ANNUAL REVIEWS

Annual Review of Biomedical Data Science

Illuminating the Virosphere Through Global Metagenomics

Lee Call, Stephen Nayfach, and Nikos C. Kyrpides

Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA; email: nckyrpides@lbl.gov, lcall@lbl.gov

Annu. Rev. Biomed. Data Sci. 2021. 4:369–91

The *Annual Review of Biomedical Data Science* is online at biodatasci.annualreviews.org

https://doi.org/10.1146/annurev-biodatasci-012221-095114

Keywords

virome, viral genomics, metagenomics, viral ecology

Abstract

Viruses are the most abundant biological entity on Earth, infect cellular organisms from all domains of life, and are central players in the global biosphere. Over the last century, the discovery and characterization of viruses have progressed steadily alongside much of modern biology. In terms of outright numbers of novel viruses discovered, however, the last few years have been by far the most transformative for the field. Advances in methods for identifying viral sequences in genomic and metagenomic datasets, coupled to the exponential growth of environmental sequencing, have greatly expanded the catalog of known viruses and fueled the tremendous growth of viral sequence databases. Development and implementation of new standards, along with careful study of the newly discovered viruses, have transformed and will continue to transform our understanding of microbial evolution, ecology, and biogeochemical cycles, leading to new biotechnological innovations across many diverse fields, including environmental, agricultural, and biomedical sciences.

WHY STUDY VIRUSES?

Bacteriophage (phage): a virus that infects bacteria or archaea

Viral particles: free viruses in the environment not actively infecting a host; also known as virions

The number of viruses on Earth is estimated at 10^{31} (1–3). Viruses infect all domains of life (4) and contribute significantly to global biogeochemical cycles (5). Viral infections drive evolutionary changes in host genomes, such as the development of virus resistance mechanisms, and can even lead to hosts co-opting viral genes. For example, nearly one-tenth (8%) of the human genome sequence is estimated to be of viral origin (6), including the syncytin gene, which plays a critical role in placental development (7). Despite such dramatic examples from our own biology, recent environmental metagenomic surveys have largely uncovered viral species targeting prokaryotes—bacteria and archaea—reflecting the likely predominance of these types of viruses throughout the world (3). Thus, while this review will touch on eukaryotic viruses, the emphasis will be on prokaryotic viruses.

THE EARLY DAYS OF VIROLOGY

Although van Leeuwenhoek amazed the seventeenth-century world with his descriptions of tiny, microscopic organisms, a major part of the microbial world remained hidden from his view. In the 1890s, Ivanovski and Beijerinck both claimed that something smaller than bacteria was infecting tobacco plants (see **Figure 1**) (8). Later, in 1915 and 1917, Twort and d'Herelle independently described infectious agents that could kill bacteria, dubbing them "bacteriophages" (9). In the years that followed, there were great debates over the nature of such filterable infectious agents: whether they were alive or not, and whether they were liquid or particulate in form (10). In the 1930s, one of the major motivations for the development of electron microscopy (EM) was the desire to observe infectious viral particles (11).

Reaching magnification levels as high as 1,000,000× and resolving structures measuring only a few nanometers (much smaller than light microscopes, whose specimens are typically measured in micrometers), EM thus allowed individual viruses to be distinguished based on their sizes and overall viral structures (12). The shape of a virus particle is primarily determined by repeating protein subunits combined in specific geometric patterns to form the viral capsid (see the sidebar titled What Is a Virus?). Further divisions could be made based on the presence of other defining structures such as a tail-like appendage in some bacteriophages.

GLIMPSING EARTH'S VIRAL DIVERSITY

The detailed descriptions made possible by EM drove efforts to classify viruses into groups with related structures. For example, in 1967 Eisenstark published morphological descriptions, including EM images, of the 111 known bacteriophages; he was followed by Ackermann, who reviewed phages as a group in a series of publications spanning several decades (13, 14). In the 1980s, EM was used to observe that the abundance and diversity of bacteriophages in samples of seawater were particularly high (15, 16). Following these reports, researchers looked for viruses in marine samples from all over the world, and the observed numbers, both in quantity and diversity, were orders of magnitude greater than previously thought (17–19). Bacteriophages were the most abundant, estimated at millions of bacteriophages per milliliter of water (20).

Although these findings brought increased interest in looking for viruses throughout different environments, the sheer amount of viral diversity posed a challenge, as manual curation of EM images can be meticulous and labor-intensive. So researchers turned to other techniques like gel electrophoresis, which could create a molecular fingerprint of all the viruses in a single sample

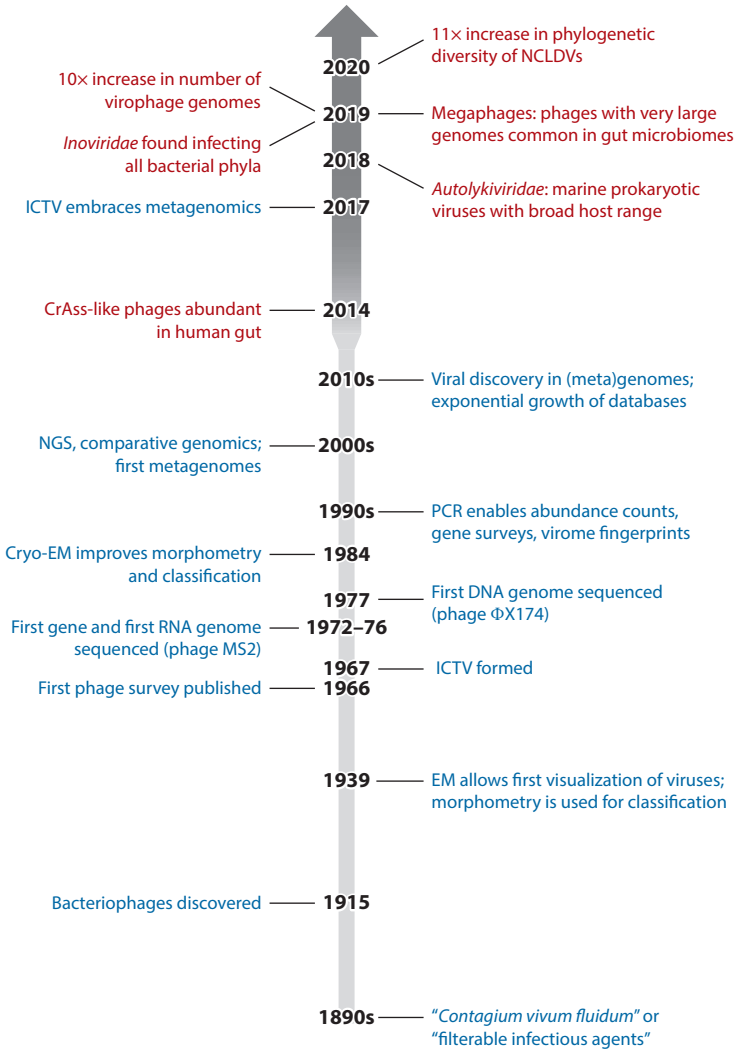


Figure 1

Major developments in viral discovery. A timeline of some of the most significant discoveries and events in virology leading to the current era of metagenomics methods. "*Contagium vivum fluidum*" (contagious living fluid) was the description Beijerinck used to describe tobacco mosaic virus. The scale of the timeline has been expanded following 2010. Items in red text highlight prominent examples of recent publications using metagenomic approaches (114–119). Abbreviations: cryo-EM, cryogenic EM; EM, electron microscopy; ICTV, International Committee on Taxonomy of Viruses; NCLDVs, nucleocytoplasmic large DNA viruses; NGS, next-generation sequencing; PCR, polymerase chain reaction.

based on the length of their genomes, allowing researchers to compare the composition of many viral communities across multiple samples (21–24).

Another common technique made use of the growing number of viral gene and genome sequences available in public data repositories. Using this information, researchers developed PCR (polymerase chain reaction)-based assays targeting genes unique to specific groups of viruses, which could then be used to detect those viruses in a sample. This technique enabled some of the

WHAT IS A VIRUS?

Viral particles are not only much smaller than most cellular life but also are much simpler, typically composed of a short piece of DNA or RNA surrounded by a layer of protein called a capsid and, in some viruses, of a lipid membrane called the viral envelope (120). Viruses have no metabolism of their own, so in order to replicate they must infect a host cell and take advantage of its metabolic machinery, followed by the release of new copies of the virus into the environment. Viruses lead one of three different lifestyles: (*a*) lytic, where host infection is followed directly by replication and release of new viral particles, killing the host in the process; (*b*) chronic, where viral replication takes place over long periods of time and release of new virions is less than lethal to the host; or (*c*) lysogenic, where the viral genome is integrated into the host's and remains dormant for a period of time before being activated and lysing the cell (121).

first quantitative comparisons of the distribution of specific viruses across space and time (25). It was also possible to compare the sequences of PCR-amplified genes, further increasing resolution and demonstrating that even among morphologically similar viruses, diversity at the sequence level is quite high (26, 27). As these techniques were applied to a wide variety of viruses, one of the striking findings was that in many cases highly similar viral sequences could be detected across very broad geographic ranges (28–30). The major limitation of these approaches, however, is that each assay must target an individual viral group, as there is no universal marker gene for all viruses, unlike the 16S and 18S ribosomal genes widely used for PCR-based surveys of prokaryotes and eukaryotes.

