Trends in Microbiology

CellPress OPEN ACCESS

Review

Challenges of Studying the Human Virome – Relevant Emerging Technologies

Mohammadali Khan Mirzaei,[1] Jinling Xue,[1] Rita Costa,[1] Jinlong Ru,[1] Sarah Schulz,[1] Zofia E. Taranu,[2] and Li Deng[1,*]

In this review we provide an overview of current challenges and advances in bacteriophage research within the growing field of viromics. In particular, we discuss, from a human virome study perspective, the current and emerging technologies available, their limitations in terms of *de novo* discoveries, and possible solutions to overcome present experimental and computational biases associated with low abundance of viral DNA or RNA. We summarize recent breakthroughs in metagenomics assembling tools and single-cell analysis, which have the potential to increase our understanding of phage biology, diversity, and interactions with both the microbial community and the human body. We expect that these recent and future advances in the field of viromics will have a strong impact on how we develop phage-based therapeutic approaches.

Highlights

Shotgun sequencing bypasses the need for metabarcoding in viromics, although it is prone to high background noises and biases towards double-stranded DNA viruses.

Protein-level assembly can be a better tool to use on virome data as they predict more protein sequences from complex unknown metagenomes.

Using viral discovery methods can help to resolve the full diversity of viral fraction of microbiome data.

Culture-independent methods such as viral-tagging can be used to measure the phage host range in the human body.

Introduction

The human body is an ecosystem, a home to a complex network of microbial organisms comprised of bacteria, archaea, eukarya, and viruses. The majority (sometimes >90%) of viruses present in the human gut are those that infect bacterial hosts; these viruses are known as phages [1–3]. Phages can replicate through two major replication cycles: lysogenic or lytic. Other replication cycles, including pseudolysogeny and chronic infection, also exist. In the lytic cycle, phages infect the host and kill it shortly afterwards. The lysogenic cycle involves phages that stay dormant as part of the host genome; when inserted into the host genome the phage is known as a prophage [1,2]. The dynamics of bacteria-phages interactions varies between ecosystems [4,5], with phages of the human gut persisting for prolonged periods of time and seemingly promoting a stable, healthy gut microbiome [6–8]. Due to the rise in multidrug-resistant bacterial infections, there has been a renewed interest in phage-based therapies as an alternative antibacterial approach. However, despite the therapeutic use of phages being over a century old [9,10], and their high abundance in the body, they are among the least described components of the human microbiota, especially when compared with bacteria [11–13].

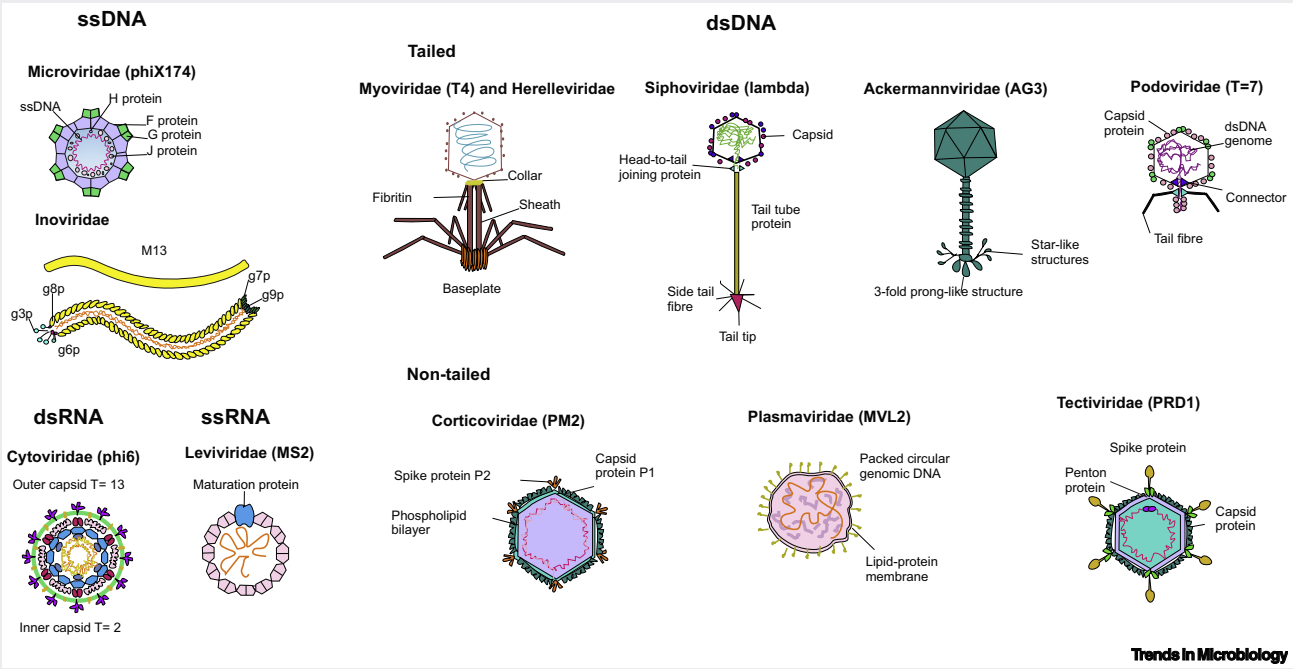
Of the estimated 10^{31} viruses on earth, the majority are phages, but only 2640 of their genomes are closed or fully sequenced (https://www.ncbi.nlm.nih.gov/refseq/), an example of how little we know about phage diversity (Box 1). The double-stranded DNA (dsDNA) tailed phages account for the majority of those characterized by electron microscopy and culture-based methods [14,15]. However, recent studies contradict the earlier ones and suggest that this dominance may be biased by the applied methods, rather than being a true representation of the human phageome, and it shows our limited understanding of phage diversity [14–17]. For example, a recent study identified more than 15 000 ssRNA phage sequences from public datasets, including over 1000 near-complete genomes, by optimizing a Hidden Markov Model (HMM)-based pipeline for the ssRNA phages' discovery. This suggests that ssRNA phages have been overlooked within microbiome studies, and the current studies may have underestimated their contribution to phage diversity [17].

