et al., 2016;Roux et al., 2017). This bias can occur in a number of ways that are of particular importance to virome samples. “Selection bias” refers to preferential amplification of certain templates in a multi-template pool, such as that of a metagenome or virome. MDA primer binding is sensitive to GC content of the priming region as high GC content regions cause problems with denaturation and primer annealing resulting in underrepresentation of high-GC regions (Ishii and Fukui, 2001). As Phi29 polymerase is not capable of strand switching, it also tends to underrepresent sequences near the beginning and end of templates(Sabina and Leamon, 2015). This also means that as the number of termini in templates increases, as is the case in fragmented or multiple short genomes such as those seen in viromes, the degree of underrepresentation also increases (Lage et al., 2003). As MDA reactions preferentially amplify small circular ssDNA genomes such as those in the family *Microviridae* their abundance in the sample can be greatly overrepresented (Dean et al., 2002;Kim and Bae, 2011). Furthermore, initial log-fold differences in coverage in virome sequences (Dutilh et al., 2014;Shkoporov et al., 2018a) are exaggerated by MDA resulting in extremes in both high and low coverage, which hamper downstream assembly and diversity estimates (Sutton et al., 2019a). The low initial yield of faecal virome samples is the primary reason MDA is required to generate sufficient DNA for library prep protocols. However this initial low yield further contributes to the bias introduced by MDA reactions. In low yield samples a small number of early amplification reactions can often determine the composition of the final amplification products (Blainey, 2013) . This results in a stochastic loss of template information referred to as “drift bias” (Sabina and Leamon, 2015).

Another significant challenge of MDA amplified virome samples which became particularly evident in the viromes sequenced on long read platforms was chimeric MDA artefacts. The formation of chimeras or rearrangements which are absent in the template DNA, is closely linked to the strand displacement ability of Phi29 polymerase (the MDA used in this study) (Lasken and Stockwell, 2007) (Supp. Figure. 1). The highly-branched amplification products of MDA can form a number of intermediate secondary structures. DNA strands extending from an initial template can become displaced and are available to prime on a separate template creating chimeric amplification products. The mechanism proposed by Lasken and Stockwell (Lasken and Stockwell, 2007) suggests that 3’ termini displaced by branch migration are available to reanneal with nearby 5’ strands that have been displaced by the Phi29 polymerase itself. This results in deletion of part of the template sequence and sequences directly flanking this region becoming joined in an inverted orientation. As these amplification products are often sheared as part of a short read library preparation and assembled in downstream processing, they are less evident in short read viromes. However, when MDA is used to generate long read libraries, these chimeric sequences remain intact and can cause serious problems with downstream analysis.



Platform	Library prep	Final Sequence type	Assembly/Long read correction	identifier	Sample
Illumina HiSeq	Illumina TruSeq	Scaffolds	SPAdes v3.11.1 (meta)	TS	919, 922, 923, 925,
Illumina HiSeq	Swift Biosciences Accel 1S Plus	Scaffolds	SPAdes v3.11.1 (meta)	ACC	919, 922, 923, 925,
Oxford Nanopore MinIon	ONT Rapid Barcoding Kit	Corrected Nanopore reads	FMLRC v1.0.0 (corrected with Illumina TruSeq reads)	NPCor	919, 922, 923, 925,
Oxford Nanopore MinIon + Illumina HiSeq	Illumina TruSeq + ONT Rapid Barcoding Kit	Scaffolds	SPAdes v3.11.1 (meta + Hybrid)	NPHySPA	919, 922, 923, 925,
Oxford Nanopore MinIon + Illumina HiSeq	Illumina TruSeq + ONT Rapid Barcoding Kit	Scaffolds	OPERA-MS v0.8.2	NPHyOPMS	919, 922, 923, 925,
Pacific Biosciences Sequel	BluePippin	Corrected PacBio reads	FMLRC v1.0.0 (corrected with Illumina TruSeq reads)	PB	919 only
Pacific Biosciences Sequel + Illumina HiSeq	Illumina TruSeq + BluePippin	Scaffolds	SPAdes v3.11.1 (meta + Hybrid)	PBHySPA	919 only
Pacific Biosciences Sequel + Illumina HiSeq	Illumina TruSeq + BluePippin	Scaffolds	OPERA-MS v0.8.2	PBHyOPMS	919 only

Table 1. Sequencing platforms, assembly methods and library prep kits which made up the nine “sequencing approaches” used in this study (see identifier column). PacBio sequencing was carried out on one sample only. Following downstream pooling and redundancy steps, this sample was treated as a separately to give five non-redundant multi-platform viromes (NR-MPV) in total.

Due to the large input requirements of long read libraries and the low DNA yields associated with faecal virome samples, initial attempts to avoid MDA by pooling multiple extractions of each sample did not to yield sufficient DNA for each of the library prep methods. Additionally, in order to examine the effect of the library prep methods and sequencing platform on the final virome it was crucial that identical DNA samples were used for each approach. For these reasons pooled DNA extracts from each sample were amplified using Phi29 MDA (illustra GenomiPhi V2 DNA Amplification Kit, GE Healthcare Life Sciences). However, in an effort to minimize the impact of drift bias (discussed above) three individual MDA reactions were pooled for each sample as described by Raghunathan et al. (Raghunathan et al., 2005). This pooled amplification product was used for all sequencing approaches except the Swift Biosciences Accel 1S Plus library, which used the unamplified pooled virome from each sample, allowing for comparison of this amplification bias across both short-read Illumina platforms. It should be noted that the previously discussed VirION pipeline (Warwick-Dugdale et al., 2019) does not use MDA, and may have avoided some of the MDA-associated issues encountered by this study. However, with all amplification protocols, the Long-Read Linker-Amplified Shotgun Library (LASL) approach (Duhaime et al., 2012) used in the VirION protocol is known to introduce its own bias to the resulting virome (e.g omission of ssDNA viruses (Kim and Bae, 2011)) but this bias has not been characterised to the same extent as that of MDA.

While there was significant overlap between analysis approaches, almost all approaches gave novel insights into the virome composition, which built upon what would be detected by the standard Truseq analysis. We report numerous cases where alternative sequencing approaches resolved large viral genomes which would have otherwise been fragmented by TruSeq assemblies. Furthermore, we observed instances where long-read and unamplified short-read sequencing approaches detected viral sequences which had been missed entirely by TruSeq approaches. However, a number of issues which are particularly pronounced in long-read sequencing approaches were also observed and are crucial considerations for future long read virome pipelines

Methods

Sample recruitment.

Faecal samples were donated by four healthy volunteers who had previously featured in the longitudinal study by Shkoporov et al. (Shkoporov et al., 2019). Donors that had been consistently low crAssphage and which had consistently high diversity across the previous longitudinal study were selected as these viromes can be some of the most challenging to assemble using short-read approaches. Furthermore viromes of crAss-rich individuals can be dominated up to 90% by crAssphage (Dutilh et al., 2012;Guerin et al., 2018;Shkoporov et al., 2018a) and have been successfully sequenced using short-read platforms. Therefore they may not benefit from the addition of long-reads to the same degree as more diverse viromes.

Faecal virome extraction

As mentioned above, in an effort to maximize DNA yield from each sample and to avoid downstream MDA steps, multiple faecal virome extractions were carried out on each sample using an up-scaled version of a previously described protocol outlined (Shkoporov et al., 2018b). Briefly eight 2.5g aliquots of each faecal sample were processed for each round of extraction. These were resuspended in 20ml SM buffer and homogenised by vortexing for five minutes. A further 25ml of SM buffer was added to each aliquot and was chilled on ice for five minutes. The aliquots were then centrifuged at 4,700 rpm in a swing bucket rotor for ten minutes at 4 °C, supernantants transferred to new tubes and centrifugation repeated. Supernatants were then filtered twice through a 0.45um pore diameter filter, with the second filtrate of each aliquot pooled into a sterile 500ml bottle. 44g (10% w/v) of polyethylene glycol (PEG) -8000 was dissolved in the faecal filtrates and placed on ice overnight. Pooled filtrates were decanted into 45ml volumes and spun at 4,700 rpm in a swing bucket rotor for 20min at 4 °C. Supernatant was decanted and the pellet was left to dry inverted. Pellets were resuspended in 1ml SM buffer and the remaining steps in the protocol carried out as outlined by Shkoporov et al. (Shkoporov et al., 2018b). Depending on the quantity of the initial sample, this extraction protocol was carried out 11, 7, 4 and 8 times for samples 919, 922, 923 and 925 respectively. The resulting viral DNA extracts were pooled using the Zymo research “DNA Clean and Concentrator” and eluted in into 40ul of elution buffer. Despite maximizing the starting material and pooling and concentrating multiple DNA extracts per sample, the final DNA yields were insufficient to generate unamplified sequencing libraries for all platforms.

Virome DNA amplification, library preparation and sequencing

Consequently, three 1ul aliquots of each sample were amplified using Phi29-based MDA (illustra GenomiPhi V2 DNA Amplification kit) and subsequently pooled per sample to reduce the impact of "drift bias" as discussed above. Despite the modified extraction protocol not yielding sufficient DNA to create all required sequencing libraries, having a larger amount of starting material for each amplification reaction will have reduced the impact of "drift bias" as previously discussed. The resulting pooled amplification products for each of the four samples were prepared for sequencing as follows.

TruSeq libraries were prepared with the Illumina TruSeq Nano DNA HT Library Prep Kit and the Accel-NGs libraries were prepared with the Swift Biosciences Accel 1S Plus kit as per the methods described by Shkoporov et al. (Shkoporov et al., 2019). Both libraries were normalised as per the manufacturer's protocol and sequenced on the were sequenced using 2 × 150 bp paired-end chemistry on an Illumina HiSeq 2500 platform (Eurofins Genomics Germany).

ONT libraries were prepared with the Oxford Nanopore Rapid Barcoding Sequencing kit (SQK-RBK004) as per the manufacturer's protocol, eluting with nuclease free water which had been pre-warmed to 65 °C. ONT libraries sequenced on the MinIon platform using a v.9.4.1 flowcell at APC Microbiome Ireland.

PacBio library preparation and sequencing was performed by our collaborators (Trevor Lawley and Ana Zhu) at the Wellcome Sanger institute. The library was prepared according to the BluePippin Size-Selection System for a 7kb fragment size and was followed by a phenol DNA extraction protocol. The final library was sequenced using the PacBio Sequel platform.

Read processing and assembly

Short-reads

Short-read quality was assessed using FastQC v0.11.5. Adapter removal was carried out with cutadapt v1.9.1(Martin, 2011) and trimming and removal of low quality reads was carried out using Trimmomatic v0.36 (Bolger et al., 2014) with the following parameters SLIDINGWINDOW:4:20 MINLEN:60 HEADCROP:10. High-quality reads from each sample and sequencing library were assembled separately using SPAdes v3.11.1 (Nurk et al., 2017)in metagenomic mode and default parameters, based on the findings of our recent assembly comparison for virome data (Sutton et al., 2019a). Assemblies were then size filtered to 1kb. Redundancy was then removed within samples and sequencing approach as described below.

Oxford Nanopore read processing and error correction

Nanopore reads were basecalled with Albacore (v2.2.7) and filtered by NanoFilt v2.2.0 with the following parameters -q 8 -l 1000 --headcrop 50. Porechop v0.2.3 was used to remove terminal adapters and split reads containing middle adapters. Trimmed quality filtered reads were then size filtered to 1kb. Nanopore reads were then error corrected using the trimmed high-quality TruSeq reads from the same sample using FMLRC v1.0.0. (Wang et al., 2018) (i.e. sample 919 TruSeq reads were used for the error correction of 919 Nanopore reads).

PacBio read processing and assembly

PacBio subreads were converted from bam to fastq and as with the Nanopore reads, corrected using high-quality TruSeq reads from the same sample using FMLRC v1.0.0 (Wang et al., 2018). However due to the number and size of the corrected PacBio reads (10.1 Gigabases over 9.75 x 10^5 reads) and the computational limitations of our server, corrected PacBio reads were size filtered to 20kb. This brought the number of reads to a similar level to those in the corrected Nanopore libraries while maximising the length of the retained sequences. Redundancy within the error corrected, size filtered PacBio reads was then removed on a per sample basis as described below.

Removal of long-read sequencing artefacts

Both Nanopore and PacBio reads featured sequencing artefacts which needed additional filtering steps to remove. These included long palindromic repeats or regions of repeated single or double nucleotides (e.g AAAA or ATAT). These repeated single and double nucleotide repeats are likely to be errors in either base-calling software, or the sequencing platform itself. Palindromic repeats (Supp Figure. 1) are likely to be the result of MDA chimeras as described above and were identified by finding reads which aligned back onto themselves. This involved carrying out an all vs all BLASTn, filtering for self-hits and flagging long-reads which had aligned within themselves with an identity of at least 98%, where the query start was not the same as the subject start, that the self-hit was over 1kb and made up at least 5% of the total length.

Long-reads which featured extended single and double nucleotide repeats were flagged by calculating the nucleotide frequency within the reads. If reads were dominated by only two or three nucleotides which made up an equal percentage of the total nucleotides, they were removed. Similarly if any nucleotide made up more than 80% of all nucleotides in a given a long-read, the read was removed. It was particularly important to remove these sequences as they would have caused issues in downstream redundancy removal and identification of viral sequences. Both of these steps are based around BLAST nucleotide alignments which masks low complexity regions such as single and double nucleotide repeats, preventing alignment. Furthermore, while long palindromic repeats may be duplicates of legitimate viral sequences (Supp. Figure 2.), they would also be twice the length of any non-chimeric instance instances of the sequence. This would have made the legitimate instance of the sequence redundant and retained the chimera. Following error correction, artefact removal and size filtering, redundancy within Nanopore and PacBio reads was removed within samples and sequencing approach as described below.

Hybrid Assembly

Hybrid assembly was carried out on a per sample basis using high-quality TruSeq reads and corrected Nanopore reads. For the sample which was also sequenced on the PacBio platform (919), hybrid assembly was also carried out using high-quality TruSeq reads and corrected, size-filtered PacBio reads. Both SPAdes v3.11.1 (Nurk et al., 2017) (using the --meta and the --nanopore or --pacbio flag in each case) and OPERA-MS v0.8.2 (Bertrand et al., 2019) (using the --no-ref-clustering --no-strain-clustering flags) were used as a means to compare hybrid assembly options across each sample. Redundancy was then removed within samples and assembly method as described below.

Sequence redundancy removal

Redundancy was removed within each sample and sequencing approach individually (e.g. the 919 corrected Nanopore reads and within the 919 Nanopore Hybrid assemblies were treated as individual samples and redundancy was removed within each independently). Redundancy was removed at 90% identity across 90% of the length of the shorter sequence, retaining the longer sequence in each case. In cases where sequences are equal in length, one representative was kept at random as per the method outlined previously ((Shkoporov et al., 2018b;Clooney et al., 2019;Shkoporov et al., 2019)). This involved carrying out an "all versus all" BLASTn and parsing resulting alignments with an in-house script. In summary, all local alignments between two sequences above an identity threshold of 90% and an e-value threshold of 1e-5 were summed, removing overlaps between consecutive alignments. The length of the combined local alignments was then given as a percentage of the length shorter sequence. This was then filtered to retain the longer sequence, should the summed alignments make up 90% of the length of the shorter sequence.

Prediction of viral sequences in assemblies and corrected long-reads

Open reading frames (ORFs) were predicted in the assemblies and corrected long reads of each sequencing approach using Prodigal v2.6.3 (Hyatt et al., 2010) in metagenomic mode. Viral sequences within each sequencing approach and sample were identified using the criteria outlined by Clooney, Sutton et al. and Shkoporov et al.(Clooney et al., 2019;Shkoporov et al., 2019). In summary, sequenced which met any one of the following criteria were deemed viral and retained for further analysis; 1) Categories 1-6 from VirSorter (Roux et al., 2015a)when run with default parameters and Refseqdb (–db 1) (virsorter ref), 2) circular, 3) a minimum of two pVogs with at least 3 per 1kb (pvogs ref) 4) BLASTn alignment to an in-house crAssphage database (e-value threshold: 1e^{-10})(Guerin et al., 2018), 5) greater than 3kb having no BLASTn alignments to the NT database (January '19) (e-value threshold: 1e^{-10}), 6) BLASTn alignments to viral RefSeq database (v.89) (e-value threshold: 1e^{-10}). HMMscan from HMMER v3.1b2 was used to search the pVOGs (Grazziotin et al., 2016) HMM profile database using predicted protein sequences on assemblies and corrected long-reads with an e-value filter of 1e^{-5}, retaining the top hit in each case. Two additional filters were applied to sequences that were flagged as "viral dark matter" (i.e. criterion 5 described above) to ensure nonsense sequencing artefacts were not being incorrectly classified as viral. First, dark-matter sequences that were masked by the built-in low-complexity filter in BLASTn "DUST" (R. L. Tatusov and D. J. Lipman, unpublished NCBI/Toolkit) were removed from the dataset. Second, the length distribution of ORFs in dark-matter sequences was calculated and sequences with a coding density below 1 ORF kb^{-1} were also removed. This cut-off is slightly more lenient than those referenced in the literature (1.4 ORF kb^{-1}) (Mahmoudabadi and Phillips, 2018) to reflect the loss of viral genomes which use alternative genetic codes, such as members of the crAss family (Guerin et al., 2018) and possible megaphage with below-average coding densities (Devoto et al., 2019).

Generating the Multi-Platform Virome (MPV).

The non-redundant, virus like sequences (VLS) from each of these sequencing approaches were pooled within each sample to make a single multi-platform virome (MPV) per individual. This consisted of VLS from Corrected Nanopore reads, TruSeq assemblies, Accel-NGS assemblies, SPAdes hybrid assemblies (Nanopore and TruSeq) and OPERA-MS hybrid assemblies (Nanopore and TruSeq) pooled within each of the four samples (919, 922, 923 and 925). The MPV of 919 which included the additional corrected PacBio reads, SPAdes hybrid assemblies (PacBio and TruSeq) and OPERA-MS hybrid assemblies (PacBio and TruSeq) was named PB_919 and was analysed independently to sample 919, giving 5 samples in total (919, 922, 923 , 925 and PB_919).

The MPV was made non-redundant within each sample to investigate if TruSeq VLS could be extended by alternative sequencing approaches, or if VLS had been missed entirely by the TruSeq approach and were only present in alternative approaches. This redundancy removal step differed from the step described above by keeping the TruSeq assembly in each case where the sequences were redundant, but equal in length. In this way a TruSeq assembly was only made redundant when it was extended by another sequencing approach and not when it had performed equally well, allowing for a fairer comparison.

Examining the breakdown of the MPV

From this non-redundant MPV (NR-MPV) it was possible to examine the contribution of each sequencing approach to the final virome, which was counted and then plotted in R using ggplot2. Next was to investigate whether a given VLS from a given sequencing approach had been included in the NR-MPV because it had resolved or extended a TruSeq assembly, or if it had been missed by the TruSeq reads themselves. This was done by mapping the high-quality TruSeq reads from each sample back to the NR-MPV using Bowtie2 v2.3.4.1 (Langmead and Salzberg, 2012) and calculating the breadth of coverage of each sequence using mpileup feature of samtools v1.7 (Li et al., 2009). However, when given two equally suitable alignment targets for a given read, bowtie2 will chose one target sequence at random, which could potentially reduce the breadth of coverage of a given sequence artificially. To minimize the impact of this issue across sequencing approaches within each sample, high-quality TruSeq reads were aligned to the VLS of each sequencing approach in the NR-MPV separately (i.e. 919 TruSeq reads vs. NR-VLS from Accel NGS in sample 919 and then vs. NR-VLS from corrected Nanopore reads in 919). From here it was possible to calculate the proportion of each sequencing approach in the NR-MPV which had been successfully sequenced by TruSeq reads and had therefore extended or resolved TruSeq assemblies. Subsequently, it was also possible to identify the proportion of each sequencing approach in the NR-MPV which had been omitted by the TruSeq reads entirely. Within each sequencing approach, VLS were broken down into 3 categories based on the breadth of coverage thresholds as previously described (Roux et al., 2017;Clooney et al., 2019;Shkoporov et al., 2019). These were; VLS with coverage >75% of the total length (i.e. successfully detected by TruSeq reads), between 75% and 30% (i.e. may have been partially detected, but would not pass the breadth of coverage thresholds described in the literature) and <30% (i.e. has been very poorly sequenced by TruSeq reads or missed entirely).

Validating non-TruSeq VLS and their long-term detection

Each of the four samples in this study had featured in a previous study which used the TruSeq approach across 12 monthly timepoints (Shkoporov et al., 2019). This made it possible to validate and examine the longitudinal stability the VLS generated by non-TruSeq approaches. Monthly timepoints (12) from each of the four samples were trimmed and quality filtered as per the parameters and methods described by Shkoporov et al. (Shkoporov et al., 2019) and aligned to each sequencing approach within the NR-MPV separately as described above. In this way it was possible to validate VLS from other sequencing approaches which had been missed by TruSeq reads in this study, and to determine their rate of detection across one year within the same individual.

Visualising VLS of interest.