Assembly: the bioinformatic method for combining the relatively short reads produced by the sequencer into longer sequences called contigs and scaffolds

IMG/VR: a comprehensive viral database, including over two million high-quality genomes and genome fragments derived from metagenomes

Virome: the entire collection of viruses in a given environment or sample or a sample enriched for viral particles

Metagenome: a collection of sequences representing the genomes from multiple organisms found together in a single sample

METAGENOMICS AND THE MODERN APPROACH TO VIRAL DETECTION

Starting in the mid-2000s and continuing until today, sequencing technologies have advanced and costs have come down by several orders of magnitude (31). These advancements have fueled the rise of viral metagenomic sequencing, which involves sequencing the total viral DNA or RNA from individual samples (e.g., soil, seawater, or host-associated), bypassing any culturing in the laboratory. Typically, DNA or RNA is extracted from an environmental sample, fragmented, and then sequenced, generating millions of short reads (e.g., 100–200 bp) that are assembled into contigs. Metagenomic viral contigs are then identified using computational tools and algorithms that use a variety of viral-specific sequence features and signatures, providing unprecedented resolution on viral genomic diversity (32–36). However, metagenomic assembly is challenging, particularly for viruses with repetitive genomic elements, viruses from diverse subpopulations, or viruses present at low abundance (37). To address these challenges, researchers have found long-read sequencing (e.g., Oxford Nanopore and PacBio technologies) to be useful for sequencing viral genomes and transcriptomes from environmental samples without the need for assembly (38–40). These technological advances have fueled an unprecedented explosion in the amount of viral sequence data generated by various labs around the world. The largest database of viral genomes is IMG/VR (Integrated Microbial Genomes/Virus) at the Department of Energy's Joint Genome Institute (41), which houses over two million sequences derived from mostly uncultivated viruses.

There are three main strategies for sequencing viruses from the environment: virome sequencing, bulk metagenome sequencing, and single-cell sequencing. The majority of viral studies to date perform virome sequencing, which involves size filtration to enrich for the viral fraction before sequencing, similar to what is typically done before a microscopy-based analysis. Enriching for the viral fraction (the so-called virome) results in improved sequence coverage of viruses by

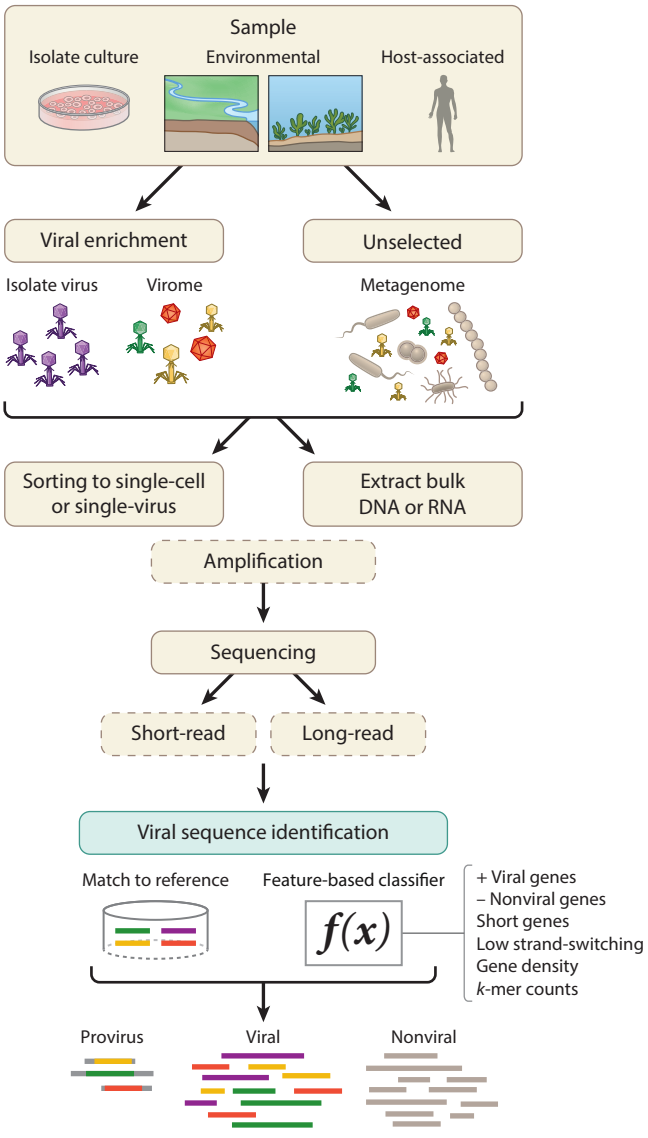


Figure 2

Workflow for identifying viral sequences in most common sample types. The principal steps required for sequence-based discovery of viruses from cultures, environmental samples, and host-associated samples (skin, gut, etc.). Bifurcating arrows indicate decisions based on different detection strategies: whether to exclude cells and whether to extract nucleic acids from single viruses/cells or from the entire sample population. The amplification step is surrounded by dashed lines to indicate that it is optional for most workflows, depending on the amount of starting material and the sequencing requirements. Likewise, there are options when choosing the type of sequencing platform.

minimizing the number of reads wasted on the sample's cellular fraction (see **Figure 2**). For researchers primarily interested in the virome, this approach increases the viral signal-to-noise ratio in the resulting sequence data, although some contaminating cellular sequences and plasmids often remain and must be removed for most analyses (42, 43). However, due to low sample

biomass, genome amplification techniques (e.g., multiple displacement amplification) are often employed, which can distort abundance and result in marked biases, including overamplification of small circular single-stranded DNA viruses (44–46). This approach may also exclude viruses that lie outside the standard fractionation size and weight cutoffs used (47), viruses that are attached to cells, viruses replicating inside of cells, and temperate viruses that have integrated into the host's genome.

Provirus: a viral sequence that has been integrated into the host's genome and is not actively being replicated to produce new viral particles

An alternative to virome sequencing is to skip the viral enrichment step altogether and sequence the bulk DNA or RNA found in a sample (see **Figure 2**), followed by computational separation of viral and cell-derived sequences. This approach greatly expands the number of samples that can be mined for viruses, as it includes metagenomics datasets collected to address other scientific objectives and is primarily responsible for the exponential growth in viral sequence databases over the last few years. Additionally, this approach allows for identifying proviruses that have integrated into a host genome, although it remains challenging to identify the sequence boundary between virus and host and to distinguish between viable proviruses and ancient or degraded remnants of proviruses (48). Lastly, since sequences from both viral and cellular origins are produced from the same sample, this approach allows for additional analysis based on associating the viruses identified in the sample together with their presumptive hosts. The major downside of this approach is that since the majority of reads derive from cellular organisms, it is considerably more challenging to assemble low-abundance viruses or viruses with large genomes.

A third approach for obtaining viral sequences from environmental samples involves using a flow cytometer to isolate viruses associated with single cells or even individual viral particles. This approach usually requires an amplification step due to the small amounts of nucleic acid found in the individually isolated cells or virions (49, 50). Single-cell approaches can provide very high resolution of virus–host interactions and can quantify and characterize the dynamics of the interaction, e.g., the number of lytic versus lysogenic infections in a community (51, 52). Single-virus approaches are especially useful in addressing some of the challenges of reconstructing complete viral genomes from metagenomes (53). A variation on single-cell sequencing involves combining fluorescently stained viral particles with either cultured bacterial hosts or even uncultured cells from the environment and then sorting and collecting the cells that the viruses target (54, 55). This increases the power of single-cell sequencing by enriching the sequenced portion for cells with attached viral particles, although perhaps at the expense of missing many of the sample's proviruses (which are contained in cells not tagged by a virus). However, with all single-cell approaches, DNA amplification can result in highly fragmented assemblies, and it can also be challenging to discriminate between viruses that were attached to the cell and viruses integrated into the host's genome.

COMPUTATIONAL METHODS HAVE FUELED AN ACCELERATED PACE OF DISCOVERY

Regardless of the environmental sequencing approach (e.g., virome, bulk metagenome, or single cell), it is essential to apply a computational method for separating viral and nonviral (e.g., cellular, plasmid) sequences in silico. Existing computational methods follow one of two broadly defined approaches: (*a*) matching the sample sequences directly to a set of reference sequences, or (*b*) using a classification algorithm to label all of the sample's sequences as either viral or nonviral. In general terms, the trade-off between the approaches is that the former is typically applied on the unassembled reads, and therefore can be computationally fast, but has a higher risk of false negatives. The latter approach requires assembly and therefore is more expensive, but it allows for far greater flexibility in identifying novel, uncharacterized viruses not found in the reference

CHARACTERISTICS OF VIRAL GENOMES

Viral genomes exhibit a wide range of sizes. The smallest circoviruses, which infect mostly birds and pigs, have single-stranded DNA genomes less than 2,000 nucleotides long and only encode three or four proteins (122). Pandoraviruses, double-stranded DNA viruses that infect amoebas, have the largest known viral genomes—over two million nucleotides encoding more than 2,000 proteins (123). These two genera are the extremes, of course, and most viruses fall somewhere in between. The majority of bacteriophages, for example, have genome lengths of around 5–10 kb up to 30–50 kb (124).

Despite lacking a universal viral marker gene, viruses have many hallmark genes that are unique to viruses and not found in any cellular genomes. These include genes such as those encoding capsid, portal, and terminase proteins, all of which are involved in the formation and packaging of viral particles (125). In addition, other features common to many viral genomes include a lower rate of strand switching (long stretches of genes encoded on the same strand of a double-stranded genome), smaller average gene size, and an enrichment in genes of unknown function (49). These characteristics of viral genomes can be useful in distinguishing sequences obtained from environmental samples as either viral or cellular (56).

set but that exhibit key viral features (see the sidebar titled Characteristics of Viral Genomes), although at a higher risk of false positives (56).