[1]Institute of Virology, Helmholtz Centre Munich and Technical University of Munich, Neuherberg, Bavaria 85764, Germany
[2]Aquatic Contaminants Research Division (ACRD), Environment and Climate Change Canada (ECCC), Montréal, QC H2Y 2E7, Canada

*Correspondence: li.deng@helmholtz-muenchen.de (L. Deng).

Box 1. Phage Taxonomy

The classification of phages has changed significantly since their discovery in 1917, and the responsibility lies currently with the International Committee on the Taxonomy of Viruses (ICTV) [97,98], which published their first report in 1971 [72,97]. Historically, phages were categorized by their morphology, host range, storage stability, and genome structure (dsDNA, dsRNA, ssDNA, ssRNA), classification becoming more and more sophisticated as new techniques became available [75,97]. With the rise of bioinformatics, this system has seen a major overhaul in recent years [75,97]. Tailed, double-stranded DNA phages are the most commonly isolated phages, with most of them belonging to the order Caudovirales, which is currently grouped into five families: Myoviridae, Siphoviridae, Podoviridae, Ackermannviridae and Herelleviridae [97,98].



Phages not belonging to the order Caudovirales belong to seven different families that have currently no order assigned to them yet [97]. Non-tailed phages comprise roughly 4% of all currently known phages and come in three distinct morphologies: polyhedral capsids, filamentous, and pleomorphic [75,97,98]. With the advance of new techniques such as metagenomics, and renewed clinical interest in phages, it is only a matter of time until the current classification becomes as outdated as the first one made in 1933 by Sir Frank Mcfarlane Burnet based on filtration size [75,98].

The lack of knowledge of phage diversity, lifestyle, and dynamics in the human body stems in part from a limited toolkit which, until recently, was restricted to classical microbiology methods, including microscopy and culture-dependent approaches [18], as well as a tendency to extrapolate data from one ecosystem to another [4,7]. Unlike classical microbiology, which isolates components of an ecosystem to explain them individually, the multi-omics approach allows the study of the organisms within a complex network of interactions [19]. However, even the advances of new high-throughput multi-omics technologies come with their own challenges and limitations related to sample and downstream processing, sequencing annotation, and *in silico* predictions [7,11,20]. In this review we describe current advances in the growing field of viromics. Our main focus is to address the challenges that phageome research is facing in the omics era, the emerging technologies, and the technical improvements that are required to overcome these challenges.

Sample Processing and Downstream Analysis

Some clinical samples, such as skin swabs, are typically limited in volume and have a low abundance of viruses as well as a high background from the host microbiome [21,22]. The entire

process, from sample collection to sequencing, will impact the detection of viral sequences and needs to be carefully tailored to sample type, origin, and volume [23,24]. Existing sampling techniques have their advantages but can be biased toward recovering the most abundant community members. For example, the use of 0.2 μm filters, a common approach for removing large particles, such as host and bacterial cells from a sample, has been shown to also deplete large viruses [25] and reduce the amount of recovered viral DNA by half [26]. Similarly, CsCl gradient ultracentrifugation purification, depending on how the method is performed, can be biased toward isolating specific phage types and those with atypical buoyancy, but it results in very pure samples [27,28]. Viral quantification methods, such as epifluorescence microscopy, can underestimate the actual number of virus-like particles (VLPs) in human samples (Feichtmeyer *et al*., unpublished observations). Automated extraction platforms are now frequently used for virus detection in combination with qPCR or droplet digital PCR due to their higher sensitivity and high-throughput work capacity [29,30], while commercial kits work better with higher viral loads (>10^6 copies/ml) and longer DNA fragments (>200 bp) [31]. If an amplification step of viral nucleic acids is required, the most commonly used methods are: (i) random amplified shotgun library (RASL) in which the template is restricted to dsDNA; (ii) linker-amplified shotgun library (LASL), which requires a high template concentration [11,29]; and (iii) multiple displacement amplification (MDA), which tends to overamplify circular single-stranded DNA (ssDNA) and unevenly amplifies linear genomes [32,33]. As illustrated by these examples, the most commonly used virus isolation methods have their specific drawbacks and/or biases.

The majority of the unpurified viral metagenome sequences are assigned to bacteria and eukaryotic DNA [12,34]. Thus, removal of background contamination using VLP purification methods is essential to get a clear image of phage abundance in the human body [34]. Recently developed flow-cytometry-based methods allow for the separation of VLPs from the background microbiota by labelling phages with a fluorescent dye [35]. VLPs are then selected based on their size and fluorescence level, and are removed from the sample using fluorescence-activated cell sorting [35]. While this method still leads to the loss of VLPs, and decreases sensitivity of viral detection [36], it significantly reduces background contamination and can eliminate the need for whole-genome amplification before sequencing [35]. The classical VLP concentration and purification methods are described in depth elsewhere [34,37,38]. However, given that background contamination is currently unavoidable [31,34], viral sequences should be checked against, and purified from, contaminating host sequences before any further analysis. Because every available sample-processing method has its limitations and biases, the study of less described phages is dependent on a bioinformatics approach, which comes with its own set of advantages and challenges (Figure 1).

Current Tools and Viral Databases

As phages lack a universal marker, such as the 16S ribosomal RNA genes in bacteria, they can be hard to identify in a mixed sample [1,7]. Shotgun sequencing of VLP-derived DNA or RNA is one solution to the issues of metabarcoding that relies on species- or group-specific markers. Metagenomics allows for culture-independent sequencing of a complex microbial sample without needing group-specific primers and can distinguish between the different species contained within the sample. Metagenomic data are, however, prone to high background noise that confounds current methods used for viral taxonomic characterization [39].

