Viromics approaches for the study of viral diversity and ecology in microbiomes

Simon Roux[1 †], Clement Coclet[1]
[1] DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

[†]e-mail: sroux@lbl.gov

This preprint has not undergone peer review or any post-submission improvements or corrections. The Version of Record of this article is published in Nature Review Genetics, and is available online at https://doi.org/10.1038/s41576-025-00871-w

The vast majority of viruses on Earth infect microbial hosts. These viruses of microbes can act as crucial regulators of microbial populations across ecosystems, from soils to seas, plants and animals. Traditional culture-based methods revealed insufficient to explore this viral diversity at scale, driving the development of viromics, i.e. the sequence-based analysis of uncultivated viruses. Viromics studies already revealed the broad geographic distribution and strong partitioning of viral communities by ecosystem type and host community composition. The expansive genetic and functional diversity of the global virome is also progressively being uncovered, challenging current approaches for functional annotation and viral taxonomy. Moving forward, large-scale viromics studies combined with new experimental and computational approaches to infer virus activity and host interactions will enable a more complete characterization of global viral diversity and impacts.

Introduction

Viruses are found everywhere on Earth, across all ecosystems, and to the best of our knowledge infect every type of organism[1]. The global virome, meant here as the entire diversity of viruses on the planet, includes viruses with a broad range of genome sizes, types, and complexity, as well as a wide diversity of virion structures[2]. Because microbes are the primary form of (cellular) life on the planet, this global virome is mainly composed of viruses of microbes, spanning viruses of bacteria, archaea, and micro-eukaryotes[3]. Despite their microscopic size, viral infections of microorganisms can result in ecosystem-scale impacts including the collapse of algal blooms in the ocean[4,5] or the redirection of key microbial metabolisms[6,7]. Viral infections in microbiomes can thus disrupt critical ecological and human-driven processes such as greenhouse gas emissions or wastewater decontamination[7–10]. Meanwhile, viruses of microbes hold incredible promise for biotechnological applications including for instance the discovery of new molecular tools for targeted manipulation of microbiomes[11–13].

Over the past 20 years, metagenomics profoundly transformed our view and understanding of this global virome. Since pioneering studies in the early 2000s[14–16], a robust conceptual and technical framework has been developed enabling the establishment of large genome catalogues for uncultivated viruses[17–19]. Comparative analyses of the diversity, abundance, and gene content of these uncultivated virus genomes, typically designated as "viromics" analyses, already revealed

several key features of the global virome. First, viromics studies identified some of the most abundant virus groups that had not been previously detected via traditional cultivation assays. These include viruses such as the "crAssphage" bacteriophages in human gut microbiomes now established as a formal *Crassvirales* order[20–22], as well as several viral groups dominating ocean or soil viral communities[23,24]. In terms of gene content, studies revealed a broad array of viral-encoded genes with the potential to alter and redirect key cellular processes[25]. These include viral-encoded genes involved in photosynthesis[26,27], carbon, methane, nitrogen, and sulphur cycling[28–32], but also genes regulating bacterial sporulation[33], highlighting the breadth of cellular processes potentially impacted by viruses. Some of these viral-encoded genes discovered via viromics also hold biotechnological potential, such as new CRISPR-Cas systems with unique properties for genome editing[34] or novel DNA polymerase enzymes with unusual accessory functions[35]. Several viromics studies identified potential new viruses for microbial hosts of interest such as key ecosystem players or evolutionary models[36–39], and more generally can provide an initial reconstruction of potential virus-host networks in microbiomes that can be used to refine fundamental eco-evolutionary models of virus-host interactions[29,40–43,43,44]. Finally, viromics approaches have also been used in a human health and microbiome context, for instance to evaluate viral transmission between mother and infant microbiomes[45], or monitor changes in the viral community during interventions such as faecal microbiota transplantation[46,47].

Considering the ubiquity and multiple uses of these methods and data, we review here the current state of viromics approaches when used to explore and interrogate viral genomic diversity. We focus primarily on uncultivated viruses of microbes, most of them bacteriophages, and on eco-evolutionary rather than clinical or epidemiological research. After describing the current methods used in the viromics field and the corresponding standard protocols and pipelines now emerging, we highlight how viromics already changed our understanding of global virus ecology and distribution, virus-host interactions, and viral impacts on microbiome processes, before considering potential innovations that could address, in the near future, some of the pressing challenges in the field.