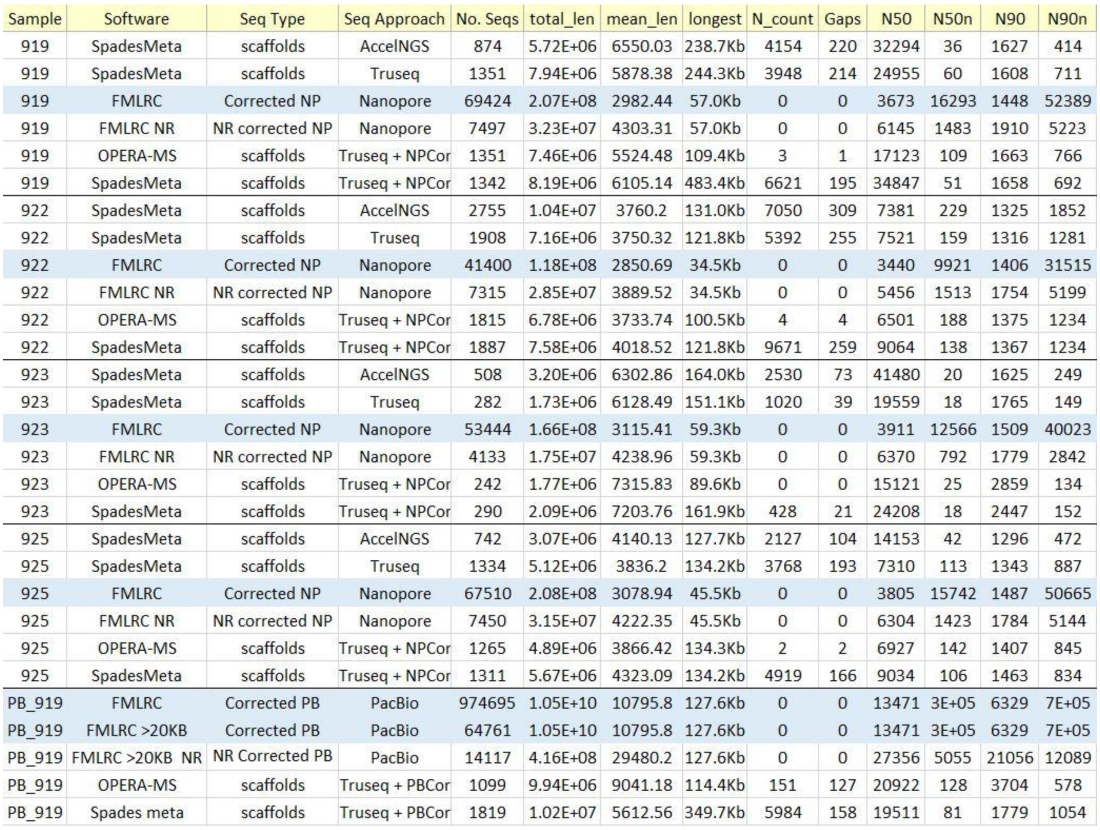
The aforementioned processing steps have been designed for large scale sequence analysis and have been based around sequences passing particular thresholds. While this facilitates large-scale sequence processing and can reveal overall patterns across the dataset, artefacts of the analysis methods can go unnoticed and skew findings (Sutton et al., 2019b;Sutton and Hill, 2019). For this reason, a selection of VLS from the NR-MPV of each sample were characterised and visualised in detail. VLS selected for visualisation included the longest VLS in the entire dataset, VLS which were poorly detected by TruSeq reads in this study but were detected in other longitudinal samples and VLS which resolved multiple fragmented TruSeq assemblies, despite having been sequenced fully by TruSeq reads.

VLS were annotated by aligning predicted protein sequences to the pVOGs (Grazziotin et al., 2016) database as described above. This approach has been found to give a greater number of functional annotations than automated tools such as RASTtk using default settings, which is likely due to use of HMM-based alignments which are more sensitive to distant protein homologies such as those seen in viral genomes (Karplus et al., 1998). The pVOGs hmm profile database features multiple functional annotations within each pVOG, which makes it difficult to assign a single function to a given query protein. By filtering annotations within each pVOG it was possible to generate a single “most common” annotation. This was done by counting the occurrences of all non-hypothetical annotations and assigning the function of the entire pVOG to that which was most common. However, as there are discrepancies within the naming scheme of pVOG functions. (e.g. predicted hypothetical protein, predicted gene product1), assigned functions were first curated manually. Using this tidied pVOG function database (featuring one function per pVOG), it was possible to carry out large-scale annotation of VLS sequences. In-house scripts were used assign function to each VLS protein using the top pVOG hit (HMMscan as described above) and convert this modified HMMscan output to GFF and GBK formats. TruSeq read coverage from both this study and longitudinal samples was then converted from samtools mpileup output (as described above) to a csv file giving the depth of aligned reads at each nt position in the genome. For particularly large VLS depth was averaged across each 100nt. From here it was possible to visualise annotated VLS and TruSeq read recruitment using GView(Petkau et al., 2010). Alignments between selected VLS and TruSeq assemblies were visualised using BRIG v0.95 (Alikhan et al., 2011)and were added to the centre of existing selected VLS plots with Adobe Illustrator CS5. tRNA sequences were predicted on VLS using ARAGORN v1.2.38 (Laslett and Canback, 2004) and aligned to Bacterial RefSeq (v.89) and NT using BLASTn. CRISPR protospacers were predicted from the 3,055 draft and complete bacterial genome assemblies in the Human Microbiome Project (HMP) database using PILR-CR v1.06 (Edgar, 2007), size filtered to between 20 and 70 nt long and aligned to VLS using "blastn-short" mode preset, e-value < 10^{-5}.

Results

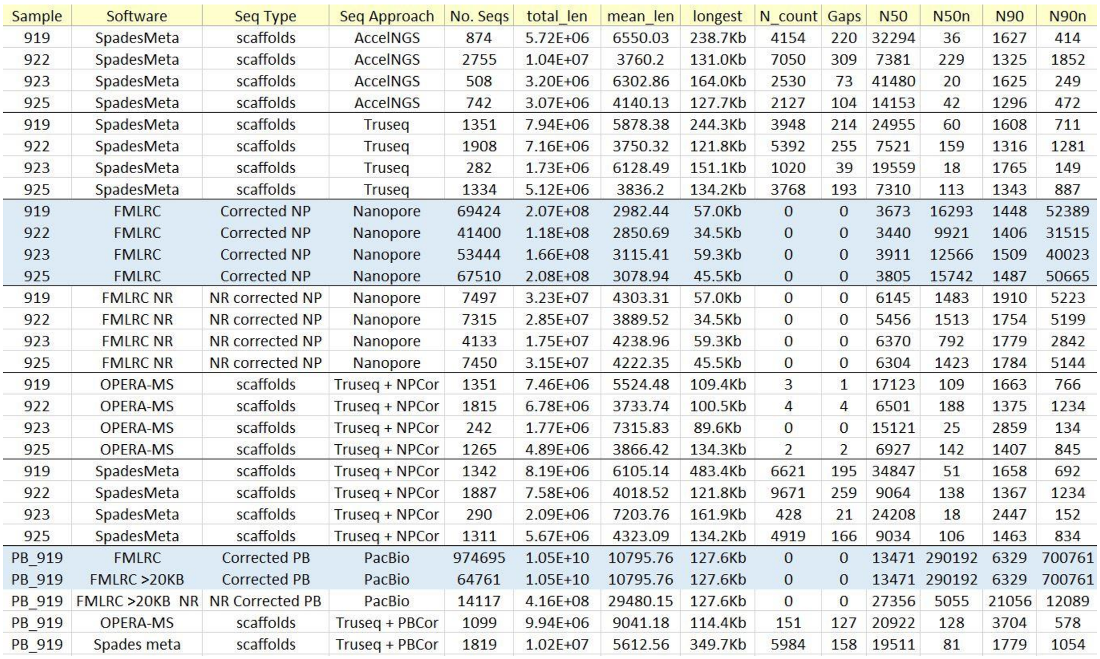
Read and final assembly counts

TruSeq libraries produced on average 2.7 x10^6 (± 1.43 x10^6, mean ± std.dev) high-quality reads across the four samples (4.7 x 10^6, 2.9 x 10^6 , 6.9 x 10^5 and 2.4 x 10^6 across samples 919, 922, 923, and 925 respectively). Accel-NGS libraries produced less high-quality reads on average at 1.9 x10^6 ± 5.5 x 10^5 (1.6 x 10^6, 2.7 x 10^6, 2.1 x 10^6, and 1.2 x 10^6 across 919, 922, 923, and 925 respectively). Counts, and length statistics for the sequencing approaches used in the study (e.g. short-read assemblies, corrected Long reads and hybrid assemblies) are outlined in Table 2 (intermediate files highlighted in blue). Table 2 is sorted by sample and for comparability, Supplementary Table 1 is the same data in Table 2 but sorted by sequencing approach. Significant redundancy (see methods) was observed in the corrected long reads of both platforms (Table 2, intermediate files) with 88±4% of corrected Nanopore reads over 1kb (5.13 ± 1.1 x 10^4, reads on average) and 78% of corrected PacBio reads over 20Kb (5 x 10^4) being made redundant. Despite this redundancy, there were on average 7.2 (±4) times more non-redundant corrected Nanopore reads and 19 (±16) times more corrected PacBio reads than all short-read and hybrid assemblies across all samples. Excluding the Corrected PacBio reads (as they had been size filtered to >20kb) the Accel-NGS library had the longest N50 across all samples (23.8kb) and the corrected Nanopore reads had the lowest (6kb). Fluctuations in total length and counts of long reads/assemblies across samples were consistent across platforms (i.e. subject 922 consistently generated the most long reads/assemblies and 923 the least) suggesting that discrepancies in DNA samples themselves (i.e. quantity, fragmentation, diversity etc.) impact all sequencing and assembly approaches.



Sample	Software	Seq Type	Seq Approach	No. Seqs	total_len	mean_len	longest	N_count	Gaps	N50	N50n	N90	N90n
919	SpadesMeta	scaffolds	AccelNGS	874	5.72E+06	6550.03	238.7Kb	4154	220	32294	36	1627	414
919	SpadesMeta	scaffolds	Truseq	1351	7.94E+06	5878.38	244.3Kb	3948	214	24955	60	1608	711
919	FMLRC	Corrected NP	Nanopore	69424	2.07E+08	2982.44	57.0Kb	0	0	3673	16293	1448	52389
919	FMLRC NR	NR corrected NP	Nanopore	7497	3.23E+07	4303.31	57.0Kb	0	0	6145	1483	1910	5223
919	OPERA-MS	scaffolds	Truseq + NPCor	1351	7.46E+06	5524.48	109.4Kb	3	1	17123	109	1663	766
919	SpadesMeta	scaffolds	Truseq + NPCor	1342	8.19E+06	6105.14	483.4Kb	6621	195	34847	51	1658	692
922	SpadesMeta	scaffolds	AccelNGS	2755	1.04E+07	3760.2	131.0Kb	7050	309	7381	229	1325	1852
922	SpadesMeta	scaffolds	Truseq	1908	7.16E+06	3750.32	121.8Kb	5392	255	7521	159	1316	1281
922	FMLRC	Corrected NP	Nanopore	41400	1.18E+08	2850.69	34.5Kb	0	0	3440	9921	1406	31515
922	FMLRC NR	NR corrected NP	Nanopore	7315	2.85E+07	3889.52	34.5Kb	0	0	5456	1513	1754	5199
922	OPERA-MS	scaffolds	Truseq + NPCor	1815	6.78E+06	3733.74	100.5Kb	4	4	6501	188	1375	1234
922	SpadesMeta	scaffolds	Truseq + NPCor	1887	7.58E+06	4018.52	121.8Kb	9671	259	9064	138	1367	1234
923	SpadesMeta	scaffolds	AccelNGS	508	3.20E+06	6302.86	164.0Kb	2530	73	41480	20	1625	249
923	SpadesMeta	scaffolds	Truseq	282	1.73E+06	6128.49	151.1Kb	1020	39	19559	18	1765	149
923	FMLRC	Corrected NP	Nanopore	53444	1.66E+08	3115.41	59.3Kb	0	0	3911	12566	1509	40023
923	FMLRC NR	NR corrected NP	Nanopore	4133	1.75E+07	4238.96	59.3Kb	0	0	6370	792	1779	2842
923	OPERA-MS	scaffolds	Truseq + NPCor	242	1.77E+06	7315.83	89.6Kb	0	0	15121	25	2859	134
923	SpadesMeta	scaffolds	Truseq + NPCor	290	2.09E+06	7203.76	161.9Kb	428	21	24208	18	2447	152
925	SpadesMeta	scaffolds	AccelNGS	742	3.07E+06	4140.13	127.7Kb	2127	104	14153	42	1296	472
925	SpadesMeta	scaffolds	Truseq	1334	5.12E+06	3836.2	134.2Kb	3768	193	7310	113	1343	887
925	FMLRC	Corrected NP	Nanopore	67510	2.08E+08	3078.94	45.5Kb	0	0	3805	15742	1487	50665
925	FMLRC NR	NR corrected NP	Nanopore	7450	3.15E+07	4222.35	45.5Kb	0	0	6304	1423	1784	5144
925	OPERA-MS	scaffolds	Truseq + NPCor	1265	4.89E+06	3866.42	134.3Kb	2	2	6927	142	1407	845
925	SpadesMeta	scaffolds	Truseq + NPCor	1311	5.67E+06	4323.09	134.2Kb	4919	166	9034	106	1463	834
PB_919	FMLRC	Corrected PB	PacBio	974695	1.05E+10	10795.8	127.6Kb	0	0	13471	3E+05	6329	7E+05
PB_919	FMLRC >20KB	Corrected PB	PacBio	64761	1.05E+10	10795.8	127.6Kb	0	0	13471	3E+05	6329	7E+05
PB_919	FMLRC >20KB NR	NR Corrected PB	PacBio	14117	4.16E+08	29480.2	127.6Kb	0	0	27356	5055	21056	12089
PB_919	OPERA-MS	scaffolds	Truseq + PBCor	1099	9.94E+06	9041.18	114.4Kb	151	127	20922	128	3704	578
PB_919	Spades meta	scaffolds	Truseq + PBCor	1819	1.02E+07	5612.56	349.7Kb	5984	158	19511	81	1779	1054

Table 2. Counts, and assembly statistics for each sequencing and assembly approach sorted per sample. Intermediate files highlighted in blue were not included in downstream analysis and are included for reference purposes only.



Sample	Software	Seq Type	Seq Approach	No. Seqs	total_len	mean_len	longest	N_count	Gaps	N50	N50n	N90	N90n
919	SpadesMeta	scaffolds	AccelNGS	874	5.72E+06	6550.03	238.7Kb	4154	220	32294	36	1627	414
922	SpadesMeta	scaffolds	AccelNGS	2755	1.04E+07	3760.2	131.0Kb	7050	309	7381	229	1325	1852
923	SpadesMeta	scaffolds	AccelNGS	508	3.20E+06	6302.86	164.0Kb	2530	73	41480	20	1625	249
925	SpadesMeta	scaffolds	AccelNGS	742	3.07E+06	4140.13	127.7Kb	2127	104	14153	42	1296	472
919	SpadesMeta	scaffolds	Truseq	1351	7.94E+06	5878.38	244.3Kb	3948	214	24955	60	1608	711
922	SpadesMeta	scaffolds	Truseq	1908	7.16E+06	3750.32	121.8Kb	5392	255	7521	159	1316	1281
923	SpadesMeta	scaffolds	Truseq	282	1.73E+06	6128.49	151.1Kb	1020	39	19559	18	1765	149
925	SpadesMeta	scaffolds	Truseq	1334	5.12E+06	3836.2	134.2Kb	3768	193	7310	113	1343	887
919	FMLRC	Corrected NP	Nanopore	69424	2.07E+08	2982.44	57.0Kb	0	0	3673	16293	1448	52389
922	FMLRC	Corrected NP	Nanopore	41400	1.18E+08	2850.69	34.5Kb	0	0	3440	9921	1406	31515
923	FMLRC	Corrected NP	Nanopore	53444	1.66E+08	3115.41	59.3Kb	0	0	3911	12566	1509	40023
925	FMLRC	Corrected NP	Nanopore	67510	2.08E+08	3078.94	45.5Kb	0	0	3805	15742	1487	50665
919	FMLRC NR	NR corrected NP	Nanopore	7497	3.23E+07	4303.31	57.0Kb	0	0	6145	1483	1910	5223
922	FMLRC NR	NR corrected NP	Nanopore	7315	2.85E+07	3889.52	34.5Kb	0	0	5456	1513	1754	5199
923	FMLRC NR	NR corrected NP	Nanopore	4133	1.75E+07	4238.96	59.3Kb	0	0	6370	792	1779	2842
925	FMLRC NR	NR corrected NP	Nanopore	7450	3.15E+07	4222.35	45.5Kb	0	0	6304	1423	1784	5144
919	OPERA-MS	scaffolds	Truseq + NPCor	1351	7.46E+06	5524.48	109.4Kb	3	1	17123	109	1663	766
922	OPERA-MS	scaffolds	Truseq + NPCor	1815	6.78E+06	3733.74	100.5Kb	4	4	6501	188	1375	1234
923	OPERA-MS	scaffolds	Truseq + NPCor	242	1.77E+06	7315.83	89.6Kb	0	0	15121	25	2859	134
925	OPERA-MS	scaffolds	Truseq + NPCor	1265	4.89E+06	3866.42	134.3Kb	2	2	6927	142	1407	845
919	SpadesMeta	scaffolds	Truseq + NPCor	1342	8.19E+06	6105.14	483.4Kb	6621	195	34847	51	1658	692
922	SpadesMeta	scaffolds	Truseq + NPCor	1887	7.58E+06	4018.52	121.8Kb	9671	259	9064	138	1367	1234
923	SpadesMeta	scaffolds	Truseq + NPCor	290	2.09E+06	7203.76	161.9Kb	428	21	24208	18	2447	152
925	SpadesMeta	scaffolds	Truseq + NPCor	1311	5.67E+06	4323.09	134.2Kb	4919	166	9034	106	1463	834
PB_919	FMLRC	Corrected PB	PacBio	974695	1.05E+10	10795.76	127.6Kb	0	0	13471	290192	6329	700761
PB_919	FMLRC >20KB	Corrected PB	PacBio	64761	1.05E+10	10795.76	127.6Kb	0	0	13471	290192	6329	700761
PB_919	FMLRC >20KB NR	NR Corrected PB	PacBio	14117	4.16E+08	29480.15	127.6Kb	0	0	27356	5055	21056	12089
PB_919	OPERA-MS	scaffolds	Truseq + PBCor	1099	9.94E+06	9041.18	114.4Kb	151	127	20922	128	3704	578
PB_919	Spades meta	scaffolds	Truseq + PBCor	1819	1.02E+07	5612.56	349.7Kb	5984	158	19511	81	1779	1054

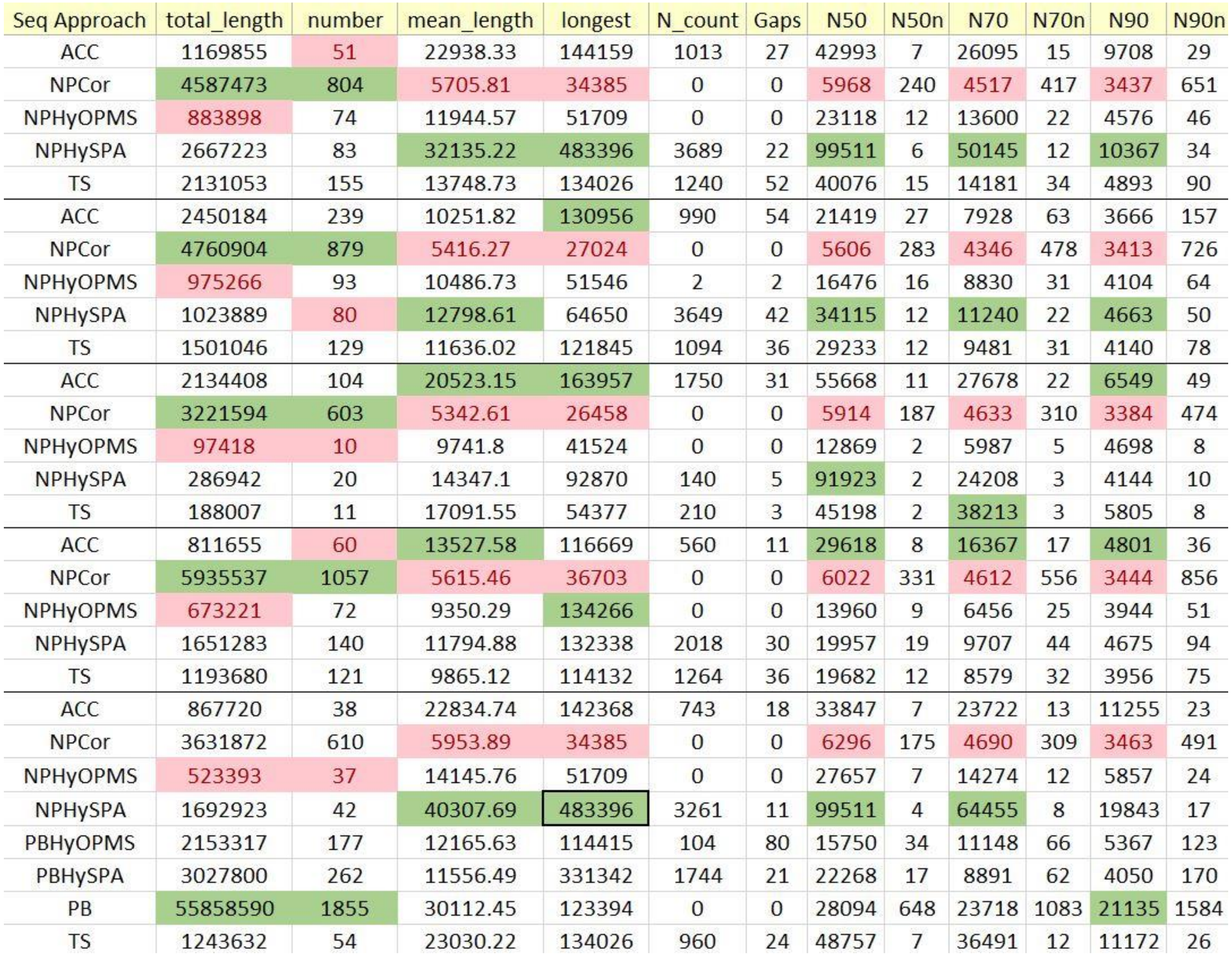
Supp.Table 1. Counts, and assembly statistics for each sequencing and assembly approach as per table 2. sorted per sequencing approach.

In the short read libraries the average N50 value relative to the number of quality filtered reads was slightly higher in the Accel-NGS library, with 1.36 x 10^{-2} in the Accel-NGS relative to 9.82 x 10^{-3} in the TruSeq. This suggests that independent of the size of the libraries, 50% of the total length of Accel -NGS assemblies was contained in longer contigs than those of the TruSeq. The longest sequence generated across all approaches and samples (483.4kb) was a SPAdes hybrid assembly using Nanopore and TruSeq reads in sample 919. This is interesting because relative to Nanopore reads, there were almost twice as many PacBio reads (1.88), the minimum length of PacBio reads in sample 919 was 20 times longer and the longest PacBio read was 70.5 kb longer than the longest Nanopore read. Despite these differences in counts and size, the 483.4kb contig was not assembled using either of the hybrid assemblers and PacBio reads. This may have been caused by the inappropriate removal of PacBio reads during read processing steps that were necessary to scaffold this contig, discrepancies in the error profile of Nanopore and PacBio reads, or biases within the methods used to prepare both libraries.