To address the poor, incorrect or insufficient, annotations present in public databases and limited homology between viral sequences to reference databases, viromics studies rely on *de novo* assembly to recover viral genomes from metagenomes [40]. However, this assembly can be challenging due to the specific characteristics of viral metagenomes: they are highly mosaic, include many repeat regions within the genomes, and show high metagenomic complexity and strain-

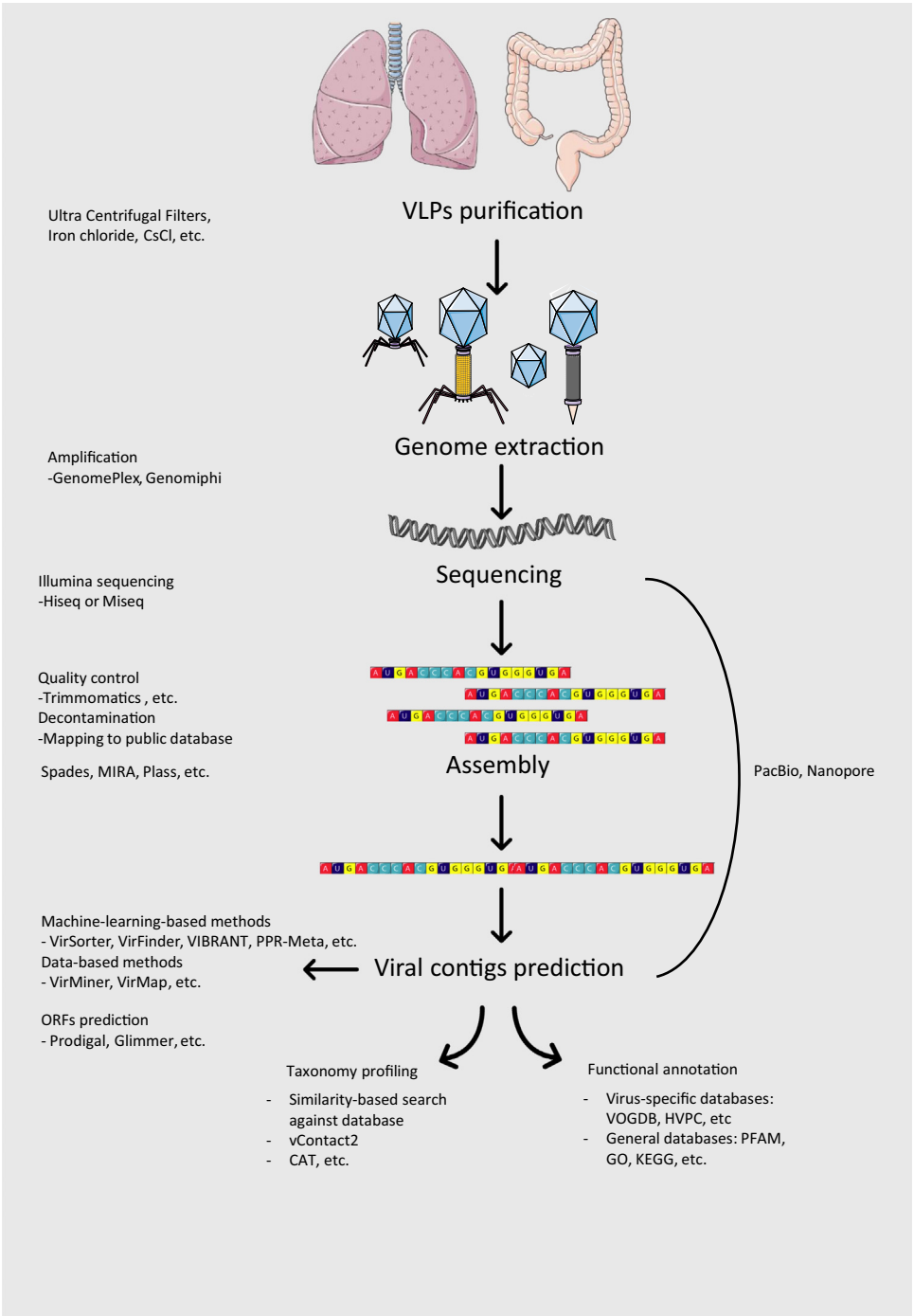


Figure 1. Experimental and Computational Approaches for the Characterization of the Free Phage Fraction of the Human Microbiota. Lung and gut illustrations represent respiratory and gastrointestinal tracts that contain the highest number of virus-like particles (VLPs) in the body. The arc indicates two sequencing platforms that do not need an assembly step. Abbreviation: ORFs, open reading frames.

level diversity [19,36,40]. The microdiversity (high level of strain evenness and nucleotide diversity) of abundant phages can also complicate *de novo* assembly [36,40,41]. Protein-level assemblers, such as Plass [42], can be better tools to use on viral metagenomic data as they predict novel proteins from nucleotide sequences, increase sequence recovery, and improve protein function prediction. They also help avoiding the mismatches from synonymous single-nucleotide polymorphisms [42]. Yet, these assemblers cannot place the assembled protein sequences into a genomic context, and they are unable to resolve homologous proteins from closely related taxa with <95% sequence identity [42]. Long-read sequencers, such as Nanopore or PacBio, could potentially be used to recover a complete phage genome within a single read without the need for assembly. However, long-read sequencers require micrograms of DNA, that is orders of magnitude more than the nanograms usually isolated from a virome sample without amplification, and they still have a relatively high error rate and operating cost [36].

Once a complete or partial phage genome sequence has been assembled, functional annotation is conducted to understand the biological meaning of the predicted genes. For protein-coding genes, open reading frames (ORFs) are predicted using Prodigal [43] and Glimmer [44] or other tools, and they are aligned to protein databases for functional annotation [45]. Virus-specific databases, such as VOGDB [46], HVPC [47], pVOGs [48], GLUVAB [49], IMG/VR [50], Virus-Host DB [51], MVP [52], or general functional annotation databases such as PFAM [53], GO [54], EggNOG [55,56], COGNIZER [57], or KEGG [58], are commonly used for functional annotation (Table 1). As most sequences in general functional annotation databases are derived from the genomes of cellular organisms, this leads to a poor coverage for viral proteins. Meanwhile, virus-specific databases require significant improvement by adding more viral sequences.