These two approaches need not be used in isolation from each other. In fact, some of the most impactful efforts, in terms of expanding the databases of known viral sequences, have utilized something resembling an iterative strategy, whereby viral sequences are identified in new datasets by matching to known viral genes, followed by the use of the novel genes discovered on those sequences as baits to identify more novel viruses and augment the reference set. Reference-matching can then be applied yet again to the sample data using the expanded reference set. Care needs to be taken, of course, to prevent adding noise to the reference set by ensuring that the novel sequences are bona fide viral, which typically means that they must meet an expertly defined set of characteristics common to viral genes and genomes. This iteration expands the reference sets with the addition of each newly identified viral sequence, increasing the power of the reference-based approach while simultaneously improving the accuracy of the classifiers due to the availability of more reference data for training. Thus, the two methods are mutually reinforcing (57, 58).

With the diversity of experimental and computational methods for identifying viruses (see **Table 1**), it is important to follow established reporting guidelines to enable the validation and replication of results. To facilitate this, a broad coalition of experts in virology and genomics established the Minimum Information About an Uncultivated Virus Genome (MIUViG) standards (59). The standards mandate reporting of various metadata categories, including the type of dataset generated (virome, bulk metagenome, single-cell, etc.), the sequence assembly method, the software used to identify viral sequences, the predicted genome characteristics (i.e., single- or double-stranded DNA or RNA, sense or antisense, segmented or nonsegmented), whether the sequence is an integrated provirus, the genome quality (finished, high-quality, or fragment), and the number of contigs that comprise the genome. Additional optional metadata that should be reported with new viral sequences include predicted taxonomic classification, predicted host, feature annotations (identifying gene-coding regions, provirus integration sites, etc.), and other experimental details such as the sorting method (for single cells) and the enrichment method (for viromes).

One particularly challenging task is assessing the quality of new viral sequences, which can range from small genome fragments to complete and near-complete genomes. For bacteria and archaea, genome quality is often estimated based on the presence and copy number of widely

Table 1 Resources useful for viral sequence discovery and analysis

Name	Type	Description	Citation
Earth's Virome VPFs	Gene database	Viral-specific protein families	Paez-Espino et al. 2017 (58)
VOGDB	Gene database	Viral-specific protein families	**http://vogdb.org/**
pVOGs	Gene database	Prokaryotic protein families	Grazziotin et al. 2017 (126)
RVDB	Gene database	Eukaryotic viral protein families	Goodacre et al. 2018 (127)
IMG/VR	Genome database	Viral sequence database, including uncultivated viruses	Roux et al. 2020 (41)
NCBI RefSeq	Genome database	Viral sequence database, including mostly isolated viruses	Brister et al. 2015 (128)
ViralZone	Knowledgebase	Fact sheets on all known virus families/genera with easy access to sequence data	Hulo et al. 2011 (129)
PhagesDB	Knowledgebase	Information related to Actinobacteria phages	Russell & Hatfull 2017 (130)
ViPR	Knowledgebase	An integrated repository of data and analysis tools for human pathogenic viruses	Pickett et al. 2012 (131)
VirSorter	Software	Viral sequence discovery	Roux et al. 2015 (56)
VirFinder	Software	Viral sequence discovery	Ren et al. 2017 (132)
DeepVirFinder	Software	Viral sequence discovery	Ren et al. 2020 (133)
VIBRANT	Software	Viral sequence discovery	Kieft et al. 2020 (62)
What the Phage	Software	Viral sequence discovery: comparing multiple pipelines	Marquet et al. 2020 (134)
PhiSpy	Software	Provirus identification	Akhter et al. 2012 (81)
Prophinder	Software	Provirus identification	Lima-Mendez et al. 2008 (135)
Prophage Finder	Software	Provirus identification	Bose & Barber 2006 (136)
vConTACT2	Software	Genome clustering and taxonomic annotation	Bin Jang et al. 2019 (66)
VICTOR	Software	Taxonomic classification	Meier-Kolthoff & Göker 2017 (68)
CCP77	Method	Phylogeny-based taxonomic classification for Caudovirales	Low et al. 2019 (67)
HostPhinder	Software	Host prediction	Villarroel et al. 2016 (83)
VirHostMatcher	Software	Host prediction	Ahlgren et al. 2017 (84)
WIsH	Software	Host prediction	Galiez et al. 2017 (85)
CheckV	Software	Estimation of quality and completeness of viral genome sequences	Nayfach et al. 2020 (61)

Abbreviations: DB, database; IMG/VR, Integrated Microbial Genomes/Virus; NCBI, National Center for Biotechnology Information; pVOGs, prokaryotic VOGs; RVDB, Reference Viral DB; VICTOR, Virus Classification and Tree Building Online Resource; ViPR, Virus Pathogen Database and Analysis Resource; VOG, virus orthologous group; VPF, viral protein family.

distributed single-copy marker genes (60). But because viruses lack such genes, it is not possible to use this approach; instead, many studies simply analyze all viruses longer than a uniform length threshold (e.g., 5 or 10 kb) (56, 58). However, this fails to account for the large variability in viral genome sizes and thus gathers sequences representing a broad range of genome completeness. To address this problem, several recent studies have introduced new tools, including CheckV, VIBRANT, and viralComplete (61–63). CheckV estimates genome completeness based on comparing the gene content of a new virus to a large database of complete viral genomes, including over 70,000 derived from environmental samples, and identifies closed genomes based on the presence

VIRAL EVOLUTION AND SYSTEMATICS

Inferring the phylogenetic relationships among viruses remains a major challenge due to the complexities inherent to viral systematics. First, horizontal gene transfer and recombination is widespread among viruses, resulting in genomic mosaics (1, 137–142). Second, many viruses evolve at such a rapid rate that there is little or no remaining information in the currently observable genomes to deduce higher-level phylogenetic relationships (143). Furthermore, it is thought that viruses may have arisen multiple times over evolutionary history, making it impossible to trace their origin to a single common ancestor (144, 145).

The International Committee on Taxonomy of Viruses (ICTV) is the authority on viral taxonomy and nomenclature. Recently, the ICTV has revised both the taxonomic rank structure used for virus classification, by increasing the number of accepted ranks to 15, and the data requirements for classifying novel viruses, by allowing the use of sequence data in the absence of EM images (146, 147). The former allows for the possibility of hierarchically connecting all viruses (although phylogenetic relationships are not necessarily implied) and the latter indicates an adaptation to the enormous amount of sequence data being generated in recent years, representing many novel viruses known only through their sequences. Finally, although the recent increase in viral sequence data generation means that it is possible, and often necessary, to describe viral taxonomies using only sequence data, it is still recommended to consider phenotypic information such as morphology whenever feasible so that taxonomic classifications reflect biologically meaningful divisions (148).

of terminal repeats or provirus integration sites. When applied to 735,106 viral sequences from the IMG/VR version 2 (v2.0) database, this approach was able to accurately estimate completeness for the majority of sequences from host-associated, marine, freshwater, and soil environments. In the case of proviruses, CheckV also predicts the host–virus sequence boundary, which allows for removal of the host region and improves the identification of bona fide auxiliary metabolic genes in the virus. For example, this approach removed numerous antibiotic resistance genes found on IMG/VR sequences, which are likely to be cellular-encoded given previous work showing that phages rarely encode resistance genes (64).

INFERRING THE EVOLUTIONARY ORIGINS AND RELATIONSHIPS AMONG VIRUSES

Assigning a taxonomic classification to newly discovered viruses is not trivial, and viral taxonomies often undergo frequent revision to reflect the latest understanding of viral evolution (see the sidebar titled Viral Evolution and Systematics). Traditional phylogenetic methods used for prokaryotes and eukaryotes do not work for viruses, which lack universally distributed marker genes (65). To address this limitation, researchers have developed several alternative methods that utilize a variety of strategies. One strategy involves clustering genomes into viral operational taxonomic units (vOTUs) based on average nucleotide identity (ANI) or shared gene content. For example, the MIUViG standards proposed a threshold of 95% ANI over 85% genome length for delineating species-level vOTUs (59), while another approach utilized gene sharing networks to delineate vOTUs at approximately the genus or subfamily ranks (66). In either case, reference genomes can be included in the clustering in order to transfer taxonomic annotations at the appropriate rank within viral OTUs. For example, using this approach, Roux et al. identified 933,352 species-level vOTUs across the 2.3 million viral genomes in IMG/VR v3 (41).

The desire to be able to define relationships for viruses across higher levels has led to other creative approaches. Low et al. performed a taxonomic classification of viruses belonging to the

Caudovirales order utilizing a phylogeny derived from 77 common, yet nonuniversal, proteins (67). Another method, which has been adapted from a framework widely used for taxonomic classification of bacteria and archaea, incorporates nucleotide and amino acid alignments, clustering, phylogenetic inference, and a flexible set of distance metrics (68). Yet another method recently described combines sequence alignments for closely related viruses with a novel mutual information metric for more distantly related viruses (69).

CRISPR: short plasmid- or viral-derived sequences interspersed by a repeating element that functions as an adaptive immune system

As the number of novel viruses discovered through metagenomic sequence data continues to increase at an accelerating rate, there is clearly a need for further development and evaluation of the various approaches for providing taxonomic annotations to uncultivated viruses. One possibility would be to establish genome-based phylogenies of all sequenced viral groups, encompassing both cultivated and uncultivated viruses, analogous to the Genome Taxonomy Database developed for bacteria and archaea (70). This approach could shed new light on the evolution of the virosphere and facilitate new methods for the automated taxonomic annotation of viral genomes. Another possibility would be to incorporate ecological properties of viruses, like host range and habitat distribution, to improve our understanding of the evolutionary relationships among viruses.