Composition of the non-redundant multiplatform virome (NR-MPV)

Pooling all VLS from each sequencing approach per sample and removing redundancy made it possible to examine the cases where TruSeq assembly had been improved upon by other sequencing approaches or where they had detected VLS which were missed by TruSeq (Figure 1). For comparability, final non-redundant VLS information as per Table 2 was presented for the MPV (Table 3). Maximum and minimum values of non-redundant VLS are highlighted across sequencing approaches per (e.g. total and mean length, counts, longest and N50, 70 and 90 values). Importantly, all sequencing approaches contributed to the final NR-MPV albeit with some approaches contributing considerably more than others (i.e. 923 Corrected Nanopore contributing 603 NR-VLS and TRuSeq assemblies only 11). Furthermore, TruSeq assemblies made up a relatively small proportion of the VLPs in the NR-MPV of each sample. This suggests that standard TruSeq approaches to virome analysis fail to represent the virome in its entirety and that a more detailed view of the virome is achievable through the addition of long-read sequencing and alternative methods of preparing short read libraries.

Table 3. Counts and assembly stats of the NR-MPV within each sample. Maximum and minimum values within each sample highlighted in green and red respectively.



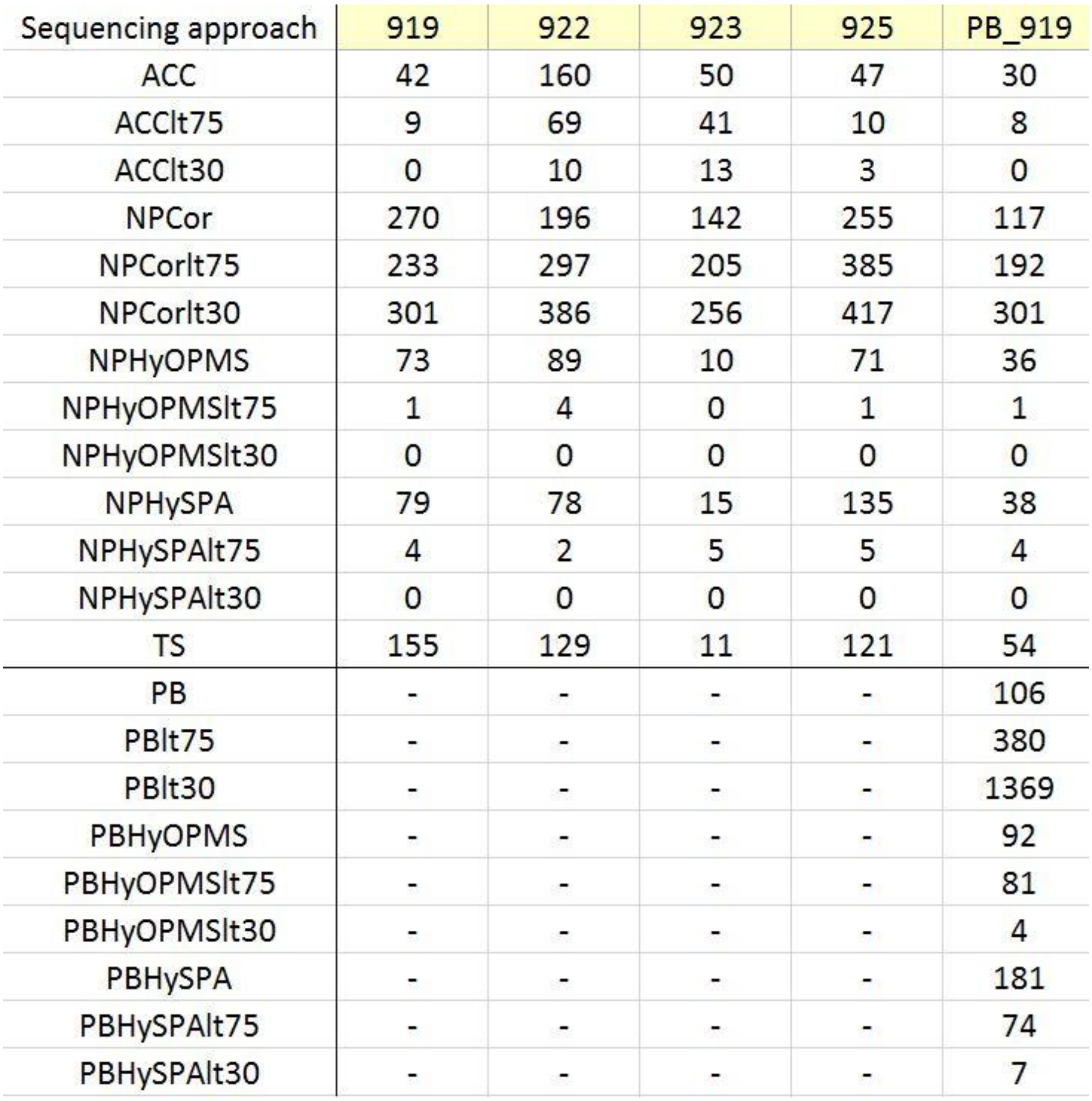
Seq Approach	total_length	number	mean_length	longest	N_count	Gaps	N50	N50n	N70	N70n	N90	N90n
ACC	1169855	51	22938.33	144159	1013	27	42993	7	26095	15	9708	29
NPCor	4587473	804	5705.81	34385	0	0	5968	240	4517	417	3437	651
NPHyOPMS	883898	74	11944.57	51709	0	0	23118	12	13600	22	4576	46
NPHySPA	2667223	83	32135.22	483396	3689	22	99511	6	50145	12	10367	34
TS	2131053	155	13748.73	134026	1240	52	40076	15	14181	34	4893	90
ACC	2450184	239	10251.82	130956	990	54	21419	27	7928	63	3666	157
NPCor	4760904	879	5416.27	27024	0	0	5606	283	4346	478	3413	726
NPHyOPMS	975266	93	10486.73	51546	2	2	16476	16	8830	31	4104	64
NPHySPA	1023889	80	12798.61	64650	3649	42	34115	12	11240	22	4663	50
TS	1501046	129	11636.02	121845	1094	36	29233	12	9481	31	4140	78
ACC	2134408	104	20523.15	163957	1750	31	55668	11	27678	22	6549	49
NPCor	3221594	603	5342.61	26458	0	0	5914	187	4633	310	3384	474
NPHyOPMS	97418	10	9741.8	41524	0	0	12869	2	5987	5	4698	8
NPHySPA	286942	20	14347.1	92870	140	5	91923	2	24208	3	4144	10
TS	188007	11	17091.55	54377	210	3	45198	2	38213	3	5805	8
ACC	811655	60	13527.58	116669	560	11	29618	8	16367	17	4801	36
NPCor	5935537	1057	5615.46	36703	0	0	6022	331	4612	556	3444	856
NPHyOPMS	673221	72	9350.29	134266	0	0	13960	9	6456	25	3944	51
NPHySPA	1651283	140	11794.88	132338	2018	30	19957	19	9707	44	4675	94
TS	1193680	121	9865.12	114132	1264	36	19682	12	8579	32	3956	75
ACC	867720	38	22834.74	142368	743	18	33847	7	23722	13	11255	23
NPCor	3631872	610	5953.89	34385	0	0	6296	175	4690	309	3463	491
NPHyOPMS	523393	37	14145.76	51709	0	0	27657	7	14274	12	5857	24
NPHySPA	1692923	42	40307.69	483396	3261	11	99511	4	64455	8	19843	17
PBHyOPMS	2153317	177	12165.63	114415	104	80	15750	34	11148	66	5367	123
PBHySPA	3027800	262	11556.49	331342	1744	21	22268	17	8891	62	4050	170
PB	55858590	1855	30112.45	123394	0	0	28094	648	23718	1083	21135	1584
TS	1243632	54	23030.22	134026	960	24	48757	7	36491	12	11172	26

Additionally, despite having been sequenced on the same platform, Accel-NGS sequences contributed considerably more VLS to the NR-MPV of certain samples than the TruSeq and more VLS on average across all samples (98.4 ± 74 Accel-NGS VLS vs. 94 ± 53 TruSeq VLS). Furthermore, when the number of VLS contributed by each short read library prep method was normalised by read count, Accel- NGS still contributed 1.5 times more VLS per read on average in each NR-MPV ($5.5x10^{-5}$ ± 2.07 $x10^{-5}$ Accel-NGS VLS/HQread vs. 3.6 ± 1.31 $x10^{-5}$ TruSeq VLS/HQread). This suggests that the differences in the numbers of VLS contributed by each short read to the NR-MPV were not caused by differences in sequencing depth alone. This also suggests that even within a given short read platform, the library prep methods and use of MDA have critical impacts on the final virome composition. Furthermore this highlights the importance of considering the limitations of any single sequencing approach when interpreting results of virome studies, as currently all but one (Warwick-Dugdale et al., 2019) have been carried out using a single approach.

Across all samples, corrected long-reads contribute the greatest number and total length of NR-VLS (corrected Nanopore reads in samples 919, 922, 923 and 925. Corrected PacBio reads in sample PB_919) Figure 2. Hybrid OPERA-MS assemblies using Nanopore and TruSeq reads contributed the least number and total length of VLS in each NR-VLS. This would suggest (somewhat counter-intuitively) that corrected long reads are more capable of detecting viral sequences than hybrid assemblies. This may reflect the novelty of metagenomic hybrid assembly and a need for optimisation in the face of challenging metagenomes such as the virome. It may also reflect high error rates in the long reads which hamper hybrid assembly, uneven coverage within template sequences or bias in library prep methods and platforms themselves skewing the amount of shared sequences between long and short read platforms. However, in agreement with the final read and assembly stats of Table 2, the longest VLS in each sample originated from either Accel-NGS assemblies or hybrid assemblies using Nanopore and TruSeq reads and not the corrected long reads themselves. These observations suggest that in order to maximise the data provided by long-read sequencing both the corrected long-reads themselves and hybrid assemblies should both be included in the final non-redundant dataset.

TruSeq read recruitment and breadth of coverage (BOC)

In order to investigate whether VLS from non-TruSeq sequencing approaches had been included in the NR-MPV because they had resolved or extended a TruSeq assembly, or if it had been missed, TruSeq reads from each sample were aligned to the NR-MPV and coverage broken into three thresholds (see methods). Table 4 represents the VLS counts in Table 3, with each sequencing approach broken down in to coverage thresholds for the five NR-MPVs (919, 922, 923, 925 and PB_919) (Figure 1.).



Sequencing approach	919	922	923	925	PB_919
ACC	42	160	50	47	30
ACClt75	9	69	41	10	8
ACClt30	0	10	13	3	0
NPCor	270	196	142	255	117
NPCorlt75	233	297	205	385	192
NPCorlt30	301	386	256	417	301
NPHyOPMS	73	89	10	71	36
NPHyOPMSlt75	1	4	0	1	1
NPHyOPMSlt30	0	0	0	0	0
NPHySPA	79	78	15	135	38
NPHySPAlt75	4	2	5	5	4
NPHySPAlt30	0	0	0	0	0
TS	155	129	11	121	54
PB	-	-	-	-	106
PBlt75	-	-	-	-	380
PBlt30	-	-	-	-	1369
PBHyOPMS	-	-	-	-	92
PBHyOPMSlt75	-	-	-	-	81
PBHyOPMSlt30	-	-	-	-	4
PBHySPA	-	-	-	-	181
PBHySPAlt75	-	-	-	-	74
PBHySPAlt30	-	-	-	-	7

Table 4. VLS counts in each NR-MPV broken down by sequencing approach and breadth of coverage (BOC) categories. The sequencing approach identifier alone (i.e. ACC, see Table 1. for identifier information) represents the number of VLS from that approach which passed BOC detection criteria (BOC > 75%). "lt75" represents those with a BOC between 30% and 75% and "lt30" represents VLS which had a TruSeq BOC < 30%.

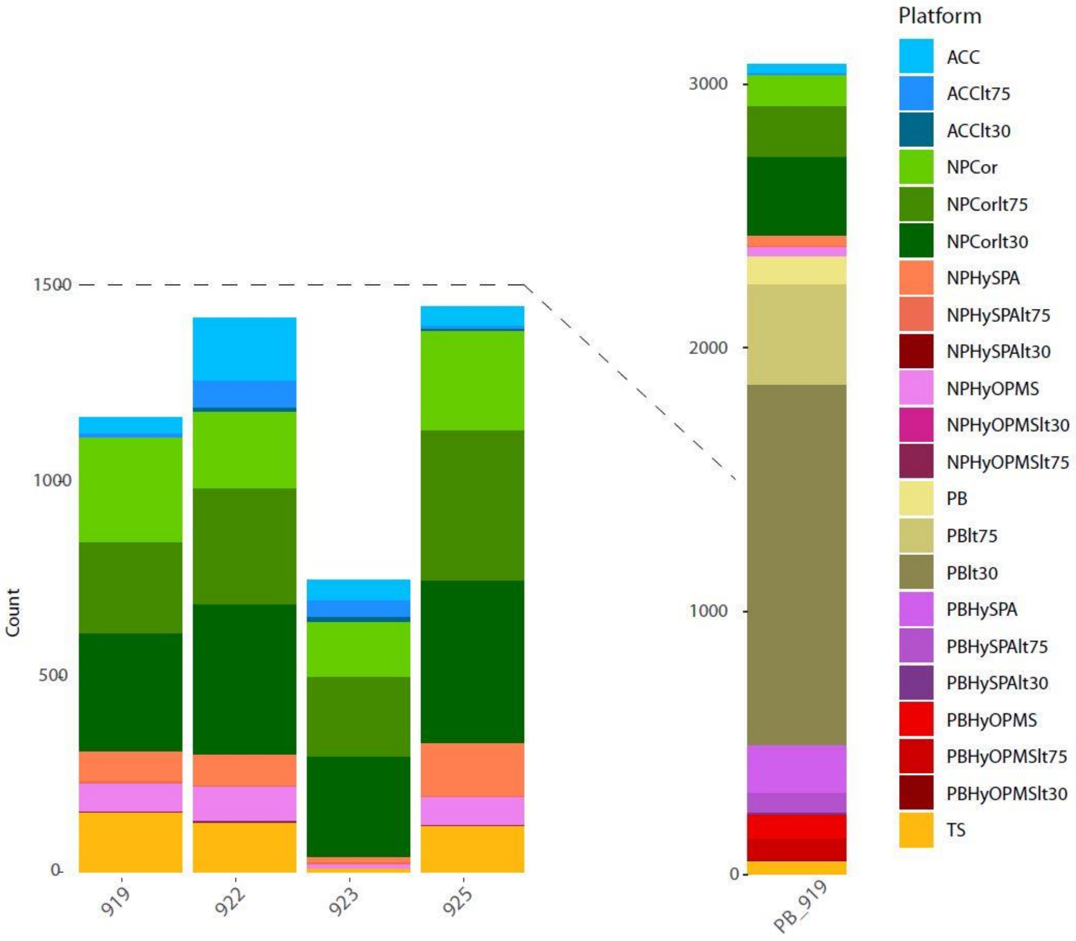


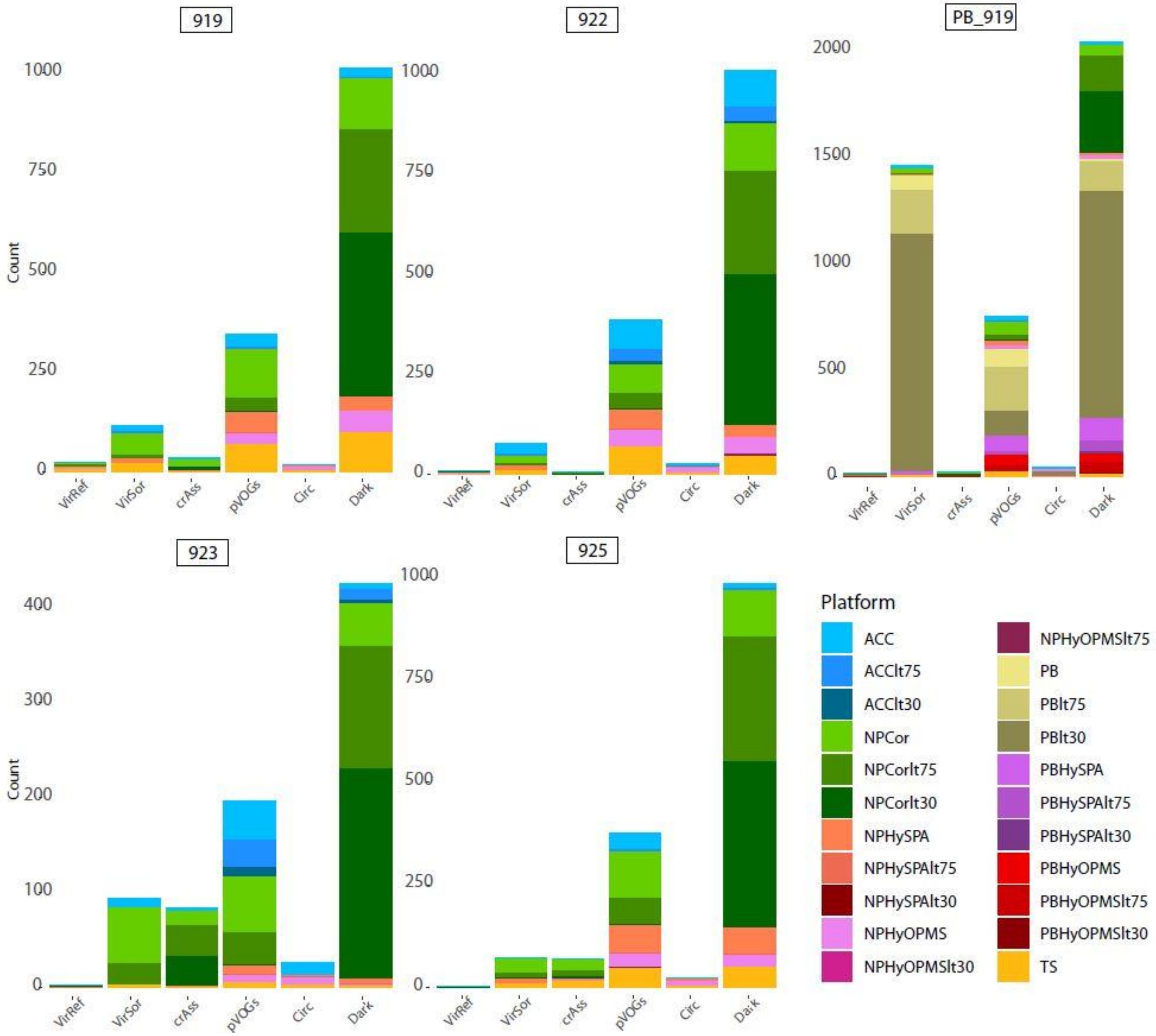
Figure 1. Composition of the NR-MPV across all samples. Note that the scale for samples 919, 922, 923, and 925 is different to that of PB_919 (highlighted by the dashed line). Each sequencing approach is broken down into (BOC) categories (lighter to darker shading within each colour), see Table 4. and "TruSeq read recruitment and breadth of coverage (BOC)" above.

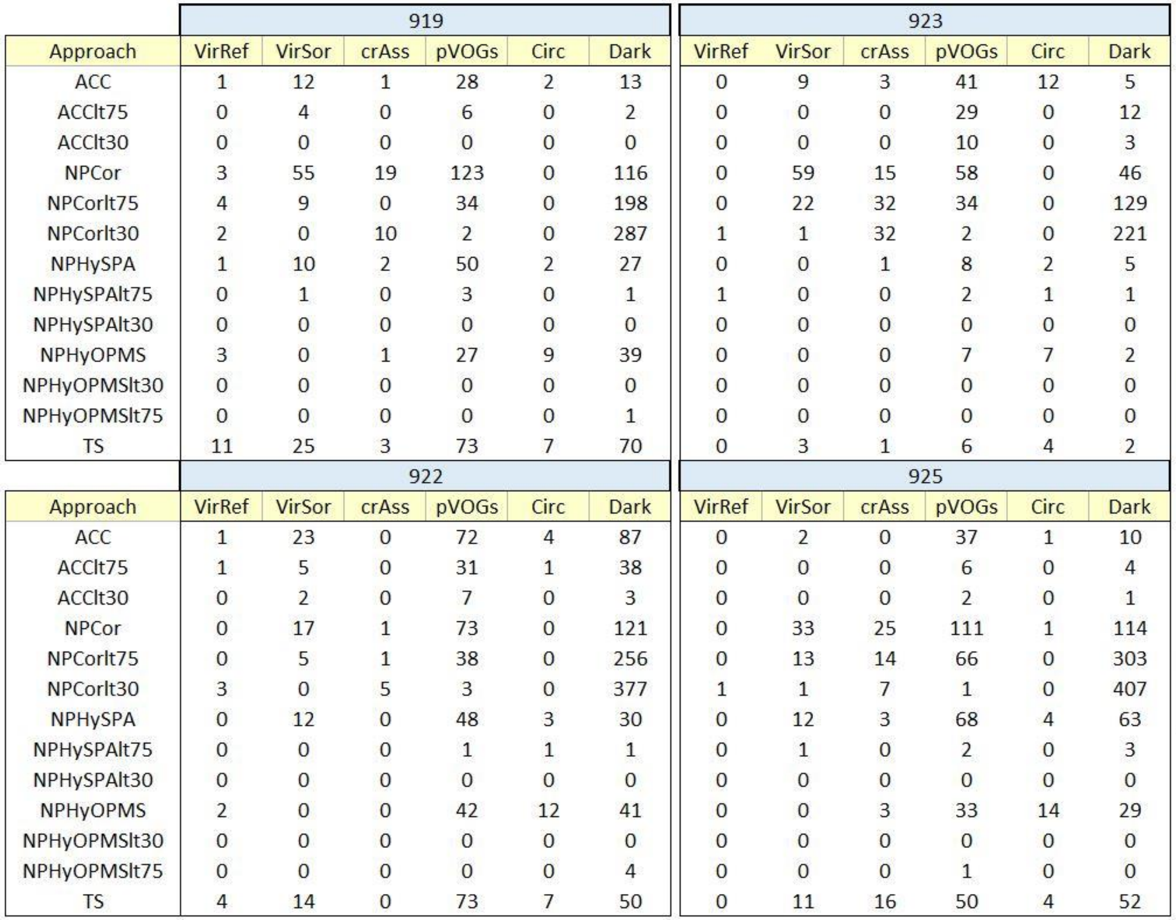
TruSeq coverage patterns were consistent across all samples and sequencing approaches, with on average 25±5 % of corrected Nanopore VLS with a breadth of coverage (BOC) of >75%, 33±3% with a BOC between 75% and 30% and 43±4% a BOC less than 30%. This suggests that the vast majority of corrected Nanopore reads would not pass the BOC filter of 75% as used in the literature and would be deemed absent from the TruSeq library. This also suggests that despite the same amplification product having been sequenced on both platforms roughly one third of Nanopore reads will pass viral inclusion criteria and yet will have been missed entirely by TruSeq libraries. A similar but far more extreme pattern was observed with the Corrected PacBio reads with 5.7% of corrected PacBio VLS with a BOC of >75%, 20.4% with a BOC between 75% and 30% and the outstanding majority of 73.8% with a BOC less than 30%.