Unknown Viruses and Discovery

With the advances in sequencing technology, the total number of uncultivated virus sequences that are identified each year is by far (e.g., five times between 2017 and 2019) more than sequences of virus isolates [59]. As a result, uncultivated viruses already represent the majority (≥95%) of the viral diversity in public databases [59]. Minimum Information about an Uncultivated Virus Genome (MIUViG) standards are being developed within the Genomic Standards Consortium framework to improve the reporting of uncultivated virus genomes. MIUViG asks for information about virus origin, genome quality, genome annotation, taxonomic classification, biogeographic distribution, and *in silico* host prediction for novel uncultivated viruses [59].

Most viral sequences show no significant homology to known reference sequences [7,60]. An alternative to database-dependent methods is clustering viral sequences by composition. For example, VirMap data processing can detect low coverage and highly divergent viruses and allows for recovery and reconstruction of viral information when closely related database entries are non-existent [39]. Similarly, virMine is not restricted by insufficient viral diversity represented in public databases, and instead scores contigs based on their comparison with both viral and nonviral sequences [61], while PHAST and PHASTER are two web server tools that use public databases for identification and annotation of prophages within bacterial genomes [62]. However, some of the detected prophages may be nonfunctional, secondary to deletions or mutations of essential genes [63].

Machine-learning methods can also be used to detect viral sequences. For example VirFinder [64] uses k-mer profiles to predict viral contigs, VIBRANT utilizes hybrid machine learning and a protein similarity approach that is independent of sequence features for viral sequence recovery [65], and Virsorter detects viral signals using a combination of reference-dependent and reference-independent approaches [66]. Likewise, MARVEL uses a random forest machine learning approach to predict dsDNA phage sequences in metagenomic bins [67]. The recently developed

Table 1. Software for Predicting Phage Hosts

Tool name	Description	Refs
Tools for sample processing		
Amicon Ultra Centrifugal Filters	Using low-binding Ultracel regenerated cellulose membranes allows for high-throughput VLP concentration and recovery	[28]
Purelink viral RNA/DNA kit	Allows simultaneous extraction of high-quality DNA and RNA from biological material	[24]
eMAG	A fully automated nucleic acid extraction platform that enables simultaneous extraction of viral genomic material from 48 specimens	[29]
GenomePlex, WGA	A whole-genome amplification (WGA) kit for the rapid and highly representative amplification of genomic DNA from minimal amounts of starting material	[12]
Iron chloride	Useful for concentrating virus particles from large-volume samples	[38]
MAF	Uses a hydrolysed macroporous epoxy-based polymer system to concentrate and purify waterborne viruses	[30]
MagNA PURE96	Another fully automated extraction system that allows simultaneous extraction of 96 specimens using magnetic bead technology	[29]
Tools for viral recovery		
DeepVirFinder	An alignment-free tool that identifies viral sequences in metagenomes using machine learning	[68]
MARVEL	A tool that uses machine learning for prediction of dsDNA phages in metagenomes	[67]
PPR-Meta	A 3-class classifier that allows identification of phages from metagenomic assemblies with enhanced performance for short fragments	[69]
PHASTER	A web-based tool for identifying and annotating prophage sequences within bacterial genomes	[62]
VIBRANT	An automated tool that uses a hybrid machine-learning and protein-similarity approach to recover and annotate viruses of microbes	[65]
VirFinder	A novel *k*-mer-based tool that identifies viral sequences from assembled metagenomic data	[64]
VirMAP	Uses a combination of nucleotide and protein signals to taxonomically classify viral sequences independently of genome coverage or read overlap	[39]
virMine	Can identify viral genomes from collective raw reads within metagenomes of different environments	[61]
Virsorter	Can recover novel viruses in metagenomic data using both reference-dependent and reference-independent approaches	[66]
Functional annotation databases		
COGNIZER	A comprehensive stand-alone annotation framework that allows functional annotation of sequences from metagenomic data	[57]
EggNOG	A public source for orthologous groups (OGs) of proteins at different taxonomic levels, with integrated functional annotations	[56]
GO	A comprehensive knowledge-based resource of gene functions	[54]
KEGG	The gold standard database for understanding functions of the different biological systems from large-scale molecular datasets	[58]
PFAM	A database of manually curated protein families, containing 14 831 Pfam-A families	[53]
Software for predicting phage hosts		
HostPhinder	Uses genomic similarity to a reference database of phages with known hosts to predict hosts for uncharacterized phages	[88]
IMFH-VH	Kernelized Logistic Matrix Factorization based on Similarity Network Fusion for predicting virus–host association	[89]
PHISDetector	Uses several interaction signals, including CRISPR and protein–protein interaction, to predict novel phage–host pairings	[86]

Table 1. (continued)

Tool name	Description	Refs
viruses_classifier	An alignment-free approach for distinction between phages and eukaryotic viruses in metagenome data	[90]
VirHostMatcher	Uses similar oligonucleotide frequency patterns between phages and bacteria to predict host range for phages on a genus level	[85]
WIsH	Fast and accurate, making it suitable for predicting phages' host range from metagenomic data	[87]
Virus-specific databases		
HVPC	A human viral protein database for diversity and functional annotation	[47]
IMG/VR	An integrated reference database of both cultured and uncultured DNA/RNA viruses	[50]
MVP	A microbe–phage interaction database with over 30 k viral clusters, gathered from public databases and microbiome sequences	[52]
pVOGS	Represents a complete set of orthologous gene families shared across multiple complete genomes of bacterial or archaeal viruses	[48]
VGDB	A public webserver that provides information about the virus orthologous groups	[46]
Virus-Host DB	Includes complete genomes of viruses and their hosts gathered from RefSeq, GenBank, UniProt, and ViralZone	[51]

DeepVirFinder improved the accuracy of viral identification for both long and short sequences using deep learning methods [68], while PPR-Meta uses deep-learning, Bi-path Convolutional Neural Network, to detect phage and plasmid sequences in metagenome assemblies simultaneously. PPR-Meta was specifically developed to improve the identification performance for short fragments [69]. However, a major drawback to viral discovery tools is that they are as efficient as the dataset they were trained on, which can lead to false positives in high confidence scored viral contigs.