EXPERIMENTAL AND COMPUTATIONAL APPROACHES FOR HOST PREDICTION

Identifying the cellular host of a virus is essential for understanding a virus's ecosystem impact and for leveraging viruses in biotechnological applications such as phage therapy. A variety of experimental and computational approaches exist for uncovering host–virus interactions (reviewed in 71). Broadly, these methods differ in terms of whether they depend on cultivation, the type of sequencing data they require (e.g., single-cell, metagenome, whole-genome, or proximity ligation, such as with Hi-C), their dependence on available reference data, the taxonomic resolution of the host prediction, and their sensitivity and specificity (see **Table 2**). The gold standard of evidence for a virus–host relationship is culturing the two together and observing lytic or lysogenic infection via a spot assay, plaque assay, or liquid assay. However, these methods require the availability of the host or virus in pure culture and are therefore not applicable for a large fraction of the virosphere. Another experimental approach involves using flow cytometry to isolate and sequence viruses attached to or replicating inside of individual cells. Single-cell techniques have been applied to samples from human gut (55) and marine (51) environments and have revealed high-resolution phage–host interactions. Another interesting approach involves sequencing cross-linked DNA from a microbial community (e.g., meta3C), which can enable associations between the host genome and proximal or integrated mobile elements, such as viruses and plasmids (72, 73).

Several computational methods utilize genomic information to predict connections between a set of viral genomes and a corresponding set of host genomes (see **Table 2**). The two most commonly used approaches are CRISPR (clustered regularly interspaced short palindromic repeats) matching and genomic similarity, which both depend on the availability of high-quality reference data. CRISPR matching is performed by identifying near-perfect alignments between CRISPR spacers and viral genomes, indicating a history of past infection. While this approach is accurate for assigning the host at low ranks (e.g., species), CRISPR-Cas systems are only found in ~40% of bacteria and 70% of archaea (74) and can be entirely absent from certain prokaryotic lineages (75), and CRISPR arrays can be challenging to assemble from short-read data (76). Additionally, CRISPR spacers rapidly turn over in the host genome, meaning that the information encoding the virus–host linkage will be quickly lost if the relationship is not actively maintained (77–79). Host–virus genomic similarity is another commonly used approach (41, 80), which is

Table 2 Methods for determining or predicting virus–host relationships

Method category	Strategy type	Description	Limitations	Literature example(s)
CRISPR spacer match	Computational	Identifies (near-) exact matches between viral genomes and CRISPR spacers found in host genome	The CRISPR-Cas system is found in only ~40% of bacteria and ~70% of archaea. CRISPR arrays rapidly turn over in the environment and many spacers fail to match any mobile element. Viruses may contain anti-CRISPR proteins that inhibit the acquisition of protospacers by the host.	Paez-Espino et al. 2016 (57)
Provirus identification	Computational	Identifies integrated proviruses in prokaryotic genomes	This approach cannot detect host associations for obligate lytic viruses. It is challenging to precisely identify host–virus boundaries. Proviruses can rapidly decay in the host genome and may no longer encode for a virus capable of entering the lytic cycle.	Akhter et al. 2012 (81)
Sequence similarity	Computational	Identifies host genomes that contain genes or genomic regions matching the virus at either the DNA or protein level	Similar to provirus identification, this method only works for lysogenic viruses. The accuracy and taxonomic resolution of this method depend on the percent identity and length of the aligned region.	Roux et al. 2020 (41)
Oligonucleotide profile	Computational	Identifies the host genome with the most similar oligonucleotide frequency profile	Viruses and hosts are often observed to display divergent nucleotide usage profiles. This approach may not be able to accurately predict the host at low taxonomic ranks.	Galiez et al. 2017 (85)
Metagenome binning	Computational	Identifies viral contigs assigned to a MAG	Proviruses and other mobile elements often display different nucleotide composition and read depth from the host genome, causing challenges for metagenome binning. Binning algorithms may incorrectly assign a viral contig to a host genome bin and these errors are challenging to detect.	Nayfach et al. 2020 (80)
Co-abundance pattern	Computational	Identifies virus and host genomes with correlated abundance patterns across samples	This method requires a large number of samples containing the same host–virus pairing to identify (lagged) correlation patterns. Whole-genome amplification techniques (e.g., MDA) can distort microbial abundances.	Coutinho et al. 2017 (82)
Host-specific marker gene	Computational	Identifies the presence of genes exclusively found in viruses that infect a specific host; these genes may be important for host recognition (e.g., a receptor-binding protein)	Identification of marker genes requires a large number of viral reference genomes where the host is known, and therefore depends heavily on prior knowledge.	Shapiro & Putonti 2018 (149)
Cultivation assay	Experimental	Experimental approach to identifying the phage infecting a cultivated bacterium; includes spot assays, plaque assays, and liquid assays	This method is relatively low throughput and time intensive. It requires the availability of the host (and sometimes virus) in pure culture.	Sullivan et al. 2003 (150)
Viral tagging	Experimental	DNA in environmental viruses is labeled nonspecifically with a fluorescent dye, viruses are mixed with a bait host, and infected cells are collected by fluorescence-activated flow cytometry	This method often requires the availability of the host in pure culture.	Deng et al. 2014 (54)

(*Continued*)

Table 2 (*Continued*)

Method category	Strategy type	Description	Limitations	Literature example(s)
Single-cell sequencing	Experimental	Utilizes flow cytometry to isolate single cells, followed by whole-genome amplification and sequencing of host and viral DNA; can be combined with viral tagging to enrich for cells with attached virions	Whole-genome amplification results in highly fragmented assemblies, making it challenging to determine if a viral sequence was derived from an integrated provirus.	Labonté et al. 2015 (51), Džunková et al. 2019 (55)
Proximity ligation sequencing	Experimental	Experimental and computational approach that exploits the physical contacts between host and virus DNA molecules to infer their proximity	This method requires physical proximity of phage and host genomes, making it most suited for associating integrated prophages with the host genome.	Marbouty et al. 2017 (72), Bickhart et al. 2019 (73)

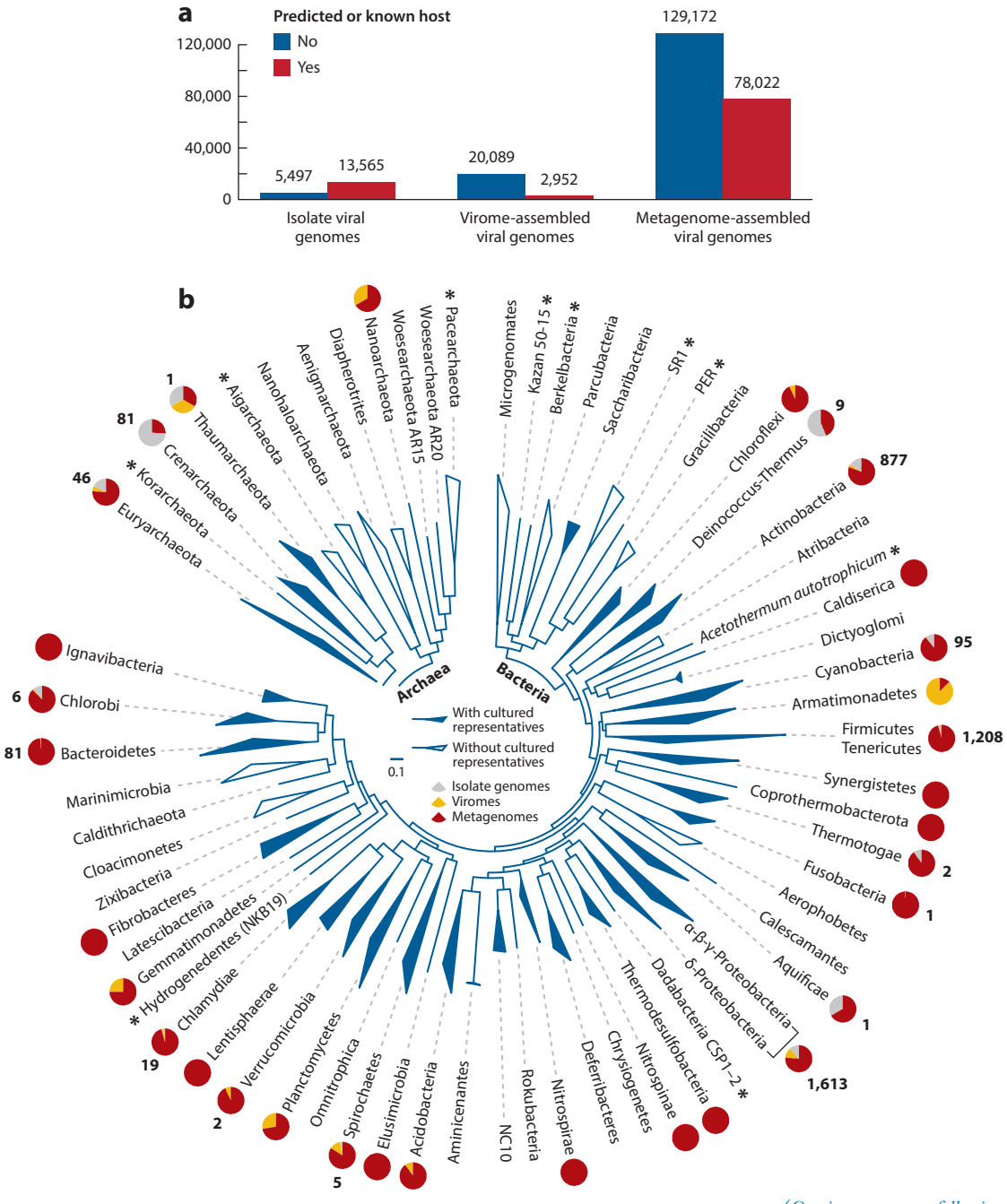
Abbreviations: CRISPR, clustered regularly interspaced short palindromic repeats; MAG, metagenome-assembled genome; MDA, multiple displacement amplification.

often a signature of either a recent or an ancient integration event by a temperate virus. The sensitivity and specificity of this approach (as well as the resolution of the taxonomic assignment) depend on several factors, including the alignment similarity, the length of the aligned region, and whether DNA or proteins are being compared (71). The main drawback of this approach is that it is not effective for obligate lytic viruses, which never integrate into the host genome, which is why it is recommended to combine this approach with CRISPR targeting. For example, using a combined approach, Roux et al. (41) resolved viral connections for the vast majority of prokaryotic phyla (see **Figure 3**). Other computational approaches include identifying integrated proviruses in microbial (meta)genomes (56, 62, 81), identifying (lagged) co-abundance patterns between viruses and hosts (82), identifying similar oligonucleotide profiles (83–85), and using computational methods that utilize viral signature genes that correspond with specific hosts (86).