Of the Nanopore hybrid assemblies (both SPAdes and OPERA-MS) in the NR-MPV, 94±5% had a BOC >75% and none were below 30%, suggesting short reads play a central role in these hybrid assemblies (Table 4.). This would also give greater confidence to the VLS generated by these approaches, having originated from two independent sequencing platforms and library prep methods. However, should the corrected Nanopore VLS represent genuine viral sequences which have been missed by the TruSeq approach, these sequences will be excluded from this assembly approach. Hybrid assemblies of the PacBio and TruSeq reads appear to make greater use of the PacBio reads that were not shared by TruSeq libraries. 62% of hybrid assemblies using PacBio and TruSeq (both SPAdes and OPERA-MS) had a BOC >75%, 35% of assemblies had a BOC between 75% and 30% and 3% of assemblies had a BOC below 30% (Table 4.).

The majority of VLS generated by the Accel-NGS approach appear to have also been sequenced using TruSeq (i.e. 71 ±13% of Accel-NGS with a BOC >75%) 25±9% of Accel-NGS VLS fell below the 75% BOC threshold and in three of the four samples a small minority of Accel-NGS VLS were missed by the TruSeq library entirely (7±4% within samples 922,923 and 925). Therefore the Accel-NGS reads supplement TruSeqs viromes through improved detection of VLS, but predominantly as a result of improved assembly.

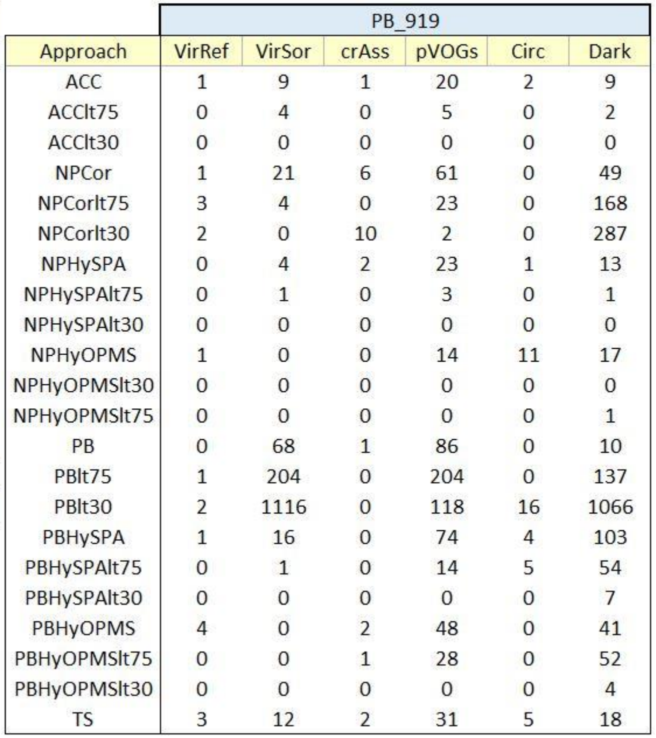
Figure 2. The NR-MPV of each sample broken down by BOC category (lighter to darker shading within each colour) and the inclusion criteria used to categorise each VLS in the NR-MPV as viral.





	919						923					
Approach	VirRef	VirSor	crAss	pVOGs	Circ	Dark	VirRef	VirSor	crAss	pVOGs	Circ	Dark
ACC	1	12	1	28	2	13	0	9	3	41	12	5
ACClt75	0	4	0	6	0	2	0	0	0	29	0	12
ACClt30	0	0	0	0	0	0	0	0	0	10	0	3
NPCor	3	55	19	123	0	116	0	59	15	58	0	46
NPCorlt75	4	9	0	34	0	198	0	22	32	34	0	129
NPCorlt30	2	0	10	2	0	287	1	1	32	2	0	221
NPHySPA	1	10	2	50	2	27	0	0	1	8	2	5
NPHySPAlt75	0	1	0	3	0	1	1	0	0	2	1	1
NPHySPAlt30	0	0	0	0	0	0	0	0	0	0	0	0
NPHyOPMS	3	0	1	27	9	39	0	0	0	7	7	2
NPHyOPMSlt30	0	0	0	0	0	0	0	0	0	0	0	0
NPHyOPMSlt75	0	0	0	0	0	1	0	0	0	0	0	0
TS	11	25	3	73	7	70	0	3	1	6	4	2

	922						925					
Approach	VirRef	VirSor	crAss	pVOGs	Circ	Dark	VirRef	VirSor	crAss	pVOGs	Circ	Dark
ACC	1	23	0	72	4	87	0	2	0	37	1	10
ACClt75	1	5	0	31	1	38	0	0	0	6	0	4
ACClt30	0	2	0	7	0	3	0	0	0	2	0	1
NPCor	0	17	1	73	0	121	0	33	25	111	1	114
NPCorlt75	0	5	1	38	0	256	0	13	14	66	0	303
NPCorlt30	3	0	5	3	0	377	1	1	7	1	0	407
NPHySPA	0	12	0	48	3	30	0	12	3	68	4	63
NPHySPAlt75	0	0	0	1	1	1	0	1	0	2	0	3
NPHySPAlt30	0	0	0	0	0	0	0	0	0	0	0	0
NPHyOPMS	2	0	0	42	12	41	0	0	3	33	14	29
NPHyOPMSlt30	0	0	0	0	0	0	0	0	0	0	0	0
NPHyOPMSlt75	0	0	0	0	0	4	0	0	0	1	0	0
TS	4	14	0	73	7	50	0	11	16	50	4	52



	PB_919					
Approach	VirRef	VirSor	crAss	pVOGs	Circ	Dark
ACC	1	9	1	20	2	9
ACClt75	0	4	0	5	0	2
ACClt30	0	0	0	0	0	0
NPCor	1	21	6	61	0	49
NPCorlt75	3	4	0	23	0	168
NPCorlt30	2	0	10	2	0	287
NPHySPA	0	4	2	23	1	13
NPHySPAlt75	0	1	0	3	0	1
NPHySPAlt30	0	0	0	0	0	0
NPHyOPMS	1	0	0	14	11	17
NPHyOPMSlt30	0	0	0	0	0	0
NPHyOPMSlt75	0	0	0	0	0	1
PB	0	68	1	86	0	10
PBlt75	1	204	0	204	0	137
PBlt30	2	1116	0	118	16	1066
PBHySPA	1	16	0	74	4	103
PBHySPAlt75	0	1	0	14	5	54
PBHySPAlt30	0	0	0	0	0	7
PBHyOPMS	4	0	2	48	0	41
PBHyOPMSlt75	0	0	1	28	0	52
PBHyOPMSlt30	0	0	0	0	0	4
TS	3	12	2	31	5	18

Supp.Table 2. The NR-MPV of each sample broken down by BOC category VLS inclusion criteria. Corresponds to Figure 2

Viral inclusion criteria

Each NR-MPV was broken down into the viral inclusion criteria used to identify VLS (Figure 2, Supp. Table 2) as a means to explore and validate the compositional patterns seen in the NR-MPV.

Across all five NR-MPVs the majority of sequences were deemed viral because they were flagged as "dark matter" (i.e. were greater than 3kb long and did not feature BLASTn alignments to the NT database, see methods). Alignment to the pVOGS database and VirSorter also contributed a considerable number of VLS, although significantly less than the "dark matter" criteria. The "dark matter" category also recruited the largest amount of TruSeq contigs (174 across all four samples). Interestingly the majority (94±5%) of corrected Nanopore reads and 78% of PacBio reads with poor coverage (less than 30 % BOC) in each sample also fell into this dark matter category and which could potentially question their validity as genuine VLS (see discussion). However, corrected long reads with poor coverage also featured in other more stringent viral categories, such as aligning to an in-house crAss database (32 corrected Nanopore reads in sample 923), being virsorter positive, (1116 corrected PacBio reads in sample PB_919) and aligning to the pVOGS database with minimum of 2 pVogs and at least 3 per 1kb (118 corrected PacBio reads in sample PB_919). This would suggest that corrected long reads do represent legitimate viral sequences which are missed by short read platforms and that they are an important addition to a virome analysis pipeline.

Recruitment of Longitudinal TruSeq viromes to the NR-MPV

The samples used in this study were donated in February 2018, 15 months after the last time-point of the longitudinal study (November 2015 to November 2016)(Shkoporov et al., 2019). In concurrence with the high levels of virome stability observed in the longitudinal study, many of the VLS in each NR-MPV were also detected in multiple longitudinal timepoints despite the 15 month gap between sampling. The BOC values for many of these VLS fluctuated above and below detection thresholds. (i.e. BOC >75%) over time which is also in concurrence with the previous longitudinal study, and referred to a transiently detected virome (TDV).

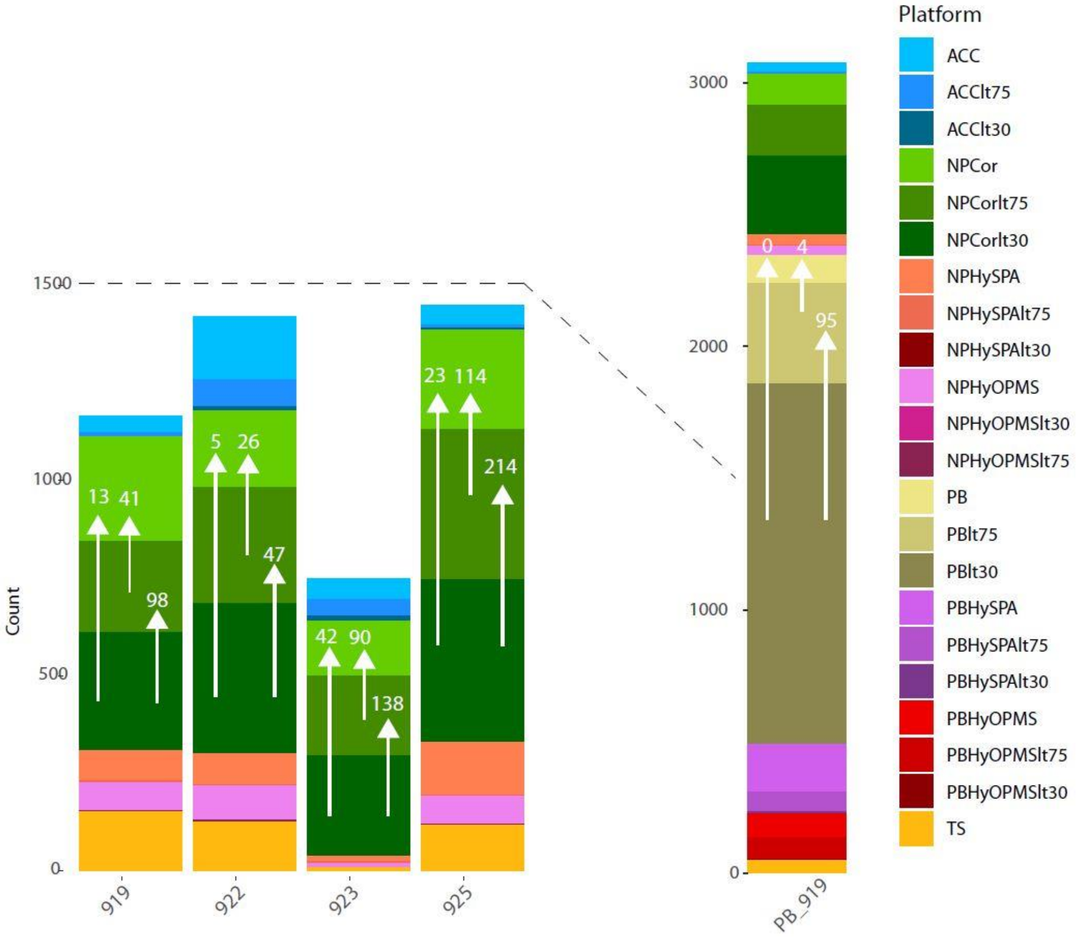


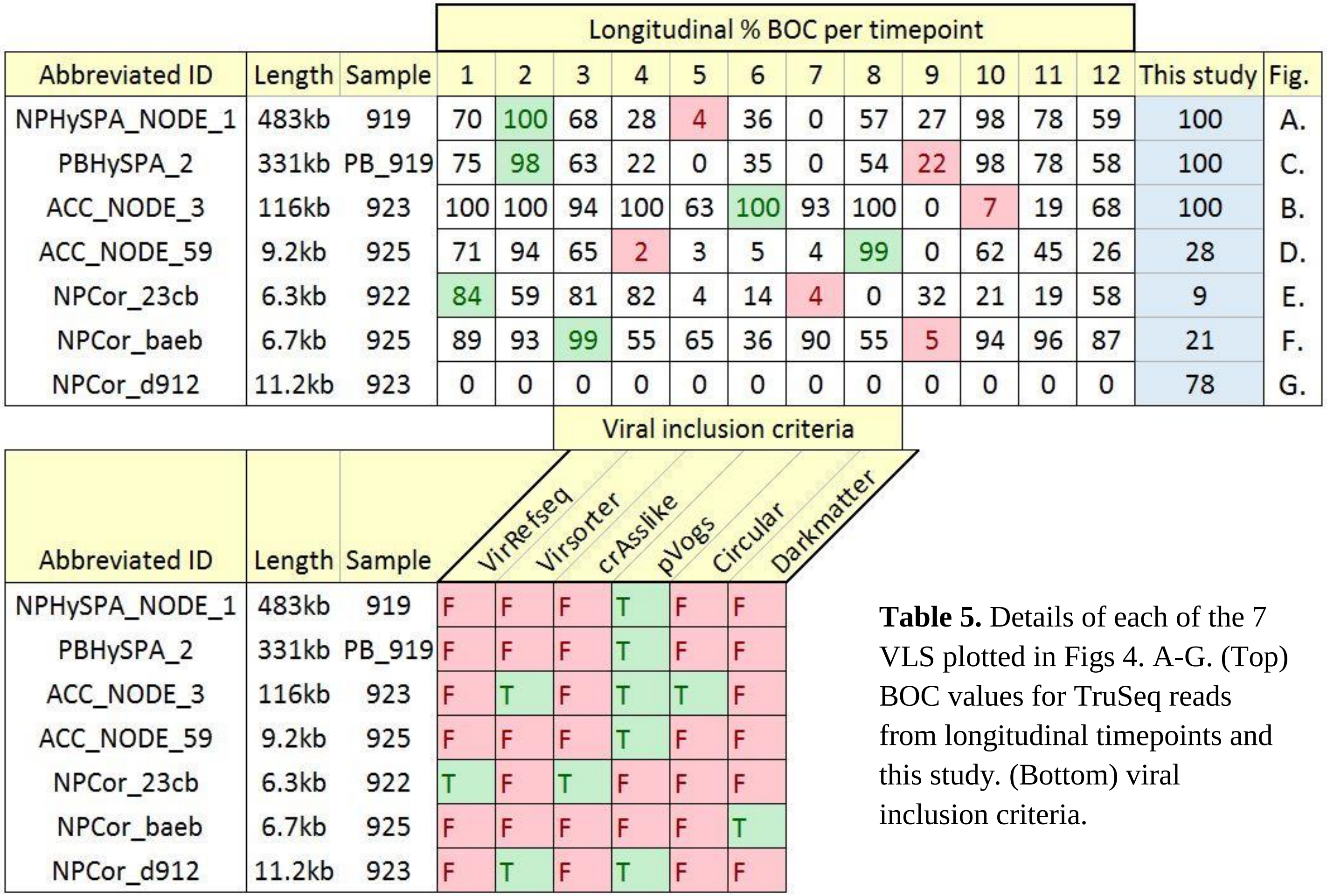
Figure 3. Composition of the NR-MPV across all samples as with Figure 1. White arrows and figures denote the number of corrected long reads which went from one BOC category when recruiting TruSeq reads in this study, to another category when recruiting TruSeq reads from another longtitudinal TruSeq library. (e.g. 19 Corrected Nanopore reads with a BOC below 30% using TruSeq reads from this study reached a BOC >75% in at least one longitudinal timepoint, and 143 reached a BOC between 30% and 75% in at least one timepoint.

To validate the VLS generated by non-TruSeq platforms and investigate if they could be detected long-term within individuals, TruSeq viromes from each individual across these 12 monthly time-points were aligned to each NR-MPV. If the coverage of a corrected long-read VLS changed from one detection threshold to another, or passed the BOC filter in a separate TruSeq library from the same individual, it would support the viral predictions within these approaches (Figure 3.). However in the majority of cases, the BOC did not increase to the point where reads would cross these BOC thresholds (i.e. from BOC below 30% to between 30% and 75%). Across all samples, 25±13% of corrected Nanopore VLS with a BOC between 75% and 30% (i.e. may be present in the TruSeq library but would not pass a BOC filter) exhibited a BOC greater than 75% in at least one longitudinal time-point from that individual. Furthermore, 23±9% of corrected Nanopore VLS with a BOC below 30% (i.e. appear to have been missed by TruSeq sequencing) yielded a BOC greater than 30% in at least one longitudinal time-point from that individual.

However, very few Nanopore VLS with A BOC below 30% passed the BOC filter in other longitudinal time-points (i.e. 7±6% went from BOC <30% to >75 in the longitudinal samples). However, even Nanopore VLS which passed BOC filters in the NR-MPV did not always remain above detection limits across longitudinal samples, with 37±16% of corrected Nanopore VLS with a BOC >75% falling below 75% in all 12 longitudinal samples. These results validate some the VLS predicted from corrected Nanopore reads which had been missed by TruSeq and suggest that TruSeq libraries fluctuate in their ability to detect certain viral sequences. Furthermore these patterns were considerably different across the corrected PacBio VLS with only 1% of VLS with a BOC below 75% reaching more than 75% in at least one longitudinal sample, 7% of VLS with a BOC below 30% reaching between 30% and 75%. VLS with a BOC below 30% did not reach detection limits of >75% BOC in at any one longitudinal sample. Similar to the corrected Nanopore VLS, 43% of corrected PacBio VLS which passed the BOC filter, and were deemed to be present in the TruSeq library of this study, did not pass the filter in any of the other 12 longitudinal samples.

Visualisation of particular VLS.

Results to this point have given a broad overview of trends within the MPV of each sample, the contribution of each sequencing approach to our final view of the virome and how they compare to the TruSeq libraries and assemblies. These trends were validated by plotting annotated genome maps of a selection of VLPs from alternative sequencing approaches in the NR-MPV and determining how they recruited TruSeq reads and assemblies. Detailed validation at the level of individual VLS is crucial in large scale virome pipelines with numerous analysis steps, as each step can introduce bias or analysis artefacts which skew findings (i.e. palindromic VLS outlined in methods). Table 5 describes the BOC values for each of the 12 longitudinal TruSeq libraries across the nine Genome maps plotted Figure 4 A – H.



			Longitudinal % BOC per timepoint													
Abbreviated ID	Length	Sample	1	2	3	4	5	6	7	8	9	10	11	12	This study	Fig.
NPHySPA_NODE_1	483kb	919	70	100	68	28	4	36	0	57	27	98	78	59	100	A.
PBHySPA_2	331kb	PB_919	75	98	63	22	0	35	0	54	22	98	78	58	100	C.
ACC_NODE_3	116kb	923	100	100	94	100	63	100	93	100	0	7	19	68	100	B.
ACC_NODE_59	9.2kb	925	71	94	65	2	3	5	4	99	0	62	45	26	28	D.
NPCor_23cb	6.3kb	922	84	59	81	82	4	14	4	0	32	21	19	58	9	E.
NPCor_baeb	6.7kb	925	89	93	99	55	65	36	90	55	5	94	96	87	21	F.
NPCor_d912	11.2kb	923	0	0	0	0	0	0	0	0	0	0	0	0	78	G.

			Viral inclusion criteria					
Abbreviated ID	Length	Sample	VirRefseq	Virsorter	crAsslike	pVogs	Circular	Darkmatter
NPHySPA_NODE_1	483kb	919	F	F	F	T	F	F
PBHySPA_2	331kb	PB_919	F	F	F	T	F	F
ACC_NODE_3	116kb	923	F	T	F	T	T	F
ACC_NODE_59	9.2kb	925	F	F	F	T	F	F
NPCor_23cb	6.3kb	922	T	F	T	F	F	F
NPCor_baeb	6.7kb	925	F	F	F	F	F	T
NPCor_d912	11.2kb	923	F	T	F	T	F	F

Table 5. Details of each of the 7 VLS plotted in Figs 4. A-G. (Top) BOC values for TruSeq reads from longitudinal timepoints and this study. (Bottom) viral inclusion criteria.

Figure 4 A. Is the genome map of a linear VLS from a hybrid SPAdes assembly using Nanopore and truSeq reads (the linear sequence is presented as circular to maximise space for annotation labels). This 483kb VLS was the longest across all samples and platforms, was the longest sequence overall (Tables 2 and 3), and is among the largest phage genomes in sequence databases. As was discussed above in "Read and final assembly counts" it is intriguing that the longest sequence did not involve a PacBio read, despite the increased sequencing depth and minimum read length of the PacBio library relative to those of the Nanopore. This sequence did not align to the viral RefSeq database but did share a 2kb region (at 72-74% identity) with *Clostridium* and *Longibaculum* species. However, top alignments of predicted tRNA sequences all aligned to *Prevotella* species, a finding which was supported by CRISPR protospacer predictions. tRNA and CRISPR host predictions and the sheer size of the VLS are supported by recent reports megaphage infecting *Prevotella* species (Devoto et al., 2019) and suggest it could be somewhat related to Lac Phage. However, BLASTn alignments to Lac Phage genomes did not yield any results. Given the lack of nucleotide homology shared across other recently discovered phage families such as the extended crAssphage family (Guerin et al., 2018) these findings are not entirely surprising. Future characterisation may be possible using techniques which are better suited to distant homology, such as those outlined by Guerin et al. and Yutin et al. (Guerin et al., 2018;Yutin et al., 2018) to investigate a possible relationship to the Lac Phage family. Despite having recruited TruSeq reads across the entire length of the sequence (BOC 100%), TruSeq assemblies were unable to resolve the VLS, generating 3 scaffolds which made up ~ 75% of the genome, and failed to generate assemblies which could span a region between ~ 130kb and ~260kb and pass viral inclusion criteria. It is possible that this gap in TruSeq assemblies represents hypervariable or low coverage regions associated with host interaction as described by Warwick-Dugdale et al.(Warwick-Dugdale et al., 2019) and highlight the benefits of combined long and short read sequencing approaches for the virome. This VLS fluctuated above and below the threshold of detection across longitudinal TruSeq libraries, with a BOC >75% in 3/12 timepoints and a BOC <30% in 4/12 timepoints and being missed entirely in one timepoint (BOC 0%). This extreme fluctuation in the level of detection may be linked to fluctuations in abundance (i.e. shifts in lytic or lysogenic replication or predator/prey dynamics) which become exaggerated by MDA selection bias and drift bias (see discussion).