For taxonomic classification, assembled contiguous sequences (contigs) are compared with annotated virus databases, either using a best-hit approach, BLAST [70], or a voting system that considers all ORFs, CAT and BAT [71]. The latter approach works best with contigs longer than 1 kb as they contain multiple ORFs [71]. For uncultivated virus sequences with no hits in reference databases, a gene-sharing network, such as vConTACT2, in which viruses are clustered together based on shared genes, can be used to automatically assign tentative taxonomy [72]. ViPTree is a web server that uses protein alignment for phylogenetic analysis and classification of viruses [73]. Concatenated protein phylogeny can also be used for classification of tailed dsDNA viruses [74]. One challenge to the taxonomic assignment of viruses is the dominance of predicted ORFs in which combined taxonomic signals may enhance the classification of unknown sequences [75].

Identifying Unculturable Phages' Host Range

How to determine a phages' host range, that is, the different bacteria it can infect, is a contentious topic of discussion, starting with the definition of infection. The phage infection cycle consists of six main stages. The first step is absorption of the phage into the bacterial cell. Second, the phage ejects its DNA into the host cell. Third, defence mechanisms are evaded. Fourth, the bacterial machinery is hijacked, turning the host into a virocell [76]. Fifth, the phage replicates and builds a new generation of phages. The sixth and final step is lysis of the bacterial cell [77]. Up to seven different types of host range determination methods have been described in depth elsewhere [78,79]. Standard methods for host determination, such as efficiency of plating (EOP), are culture-dependent and the results vary between different methods [80], which makes the host-range

determination for unculturable phages difficult (Figure 2). Alternatives to culture-dependent methods are viral-tagging or *in silico* abundance profiles, determination of tRNAs or prophages, or CRISPR recorded short phage segments [23].

There are a number of different culture-independent methods that can be used to measure phage host range. Viral-tagging uses fluorescence-activated cell sorting to separate out fluorescently labelled phages that are attached to a bacterial cell for further downstream applications and sequencing [18]. While attachment does not equal absorption or replication, it links to the first step of the phage infection cycle and has been demonstrated to successfully predict unique host–phage pairings in both the marine [18,81,82] and human [23] environments. For example, a recent study revealed a total of 363 unique bacteria–phages interactions within the faecal samples from 11 healthy volunteers [23].

Abundance profiles are another culture-independent approach to link phages to hosts by using (lagged) correlations in phage and bacterial abundance patterns. While promising in theory, the complex dynamics [7] underlying the interactions between phages and their hosts tend to defy straightforward correlation analysis, yielding low accuracy of this approach [1,83]. Genetic signatures can sometimes be used to link phages to their bacterial hosts; they are largely associated with the fifth step of the phage infection cycle. The most commonly used genetic signatures are: (i) horizontal gene transfer leading to genetic homology between phage and bacteria, though this is dependent on a comprehensive database [83]; (ii) prophage integration into host genomes, though this is limited to temperate phages [24]; (iii) the recording of a short segment of an infecting phage using CRISPRs to prevent reinfection, which can be used to identify the phage in question [83] – however, only ~10% of bacteria encode a CRISPR system in the first place which can be

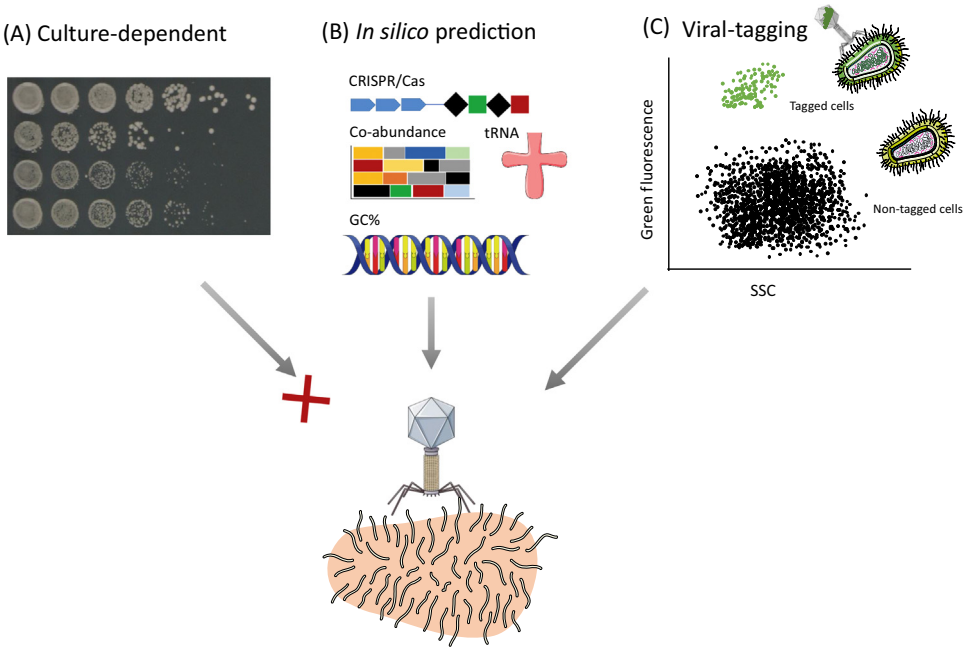


Figure 2. Overview of Different Methods Used for the Analysis of Phages' Host Range. (A) Spot test is a culture-dependent method which relies on the ability of the phage to lyse its potential host. This method fails to indicate the host range of unculturable phages. (B) *In silico* prediction and (C) viral-tagging: two culture-independent methods for determining the host range of unculturable phages. Abbreviation: SSC, side scatter.

identified using state-of-the-art algorithms [83,84]; and (iv) tracking of viral tRNA thought to originate with the host [68], though this is not specific at a species level and only 7% of known phages have tRNA sequences [23]. Due to these limitations, most *in silico* tools, such as VirHostMatcher [85], PHISDetector [86], and WIsH [87], combine multiple genetic signatures to predict phages' host range (Table 1). For more virus–host prediction tools see [88–90].