REVEALING THE ECOLOGICAL IMPACT OF THE GLOBAL VIROME

Viruses exert significant influence on the ecology of the communities in which they are found, which include most of Earth's known biomes (see **Figure 4**). While it is clear that viruses can directly affect the population of their microbial hosts, the various mechanisms involved are still being clarified. Evidence supporting the so called kill-the-winner hypothesis, whereby the organisms most successful at growing within a given biological niche become the target of a larger number of viruses, has been observed in multiple studies (87, 88). Our understanding of this dynamic has increased greatly as the genomic features of the viruses involved (along with their targeted hosts) are elucidated. The forces viruses exert on their hosts, through both lytic and lysogenic virus lifestyles, contribute greatly to the richness and diversity of the microbial communities of which they are a part (89). The fact that viral genes can be transferred to their hosts in a variety of ways has led some researchers to view viruses as an extended gene pool that contributes to the genetic diversity of their community (90). Proviruses in particular are known for providing genes that can eventually become part of the host's genetic repertoire. In addition, cases of coinfection, when a host is simultaneously infected by more than one virus, can lead to viral genome recombination and expanded genomic diversity.

Beyond direct effects on their hosts, and thus an influence on the overall microbial ecology of their communities, viruses play a key role in global nutrient cycles, especially in the surface of the world's lakes and oceans. When an aquatic virus kills its host, the cellular debris releases organic carbon, nitrogen, and phosphorus, which are then available for heterotrophic bacteria. This has been called the viral shunt, as it prevents the normal accumulation of organic carbon



(*Caption appears on following page*)

Figure 3 (*Figure appears on preceding page*)

Prokaryotic virus–host connections. (*a*) Number of viral genomes from IMG/VR v3 (Integrated Microbial Genomes/Virus, version 3) with and without host assignments. (*b*) Phylogenetic distribution of bacterial and archaeal hosts for viruses in IMG/VR v3. For each phylum, a pie chart indicates the fraction of genomes assigned to this phylum from bulk metagenomes (*red*), viromes (i.e., samples that were enriched for viruses; *orange*), or isolate viruses (*gray*). The numbers next to the pie charts indicate the number of genomes from isolate viruses assigned to each phylum, if any. For viruses from viromes and metagenomes, only high-quality genomes are shown (>90% completeness). The set of isolate viruses in IMG/VR was originally obtained from NCBI (National Center for Biotechnology Information) RefSeq and GenBank. While most of these are isolates, some may not be, but that information is not recorded in IMG/VR and all are reported as isolate viral genomes here. The number of viruses with hosts for the isolate viruses in panel *a* includes eukaryotic viruses, which are not shown in the tree of panel *b*. Asterisks denote the ten phyla for which no viruses, even of medium- or low-quality genomes, have been identified. Phyla with no asterisks or pie charts next to them do have viruses with medium- or low-quality genomes that have been identified and available through IMG/VR, but they are not shown here. The scale bar for branch lengths indicates amino acid substitutions per site. Panel *b* adapted from Reference 57.

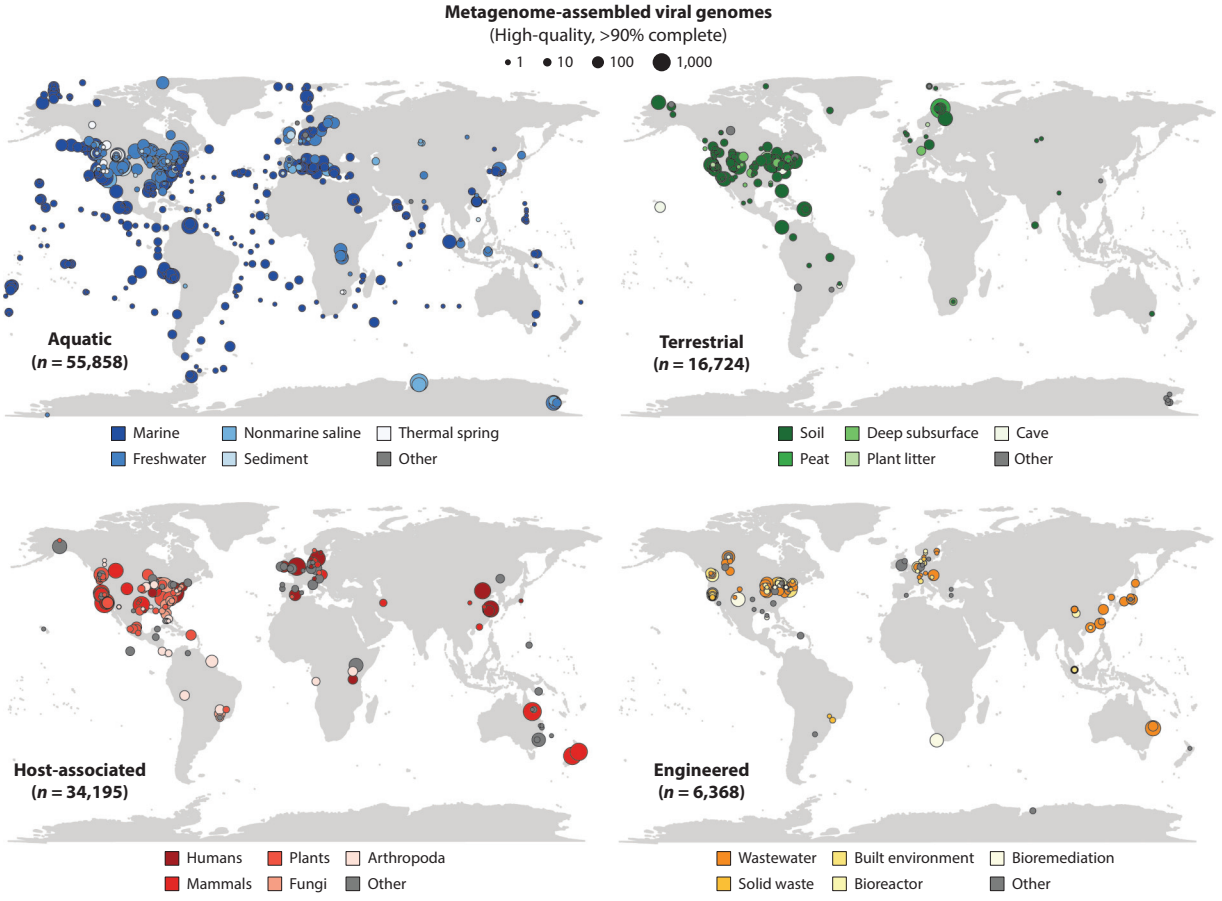


Figure 4

Distribution of metagenome-assembled viral genomes across Earth's biomes. The maps show the geographic distribution of high-quality viral genomes from IMG/VR (Integrated Microbial Genomes/Virus) across major biomes, as defined by the GOLD (General Ontology for Linguistic Description) ontology. High-quality viral genomes were identified using CheckV and contain more than 90% of the expected genome length. Figure adapted from Reference 80, which is distributed under a CC-BY license.

in larger organisms that graze on the microbes (91). Simultaneously, some of the host debris can form sticky aggregates that sink from the ocean surface (92). This so-called viral shuttle can lead to an increase in the long-term storage of carbon in the subsurface ocean layers and the seafloor (93, 94). It has also been suggested that viral lytic activity in the aquatic surface ecosystem may have a role in stimulating carbon fixation and primary production by certain photosynthetic plankton, either via reducing the phytoplankton's predators or competitors or through the lytic release of nutrients (95). These processes merit further attention given the current worldwide interest in accurately measuring and modeling carbon and other biogeochemical cycles.

Phages have been used in antibacterial therapy for nearly 100 years, although their utility in this role has been mostly surpassed by the much more common small-molecule antibiotics (96). However, this trend may begin to change as the rising incidence of multi-drug-resistant infections has increased the demand for novel therapies and metagenomic sequencing efforts have discovered many novel phages that could be harnessed in future therapeutics (97–99). The development of tools for predicting the therapeutic value of specific phages, as well as growing databases of known phage–host relationships, should prove useful in advancing phages as therapeutics. In addition to phage therapy, the genetic content of phage communities has been mined for novel antimicrobial protein-based therapeutics such as lysins with broad- and narrow-spectrum effects (100).

Alongside therapeutics, the specificity of phages' host targeting can also be exploited for diagnostic applications. A variety of methods have been devised to harness phage–host interactions for the detection of pathogens, from simple growth inhibition assays to engineered phages carrying reporter genes (101). Such technologies have the potential to be valuable clinical tools in the diagnosis and management of infection, and current and future efforts are ongoing to develop their promise (102–104).

While advances in phage-based therapies may steal the headlines, it can be argued that phage applications in food and agriculture represent a potentially much broader impact on human life. Phages have been used commercially as a replacement for or adjunct to conventional pesticides on tomato and pepper crops for over a decade (105), and they are even certified for organic production (106). There is increasing interest in applying them to other crops and aquacultures (107), and the growing catalog of known phages will be a significant resource in these efforts (108, 109). Beyond representing a replacement for chemical pesticides, phages have found additional uses as approved methods for reducing microbial contamination in food processing and supply chains (110, 111). As the development of phage-based products continues and their adoption widens, careful thought must accompany their application, including a thorough understanding of their benefits and limitations (112).