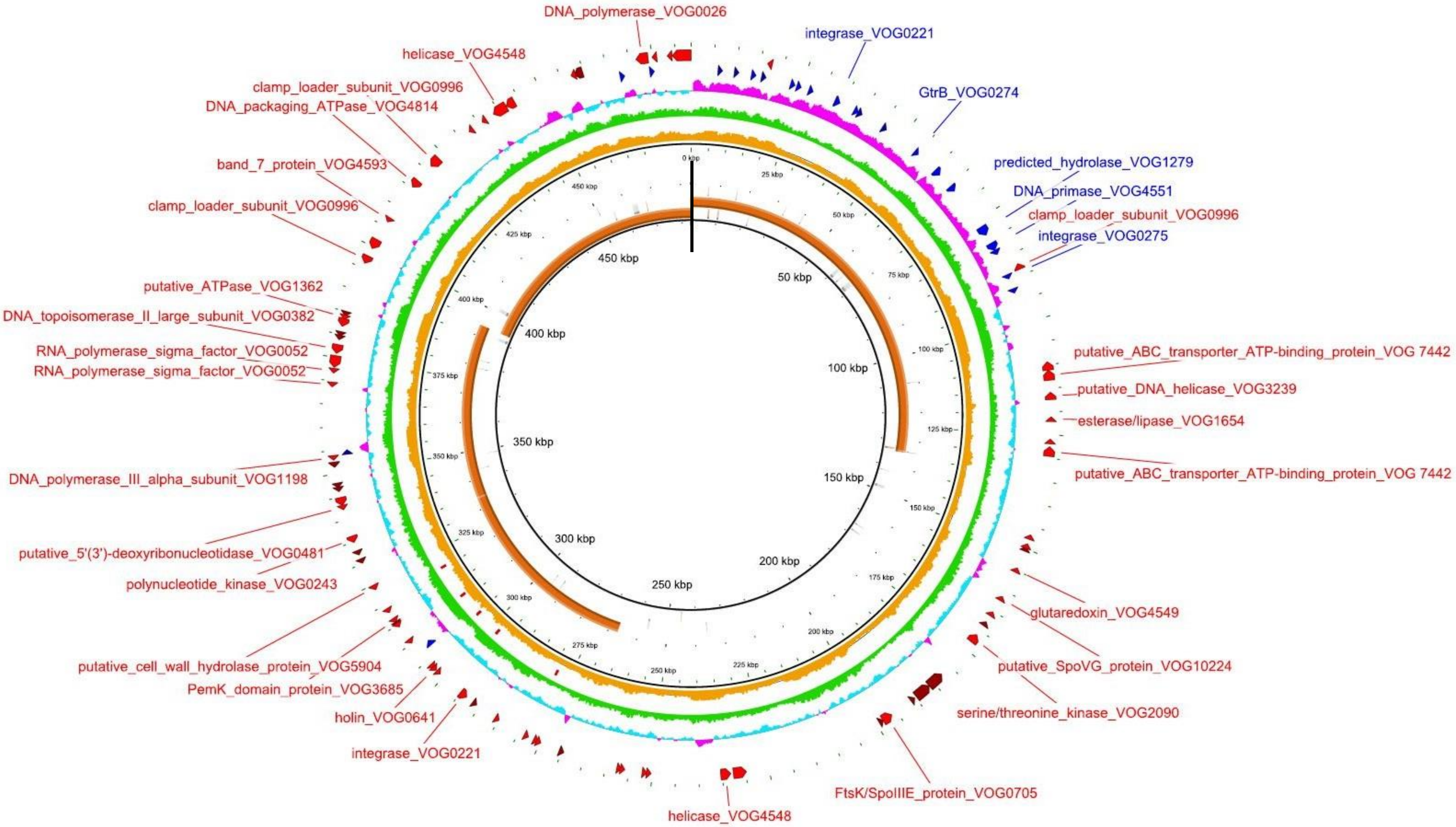

DNA_polymerase_VOG0026
integrase_VOG0221
GtrB_VOG0274
predicted_hydrolase_VOG1279
DNA_primase_VOG4551
clamp_loader_subunit_VOG0996
integrase_VOG0275
putative_ABC_transporter_ATP-binding_protein_VOG 7442
putative_DNA_helicase_VOG3239
esterase/lipase_VOG1654
putative_ABC_transporter_ATP-binding_protein_VOG 7442
glutaredoxin_VOG4549
putative_SpoVG_protein_VOG10224
serine/threonine_kinase_VOG2090
FtsK/SpoIIIE_protein_VOG0705
helicase_VOG4548
integrase_VOG0221
holin_VOG0641
PemK_domain_protein_VOG3685
putative_cell_wall_hydrolase_protein_VOG5904
polynucleotide_kinase_VOG0243
putative_5'(3')-deoxyribonucleotidase_VOG0481
DNA_polymerase_III_alpha_subunit_VOG1198
RNA_polymerase_sigma_factor_VOG0052
RNA_polymerase_sigma_factor_VOG0052
DNA_topoisomerase_II_large_subunit_VOG0382
putative_ATPase_VOG1362
clamp_loader_subunit_VOG0996
band_7_protein_VOG4593
DNA_packaging_ATPase_VOG4814
clamp_loader_subunit_VOG0996
helicase_VOG4548
0 kbp
50 kbp
100 kbp
150 kbp
200 kbp
250 kbp
300 kbp
350 kbp
400 kbp
450 kbp

Figure 4.A. (overleaf) Gview plot of the longest VLS generated by this study (hybrid assembly Nanopore + TruSeq). The VLS is linear and has been circularised for plotting purposes (black intersecting line highlights beginning and end of sequence). Outer to inner rings are as follows; Forward (blue) and reverse (red) CDS annotated with pVOGs, GC skew (Blue/Purple), max (green) and min (red) longitudinal read coverage, TruSeq read coverage from this study, TruSeq assemblies from this study.

Figure 4 B. is a genome map for the longest VLS which featured PacBio reads, and similar to Figure 4 A was linear and fully covered by TruSeq reads generated in this study (BOC 100%). In contrast to Figure 4 A, this VLS shared homology to multiple TruSeq assemblies across the entirety of its length and made all four TruSeq assemblies redundant (90% identity across 90% of the TruSeq length). This VLS also fluctuated in and out of the threshold of detection across longitudinal TruSeq libraries (four time-points below BOC 30%, four time-points above BOC 75%). Annotation and host predictions were somewhat contradictory and inconclusive. Top CRISPR protospacer predictions aligned to *Fusobacterium* species and contradicted tRNA predictions which aligned to *Bacillus* and *Staphylococcus* species. There were no hits to Viral Refseq and the longest alignments (>4kb, 74-75% ID) to NT were to the class *Mollicutes.* Both of these VLS highlight the benefits of using long reads to improve the assembly of viral sequences which were detected but fragmented using TruSeq libraries alone.

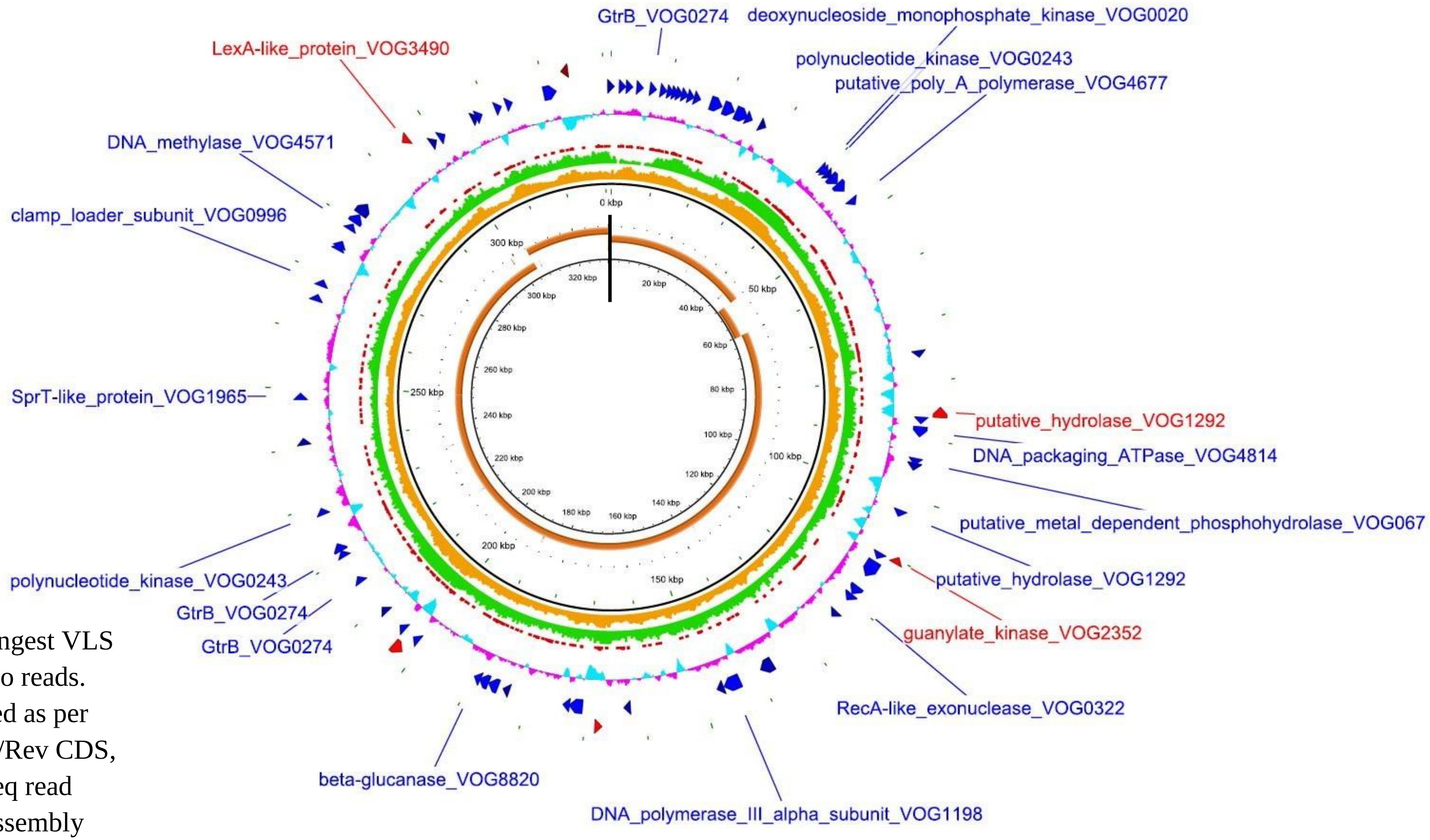

GtrB_VOG0274
deoxynucleoside_monophosphate_kinase_VOG0020
polynucleotide_kinase_VOG0243
putative_poly_A_polymerase_VOG4677
LexA-like_protein_VOG3490
DNA_methylase_VOG4571
clamp_loader_subunit_VOG0996
SprT-like_protein_VOG1965
polynucleotide_kinase_VOG0243
GtrB_VOG0274
GtrB_VOG0274
beta-glucanase_VOG8820
DNA_polymerase_III_alpha_subunit_VOG1198
RecA-like_exonuclease_VOG0322
guanylate_kinase_VOG2352
putative_hydrolase_VOG1292
putative_metal_dependent_phosphohydrolase_VOG067
DNA_packaging_ATPase_VOG4814
putative_hydrolase_VOG1292
0 kbp
50 kbp
100 kbp
150 kbp
200 kbp
250 kbp
300 kbp
20 kbp
40 kbp
60 kbp
80 kbp
100 kbp
120 kbp
140 kbp
160 kbp
180 kbp
200 kbp
220 kbp
240 kbp
260 kbp
280 kbp
300 kbp
320 kbp

Figure 4.B. Longest VLS featuring PacBio reads. Rings are plotted as per Figure. 4.A.Fw/Rev CDS, GCskew, TruSeq read coverage and assembly coverage

Figure 4 C represents the longest circular (which could be taken as a strong proxy for being complete) VLS across all sequencing approaches and samples. Interestingly, this originated from the Accel-NGS library and did not feature in the long-reads from either platform. As with the previous examples, this VLS exhibited a BOC of 100% with TruSeq reads but was resolved in six separate TruSeq assemblies across the entirety of its length. Furthermore, this VLS fluctuated fluctuated in and out of the threshold of detection across longitudinal samples, albeit with a greater number of high coverage cases (seven time-points with a BOC above 90%). This VLS did not align to Viral RefSeq and had one short (80% ID across 1.5kb) with *Anaerostipes hadrus* (order Clostridia) when aligned to NT. Predicted tRNAs did not have any alignments to either NT or bacterial RefSeq, CRISPR protospacers had top hits to *Lachnoanaerobaculum*, another genus within the order Clostridia. This VLS highlights the impact of MDA and library prep on final Virome assemblies and suggests that virome analysis pipelines can also be improved without changing sequencing platform.

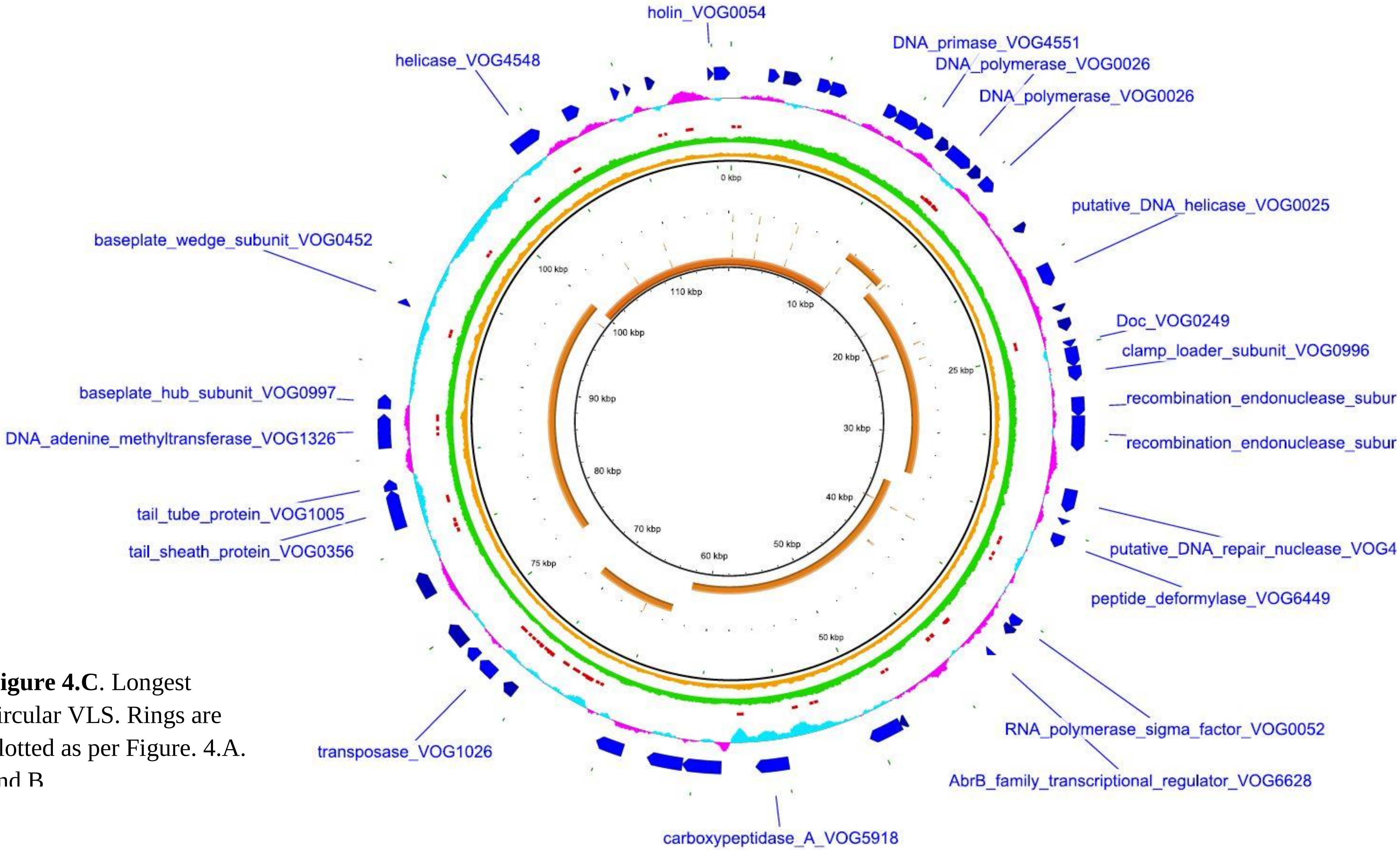

holin_VOG0054
DNA_primase_VOG4551
DNA_polymerase_VOG0026
DNA_polymerase_VOG0026
helicase_VOG4548
putative_DNA_helicase_VOG0025
baseplate_wedge_subunit_VOG0452
Doc_VOG0249
clamp_loader_subunit_VOG0996
baseplate_hub_subunit_VOG0997
recombination_endonuclease_subur
DNA_adenine_methyltransferase_VOG1326
recombination_endonuclease_subur
tail_tube_protein_VOG1005
tail_sheath_protein_VOG0356
putative_DNA_repair_nuclease_VOG4
peptide_deformylase_VOG6449
RNA_polymerase_sigma_factor_VOG0052
transposase_VOG1026
AbrB_family_transcriptional_regulator_VOG6628
carboxypeptidase_A_VOG5918
0 kbp
100 kbp
110 kbp
10 kbp
100 kbp
20 kbp
25 kbp
90 kbp
30 kbp
80 kbp
40 kbp
70 kbp
50 kbp
60 kbp
75 kbp
50 kbp

Figure 4.C. Longest circular VLS. Rings are plotted as per Figure. 4.A. and B

In Figure 4. A-C. TruSeq approaches could detect but not fully assemble VLS from alternative sequencing approaches. Figure 4. D-F. represent cases which appear to have been missed by the TruSeq libraries entirely (BOC < 30%) in this study but passed the threshold of detection in longitudinal timepoints (BOC >75%). This validates these VLS and highlights how alternative sequencing approaches not only improve truSeq assembly but can detect sequences which are missed by TruSeq. Figure 4. D. depicts a linear Accel-NGS VLS and is the longest (9.2kb) VLS with a BOC 28% of Truseq reads in this study and BOC >75% in at least one longitudinal TruSeq library. However, in half of the longitudinal timepoints (6/ 12) BOC remained below 30%, reaching above 90% in two timepoints. This case highlights the potential impact of MDA and library prep in the detection limits of TruSeq libraries. This VLS did not feature alignments to either viral RefSeq or NT, did not feature predicted tRNAs and had a top CRISPR protospacer alignment to *Bacteroides dorei.*

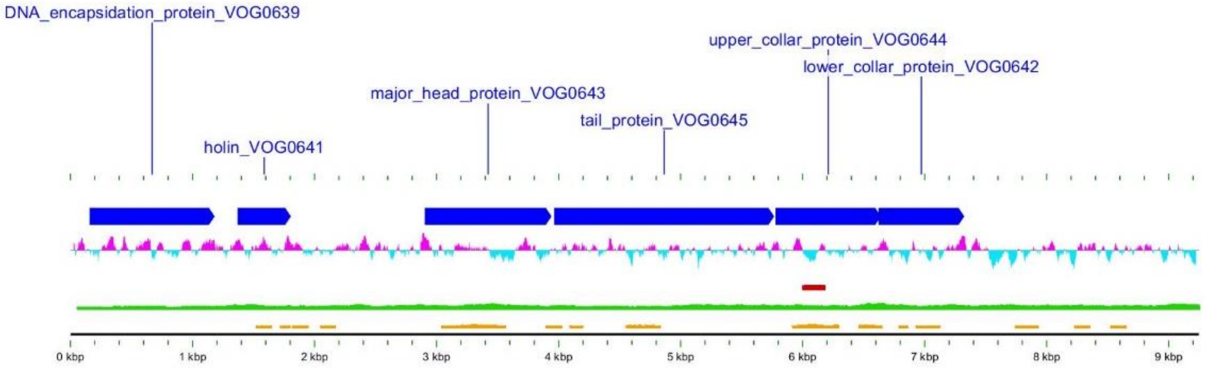


Figure 4.D. Accel-NGS VLS plotted as per Figure 4.A. Poor Coverage by TruSeq reads in this study (orange) and fluctuation from full coverage (green) to poor covergage (red) in longitudinal samples.