Statistical Analysis of Multidimensional Data

Analyzing multidimensional (-omics) data to elucidate species–environment relationships is a challenge currently faced by many disciplines [91]. Recent advances in computational and statistical approaches, such as machine learning, have helped to address this issue (including new tool kits that combine neural networks, random forests, and indicator species analysis to identify key players in driver-response relationships) [92]. These approaches, however, are data hungry, requiring a large number of observations. In cases where small sample size precludes the use of machine-learning approaches, canonical methods likewise offer a promising path forward in analyzing species–environment relationships [93,94]. For instance, Multiple Factor Analysis (MFA) [95] was applied in a recent study [2] to examine the multivariate correlation between dominant bacterial and phage species, and environmental metadata. To reduce dimensionality prior to running the MFA, dominant bacterial and phage species were first determined by conducting Principal Component Analyses (PCAs) on each community matrix and identifying the species that contributed most to significant PC axes. This dimension reduction improved the interpretation of the MFA, which in turn helped to relate changes in the gut microbiome to several environmental factors (i.e., health status, diet, age, and sex). Ultimately, the appropriateness of an approach, with respect to the data and question, is likely to vary among studies, and, echoing the conclusions made [92], the integration of multiple, complementary statistical methods will likely offer the most robust conclusions and help to untangle complex, multidimensional data.

Concluding Remarks

The human body is one of the densest and most diverse microbial habitats known. The viral fraction alone accounts for ~10^{12} VLPs (it varies from high: ~10^{9}–10^{10}/g faecal content and ~10^{8}/ml of respiratory fluids, to low: ~10^{6}/cm^{2} skin and ~10^{5}/ml blood), plus prophages in bacterial genomes [7,21,22]. Phages play critical roles in maintaining gut homeostasis through interacting with the bacterial community [96]. Understanding how phages regulate this complex microbial network will pave the way for the development of novel phage-based therapeutics to re-establish gut health in disease associated with dysbiosis such as inflammatory bowel disease. The novel advances in sequencing technology and bioinformatics have enabled rapid expansion in viral discovery. Yet, we have a way to go until the complete phage diversity (Box 1) in the environment and the human body has been revealed, and the functions of these phages have been elucidated. This is mainly due to isolation protocols and computational shortcomings, despite recent advances to better study phages. A current challenge is to develop suitable isolation methods and *in silico* analytics to better identify RNA phages. Future studies, in particular, should be adjusted to ensure that RNA phages are adequately represented.

The factors that are responsible for the bias found in phage metagenomics call for the scientific community to work together to improve the toolkits currently used in the field, in the laboratory, and *in silico*. For example, the application of single-cell technologies can significantly advance our understanding of phages in the human body by identifying their specific function and their interactions with host bacteria, and by revealing their impact on our health. We expect that the phage research field will benefit from near-future technological advancements as the world gets closer to completing the picture of global phage abundance, diversity, and distribution, as well as the interactions of phages with their bacterial hosts (see Outstanding Questions).

Outstanding Questions

What role do RNA phages play in the human microbiota? Does their nature of interaction with the bacterial community differ from DNA phages?

Do we need to first further study phages' biology to be able to improve the currently available detection methods in human samples to later understand the role of phages in human health and disease?

How can experimentalist and computationalists work together toward removing biases from virome analysis?

What roles can machine learning and big data analysis play in solving the viral dark matter in the human body?

Author Contributions

All authors listed have made a significant, direct, and intellectual contribution to the work and approved it for publication.

Acknowledgements

The authors thank Bas E. Dutlih, E. Haggard-ljungquist, and H. Foroughi-Asl for their constructive comments. Kawtar Tiamani created the box image; the image is adapted with permission from Moïra B. Dion *et al.* (10.1038/s41579-019-0311-5). Spot test image in Figure 2 is adapted from Wikimedia Commons; all other images were taken from Smart Medical Art (SMART): https://smart.servier.com. This work was funded by German Reaserch Foundation (DFG Emmy Noether program, Project No. 273124240; and SFB 1371, Project No. 395357507) and European Research Council starting grant (ERC StG 803077) awarded to L.D.