WHERE DO WE GO FROM HERE?

It is difficult to predict where research will take us, but based on past developments, it is likely there will be many impactful discoveries made by continuing our efforts to unearth novel viruses and improve our understanding of viral biology. Large-scale coordinated efforts for characterizing novel viruses from many environments are ongoing in many labs around the world, and the number of unknown viruses remaining to be discovered is predicted to be vast based on the observation that the rate of discovery of new vOTUs is not yet reaching any plateau (41). Apace with the discovery efforts, multiple classification efforts move ahead with new tools and techniques propelled by the increase in sequence data. One particular area that is expected to promote further rapid growth in the field is the development and wider adaptation of standards across all aspects of viral genomics research (59). The scientific community is strongly encouraged to adhere to

and promote the refinement of these standards, as well as to contribute to the development of new ones. Building on the discovery, classification, and functional characterization efforts, our collective understanding of viruses' effects on global ecology will continue to be refined and used for predictive modeling (113).

Efforts to improve software tools for the discovery of viral sequence data in metagenomic datasets constitute an area of very active research. These developments are fueled by the application of modern data science techniques, as well as by the continued improvement of sequencing technologies. Different types of sequencing, along with creative analytic approaches, should be exploited to fill in the gaps in our coverage of viral and virus–host space. The tools and resources for viral sequence discovery will continue to adapt and promote further participation by newcomers to the field as we increase our understanding of viruses' roles in the continuing evolution of life on Earth.

SUMMARY POINTS

1. Viruses are found everywhere on Earth, and through interactions with their hosts they are major players in all ecosystems in which they are found.
2. Nonculture, sequence-based methods have exponentially increased the discovery and study of new viral lineages and functions.
3. Despite recent technology advances, vast amounts of viral diversity remain undiscovered.
4. Full utilization of the massive amounts of sequence data being generated for research presents unique challenges and requires following principles of good data management.
5. Viral genome databases that collect, curate, and support the comparative analysis of viruses are critical for advancing our understanding of the viral world.
6. Development and implementation of standards for viral genomics are of fundamental importance for data comparability and reusability from different research groups.
7. Employment of multiple complementary approaches for the identification of viruses (including virome and metagenome studies) can lead to the most comprehensive reconstruction of viral diversity in any ecosystem.

FUTURE ISSUES

1. Researchers in the field should adhere to standards for data generation, reporting, and sharing in order to maximize scientific collaboration, rigor, and reproducibility.
2. Methods for assembly of genomes from complex metagenomes should be optimized in order to achieve high-quality viral genomes.
3. New methods for the identification of novel viral genomes should be developed, with emphasis on those with limited similarity to known viruses.
4. A genome-based taxonomy of cultivated and uncultivated viruses should be constructed and curated to enable rapid and accurate taxonomic classification of newly discovered viruses.

5. Computational methods for connecting viruses to their hosts should be improved, which is especially important given that the host is not known for the vast majority of uncultivated viruses.
6. Computational methods should be used to illuminate the evolutionary arms race between viruses and their hosts, including mechanisms for host or viral defense and mechanisms of recognition (e.g., virus–host protein–protein interactions).
7. Researchers should develop a deeper understanding of the mechanistic actions of phages and viruses in Earth's ecosystems and uncover their underlying roles in controlling health and disease.

DISCLOSURE STATEMENT

This work, conducted by the US Department of Energy (DOE) Joint Genome Institute, a DOE Office of Science user facility, was supported under contract number DE-AC02-05CH11231.

LITERATURE CITED

1. Hendrix RW, Smith MC, Burns RN, Ford ME, Hatfull GF. 1999. Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. *PNAS* 96(5):2192–97
2. Bar-On YM, Phillips R, Milo R. 2018. The biomass distribution on Earth. *PNAS* 115(25):6506–11
3. Mushegian AR. 2020. Are there 10^{31} virus particles on Earth, or more, or fewer? *J. Bacteriol.* 202(9):e00052-20
4. Nasir A, Forterre P, Kim KM, Caetano-Anollés G. 2014. The distribution and impact of viral lineages in domains of life. *Front. Microbiol.* 5:194
5. Jover LF, Effler TC, Buchan A, Wilhelm SW, Weitz JS. 2014. The elemental composition of virus particles: implications for marine biogeochemical cycles. *Nat. Rev. Microbiol.* 12(7):519–28
6. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409(6822):860–921
7. Mi S, Lee X, Li X, Veldman GM, Finnerty H, et al. 2000. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* 403(6771):785–89
8. Horzinek MC. 1997. The birth of virology. *Antonie Leeuwenhoek* 71(1–2):15–20
9. Keen EC. 2015. A century of phage research: bacteriophages and the shaping of modern biology. *BioEssays* 37(1):6–9
10. van Helvoort T. 1994. History of virus research in the twentieth century: the problem of conceptual continuity. *Hist. Sci.* 32(2):185–235
11. Kruger DH, Schneck P, Gelderblom HR. 2000. Helmut Ruska and the visualisation of viruses. *Lancet* 355(9216):1713–17
12. Ackermann HW. 1992. Frequency of morphological phage descriptions. *Arch. Virol.* 124(3–4):201–9
13. Eisenstark A. 1967. Bacteriophage techniques. In *Methods in Virology*, Vol. 1, ed. K Maramorosch, H Koprowski, pp. 449–525. New York: Academic
14. Ackermann H-W. 2007. 5500 phages examined in the electron microscope. *Arch. Virol.* 152(2):227–43
15. Torrella F, Morita RY. 1979. Evidence by electron micrographs for a high incidence of bacteriophage particles in the waters of Yaquina Bay, Oregon: ecological and taxonomical implications. *Appl. Environ. Microbiol.* 37(4):774–78
16. Frank H, Moebus K. 1987. An electron microscopic study of bacteriophages from marine waters. *Helgol. Meeresunters.* 41(4):385–414
17. Bergh O, Børsheim KY, Bratbak G, Heldal M. 1989. High abundance of viruses found in aquatic environments. *Nature* 340(6233):467–68

18. Proctor LM, Fuhrman JA. 1990. Viral mortality of marine bacteria and cyanobacteria. *Nature* 343(6253):60–62
19. Suttle CA, Chan AM, Cottrell MT. 1990. Infection of phytoplankton by viruses and reduction of primary productivity. *Nature* 347(6292):467–69
20. Børsheim KY. 1993. Native marine bacteriophages. *FEMS Microbiol. Ecol.* 11(3–4):141–59
21. Klieve AV, Swain RA. 1993. Estimation of ruminal bacteriophage numbers by pulsed-field gel electrophoresis and laser densitometry. *Appl. Environ. Microbiol.* 59(7):2299–303
22. Swain RA, Nolan JV, Klieve AV. 1996. Natural variability and diurnal fluctuations within the bacteriophage population of the rumen. *Appl. Environ. Microbiol.* 62(3):994–97
23. Wommack KE, Ravel J, Hill RT, Chun J, Colwell RR. 1999. Population dynamics of Chesapeake Bay virioplankton: total-community analysis by pulsed-field gel electrophoresis. *Appl. Environ. Microbiol.* 65(1):231–40
24. Steward GF, Montiel JL, Azam F. 2000. Genome size distributions indicate variability and similarities among marine viral assemblages from diverse environments. *Limnol. Oceanogr.* 45(8):1697–706
25. Chen F, Suttle CA. 1995. Amplification of DNA polymerase gene fragments from viruses infecting microalgae. *Appl. Environ. Microbiol.* 61(4):1274–78
26. Chen F, Suttle CA, Short SM. 1996. Genetic diversity in marine algal virus communities as revealed by sequence analysis of DNA polymerase genes. *Appl. Environ. Microbiol.* 62(8):2869–74
27. Culley AI, Lang AS, Suttle CA. 2003. High diversity of unknown picorna-like viruses in the sea. *Nature* 424(6952):1054–57
28. Short SM, Suttle CA. 2002. Sequence analysis of marine virus communities reveals that groups of related algal viruses are widely distributed in nature. *Appl. Environ. Microbiol.* 68(3):1290–96
29. Breitbart M, Miyake JH, Rohwer F. 2004. Global distribution of nearly identical phage-encoded DNA sequences. *FEMS Microbiol. Lett.* 236(2):249–56
30. Short CM, Suttle CA. 2005. Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. *Appl. Environ. Microbiol.* 71(1):480–86
31. Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17(6):333–51
32. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, et al. 2002. Genomic analysis of uncultured marine viral communities. *PNAS* 99(22):14250–55
33. Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, et al. 2003. Metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol.* 185(20):6220–23
34. Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, et al. 2006. The marine viromes of four oceanic regions. *PLOS Biol.* 4(11):e368
35. Culley AI, Lang AS, Suttle CA. 2006. Metagenomic analysis of coastal RNA virus communities. *Science* 312(5781):1795–98
36. Mizuno CM, Rodriguez-Valera F, Kimes NE, Ghai R. 2013. Expanding the marine virosphere using metagenomics. *PLOS Genet.* 9(12):e1003987
37. Smits SL, Bodewes R, Ruiz-Gonzalez A, Baumgärtner W, Koopmans MP, et al. 2014. Assembly of viral genomes from metagenomes. *Front. Microbiol.* 5:714
38. Boldogkői Z, Moldován N, Balázs Z, Snyder M, Tombácz D. 2019. Long-read sequencing—a powerful tool in viral transcriptome research. *Trends Microbiol.* 27(7):578–92
39. Warwick-Dugdale J, Solonenko N, Moore K, Chittick L, Gregory AC, et al. 2019. Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. *PeerJ.* 7:e6800
40. Beaulaurier J, Luo E, Eppley JM, Uyl PD, Dai X, et al. 2020. Assembly-free single-molecule sequencing recovers complete virus genomes from natural microbial communities. *Genome Res.* 30(3):437–46
41. Roux S, Páez-Espino D, Chen I-MA, Palaniappan K, Ratner A, et al. 2020. IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Res.* 49:D764–75
42. Roux S, Krupovic M, Debroas D, Forterre P, Enault F. 2013. Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. *Open Biol.* 3(12):130160