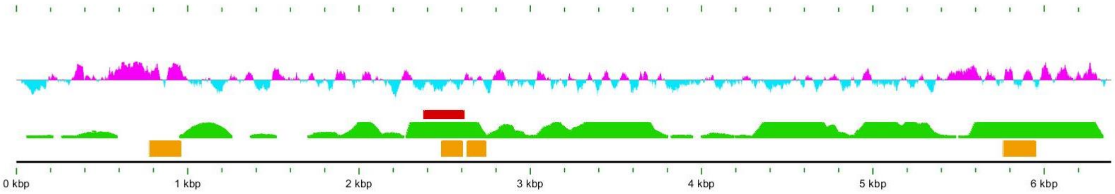


Figure 4.E. Corected Nanopore read plotted as per Figure 4.A. This sequence shared significant homology with crAssphage, but did not feature alignments to the pVOGs database. TruSeq read coverage in red, green and orange as before

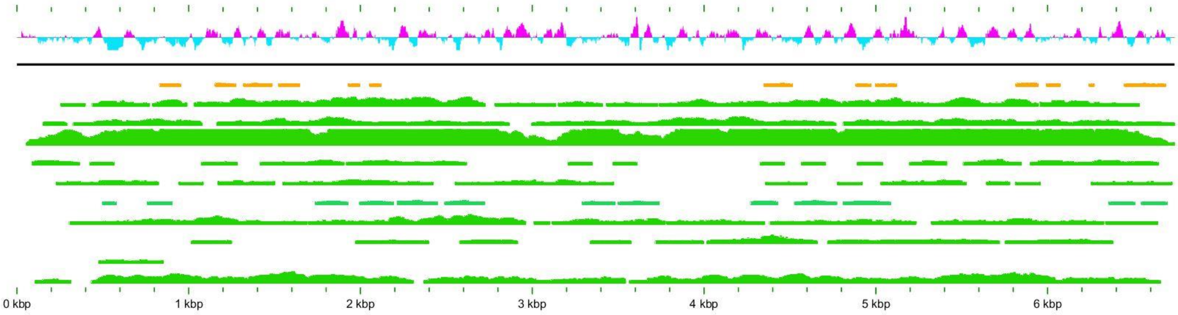


Figure 4.F. Gview plot of Corected Nanopore read. Plots (top to bottom) GC skew (blue/purple), Truseq read coverage from this study (orange). The remaining 10 plots are longitudinal TruSeq read coverage (timepoints 1-10) highlighting high BOC

In Contrast to Figure 4 D, which depicted an unamplified sample, Figure 4 E represents a VLS from the same MDA amplification as that sequenced in the TruSeq library. However this VLS was also entirely missed by the TruSeq library in this study and detected in multiple (3/12) longitudinal TruSeq libraries from the same individual. This sequence appears to be a fragment of a crAssphage genome as it shares multiple (albeit partial, longest 4.6 kb with 88% identity) alignments to prototypical crAssphage in both NT and viral RefSeq databases. It did not feature predicted tRNAs and had top CRISPR protospacer alignments to *Fusobacterium nucleatum*. This suggests that the limitations in TruSeq detection may not be due to MDA and that ubiquitous members of the viral community such as crAssphage may be undetected as a result of library prep methods and sequencing platforms. By underrepresenting potentially shared viral sequences, this also suggests that sequencing approaches themselves could exaggerate the high levels of inter-individuality which hamper virome analysis. Similar to Figure 4 E, the VLS depicted in Figure 4 F also originated from the same amplification product as that which was sequenced on the TruSeq library, was missed entirely in the TruSeq library of this study, but featured regularly in every other time-point within that individual. However, this sequence was entirely unknown and did not feature alignments to any databases, CRISPR spacers or tRNA hits. This sequence represents the numerous long read sequences which did not feature in TruSeq libraries, or reference databases, but did appear throughout multiple time-points, suggesting that these occurrences may be more numerous than we are able to characterise, due to limitations in databases, and aligning to databases which are themselves based on short read sequencing.

The VLS depicted in Figs 4 A-F are clear examples of alternative sequencing approaches improving either the assembly or detection of viral sequences in TruSeq libraries. In contrast, Figure 4 G is an example of Corrected Nanopore VLS which are harder to validate and suggest that sequencing artefacts can be carried through to final VLS by both TruSeq and/or Corrected Nanopore approaches. However, it is difficult to validate one approach over another. The VLS in Figure 4 G represents a linear corrected Nanopore read which was fully represented in the corresponding TruSeq library in this study (BOC 78%). However, this VLS could not be validated with longitudinal TruSeq reads as it did not recruit reads from any longitudinal time-points.

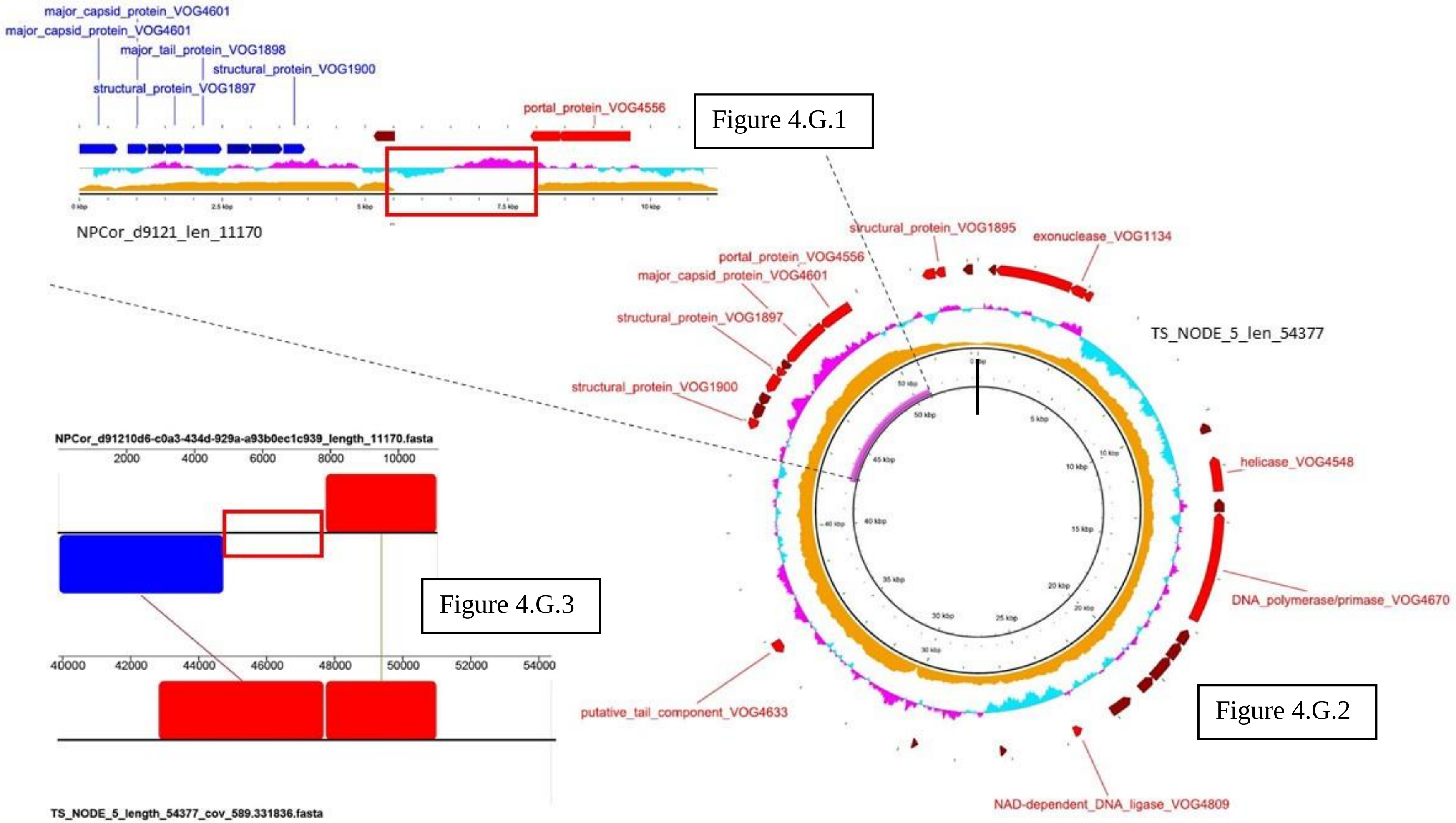

major_capsid_protein_VOG4601
major_capsid_protein_VOG4601
major_tail_protein_VOG1898
structural_protein_VOG1900
structural_protein_VOG1897
portal_protein_VOG4556
0 kbp
2.5 kbp
5 kbp
7.5 kbp
10 kbp
NPCor_d9121_len_11170
Figure 4.G.1
structural_protein_VOG1895
exonuclease_VOG1134
portal_protein_VOG4556
major_capsid_protein_VOG4601
structural_protein_VOG1897
structural_protein_VOG1900
TS_NODE_5_len_54377
helicase_VOG4548
DNA_polymerase/primase_VOG4670
putative_tail_component_VOG4633
NAD-dependent_DNA_ligase_VOG4809
Figure 4.G.2
NPCor_d91210d6-c0a3-434d-929a-a93b0ec1c939_length_11170.fasta
2000
4000
6000
8000
10000
Figure 4.G.3
40000
42000
44000
46000
48000
50000
52000
54000
TS_NODE_5_length_54377_cov_589.331836.fasta

Figure 4.G.1 Gview plot of Corected Nanopore read. Plots (top to bottom) Forward and reverse CDS (red/blue), GC skew (blue/purple), Truseq read coverage from this study (orange), no longitudinal reads were recruited. Red box highlights gap in the corrected read with no TruSeq coverage

Figure 4.G.2 Gview plot TruSeq assembly which shared significant homology with the Nanopore read in Figure 4.G.1. The VLS is linear and has been circularised for plotting purposes(black intersecting line highlights beginning and end of sequence). Plots (Outer to inner ring) Forward and reverse CDS (red/blue), GC skew (blue/purple), Truseq read coverage from this study (orange), alignment to corrected Nanopore read in Figure 4.G.1 (pink). Note no gap is present in TruSeq assembly and all CDS are in the same orientation.

Figure 4.G.3 Mauve alignment highlighting the alignment gap in the corrected Nanopore read and the inversion between the two sequences.

This VLS was VirSorter positive and passed pVOGs viral inclusion criteria (Table 5.) which could suggest that it is a legitimate viral sequences which was consistently below the limits of detection for the 12 longitudinal timepoints and only reached levels of detection in the 15 months between the study sampling dates. Alternatively, this VLS may have been newly acquired between the two studies. Examining the alignment of TruSeq reads and assemblies to this VLS in highlighted some unusual behaviour making the sequence difficult to validate. The VLS in Figure 4 G1 is the Corrected Nanopore read (abbreviated read ID "NPCor_d9212") with positive and negative CDS in red and blue and TruSeq read recruitment in orange as before. TruSeq reads do not align to a 2.5kb region between (5.5kb-8kb, red box) which also did not recruit known pVOGs and centred on a shift in GC skew. Having been included in the final NR-MPV of sample 923 this sequence was not made redundant by other sequences (90% ID over 90% length).This VLS also did not make any TruSeq assemblies redundant. However, it did align almost entirely to an 8kb region within much longer (54.3 kb) TruSeq VLS (Figure 4. G.2) between the regions of 43 and 51kb (pink bar, inner ring). In agreement with the TruSeq read recruitment, the 2.5kb region was not present in the longer TruSeq assembly. This gap was long enough (relative to the Corrected Nanopore read) for the redundancy step to deem these sequences sufficiently different to both be kept in the final NR-MPV. For this reason, the origin and validity of this gap is crucial to determining whether the Corrected Nanopore read is a more accurate representation of this VLS by including regions which TruSeq reads had missed, or if it introduced sequencing anomalies which are inflating its length and potentially skewing the representation of corrected long-reads in the final NR-MPV.

Figure 4 G3 depicts a mauve alignment of this region highlighting the gap in the Corrected Nanopore VLS which was absent in the TruSeq assembly. Furthermore this Mauve alignment highlights an inversion of the region downstream of this gap in the corrected Nanopore read (i.e. major capsid and tail proteins on the Nanopore read are in the opposite orientation to the portal protein VOG4556, but are in the same orientation on the TruSeq assembly). As described in the introduction, MDA steps that are often required to generate sufficient quantities of virome DNA for sequencing libraries can introduce a number of biases, including the introduction of chimeras (Lasken and Stockwell, 2007). Given the inversion seen here, it is possible that the

Nanopore read has carried a chimeric repeat through to the final VLS and by inflating the length of the original template sequence, has been incorrectly retained in the final NR-MPV. However, as the same amplified DNA sample was sequenced on both TruSeq and NanoPore platforms this would also suggest that Nanopore reads are more susceptible to carrying chimeric repeats from MDA than the corresponding TruSeq reads and assemblies. Contrary to this, the current proposed mechanism for the introduction of chimeric inversions in MDA reactions, suggest that a region of the template sequence is deleted and flanking regions are recombined in an inverted orientation. Given that the corrected Nanopore read contains a sequence which is absent in the TruSeq assembly, this would suggest that the inversion and deletion event may have occurred in the TruSeq assembly and that the Corrected Nanopore read represents the template sequence more accurately.

In this study, both of these cases are equally likely and cannot be resolved. It is therefore difficult to say if Figure 4 G1 represents a case where Corrected Nanopore reads have overcome an MDA artefact and given a more accurate representation of the VLS, or if they are more prone to falling victim to MDA artefacts and must be treated with caution.

Discussion

The human gut virome represents one of the biggest gaps in our understanding of human gut microbiome and poses unique analysis challenges. It is heavily dependent on sequence-based analysis methods and *de novo* bioinformatic tools which introduce bias and skew the composition viral community members. The majority of our current understanding of the virome is based on amplified short read libraries. While this technique has been shown to introduce bias, it is not fully known how this bias impacts the final virome composition. As far as we are aware this study is the first to explore long read sequencing in the gut virome and investigate the limitations of MDA Short read sequencing. We explore five combinations of sequencing approach and library prep methods including corrected long reads, hybrid assemblies and short read assemblies with and without MDA amplification. By pooling the VLS generated by each approach per individual and removing redundancy it was possible to compare the contribution of each approach to the final “consensus virome” or NR-MPV. As amplified short read libraries (TruSeq in this study) are the foundation of current virome research, comparisons were performed relative to this approach.

The majority of VLS in the pooled “consensus virome” originated from alternative sequencing approaches meaning that they either extended or scaffolded existing TruSeq assemblies, or generated entirely new VLS which had been missed by TruSeq assemblies and/or sequencing reads. Both of these cases occurred across all samples and we describe examples where alternative sequencing approaches improved both the detection and assembly of TruSeq libraries. This highlights the impact of sequencing approach on the viral sequences available for downstream analysis.

The NR-MPV for each sample was dominated by corrected long reads with poor TruSeq read coverage meaning that the long read platforms generated a large number of VLS which did not feature in TruSeq libraries, despite identical MDA amplified DNA samples having been sequenced with both approaches. This implies that either long-read platforms generate large amounts of invalid sequences or that bias within the long read and TruSeq library prep methods skew the composition of the sample in entirely different directions. Furthermore, as the Accel NGS library generated fewer VLS with poor TruSeq coverage than the long-read platforms. This suggests that the choice of platform (i.e. NanoPore vs. PacBio vs. HiSeq) has a greater impact on the final virome composition than the impact of MDA within platforms.

We validated the VLS which would have otherwise been fragmented or undetected using longitudinal samples from each individual and tracked their levels of detection across time. Many VLS fluctuated above and below the thresholds of detection across time which we suggest may reflect fluctuations in abundance that are exaggerated by MDA bias. However we also report the long-term detection (across 27 months in total) of a number of these VLS, in agreement with the previous longitudinal virome study (Shkoporov et al., 2019). It has also been established that there is an abundance threshold with short read libraries, below which all current assembly programs struggle with genome recovery and fragmentation (Sutton et al., 2019a). By visualising and annotating a selection of these VLS it was possible to further validate and describe in detail instances where alternative sequencing approaches addressed limitations in TruSeq detection and assembly. However, in some cases, the discrepancies between short and long read VLS are difficult to explain. We propose that these discrepancies are closely linked to MDA artefacts, but it is not clear whether short or long reads are more capable of overcoming these artefacts. Below we discuss the finer details of the findings in the context of the analysis methods, their implications for past and future virome studies and the future prospects of this analysis.

Detection thresholds, drift bias of MDA samples and long-term stability

In concurrence with the original longitudinal study, high coverage VLS in this study were often found at multiple time-points. However, defining detection is difficult on at a read level. Previous studies which defined detection using a single read (Manrique et al., 2016) were called into scrutiny (Clooney et al., 2019;Gregory et al., 2019;Shkoporov et al., 2019) as this criteria does not account for shared sequences (i.e. repeats or gene cassettes) across distantly related viral families (Iranzo et al., 2016). Additionally, even VLS which stand out in a given sample as a result of having recruited high numbers of reads and that may even be differentially abundant across cohorts (Zuo et al., 2019) can be the result of spurious read alignments (Sutton et al., 2019b). In these cases, recruited reads are stacked over short regions of the genome rather than evenly distributed and also represent spurious shared sequences rather than confirming the existence of the VLS in a given sample. For this reason, breadth of coverage filters have been recommended by a number of virome studies as a means to differentiate spurious read alignments from shared VLS (Roux et al., 2017;Clooney et al., 2019). However, the application of rigid filters such as 75% have limitations and ultimately determine what is defined as present or absent in a sample. Consequently, this study analysed a spectrum of breadth of coverage (BOC) values across the dataset as a means to infer the presence or absence of a VLS in a given TruSeq library rather than stating it outright. All of the VLS examined in detail (Table 5, Figure 4 A-G) fluctuated above and below the rigid detection threshold of BOC 75% which has featured in previous studies, which is in agreement with the "transiently detected virome" in the original longitudinal study (Shkoporov et al., 2019). However, given the detection limitations of amplified TruSeq libraries highlighted by this study, it is likely that aspects of this "transient detection" are linked to the sequencing approach itself. This also means that the human gut virome may well exhibit an even greater degree of longitudinal stability than we had previously thought. It is possible that this transient detection is linked to fluctuations in abundance viral community members (i.e. predator prey dynamics, or shifts in the lysogenic/ lytic replication cycles of temperate phage). These initial differences in abundance are in turn, exaggerated by MDA "drift bias" as amplification products generated early in the MDA reaction eventually dominate the final sample. When paired with bias within the TruSeq library prep or the platform itself this could lead to VLS which had been present at low abundance being underrepresented or excluded from the sequencing entirely.

The limited ability of amplified short read libraries to fully represent the virome, as highlighted in this study, also has significant implications for the high levels of inter-individual variation associated with virome data. High levels of inter-individual variation complicate virome analyses, as without shared features across cohorts, it is not possible to identify patterns in the virome associated with features of the cohort (i.e. health vs. disease) (Clooney et al., 2019). This is individuality is believed to be linked to rapid evolutionary rates in viral communities (Minot et al., 2013) and the assembly or strain level of resolution at which virome studies are carried out (Clooney et al., 2019;Sutton et al., 2019a). However, similar to the fluctuation in longitudinal detection within individuals as discussed above, the results of this study suggests that high levels of inter individuality may be exaggerated by the detection limits of amplified short read sequencing libraries themselves. Similar to the longitudinal stability, this implies that greater numbers of viral sequences may be shared across individuals than had previously been thought. The observations of this study suggest that alternative sequencing approaches are capable of addressing these detection issues. This makes alternative sequencing approaches a promising means of increasing the number of shared viral sequences across individuals in future virome studies and complementary to cluster-based methods of lowering inter-individual variation(Clooney et al., 2019).

Corrected long-reads with low levels of TruSeq coverage

The long-read error correction program used in this study (FMLRC) (Wang et al., 2018) uses a kmer-based correction method to create de Bruijn graphs from both long and short read datasets. An FM-index is then used represent all de Bruijn graphs and used to generate a consensus sequence. This means that although long read correction does not depend directly on the recruitment of TruSeq reads, it would be impacted by the amount of shared sequences between the two libraries. Subsequently, long-reads which were poorly sequenced in TruSeq libraries are likely to have been corrected to a lower degree than those which were fully represented. The majority of these long reads with poor TruSeq coverage were also classified as “dark matter”, meaning they did not share nucleotide homology with reference databases. It is possible that this “dark matter” dominance in corrected long-reads with poor TruSeq coverage is linked to uncorrected long-read errors (Myers, 2014) that prevents successful alignment to databases. Furthermore, if long reads were insufficiently corrected the high error profile may have hampered the BLASTn-based alignments which were used throughout the study to remove redundancy and predict viral sequences. If BLASTn could not find sufficient redundancy within each long read library or between long read VLS and those generated by other approaches, the long-read representation in the NR-MPV would have been artificially inflated.

However, the databases to which these corrected long reads were aligned have also been founded primarily on short read sequencing. Furthermore these databases are known to represent extremely minor fractions of the virome as a whole (Roux et al., 2015b). Therefore, it is also possible that these long “dark matter” reads represent legitimate undiscovered viral sequences that are undetected by TruSeq reads (i.e. the ‘unknown unknowns’ of the virome). This lack of detection which could be caused by low abundance or genomic features of the VLS themselves which are negatively selected by amplified short read libraries. That considered, not all low-coverage corrected long reads were “dark matter” as highlighted by Figure 4 D. This corrected Nanopore VLS represented a fragment of a crAssphage, an important member of the gut virome which had not been detected by the TruSeq library, despite the same amplified DNA sample having been sequenced on both TruSeq and Nanopore platforms. Due to the relative novelty of these long-read platforms and their use in metagenomics, we do not know the full extent their of detection limits relative to amplified short read libraries and vice versa. Pilot studies such as this are key to building and understanding of how these technologies could improve our view of the virome, despite often generating as many questions as they seem to answer.