Current viromics methods and tools

Different sample types provide complementary views of viral diversity

"Viromics" is most often meant as the study of viral diversity based on genomes of uncultivated viruses, and encompasses a variety of sample processing techniques typically adapted to a target host, ecosystem, or virus group, and/or designed to address a specific biological question. Among commonly used methods, two sample processing steps have a major influence on the downstream dataset and analysis: the use of bulk sample vs separated cellular and viral fractions, and the sequencing of DNA vs RNA (Fig. 1a). Historically, the former has been driven by sample constraints and study design: liquid samples are typically collected on filters, so that collecting and studying the smallest size fraction, typically under 0.2 μm, to study viruses was often a natural extension of existing protocols. For solid substrates, bulk metagenomes are more common, although viral particle enrichment protocols have also been proposed[48–52]. After nucleic acid extraction, the choice of library preparation is typically guided by the target virus group: methods specific for dsDNA or inclusive of both ss- and dsDNA are used when targeting DNA viruses, and RNA sequencing is used for RNA viruses. Even when the analysis of viral genomes is auxiliary to the main study goals, the type of library preparation, e.g. specific of dsDNA for microbial metagenomes or of RNA for microbial metatranscriptomes, will strongly influence which type of

viruses will be primarily recovered (Fig. 1a).

Most importantly, these different datasets provide complementary views of viral diversity. For instance, viral genomes obtained from cellular fractions will be enriched in "in-cell" viruses, i.e. viruses currently undergoing active lytic infections or latent infections such as prophages, and viral fraction metagenomes will mostly recover genomes from viral particles present in the sample, which is often an overlapping but distinct set of viruses compared to the ones present in the cellular fraction[53,54]. Direct comparisons have shown that viral fraction metagenomes could provide better resolution in the dynamics of natural viral communities, although viral size fractions can also include host-associated sequences for instance from extracellular vesicles[53–56]. A recent meta-analysis systematically evaluated the difference in virus recovery between these different types of metagenome preparation, and confirmed that metagenomes derived from cellular and viral size fractions in particular led to the identification of distinct and only partially overlapping viral communities[54]. Overall, a truly extensive and comprehensive survey of viral diversity in a given sample or ecosystem would require combining these different types of datasets[54].

Genomes from uncultivated virus genomes can also be obtained from other data types such as single-cell microbial genomes[57,58], single-virus genomes[59], MDA-amplified viral size fraction[60–63], or RNA sequencing of a viral size fraction[64,65]. All of these datasets are potential sources of uncultivated virus genomes, yet the type of sample processing must be taken into account for data analysis to avoid mis-interpretations, e.g. regarding the presence/absence or estimated abundance of a given virus or virus taxon. To that end, community standards have been developed to systematically report this sample processing information and improve future re-use of viromics data[66]. With viromics increasingly used across diverse study fields such as ecology, evolution, and epidemiology, the community has been working towards identifying which processes and data may be common enough to warrant unified approaches and databases, and will continue to update corresponding guidelines and standards [Box 1].

Box 1 | Standardisation efforts across the viromics communities.

As a field, viromics has been quickly and steadily growing over the past 20 years, both in terms of technical capabilities and interest from researchers. Nevertheless, it remains a young and developing field for which standards, guidelines, and universal nomenclatures are still being established. Community efforts towards this goal include for instance the development of the Minimum Information on Uncultivated Virus Genome checklist that highlighted key metadata and best practices to analyze and share metagenome-assembled virus genomes, or the RdRP summit that outlined a potential framework for better interoperability of RNA virus sequencing datasets[66,164]. Similarly, the International Committee for Taxonomy of Viruses (ICTV) has been actively working on integrating uncultivated viruses into the formal virus taxonomic framework[165]. Additional consultations and consensus forming between research communities, e.g. environmental virus surveillance and microbiome researchers, are still needed. Some of the most pressing needs include nomenclature clarification, with for instance viral metagenomes, viromes, metaviromes, and virion-associated nucleic acids (VANA) all being currently used in the literature to designate viral-targeted metagenomes, or additional tool benchmarking through community projects, as done for instance in the CAMI and ICTV Taxonomy Challenge projects[70]. Together, a standardization of nomenclature along with the emergence of carefully benchmarked analysis pipelines are critical to lower the barrier of entry to viromics, enable broader data re-use and robust comparisons, and fully leverage the potential of viromics for virus discovery, ecology, evolution, and surveillance.