43. Zolfo M, Pinto F, Asnicar F, Manghi P, Tett A, et al. 2019. Detecting contamination in viromes using ViromeQC. *Nat. Biotechnol.* 37(12):1408–12
44. Kim K-H, Bae J-W. 2011. Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Appl. Environ. Microbiol.* 77(21):7663–68
45. Székely AJ, Breitbart M. 2016. Single-stranded DNA phages: from early molecular biology tools to recent revolutions in environmental microbiology. *FEMS Microbiol. Lett.* 363(6):fnw027
46. Parras-Moltó M, Rodríguez-Galet A, Suárez-Rodríguez P, López-Bueno A. 2018. Evaluation of bias induced by viral enrichment and random amplification protocols in metagenomic surveys of saliva DNA viruses. *Microbiome* 6:119
47. Castro-Mejía JL, Muhammed MK, Kot W, Neve H, Franz CMAP, et al. 2015. Optimizing protocols for extraction of bacteriophages prior to metagenomic analyses of phage communities in the human gut. *Microbiome* 3:64
48. Bobay L-M, Touchon M, Rocha EPC. 2014. Pervasive domestication of defective prophages by bacteria. *PNAS* 111(33):12127–32
49. Roux S, Hawley AK, Torres Beltran M, Scofield M, Schwientek P, et al. 2014. Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics. *eLife* 3:e03125
50. Munson-McGee JH, Peng S, Dewerff S, Stepanauskas R, Whitaker RJ, et al. 2018. A virus or more in (nearly) every cell: ubiquitous networks of virus-host interactions in extreme environments. *ISME J.* 12(7):1706–14
51. Labonté JM, Swan BK, Poulos B, Luo H, Koren S, et al. 2015. Single-cell genomics-based analysis of virus-host interactions in marine surface bacterioplankton. *ISME J.* 9(11):2386–99
52. Labonté JM, Pachiadaki M, Fergusson E, McNichol J, Grosche A, et al. 2019. Single cell genomics-based analysis of gene content and expression of prophages in a diffuse-flow deep-sea hydrothermal system. *Front. Microbiol.* 10:1262
53. Martinez-Hernandez F, Fornas O, Lluesma Gomez M, Bolduc B, de la Cruz Peña MJ, et al. 2017. Single-virus genomics reveals hidden cosmopolitan and abundant viruses. *Nat. Commun.* 8:15892
54. Deng L, Ignacio-Espinoza JC, Gregory AC, Poulos BT, Weitz JS, et al. 2014. Viral tagging reveals discrete populations in *Synechococcus* viral genome sequence space. *Nature* 513(7517):242–45
55. Džunková M, Low SJ, Daly JN, Deng L, Rinke C, Hugenholtz P. 2019. Defining the human gut host-phage network through single-cell viral tagging. *Nat. Microbiol.* 4(12):2192–203
56. Roux S, Enault F, Hurwitz BL, Sullivan MB. 2015. VirSorter: mining viral signal from microbial genomic data. *PeerJ.* 3:e985
57. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, et al. 2016. Uncovering Earth's virome. *Nature* 536:425
58. Paez-Espino D, Pavlopoulos GA, Ivanova NN, Kyrpides NC. 2017. Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data. *Nat. Protoc.* 12(8):1673–82
59. Roux S, Adriaenssens EM, Dutilh BE, Koonin EV, Kropinski AM, et al. 2019. Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nat. Biotechnol.* 37:29–37
60. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25(7):1043–55
61. Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides NC. 2020. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* 39:578–85
62. Kieft K, Zhou Z, Anantharaman K. 2020. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* 8:90
63. Antipov D, Raiko M, Lapidus A, Pevzner PA. 2020. Metaviral SPAdes: assembly of viruses from metagenomic data. *Bioinformatics* 36(14):4126–29
64. Enault F, Briet A, Bouteille L, Roux S, Sullivan MB, Petit M-A. 2017. Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. *ISME J.* 11:237–47
65. Rohwer F, Edwards R. 2002. The Phage Proteomic Tree: a genome-based taxonomy for phage. *J. Bacteriol.* 184(16):4529–35
66. Bin Jang H, Bolduc B, Zablocki O, Kuhn JH, Roux S, et al. 2019. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* 37(6):632–39

57. First large-scale study to identify viral sequences across all public metagenomic datasets.

59. Viral genomics standards developed within the Genomic Standards Consortium framework.

67. Low SJ, Džunková M, Chaumeil P-A, Parks DH, Hugenholtz P. 2019. Evaluation of a concatenated protein phylogeny for classification of tailed double-stranded DNA viruses belonging to the order Caudovirales. *Nat. Microbiol.* 4(8):1306–15
68. Meier-Kolthoff JP, Göker M. 2017. VICTOR: genome-based phylogeny and classification of prokaryotic viruses. *Bioinformatics* 33(21):3396–404
69. Dougan TJ, Quake SR. 2019. Viral taxonomy derived from evolutionary genome relationships. *PLOS ONE* 14(8):e0220440
70. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, et al. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* 36(10):996–1004

71. Systematic statistical evaluation of various computational approaches to viral host prediction.

71. Edwards RA, McNair K, Faust K, Raes J, Dutilh BE. 2016. Computational approaches to predict bacteriophage–host relationships. *FEMS Microbiol. Rev.* 40(2):258–72

72. Marbouty M, Baudry L, Cournac A, Koszul R. 2017. Scaffolding bacterial genomes and probing host-virus interactions in gut microbiome by proximity ligation (chromosome capture) assay. *Sci. Adv.* 3(2):e1602105
73. Bickhart DM, Watson M, Koren S, Panke-Buisse K, Cersosimo LM, et al. 2019. Assignment of virus and antimicrobial resistance genes to microbial hosts in a complex microbial community by combined long-read assembly and proximity ligation. *Genome Biol.* 20:153
74. Godde JS, Bickerton A. 2006. The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. *J. Mol. Evol.* 62(6):718–29
75. Burstein D, Sun CL, Brown CT, Sharon I, Anantharaman K, et al. 2016. Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. *Nat. Commun.* 7:10613
76. Skennerton CT, Imelfort M, Tyson GW. 2013. Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. *Nucleic Acids Res.* 41(10):e105
77. Deveau H, Barrangou R, Garneau JE, Labonté J, Fremaux C, et al. 2008. Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J. Bacteriol.* 190(4):1390–400
78. Andersson AF, Banfield JF. 2008. Virus population dynamics and acquired virus resistance in natural microbial communities. *Science* 320(5879):1047–50
79. Tyson GW, Banfield JF. 2008. Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ. Microbiol.* 10(1):200–7
80. Nayfach S, Roux S, Seshadri R, Udwary D, Varghese N, et al. 2020. A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* 39:499–509
81. Akhter S, RK Aziz, Edwards RA. 2012. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Res.* 40(16):e126
82. Coutinho FH, Silveira CB, Gregoracci GB, Thompson CC, Edwards RA, et al. 2017. Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. *Nat. Commun.* 8:15955
83. Villarroel J, Kleinheinz KA, Jurtz VI, Zschach H, Lund O, et al. 2016. HostPhinder: a phage host prediction tool. *Viruses* 8(5):116
84. Ahlgren NA, Ren J, Lu YY, Fuhrman JA, Sun F. 2017. Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res.* 45(1):39–53
85. Galiez C, Siebert M, Enault F, Vincent J, Söding J. 2017. WIsH: Who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics* 33(19):3113–14
86. Baláž A, Kajsík M, Budiš J, Szemeš T, Turňa J. 2020. PHERI—phage host exploration pipeline. bioRxiv 2020.05.13.093773. **https://doi.org/10.1101/2020.05.13.093773**
87. Thingstad TF, Lignell R. 1997. Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. *Aquat. Microb. Ecol.* 13:19–27
88. Bouvier T, del Giorgio PA. 2007. Key role of selective viral-induced mortality in determining marine bacterial community composition. *Environ. Microbiol.* 9(2):287–97
89. Pal C, Maciá MD, Oliver A, Schachar I, Buckling A. 2007. Coevolution with viruses drives the evolution of bacterial mutation rates. *Nature* 450(7172):1079–81
90. Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, et al. 2008. Functional metagenomic profiling of nine biomes. *Nature* 452(7187):629–32