Future Prospects and limitations

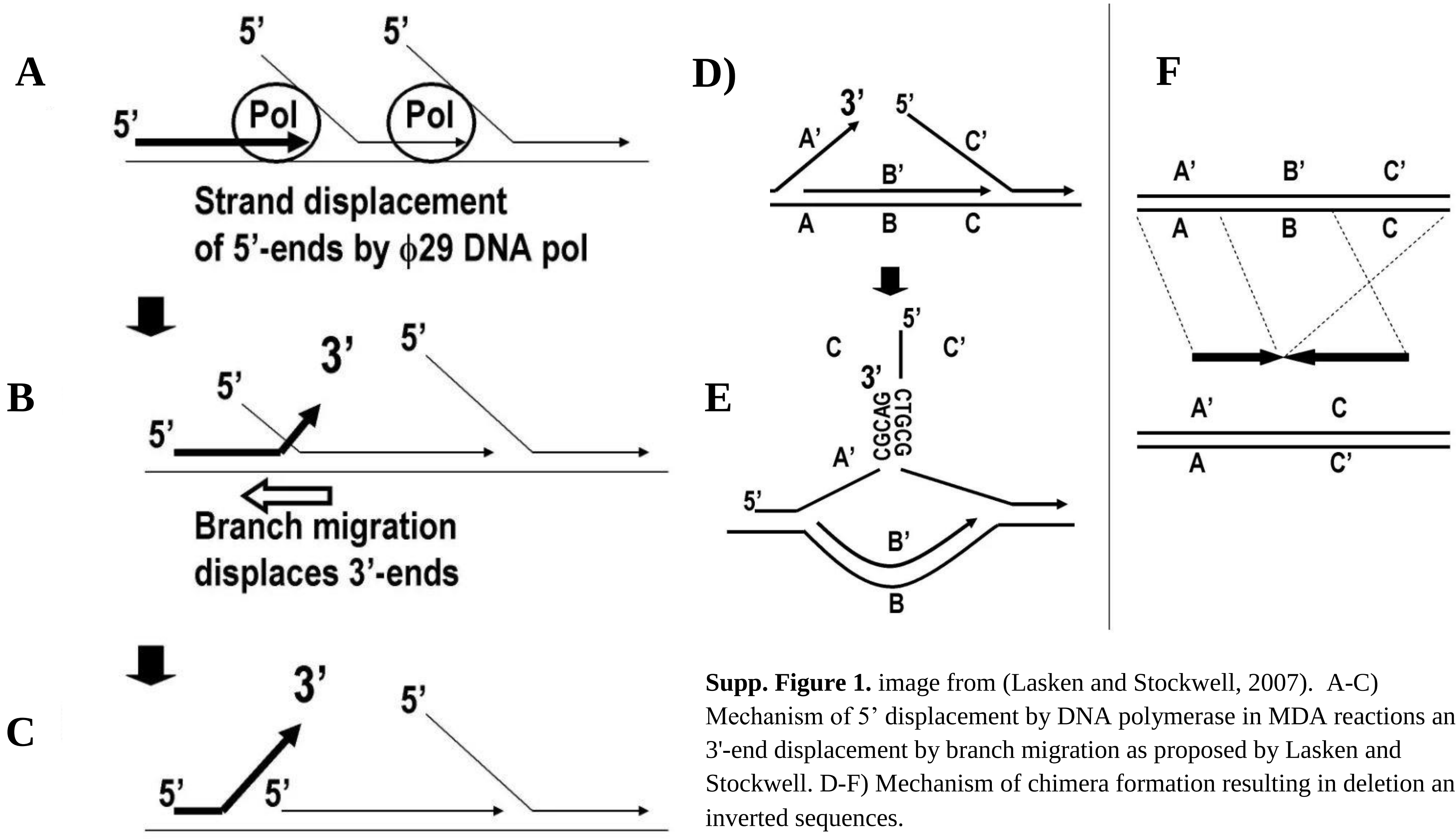
In this study comparisons were made primarily to TruSeq libraries as they represented the amplified short read libraries which make up the majority of virome research. These libraries were shown to underrepresent aspects of the virome and were outperformed (i.e. total VLS counts and other assembly stats) by unamplified short read libraries using the Accel-NGS library prep. However, as both TruSeq and Long-read libraries used identical MDA-amplified DNA samples, TruSeq reads were also used in the error correction and hybrid assembly steps. If the Accel-NGS library prep kit could avoid some of the detection limits of the TruSeq it is possible that long-read error correction and hybrid assembly could be improved by correcting and performing hybrid assemblies with the Accel NGS library.

Given the improved levels of assembly and detection observed in the Accel-NGS approach relative to the TruSeq, it is a promising alternative to virome analysis pipelines. However, as this study sought to compare alternative sequencing approaches to amplified short read libraries it highlights sequences which were elongated or scaffolded TruSeq assemblies, or were undetected in TruSeq libraries. The extent to which these alternative platforms missed VLS successfully sequenced and assembled is therefore not known. This is an obvious direction for future iterations of this study and is particularly important for future studies which may use the Accel-NGS as a replacement to the TruSeq library prep rather than to supplement it as was carried out here.

During the long-read processing, reads which featured palindromic repeats were removed. However, as these sequences may represent legitimate VLS which became chimeric due to MDA it would be more accurate to identify the repeat region and split the palindrome and remove redundancy of the entire library. One program has been designed to address this issue (Warris et al., 2018) and will be included in future iterations of this study as currently the long reads may not be being used to their full extent. Despite this limitation, we feel that the main findings of the study would remain consistent although potentially increasing the number of corrected long-read VLS that would likely exaggerate the current findings. Similarly, an aggressive size filter was applied to the PacBio reads due to the sheer size of the dataset and computational limitations. Future iterations of this study could benefit from cloud computing facilities in order to make use of this dataset in its entirety.

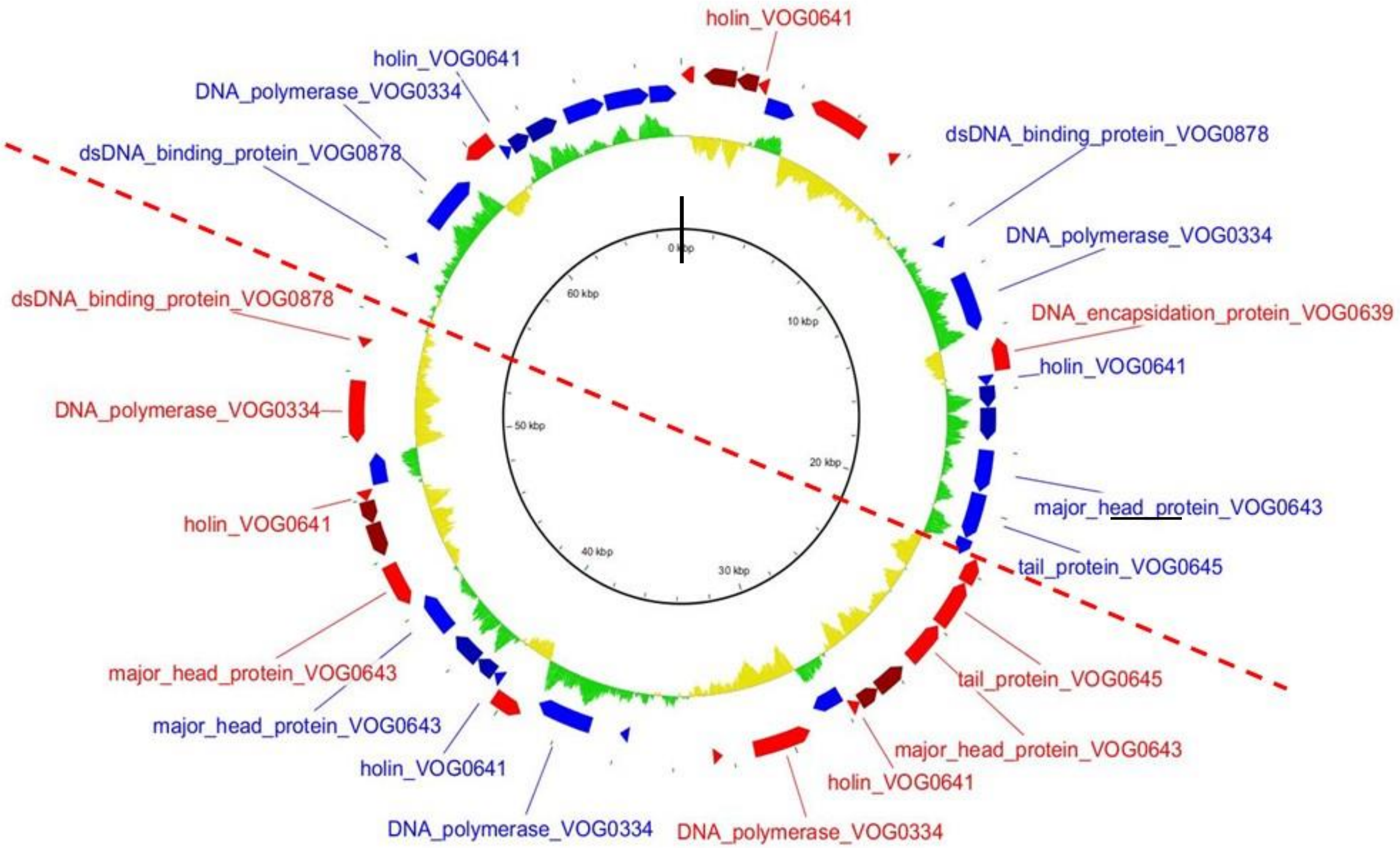
Conclusions

Here we present what we believe to be the first long-read sequencing study of the human gut virome and comparison of sequencing and assembly approaches. We highlight limitations in the ability of amplified short read libraries to accurately represent the human gut virome and advocate the use of long-read sequencing and alternative library prep methods as a means to address these challenges. This study highlights the need to consider the impact of sequencing approach when interpreting results from virome studies and has considerable implications for our current understanding of the human gut virome. We propose that amplified short read sequencing approaches mask the detection of viral sequences and may therefore exaggerate the levels of inter-individual variation associated with the virome. Furthermore, we suggest the virome may be more stable within individuals than had been previously thought and that transient detection is also exaggerated by the choice of sequencing approach. We propose that long-read sequencing and alternative library prep methods have an important role in virome analysis and can resolve members of the viral community which would have been fragmented or undetected using standard amplified short reads libraries.



Supp. Figure 1. image from (Lasken and Stockwell, 2007). A-C) Mechanism of 5’ displacement by DNA polymerase in MDA reactions and 3'-end displacement by branch migration as proposed by Lasken and Stockwell. D-F) Mechanism of chimera formation resulting in deletion and inverted sequences.

Supp.Figure 2. Example of a Palindromic corrected PacBio VLS annotated using pVOGS. VLS is linear and circularised for plotting purposes, black vertical line represents sequence start and end Forward CDS in blue, reverse CDS in red, GC skew in yellow and green, dashed red line showing the axis of symmetry.



References

Abeles, S.R., Robles-Sikisaka, R., Ly, M., Lum, A.G., Salzman, J., Boehm, T.K. et al. (2014). Human oral viruses are personal, persistent and gender-consistent. *The ISME journal* 8**,** 1753.

Alikhan, N.-F., Petty, N.K., Zakour, N.L.B., and Beatson, S.A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC genomics* 12**,** 402.

Bertrand, D., Shaw, J., Kalathiyappan, M., Ng, A.H.Q., Kumar, M.S., Li, C. et al. (2019). Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nature biotechnology* 37**,** 937-944.

Blainey, P.C. (2013). The future is now: single-cell genomics of bacteria and archaea. *FEMS microbiology reviews* 37**,** 407-427.

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

Clooney, A.G., Sutton, T.D.S., Shkoporov, A.N., Holohan, R.K., Daly, K.M., O'regan, O. et al. (2019). Whole-Virome Analysis Sheds Light on Viral Dark Matter in Inflammatory Bowel Disease. *Cell Host & Microbe* 26**,** 764-778.e765. doi:https://doi.org/10.1016/j.chom.2019.10.009.

Dean, F.B., Hosono, S., Fang, L., Wu, X., Faruqi, A.F., Bray-Ward, P. et al. (2002). Comprehensive human genome amplification using multiple displacement amplification. *Proceedings of the National Academy of Sciences* 99**,** 5261-5266.

Devoto, A.E., Santini, J.M., Olm, M.R., Anantharaman, K., Munk, P., Tung, J. et al. (2019). Megaphages infect Prevotella and variants are widespread in gut microbiomes. *Nature microbiology* 4**,** 693.

Driscoll, C.B., Otten, T.G., Brown, N.M., and Dreher, T.W. (2017). Towards long-read metagenomics: complete assembly of three novel genomes from bacteria dependent on a diazotrophic cyanobacterium in a freshwater lake co-culture. *Standards in genomic sciences* 12**,** 9.

Duhaime, M.B., Deng, L., Poulos, B.T., and Sullivan, M.B. (2012). Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization of the linker amplification method. *Environmental microbiology* 14**,** 2526-2537.

Dutilh, B.E., Cassman, N., Mcnair, K., Sanchez, S.E., Silva, G.G., Boling, L. et al. (2014). A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nature communications* 5**,** ncomms5498.

Dutilh, B.E., Schmieder, R., Nulton, J., Felts, B., Salamon, P., Edwards, R.A. et al. (2012). Reference-independent comparative metagenomics using cross-assembly: crAss. *Bioinformatics* 28**,** 3225-3231.

Edgar, R.C. (2007). PILER-CR: fast and accurate identification of CRISPR repeats. *BMC bioinformatics* 8**,** 18.

Grazziotin, A.L., Koonin, E.V., and Kristensen, D.M. (2016). Prokaryotic virus orthologous groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic acids research***,** gkw975.

Gregory, A.C., Zablocki, O., Howell, A., Bolduc, B., and Sullivan, M.B. (2019). The human gut virome database. BioRxiv [Preprint]. Available at: https://www.biorxiv.org/content/10.1101/655910v1.full (Accessed August 09, 2019). *BioRxiv*.

Guerin, E., Shkoporov, A., Stockdale, S.R., Clooney, A.G., Ryan, F.J., Sutton, T.D.S. et al. (2018). Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. *Cell Host & Microbe* 24**,** 653-664.e656. doi:https://doi.org/10.1016/j.chom.2018.10.002.

Hesse, U., Van Heusden, P., Kirby, B.M., Olonade, I., Van Zyl, L.J., and Trindade, M. (2017). Virome Assembly and Annotation: A Surprise in the Namib Desert. *Frontiers in Microbiology* 8. doi:10.3389/fmicb.2017.00013.

Hyatt, D., Chen, G.-L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics* 11**,** 119.

Iranzo, J., Krupovic, M., and Koonin, E.V. (2016). The Double-Stranded DNA Virosphere as a Modular Hierarchical Network of Gene Sharing. *mBio* 7**,** e00978-00916. doi:10.1128/mBio.00978-16.

Ishii, K., and Fukui, M. (2001). Optimization of annealing temperature to reduce bias caused by a primer mismatch in multitemplate PCR. *Appl. Environ. Microbiol.* 67**,** 3753-3755.

Kang, D.-W., Adams, J.B., Gregory, A.C., Borody, T., Chittick, L., Fasano, A. et al. (2017). Microbiota Transfer Therapy alters gut ecosystem and improves gastrointestinal and autism symptoms: an open-label study. *Microbiome* 5**,** 10.

Karplus, K., Barrett, C., and Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics (Oxford, England)* 14**,** 846-856.

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***,** AEM. 00289-00211.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research* 27**,** 722-736.

Lage, J.M., Leamon, J.H., Pejovic, T., Hamann, S., Lacey, M., Dillon, D. et al. (2003). Whole genome analysis of genetic alterations in small DNA samples using hyperbranched strand displacement amplification and array–CGH. *Genome research* 13**,** 294-307.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods* 9**,** 357.

Lasken, R.S., and Stockwell, T.B. (2007). Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC biotechnology* 7**,** 19.

Laslett, D., and Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic acids research* 32**,** 11-16.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N. et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25**,** 2078-2079.

Manrique, P., Bolduc, B., Walk, S.T., Van Der Oost, J., De Vos, W.M., and Young, M.J. (2016). Healthy human gut phageome. *Proceedings of the National Academy of Sciences* 113**,** 10400-10405.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* 17**,** pp. 10-12.

Minot, S., Bryson, A., Chehoud, C., Wu, G.D., Lewis, J.D., and Bushman, F.D. (2013). Rapid evolution of the human gut virome. *Proceedings of the National Academy of Sciences* 110**,** 12450-12455.

Myers, G. (Year). "Efficient local alignment discovery amongst noisy long reads", in: *International Workshop on Algorithms in Bioinformatics*: Springer), 52-67.

Nicholls, S.M., Quick, J.C., Tang, S., and Loman, N.J. (2019). Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *GigaScience* 8. doi:10.1093/gigascience/giz043.

Norman, J.M., Handley, S.A., Baldridge, M.T., Droit, L., Liu, C.Y., Keller, B.C. et al. (2015). Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* 160**,** 447-460.

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome research***,** gr. 213959.213116.

Olson, N.D., Treangen, T.J., Hill, C.M., Cepeda-Espinoza, V., Ghurye, J., Koren, S. et al. (2017). Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Briefings in bioinformatics*.

Petkau, A., Stuart-Edwards, M., Stothard, P., and Van Domselaar, G. (2010). Interactive microbial genome visualization with GView. *Bioinformatics* 26**,** 3125-3126.

Raghunathan, A., Ferguson, H.R., Bornarth, C.J., Song, W., Driscoll, M., and Lasken, R.S. (2005). Genomic DNA amplification from a single bacterium. *Appl. Environ. Microbiol.* 71**,** 3342-3347.

Roux, S., Emerson, J.B., Eloe-Fadrosh, E.A., and Sullivan, M.B. (2017). Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ* 5**,** e3817.

Roux, S., Enault, F., Hurwitz, B.L., and Sullivan, M.B. (2015a). VirSorter: mining viral signal from microbial genomic data. *PeerJ* 3**,** e985.

Roux, S., Hallam, S.J., Woyke, T., and Sullivan, M.B. (2015b). Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *Elife* 4**,** e08490.

Roux, S., Solonenko, N.E., Dang, V.T., Poulos, B.T., Schwenck, S.M., Goldsmith, D.B. et al. (2016). Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. *PeerJ* 4**,** e2777.

Sabina, J., and Leamon, J. (2015). Bias in whole genome amplification: causes and considerations. Whole genome amplification. *Methods in Molecular Biology. Springer New York, New York, NY***,** 15-41.

Sevim, V., Lee, J., Egan, R., Clum, A., Hundley, H., Lee, J. et al. (2019). Shotgun metagenome data of a defined mock community using Oxford Nanopore, PacBio and Illumina technologies. *Scientific data* 6**,** 1-9.

Shkoporov, A.N., Clooney, A.G., Sutton, T.D.S., Ryan, F.J., Daly, K.M., Nolan, J.A. et al. (2019). The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. *Cell Host & Microbe* 26**,** 527-541.e525. doi:https://doi.org/10.1016/j.chom.2019.09.009.

Shkoporov, A.N., Khokhlova, E.V., Fitzgerald, C.B., Stockdale, S.R., Draper, L.A., Ross, R.P. et al. (2018a). ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. *Nature Communications* 9**,** 4781. doi:10.1038/s41467-018-07225-7.

Shkoporov, A.N., Ryan, F.J., Draper, L.A., Forde, A., Stockdale, S.R., Daly, K.M. et al. (2018b). Reproducible protocols for metagenomic analysis of human faecal phageomes. *Microbiome* 6**,** 68.

Slaby, B.M., Hackl, T., Horn, H., Bayer, K., and Hentschel, U. (2017). Metagenomic binning of a marine sponge microbiome reveals unity in defense but metabolic specialization. *The ISME journal* 11**,** 2465.

Somerville, V., Lutz, S., Schmid, M., Frei, D., Moser, A., Irmler, S. et al. (2019). Long-read based de novo assembly of low-complexity metagenome samples results in finished genomes and reveals insights into strain diversity and an active phage system. *BMC microbiology* 19**,** 143.

Sutton, T.D., Clooney, A.G., Ryan, F.J., Ross, R.P., and Hill, C. (2019a). Choice of assembly software has a critical impact on virome characterisation. *Microbiome* 7**,** 12.

Sutton, T.D.S., Clooney, A.G., and Hill, C. (2019b). Giant oversights in the human gut virome. *Gut***,** gutjnl-2019-319067. doi:10.1136/gutjnl-2019-319067.

Sutton, T.D.S., and Hill, C. (2019). Gut bacteriophage: Current understanding and challenges. *Frontiers in Endocrinology* 10**,** 784.

Tsai, Y.-C., Conlan, S., Deming, C., Segre, J.A., Kong, H.H., Korlach, J. et al. (2016). Resolving the complexity of human skin metagenomes using single-molecule sequencing. *MBio* 7**,** e01948-01915.

Wagner, J., Maksimovic, J., Farries, G., Sim, W.H., Bishop, R.F., Cameron, D.J. et al. (2013). Bacteriophages in gut samples from pediatric Crohn's disease patients: metagenomic analysis using 454 pyrosequencing. *Inflammatory bowel diseases* 19**,** 1598-1608.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S. et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one* 9**,** e112963.

Wang, J.R., Holt, J., Mcmillan, L., and Jones, C.D. (2018). FMLRC: Hybrid long read error correction using an FM-index. *BMC bioinformatics* 19**,** 50.

Warris, S., Schijlen, E., Van De Geest, H., Vegesna, R., Hesselink, T., Te Lintel Hekkert, B. et al. (2018). Correcting palindromes in long reads after whole-genome amplification. *BMC genomics* 19**,** 798.

Warwick-Dugdale, J., Solonenko, N., Moore, K., Chittick, L., Gregory, A.C., Allen, M.J. et al. (2019). Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. *PeerJ* 7**,** e6800.

Watson, M., and Warr, A. (2019). Errors in long-read assemblies can critically affect protein prediction. *Nature biotechnology* 37**,** 124.

Weirather, J.L., De Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.-J. et al. (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research* 6.

Yutin, N., Makarova, K.S., Gussow, A.B., Krupovic, M., Segall, A., Edwards, R.A. et al. (2018). Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nature microbiology* 3**,** 38.

Zuo, T., Lu, X.-J., Zhang, Y., Cheung, C.P., Lam, S., Zhang, F. et al. (2019). Gut mucosal virome alterations in ulcerative colitis. *Gut***,** gutjnl-2018-318131.

Thesis Summary

The phages and bacteria of the human microbiome are two sides of the same coin and the composition of each community is intrinsically linked to the other. Understanding the forces which shape the human gut microbiome is key to understanding its role in the maintenance of human health, yet arguably the most important of these forces, the human gut virome, remains almost entirely unknown. The majority of current virome research is built on nearly two decades of sequence-based studies. These have given us new insights into this missing part of the microbiome puzzle but it appears we have only scratched the surface and the virome remains dominated by unknown sequences encoding as yet unknown functions. This unknown majority is also one of the biggest challenges of virome data, as it not only determines which analysis approaches are possible, but also makes these approaches very difficult to validate. As a result, virome data is particularly sensitive to methodological artefacts and the final conclusions which are drawn from virome studies can depend heavily on the analysis approach used.

Chapter One of this thesis discussed major findings in the gut virome field and highlighted a number of inconsistencies. We recommend that virome researchers consider and acknowledge the distortion every particular method may have on a given result and report it accordingly. Too often virome studies take results at face value and draw incorrect biological conclusions. Doing so leads to the propagation of certain analytical methods and could lead to bias in arriving at expected findings. A prime example is the frequent report of changes in alpha diversity of known *Caudovirales* in a number of diseases, which remains a regular feature of more recent virome studies. While it is possible that this reflects an underlying biological signal, it gives very little insight into the role of the virome in disease as discussed at length in Chapter One. This repetition of limited analysis methods and results which give little insight into the composition and function of the human gut virome highlights a real need for studies which validate analysis pipelines and develop new approaches. For these reasons, the work presented in this thesis is of critical importance to the progression of the virome field and understanding its role in the microbiome.

This thesis sought to validate some of the major steps in sequence-based virome analysis pipelines from the choice of sequencing platform and assembler, to how the resulting data is analysed. Interestingly and somewhat alarmingly every step

that we analysed had a significant impact on the final output, highlighting the fragility of our understanding of the gut virome. While the challenges of virome data may be exaggerated relative to other metagenomes (dominance of unknown sequences, extremes in high and low sequencing coverage etc.) they are not unique to virome datasets. This suggests that many of the challenges and limitations of viral metagenomes in this thesis may highlight limitations of microbial metagenomics in general.