References

1. Khan Mirzaei, M. and Maurice, C.F. (2017) Ménage à trois in the human gut: interactions between host, bacteria and phages. *Nat. Rev. Microbiol.* 15, 397–408
2. Khan Mirzaei, M. *et al.* (2020) Bacteriophages isolated from stunted children can regulate gut bacterial communities in an age-specific manner. *Cell Host Microbe* 27, 199–212.e5
3. Gregory, A.C. The human gut virome database. *bioRxiv*. Published online May 31, 2019. https://www.biorxiv.org/content/10.1101/655910v1.full.
4. Edwards, R.A. *et al.* (2019) Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. *Nat. Microbiol.* 4, 1727–1736
5. Al-Shayeb, B. *et al.* (2020) Clades of huge phages from across Earth's ecosystems. *Nature* 578, 425–431
6. Minot, S. *et al.* (2013) Rapid evolution of the human gut virome. *PNAS* 110, 12450–12455
7. Shkoporov, A.N. and Hill, C. (2019) Bacteriophages of the human gut: The 'known unknown' of the microbiome. *Cell Host Microbe* 25, 195–209
8. Reyes, A. *et al.* (2010) Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* 466, 334–338
9. Dąbrowska, K. and Abedon, S.T. (2019) Pharmacologically aware phage therapy: pharmacodynamic and pharmacokinetic obstacles to phage antibacterial action in animal and human bodies. *Microbiol. Mol. Biol. Rev.* 83, e00012-19
10. Altamirano, F.L.G. and Barr, J.J. (2019) Phage therapy in the postantibiotic era. *Clin. Microbiol. Rev.* 32, e00066-18
11. Reyes, A. *et al.* (2012) Going viral: next-generation sequencing applied to phage populations in the human gut. *Nat. Rev. Microbiol.* 10, 607–617
12. Moreno-Gallego, J.L. *et al.* (2019) Virome diversity correlates with intestinal microbiome diversity in adult monozygotic twins. *Cell Host Microbe* 25, 261–272.e5
13. Khan Mirzaei, M. and Maurice, C.F. (2017) The mammalian gut as a matchmaker. *Cell Host Microbe* 22, 726–727
14. Brum, J.R. *et al.* (2013) Global morphological analysis of marine viruses shows minimal regional variation and dominance of non-tailed viruses. *ISME J.* 7, 1738–1751
15. Hopkins, M. *et al.* (2014) Diversity of environmental single-stranded DNA phages revealed by PCR amplification of the partial major capsid protein. *ISME J.* 8, 2093–2103
16. Sutton, T.D.S. and Hill, C. (2019) Gut bacteriophage: current understanding and challenges. *Front. Endocrinol.* 10, 784
17. Callanan, J. *et al.* (2020) Expansion of known ssRNA phage genomes: From tens to over a thousand. *Sci. Adv.* 6, eaay5981
18. Deng, L. *et al.* (2014) Viral tagging reveals discrete populations in *Synechococcus* viral genome sequence space. *Nature* 513, 242–245
19. Ye, S.H. *et al.* (2019) Benchmarking metagenomics tools for taxonomic classification. *Cell* 178, 779–794
20. Carding, S.R. *et al.* (2017) Review article: the human intestinal virome in health and disease. *Aliment. Pharmacol. Ther.* 46, 800–815
21. Haynes, M. and Rohwer, F. (2010) The human virome. In *Metagenomics of the Human Body* (Nelson, K.E., ed.), pp. 63–77, Springer
22. Rascovan, N. *et al.* (2016) Metagenomics and the human virome in asymptomatic individuals. *Annu. Rev. Microbiol.* 70, 125–141
23. Džunková, M. *et al.* (2019) Defining the human gut host–phage network through single-cell viral tagging. *Nat. Microbiol.* 4, 2192–2203
24. Sathiamoorthy, S. *et al.* (2018) Selection and evaluation of an efficient method for the recovery of viral nucleic acids from complex biologicals. *npj Vaccines* 3, 1–6
25. Conceição-Neto, N. *et al.* (2015) Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. *Sci. Rep.* 5, 1–14
26. Hoyles, L. *et al.* (2014) Characterization of virus-like particles associated with the human faecal and caecal microbiota. *Res. Microbiol.* 165, 803–812
27. Callanan, J. *et al.* (2018) RNA phage biology in a metagenomic era. *Viruses* 10, 386
28. Shkoporov, A.N. *et al.* (2018) Reproducible protocols for metagenomic analysis of human faecal phageomes. *Microbiome* 6, 68
29. Hindiyeh, M. *et al.* (2019) Comparison of the new fully automated extraction platform eMAG to the MagNA PURE 96 and the well-established easyMAG for detection of common human respiratory viruses. *PLoS ONE* 14, e0211079
30. Pei, L. *et al.* (2012) Combination of crossflow ultrafiltration, monolithic affinity filtration, and quantitative reverse transcriptase PCR for rapid concentration and quantification of model viruses in water. *Environ. Sci. Technol.* 46, 10073–10080
31. Cook, L. *et al.* (2018) Does size matter? Comparison of extraction yields for different-sized DNA fragments by seven different routine and four new circulating cell-free extraction methods. *J. Clin. Microbiol.* 56, e01061-18
32. Chen, M. *et al.* (2014) Comparison of multiple displacement amplification (MDA) and multiple annealing and looping-based amplification cycles (MALBAC) in single-cell sequencing. *PLoS One* 10, 1371
33. Roux, S. *et al.* (2016) Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. *PeerJ* 4, e2777
34. Kleiner, M. *et al.* (2015) Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes. *BMC Genom.* 16, 7
35. Džunková, M. *et al.* (2015) Direct sequencing of human gut virome fractions obtained by flow cytometry. *Front. Microbiol.* 6, 955
36. Warwick-Dugdale, J. *et al.* (2019) Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. *PeerJ* 7, e6800
37. Castro-Mejía, J.L. *et al.* (2018) Extraction and purification of viruses from fecal samples for metagenome and morphology analyses. *Methods Mol. Biol.* 1838, 49–57
38. Poulos, B.T. *et al.* (2018) Iron chloride flocculation of bacteriophages from seawater. *Methods Mol. Biol.* 1681, 49–57
39. Ajami, N.J. *et al.* (2018) Maximal viral information recovery from sequence data using VirMAP. *Nat. Commun.* 9, 1–9
40. Sutton, T.D.S. *et al.* (2019) Choice of assembly software has a critical impact on virome characterisation. *Microbiome* 7, 12
41. Willner, D. *et al.* (2009) Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS ONE* 4, e7370