91. Wilhelm SW, Suttle CA. 1999. Viruses and nutrient cycles in the sea. *Bioscience* 49(10):781–88
92. Proctor LM, Fuhrman JA. 1991. Roles of viral infection in organic particle flux. *Mar. Ecol. Prog. Ser.* 69(1/2):133–42
93. Danovaro R, Dell'Anno A, Corinaldesi C, Magagnini M, Noble R, et al. 2008. Major viral impact on the functioning of benthic deep-sea ecosystems. *Nature* 454(7208):1084–87
94. Weinbauer MG. 2004. Ecology of prokaryotic viruses. *FEMS Microbiol. Rev.* 28(2):127–81
95. Staniewski MA, Short SM. 2014. Potential viral stimulation of primary production observed during experimental determinations of phytoplankton mortality. *Aquat. Microb. Ecol.* 71(3):239–56
96. Summers WC. 2012. The strange history of phage therapy. *Bacteriophage* 2(2):130–33
97. Divya Ganeshan S, Hosseinidoust Z. 2019. Phage therapy with a focus on the human microbiota. *Antibiotics* 8(3):131
98. Schmidt C. 2019. Phage therapy's latest makeover. *Nat. Biotechnol.* 37(6):581–86
99. Duplessis CA, Stockelman M, Hamilton T, Merril G, Brownstein M, et al. 2019. A case series of emergency investigational new drug applications for bacteriophages treating recalcitrant multi-drug resistant bacterial infections: confirmed safety and a signal of efficacy. *J. Intensive Crit. Care* 5(2):11
100. Kim B-O, Kim ES, Yoo Y-J, Bae H-W, Chung I-Y, Cho Y-H. 2019. Phage-derived antibacterials: harnessing the simplicity, plasticity, and diversity of phages. *Viruses* 11(3):268
101. Griffiths MW. 2014. Phage-based methods for the detection of bacterial pathogens. In *Bacteriophages in the Control of Food- and Waterborne Pathogens*, pp. 31–59. Washington, DC: ASM Press
102. Schofield DA, Sharp NJ, Westwater C. 2012. Phage-based platforms for the clinical detection of human bacterial pathogens. *Bacteriophage* 2(2):105–283
103. Paczesny J, Richter Ł, Hołyst R. 2020. Recent progress in the detection of bacteria using bacteriophages: a review. *Viruses* 12(8):845
104. Meile S, Kilcher S, Loessner MJ, Dunne M. 2020. Reporter phage-based detection of bacterial pathogens: design guidelines and recent developments. *Viruses* 12(9):944
105. Balogh B, Jones JB, Momol MT, Olson SM, Obradovic A, et al. 2003. Improved efficacy of newly formulated bacteriophages for management of bacterial spot on tomato. *Plant Dis.* 87(8):949–54
106. Buttimer C, McAuliffe O, Ross RP, Hill C, O'Mahony J, Coffey A. 2017. Bacteriophages and bacterial plant diseases. *Front. Microbiol.* 8:34
107. Plaza N, Castillo D, Pérez-Reytor D, Higuera G, García K, Bastías R. 2018. Bacteriophages in the control of pathogenic vibrios. *Electron. J. Biotechnol.* 31:24–33
108. Svircev A, Roach D, Castle A. 2018. Framing the future with bacteriophages in agriculture. *Viruses* 10(5):218
109. Vu NT, Oh C-S. 2020. Bacteriophage usage for bacterial disease management and diagnosis in plants. *Plant Pathol. J.* 36(3):204–17
110. Perera MN, Abuladze T, Li M, Woolston J, Sulakvelidze A. 2015. Bacteriophage cocktail significantly reduces or eliminates *Listeria monocytogenes* contamination on lettuce, apples, cheese, smoked salmon and frozen foods. *Food Microbiol.* 52:42–48
111. Gutiérrez D, Rodríguez-Rubio L, Fernández L, Martínez B, Rodríguez A, García P. 2017. Applicability of commercial phage-based products against *Listeria monocytogenes* for improvement of food safety in Spanish dry-cured ham and food contact surfaces. *Food Control.* 73:1474–82
112. Połaska M, Sokołowska B. 2019. Bacteriophages—a new hope or a huge problem in the food industry. *AIMS Microbiol.* 5(4):324–46
113. Sullivan MB, Weitz JS, Wilhelm S. 2017. Viral ecology comes of age. *Environ. Microbiol. Rep.* 9(1):33–35
114. Schulz F, Roux S, Paez-Espino D, Jungbluth S, Walsh DA, et al. 2020. Giant virus diversity and host interactions through global metagenomics. *Nature* 578(7795):432–36
115. Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GGZ, et al. 2014. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* 5:4498
116. Kauffman KM, Hussain FA, Yang J, Arevalo P, Brown JM, et al. 2018. A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria. *Nature* 554(7690):118–22
117. Roux S, Krupovič M, Daly RA, Borges AL, Nayfach S, et al. 2019. Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. *Nat. Microbiol.* 4(11):1895–906

114. Significant expansion of the global diversity of nucleocytoplasmic large DNA viruses (NCLDVs) from metagenomes.

118. Devoto AE, Santini JM, Olm MR, Anantharaman K, Munk P, et al. 2019. Megaphages infect *Prevotella* and variants are widespread in gut microbiomes. *Nat. Microbiol.* 4(4):693–700
119. Paez-Espino D, Zhou J, Roux S, Nayfach S, Pavlopoulos GA, et al. 2019. Diversity, evolution, and classification of virophages uncovered through global metagenomics. *Microbiome* 7:157
120. Lucas W. 2010. Viral capsids and envelopes: structure and function. *eLS*. **https://doi.org/10.1002/9780470015902.a0001091.pub2**
121. Howard-Varona C, Hargreaves KR, Abedon ST, Sullivan MB. 2017. Lysogeny in nature: mechanisms, impact and ecology of temperate phages. *ISME J.* 11(7):1511–20
122. Rosario K, Duffy S, Breitbart M. 2009. Diverse circovirus-like genome architectures revealed by environmental metagenomics. *J. Gen. Virol.* 90(Part 10):2418–24
123. Philippe N, Legendre M, Doutre G, Couté Y, Poirot O, et al. 2013. Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science* 341(6143):281–86
124. Hatfull GF. 2008. Bacteriophage genomics. *Curr. Opin. Microbiol.* 11(5):447–53
125. Koonin EV, Senkevich TG, Dolja VV. 2006. The ancient Virus World and evolution of cells. *Biol. Direct.* 1:29
126. Grazziotin AL, Koonin EV, Kristensen DM. 2017. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Res.* 45(D1):D491–98
127. Goodacre N, Aljanahi A, Nandakumar S, Mikailov M, Khan AS. 2018. A reference viral database (RVDB) to enhance bioinformatics analysis of high-throughput sequencing for novel virus detection. *mSphere* 3(2):e00069-18
128. Brister JR, Ako-Adjei D, Bao Y, Blinkova O. 2015. NCBI viral genomes resource. *Nucleic Acids Res.* 43:D571–77
129. Hulo C, de Castro E, Masson P, Bougueleret L, Bairoch A, et al. 2011. ViralZone: a knowledge resource to understand virus diversity. *Nucleic Acids Res.* 39:D576–82
130. Russell DA, Hatfull GF. 2017. PhagesDB: the actinobacteriophage database. *Bioinformatics* 33(5):784–86
131. Pickett BE, Sadat EL, Zhang Y, Noronha JM, Squires RB, et al. 2012. ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res.* 40:D593–98
132. Ren J, Ahlgren NA, Lu YY, Fuhrman JA, Sun F. 2017. VirFinder: a novel *k*-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* 5:69
133. Ren J, Song K, Deng C, Ahlgren NA, Fuhrman JA, et al. 2020. Identifying viruses from metagenomic data using deep learning. *Quant. Biol.* 8:64–77
134. Marquet M, Hölzer M, Pletz MW, Viehweger A, Makarewicz O, et al. 2020. What the Phage: a scalable workflow for the identification and analysis of phage sequences. bioRxiv 2020.07.24.219899. **https://doi.org/10.1101/2020.07.24.219899**
135. Lima-Mendez G, Van Helden J, Toussaint A, Leplae R. 2008. Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics* 24(6):863–65
136. Bose M, Barber RD. 2006. Prophage Finder: a prophage loci prediction tool for prokaryotic genome sequences. *In Silico Biol.* 6(3):223–27
137. Chare ER, Holmes EC. 2006. A phylogenetic survey of recombination frequency in plant RNA viruses. *Arch. Virol.* 151(5):933–46
138. Lefeuvre P, Lett J-M, Varsani A, Martin DP. 2009. Widely conserved recombination patterns among single-stranded DNA viruses. *J. Virol.* 83(6):2697–707
139. Martin DP, Biagini P, Lefeuvre P, Golden M, Roumagnac P, Varsani A. 2011. Recombination in eukaryotic single stranded DNA viruses. *Viruses* 3(9):1699–738
140. Cumby N, Davidson AR, Maxwell KL. 2012. The moron comes of age. *Bacteriophage* 2(4):225–28
141. Krupovic M, Koonin EV. 2014. Evolution of eukaryotic single-stranded DNA viruses of the Bidnaviridae family from genes of four other groups of widely different viruses. *Sci. Rep.* 4:5347
142. Gregory AC, Solonenko SA, Ignacio-Espinoza JC, LaButti K, Copeland A, et al. 2016. Genomic differentiation among wild cyanophages despite widespread horizontal gene transfer. *BMC Genom.* 17:930
143. Krupovič M, Bamford DH. 2010. Order to the viral universe. *J. Virol.* 84(24):12476–79
144. Krupovič M, Koonin EV. 2017. Multiple origins of viral capsid proteins from cellular ancestors. *PNAS* 114(12):E2401–10

145. Krupovič M, Dolja VV, Koonin EV. 2019. Origin of viruses: primordial replicators recruiting capsids from hosts. *Nat. Rev. Microbiol.* 17(7):449–58
146. Int. Comm. Taxon. Viruses Exec. Comm. 2020. The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. *Nat. Microbiol.* 5(5):668–74
147. Simmonds P, Adams MJ, Benkő M, Breitbart M, Brister JR, et al. 2017. Consensus statement: virus taxonomy in the age of metagenomics. *Nat. Rev. Microbiol.* 15(3):161–68
148. Simmonds P. 2015. Methods for virus classification and the challenge of incorporating metagenomic sequence data. *J. Gen. Virol.* 96(6):1193–206
149. Shapiro JW, Putonti C. 2018. Gene co-occurrence networks reflect bacteriophage ecology and evolution. *mBio* 9(2):e01870-17
150. Sullivan MB, Waterbury JB, Chisholm SW. 2003. Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus*. *Nature* 424(6952):1047–51

147. Consensus from ICTV workshop stating that uncultivated viruses can be incorporated into the official taxonomy.

148. A detailed discussion of the hazards of modern viral taxonomic classification.