Chapter Two highlighted a central part of every virome analysis pipeline, the assembly step. As virome studies are dominated by unknown sequences, alignment of sequencing reads to sequence databases gives a very limited view of virome composition. As a result, almost all sequence-based virome studies assemble sequencing reads to reconstruct the genomes of the viral community. This also means that assembly performance ultimately determines the amount of sequencing data which can be used in a given virome study. We compared the performance of all short-read assembly programs used in virome studies to date and found that the choice of assembler significantly varied the composition of the final virome. Furthermore, extremes in both high and low coverage resulted in fragmented assemblies and poor genome recovery. Given that these extremes are common in virome datasets, the limits of assembly must be considered before drawing conclusions. Furthermore, the poor performance of some assemblers suggests that not only should they be avoided in future virome studies, but that the findings of studies that had used them originally should be treated with caution. Furthermore, studies which had used poor assemblers present unique opportunities to gain new insight into the virome simply by re-assembling and reanalysing existing data.

In Chapter Three, we performed exactly that. Current understanding of the gut virome in IBD was based primarily in the findings and methods of one keystone study by Norman *et al*. (Norman et al., 2015). In that study, disease-specific differences in viral richness had been reported between CD, UC and healthy cohorts. However, these findings were based on a minor subset of identifiable *Caudovirales*. This database-dependent analysis approach has been since used by many subsequent virome studies and the findings and methods are regularly referred to and replicated. We sought to develop a database-independent analysis approach using this dataset to investigate if the patterns seen across the identifiable minority were truly representative of the whole-virome. This whole-virome analysis approach not only highlighted the limitations of database-dependent approaches but gave new insights into the microbiome in IBD. However, by including the entire dataset and not just the identifiable minority, we encountered issues with high levels of inter-individual variation that masked any compositional changes across cohorts. This was addressed by clustering viral sequences at a gene-level to increase the number of shared sequences across cohorts. By doing so, we observed a healthy core virome of virulent phage which was absent in disease. Furthermore, this healthy core appeared to be replaced by an individual-specific shift towards temperate phage in disease. This provided the first functional insight into the gut virome in inflammatory bowel disease and new insights into viral dark matter in the gut. This study also highlighted the importance of maximising the data used in a given study and how by changing the analysis approach we can drastically change our understanding of the human gut virome.

Chapter Four of this thesis went to the very core of a sequence-based study and examined both sequencing platforms and library prep methods. We performed a pilot study using long and short read sequencing and as with the previous chapters highlighted the sensitivity of virome studies to methodological bias. Our current understanding of the human gut virome is built on a foundation of MDA amplified short-read sequencing and while MDA is known to skew the composition of DNA samples, it has not been fully characterised in the virome. Somewhat alarmingly, we observed significant limitations in the ability of these amplified short-read libraries to fully recover the human gut virome. This has serious implications for how we perform virome studies, and the conclusions we have drawn from them to-date. However, we also see promise in alternative sequencing approaches (i.e. alternative library prep and long-read platforms) as a means of addressing these issues and suggest that future virome studies would benefit from the addition of these approaches.

These four chapters provide an important resource to virome researchers by highlighting the importance of considering the analysis approach when drawing conclusions from virome data. We describe some of the limitations of previous approaches and suggest that the findings presented by studies using these methods should be treated with caution. Furthermore, these studies are promising candidates for reanalysis, which as shown in Chapter Three, can provide useful insight into the structure and function of the virome. The contents of this thesis will hopefully contribute to future virome studies by validating existing sequence-based tools and describing new approaches to analyse the virome. As Melvin James “Sy” Oliver said in 1939, “*'tain't what you do, it's the way that you do it*”. Perhaps he too was considering the impact of analysis methods in virome studies in the future?

Appendix 1

Giant oversights in the human virome

Thomas DS Sutton, Adam G Clooney, Colin Hill

This letter has been published as the following:

Sutton, T.D.S., Clooney, A.G., and Hill, C. (2019a). Giant oversights in the human gut virome. *Gut*. doi:10.1136/gutjnl-2019-319067.

In response to following publication by Zuo et al.

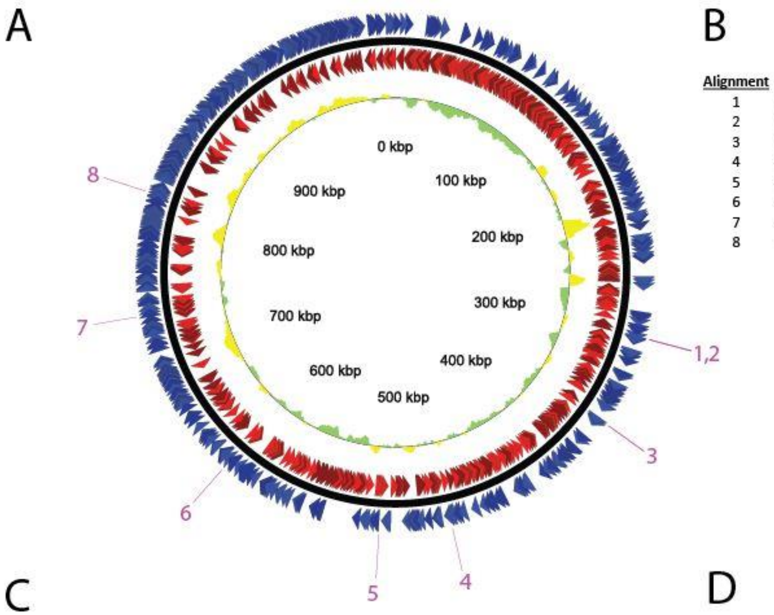
Zuo, T., Lu, X.-J., Zhang, Y., Cheung, C.P., Lam, S., Zhang, F. et al. (2019). Gut mucosal virome alterations in ulcerative colitis. Gut, gutjnl-2018-318131.

We read with interest the paper "Gut mucosal virome alterations in ulcerative colitis" by Zuo et al.(Zuo et al., 2019) which used deep sequencing to identify gut mucosal virome alterations in individuals with ulcerative colitis (UC). One of many interesting findings reported by the authors was the detection of giant viruses infecting algae and amoeba (*Mimivirus* and *Chrysochromulina ericina* virus). The authors suggested an association with the geographical distribution of individuals and concluded that they were more abundant in UC patients than controls. We reanalysed the data and propose that issues related to the virome analysis pipeline led to the incorrect identification of these viruses.

Firstly, the DNA extraction method states that 0.22 µm filters were used to remove bacterial and eukaryotic cells, followed by chemical and enzymatic degradation of DNA unprotected by viral capsids. *Mimivirus* has capsid with a diameter of 0.5 µm, surrounded by a 0.125 µm thick layer of closely packed fibres (Klose et al., 2010), making its presence in the final DNA extract extremely unlikely.

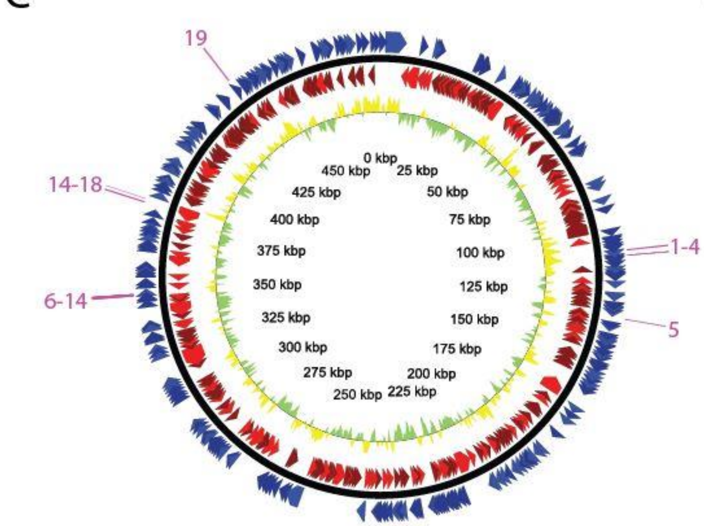
Secondly, we aligned sequencing reads provided by the authors to a database of all representative *Mimivirus* and *C. ericina* virus sequences present in the UniProt TrEMBL database (228 as of April 2019; see supplementary methods). *Mimivirus* sequences recruited at most 82 reads from any given sample, all of which aligned to short "AT" repeats covering less than 400 nt of the 1.2 Mb reference sequence (Figure 1A). *C. ericina* virus recruited significantly more reads on average (934; max 8,205), but again all reads aligned to three short intergenic "AT" repeats at coordinates 91099-91267, 260723-260745 and 267717-267744. Individual samples covered, at most, 0.02% of the genome.

The annotation of viral assemblies as *Mimivirus* and *C. ericina* virus highlights an issue in the taxonomic assignment method used in this (and other) virome studies. We used the assembled sequencing reads to repeat the Zuo et al. method by predicting open reading frames from viral contigs, aligning them to the UniProt TrEMBL database to assign viral taxonomy to each ORF and finally using a voting system to assign a consensus taxonomy across the contig.



B

Alignment	%ID	Alignment Length	contig length	% Query contig	Start Coord	End Coord	Target annotation
1	67.1	503	7783	6.5	297121	297624	isoleucyl-tRNA synthetase
2	68.8	140	18281	0.8	297334	297474	isoleucyl-tRNA synthetase
3	65.8	373	3141	11.9	356460	356833	DNA topoisomerase 2
4	72.8	135	3734	3.6	474736	474871	HSP70-like protein
5	74.8	150	5953	2.5	520770	520920	NAD-dependent epimerase/dehydratase
6	71.3	134	62204	0.2	627768	627902	DEAD/SNF2-like helicase
7	72.1	107	8639	1.2	738218	738325	glucose-methanol-choline oxidoreductase
8	68.6	155	2029	7.6	817100	817255	arginyl-tRNA synthetase



D

Alignment	%ID	Alignment Length	contig length	% Query contig	Start Coord	End Coord	Target annotation
1	65.8	319	4128	7.7	109983	110302	Intein containing GDP-mannose 4,6-dehydratase
2	74.3	100	30903	0.3	111747	111847	Intein containing GDP-mannose 4,6-dehydratase
3	74.3	100	23268	0.4	111747	111847	Intein containing GDP-mannose 4,6-dehydratase
4	73.2	96	14315	0.7	111748	111844	Intein containing GDP-mannose 4,6-dehydratase
5	66.8	189	3428	5.5	131252	131441	guanosine 5'-monophosphate oxidoreductase
6	66.4	721	7753	9.3	349202	349923	Hsp70 protein
7	68.9	150	2183	6.9	349266	349416	Hsp70 protein
8	72.2	248	8943	2.8	349562	349810	Hsp70 protein
9	73.8	248	12431	2.0	349562	349810	Hsp70 protein
10	69.9	325	42462	0.8	349562	349887	Hsp70 protein
11	69.4	249	5416	4.6	349602	349851	Hsp70 protein
12	66.1	277	5741	4.8	349610	349887	Hsp70 protein
13	68.1	134	12577	1.1	349610	349744	Hsp70 protein
14	68.1	134	15973	0.8	349610	349744	Hsp70 protein
15	70.9	314	1042	30.1	378192	378506	DNA-directed RNA polymerase II, subunit RpB1
16	77.4	61	11617	0.5	379387	379448	DNA-directed RNA polymerase II, subunit RpB1
17	78.1	71	24522	0.3	423765	423836	Non-coding region
18	79.7	58	5898	1.0	423968	424026	Non-coding region
19	82.4	50	12671	0.4	424079	424129	Non-coding region

Figure 1. (overleaf) (A) Graphical representation of *Moumouvirus* displaying GC skew (green/yellow), reverse strand CDS (red) forward strand CDS (blue) and alignments of all classified *Mimivirus* contigs to *Moumouvirus* genome (pink). (B) Details of alignments highlighted in (A) with percentage identity, alignment length, query contig length, percentage of query contig aligning to the reference, start and end coordinates of the alignment to the reference genome and reference annotation at these coordinates. (C) Graphical representation of *C. ericina* virus as was plotted for (A) with alignments of all classified *Phycodnaviridae* contigs (pink). (D) Details of alignments highlighted in (C) as was described for (B).

The only alignments of contigs classified as *Mimivirus* and *C. ericina* virus to the *Mimivirus* and *C. ericina* virus database were short (4.3% of the query contig length on average), low identity hits to proteins such as heat-shock protein and t-RNA synthetase (Figure 1 A-D), which are conserved across domains of life (Fujishima and Kanai, 2014) (Feder and Hofmann, 1999). When aligned to the NT database these same contigs displayed high identity and often full length alignments to bacterial and fungal sequences and none to *C. ericina* or *Mimivirus* genomes. We appreciate that these viruses encode similar heat-shock and DNA metabolism genes, but we believe the observed alignments reflect distant similarities between proteins conserved across all domains of life. There were no alignments to any genes unique to viruses such as the major capsid protein. We conclude that there is no evidence for the presence of *Mimivirus* or *C. ericina* virus in the human mucosal virome and any suggestion of an association with ulcerative colitis should be treated with caution.

The authors also found PhiX174 (incorrectly assigned to the order *Caudovirales*) to be significantly increased in subjects with UC, but it should also be noted that *PhiX174microvirus* is an internal control used in Illumina sequencing and is a common contaminant of microbial sequencing studies (Mukherjee et al., 2015) (Zaheer et al., 2018). The published protocol did not describe any steps to remove *Phix174* sequencing controls and so it is possible that the presence and differential abundance of *Phix174* is a result of sequencing artefacts rather than biological changes.

Supplementary materials and methods

All nucleotide sequences associated with all *C. ericina* virus and *Mimivirus* UniProtKB/TrEMBL taxon IDs were downloaded from GenBank to form a database consisting of all full length and genomic fragments (n=228) with *C. ericina* virus and *Mimivirus* genes in the UniProt TrEMBL database. In order to maximise the chances of detection, all relevant entries were added to this database, including those which were partial or not yet reviewed. This database was used to recruit reads which were downloaded from the accession numbers provided in the paper (PRJNA504921 and PRJNA506811) and quality filtered with Trimmomatic v0.36 (Bolger et al., 2014) and the following parameters (HEADCROP:20 CROP:120 SLIDINGWINDOW:4:20 MINLEN:20). Read alignments were carried out using Bowtie2 v2.1.0 (Langmead and

Salzberg, 2012) and corresponding read recruitment and coverage statistics were calculated using SAMtools v0.1.19 (Li et al., 2009). Regions which featured a read coverage greater than 1 were extracted and investigated in detail using EMBOSS v6.6.0.0 extractseq (Rice et al., 2000).

Trimmed high quality reads were assembled using SPAdes v3.10.0 (Nurk et al., 2017) metagenomic mode. Assemblies were clustered at 95% as outlined in (Shkoporov et al., 2018) and open reading frames predicted using Prodigal v2.6.3.(Hyatt et al., 2010) We appreciate that our choice of assembly software differs from those outlined in the materials and methods of Zuo et al., however, based on a recent assembly comparison (Sutton et al., 2019) we observed significant improvements in SPAdes meta contiguity over IDBA with virome data and felt it would maximise the detection of giant virus genomes. We also use Prodigal opposed to Glimmer as was outlined in Zuo et al., as its accuracy has been found to be superior according to ORF caller benchmarking studies (Tripp et al., 2015).

Taxonomic assignment was carried out as described in Zuo et al., by aligning predicted ORFs to a database of all sequences in the UniProtKB/TrEMBL with viral taxonomic assignments, using Blastx (e-value < $1e^{-5}$). Contigs with one ORF per 10Kb were excluded and best hit viral taxonomy was applied to each ORF. Contigs were then classified based on the majority taxonomic assignment across all ORFs, or labelled "unclassified" if a majority was not found. All contigs annotated as *Phycodnaviridae* (n=686) and *Mimivirus* (n=111) were aligned to the aforementioned *C. ericina* virus and *Mimivirus* database and the NT (March 2019) database using Blastn (e-value < $1e^{-5}$) –task Blastn as a means to maximise the likelihood of high-quality alignments. Blastn alone did not yield any alignments. Alignments of these contigs were visualised using Gview (Petkau et al., 2010), using *C. ericina* virus (Accession no. NC_028094.1) and *Moumouvirus* (Accession no. JX962719.1) genomes as references and plotting alignment coordinates as the outermost ring. *Moumouvirus* was used to represent all *Mimivirus* alignments as it recruited the greatest number of *Mimivirus*-classified contigs from the dataset.

References

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

Devoto, A.E., Santini, J.M., Olm, M.R., Anantharaman, K., Munk, P., Tung, J. et al. (2019). Megaphages infect Prevotella and variants are widespread in gut microbiomes. *Nature microbiology* 4**,** 693.

Feder, M.E., and Hofmann, G.E. (1999). Heat-shock proteins, molecular chaperones, and the stress response: evolutionary and ecological physiology. *Annual review of physiology* 61**,** 243-282.

Fujishima, K., and Kanai, A. (2014). tRNA gene diversity in the three domains of life. *Frontiers in genetics* 5**,** 142.

Guerin, E., Shkoporov, A., Stockdale, S.R., Clooney, A.G., Ryan, F.J., Sutton, T.D.S. et al. (2018). Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. *Cell Host & Microbe* 24**,** 653-664.e656. doi:https://doi.org/10.1016/j.chom.2018.10.002.

Hyatt, D., Chen, G.-L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics* 11**,** 119.

Klose, T., Kuznetsov, Y.G., Xiao, C., Sun, S., Mcpherson, A., and Rossmann, M.G. (2010). The three-dimensional structure of Mimivirus. *Intervirology* 53**,** 268-273.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods* 9**,** 357.

Lasken, R.S., and Stockwell, T.B. (2007). Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC biotechnology* 7**,** 19.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N. et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25**,** 2078-2079.

Mahmoudabadi, G., and Phillips, R. (2018). A comprehensive and quantitative exploration of thousands of viral genomes. *eLife* 7**,** e31955. doi:10.7554/eLife.31955.

Mukherjee, S., Huntemann, M., Ivanova, N., Kyrpides, N.C., and Pati, A. (2015). Large-scale contamination of microbial isolate genomes by Illumina PhiX control. *Standards in genomic sciences* 10**,** 18.

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome research***,** gr. 213959.213116.

Petkau, A., Stuart-Edwards, M., Stothard, P., and Van Domselaar, G. (2010). Interactive microbial genome visualization with GView. *Bioinformatics* 26**,** 3125-3126.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European molecular biology open software suite. *Trends in genetics* 16**,** 276-277.

Shkoporov, A.N., Ryan, F.J., Draper, L.A., Forde, A., Stockdale, S.R., Daly, K.M. et al. (2018). Reproducible protocols for metagenomic analysis of human faecal phageomes. *Microbiome* 6**,** 68.

Sutton, T.D., Clooney, A.G., Ryan, F.J., Ross, R.P., and Hill, C. (2019). Choice of assembly software has a critical impact on virome characterisation. *Microbiome* 7**,** 12.

Tripp, H.J., Sutton, G., White, O., Wortman, J., Pati, A., Mikhailova, N. et al. (2015). Toward a standard in structural genome annotation for prokaryotes. *Standards in genomic sciences* 10**,** 45.

Zaheer, R., Noyes, N., Polo, R.O., Cook, S.R., Marinier, E., Van Domselaar, G. et al. (2018). Impact of sequencing depth on the characterization of the microbiome and resistome. *Scientific reports* 8**,** 5890.

Zuo, T., Lu, X.-J., Zhang, Y., Cheung, C.P., Lam, S., Zhang, F. et al. (2019). Gut mucosal virome alterations in ulcerative colitis. *Gut***,** gutjnl-2018-318131.

Appendix 2

Minor contributions were made to the following papers (front pages attached) as follows

Shkoporov, A.N., Clooney, A.G., **Sutton, T.D.S.**, Ryan, F.J., Daly, K.M., Nolan, J.A. et al. (2019). The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. Cell Host & Microbe 26, 527-541.e525.

> Assisted the design and implementation of the analysis approach and assisted in drafting and editing of the final manuscript. Carried out read processing and assembly of the whole community metagenomic samples and analysed virome-bacteriome interactions.

Guerin, E., Shkoporov, A., Stockdale, S.R., Clooney, A.G., Ryan, F.J., **Sutton, T.D.S.** et al. (2018). Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. Cell Host & Microbe 24, 653-664.e656.

> Assisted in the design and implementation of analysis approaches and reviewed the final manuscript. Carried out read processing, metadata processing and assembly for the Norman et al. dataset which was used to investigate crAss families across studies.

Fitzgerald, C.B., Shkoporov, A.N., **Sutton, T.D.S.**, Chaplin, A.V., Velayudhan, V., Ross, R.P. et al. (2018). Comparative analysis of *Faecalibacterium prausnitzii* genomes shows a high level of genome plasticity and warrants separation into new species-level taxa. BMC genomics 19, 931.

> Investigated prophage regions within *Faecalibacterium prausnitzii* genomes by recruiting induced VLP reads to the host genomes and characterising the coverage patterns near predicted prophage coordinates.

Shkoporov, A.N., Ryan, F.J., Draper, L.A., Forde, A., Stockdale, S.R., Daly, K.M. McDonnell, S.A., **Sutton T.D.S,** et al. (2018). Reproducible protocols for metagenomic analysis of human faecal phageomes. Microbiome 6, 68

> Developed the script used to remove redundancy across virome assemblies which has featured in all subsequent virome studies from the group.

Acknowledgements

I would like to thank Prof. Colin Hill for giving me the opportunity in his lab, for his continued support and guidance and pushing me to perform to the best of my abilities. I would also like to thank Paul Ross for helping to set up the lab and for his continuous support through the project. To Adam, Feargal, and Hugh (listed alphabetically to avoid fallout!) I owe the vast majority of my bioinformatics to you. The work in this thesis would not have been possible without your patience and support. Thank you to Lorraine who was always to hand to put out fires and steer things back on course. To “the lads” David, Maurice, Mrinmoy, Ross and Sidney (alphabetical again!) the past four years would have been much harder without ye, thank you for always having the necessary antidotes to academia to hand. To all in the Phage lab, thank you for always making me feel welcome and supported.

To Mary and Gerry, thank you for nurturing an inquisitive mind and for pushing and guiding me to where I am now. Perhaps this all started with BSE! To Hugh (Sutton this time), when times were tough you inspired me to carry on. Another few assemblies are far easier than another 20km on the water. Finally to Heather, thank you for making all of this possible, for sharing both the highs and lows and for putting up with me. Nobody manages an undertaking like this alone and you were central to every step. I’ll have to ask UCC if they have an honorary doctorate available. To all friends and family who helped make this happen, thank you.