42. Steinegger, M. *et al.* (2019) Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat. Methods* 16, 603–606
43. Hyatt, D. *et al.* (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformat.* 11, 119
44. Kelley, D.R. *et al.* (2012) Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic Acids Res.* 40, e9
45. McNair, K. *et al.* (2018) Phage genome annotation using the RAST pipeline. In *Bacteriophages: Methods and Protocols* (vol. 3) (Clokie, M.R.J. *et al*., eds), pp. 231–238, Springer
46. Thannesberger, J. *et al.* (2017) Viruses comprise an extensive pool of mobile genetic elements in eukaryote cell cultures and human clinical samples. *FASEB J.* 31, 1987–2000
47. Elbehery, A.H.A. *et al.* (2018) The human virome protein cluster database (HVPC): a human viral metagenomic database for diversity and function annotation. *Front. Microbiol.* 9, 1110
48. Grazziotin, A.L. *et al.* (2017) Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Res.* 45, D491–D498
49. Coutinho, F.H. *et al.* (2019) Charting the diversity of uncultured viruses of Archaea and Bacteria. *BMC Biol.* 17, 109
50. Paez-Espino, D. *et al.* (2019) IMG/VR v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes. *Nucleic Acids Res.* 47, D678–D686
51. Mihara, T. *et al.* (2016) Linking virus genomes with host taxonomy. *Viruses* 8, 66
52. Gao, N.L. *et al.* (2018) MVP: a microbe-phage interaction database. *Nucleic Acids Res.* 46, D700–D707
53. Finn, R.D. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.* 42, D222–D230
54. The Gene Ontology Consortium (2019) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* 47, D330–D338
55. Jensen, L.J. *et al.* (2008) eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res.* 36, D250–D254
56. Huerta-Cepas, J. *et al.* (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 47, D309–D314
57. Bose, T. *et al.* (2015) COGNIZER: A framework for functional annotation of metagenomic datasets. *PLoS ONE* 10, e0142102
58. Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 28, 27–30
59. Roux, S. *et al.* (2019) Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nat. Biotechnol.* 37, 29–37
60. Paez-Espino, D. *et al.* (2016) Uncovering Earth's virome. *Nature* 536, 425–430
61. Garretto, A. *et al.* (2019) virMine: automated detection of viral sequences from complex metagenomic samples. *PeerJ* 7, e6695
62. Arndt, D. *et al.* (2016) PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* 44, W16–W21
63. Bobay, L.-M. *et al.* (2014) Pervasive domestication of defective prophages by bacteria. *PNAS* 111, 12127–12132
64. Ren, J. *et al.* (2017) VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* 5, 69
65. VIBRANT: Automated recovery, annotation and curation of microbial viruses, and evaluation of virome function from genomic sequences. *bioRxiv*. Published online November 26, 2019. https://www.biorxiv.org/content/10.1101/855387v1.
66. Roux, S. *et al.* (2015) VirSorter: mining viral signal from microbial genomic data. *PeerJ* 3, e985
67. Amgarten, D. *et al.* (2018) MARVEL, a tool for prediction of bacteriophage sequences in metagenomic bins. *Front. Genet.* 9, 304
68. Ren, J. *et al.* (2020) Identifying viruses from metagenomic data using deep learning. *Quant. Biol.* 8, 64–77
69. Fang, Z. *et al.* (2019) PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. *Gigascience* 8, 1–14
70. Altschul, S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410
71. von Meijenfeldt, F.A.B. *et al.* (2019) Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol.* 20, 217
72. Bin Jang, H. *et al.* (2019) Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* 37, 632–639
73. Nishimura, Y. *et al.* (2017) ViPTree: the viral proteomic tree server. *Bioinformatics* 33, 2379–2380
74. Low, S.J. *et al.* (2019) Evaluation of a concatenated protein phylogeny for classification of tailed double-stranded DNA viruses belonging to the order Caudovirales. *Nat. Microbiol.* 4, 1306–1315
75. Hatfull, G.F. (2015) Dark matter of the biosphere: the amazing world of bacteriophage diversity. *J. Virol.* 89, 8107–8110
76. Forterre, P. (2013) The virocell concept and environmental microbiology. *ISME J.* 7, 233–236
77. Ross, A. *et al.* (2016) More is better: selecting for broad host range bacteriophages. *Front. Microbiol.* 7, 1352
78. Hyman, P. (2019) Phages for phage therapy: isolation, characterization, and host range breadth. *Pharmaceuticals (Basel)* 12, 35
79. Hyman, P. and Abedon, S.T. (2010) Bacteriophage host range and bacterial resistance. *Adv. Appl. Microbiol.* 70, 217–248
80. Khan Mirzaei, M. and Nilsson, A.S. (2015) Isolation of phages for phage therapy: a comparison of spot tests and efficiency of plating analyses for determination of host range and efficacy. *PLoS One* 10, 1371
81. Roux, S. *et al.* (2014) Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics. *eLife* 3, e03125
82. Labonté, J.M. *et al.* (2015) Single-cell genomics-based analysis of virus–host interactions in marine surface bacterioplankton. *ISME J.* 9, 2386–2399
83. Edwards, R.A. *et al.* (2016) Computational approaches to predict bacteriophage–host relationships. *FEMS Microbiol. Rev.* 40, 258–272
84. Burstein, D. *et al.* (2016) Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. *Nat. Commun.* 7, 10613
85. Ahlgren, N.A. *et al.* (2017) Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res.* 45, 39–53
86. Zhang, F. *et al.* (2019) PHISDetector: a web tool to detect diverse in silico phage–host interaction signals. *bioRxiv* Published online June 10, 2019. https://doi.org/10.1101/661074
87. Galiez, C. *et al.* (2017) WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics* 33, 3113–3114
88. Villarroel, J. *et al.* (2016) HostPhinder: a phage host prediction tool. *Viruses* 8, 116
89. Liu, D. *et al.* (2019) Predicting virus–host association by Kernelized logistic matrix factorization and similarity network fusion. *BMC Bioinformat.* 20, 594
90. Galan, W. *et al.* (2019) Host Taxon Predictor – a tool for predicting taxon of the host of a newly discovered virus. *Sci. Rep.* 9, 3436
91. Pavlopoulos, G.A. *et al.* (2013) Unraveling genomic variation from next generation sequencing data. *BioData Min.* 6, 13
92. Thompson, J. *et al.* (2019) Machine learning to predict microbial community functions: An analysis of dissolved organic carbon from litter decomposition. *PLoS ONE* 14, e0215502
93. Buttigieg, P.L. and Ramette, A. (2014) A guide to statistical analysis in microbial ecology: a community-focused, living review of multivariate data analyses. *FEMS Microbiol. Ecol.* 90, 543–550
94. Dhariwal, A. *et al.* (2017) MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic Acids Res.* 45, W180–W188
95. Borcard, D. *et al.* (2018) *Numerical Ecology with R* (2nd edn), Springer International
96. Koskella, B. (2019) New approaches to characterizing bacteria–phage interactions in microbial communities and microbiomes. *Environ. Microbiol. Rep.* 11, 15–16
97. Adriaenssens, E.M. *et al.* (2020) Taxonomy of prokaryotic viruses: 2018–2019 update from the ICTV Bacterial and Archaeal Viruses Subcommittee. *Arch. Virol.* 165, 1253–1260
98. Dion, M.B. *et al.* (2020) Phage diversity, genomics and phylogeny. *Nat. Rev. Microbiol.* 18, 125–138