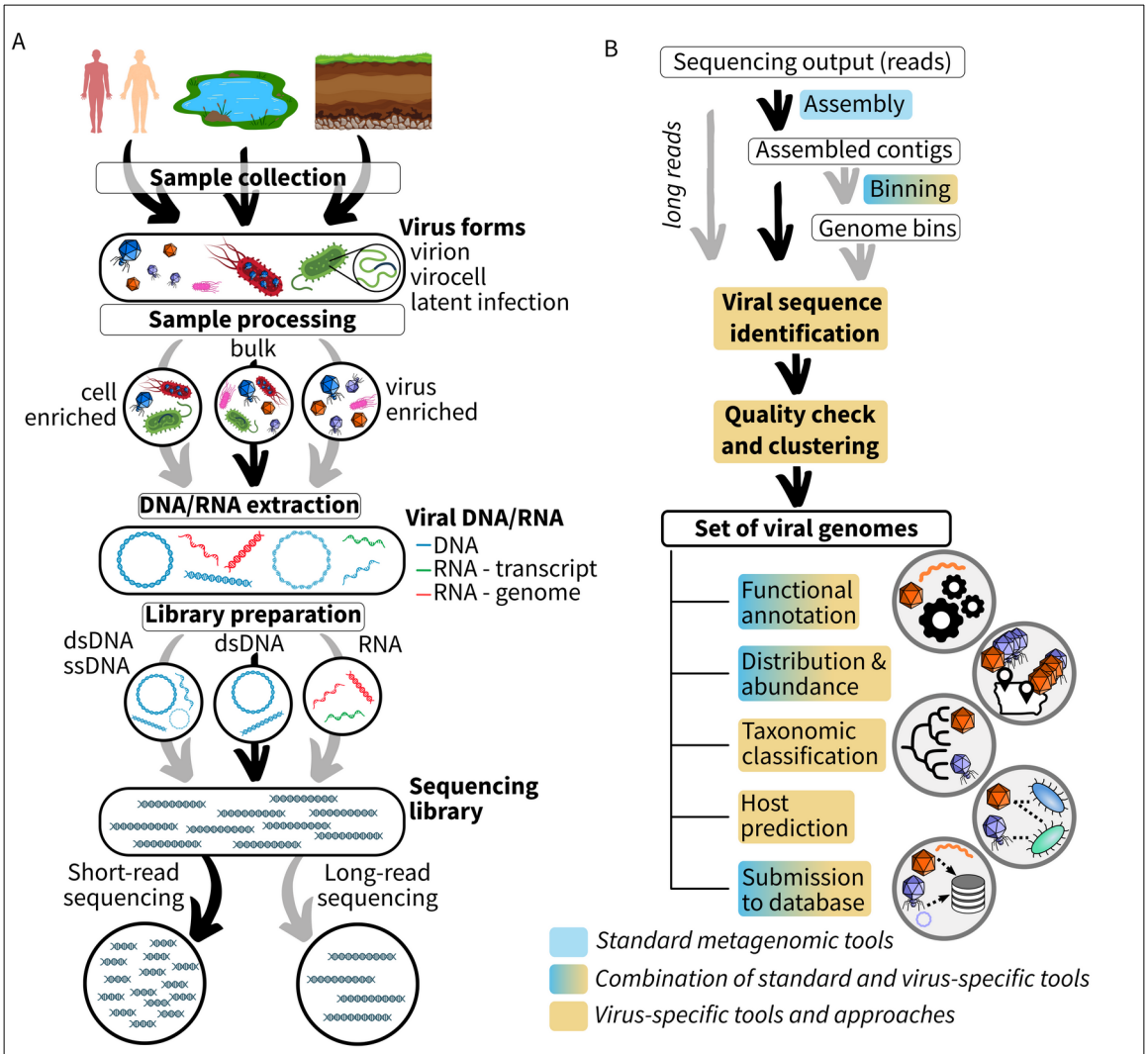

Figure 1: **Overview of the viromics sample processing and pipelines.** A. Types of samples and sample processing methods typically used in viromics. Viral nucleic acids can originate from different sources including virions, virocells in which active infections are ongoing, or microbial cells under latent infection by e.g. an integrated prophages. The different methods for sample processing, nucleic acids extraction, library preparation, and sequencing will determine which part of the original viral nucleic acids will be preferentially sequenced and most represented in the final dataset. The most common approaches used currently in viromics are highlighted in black, while others are indicated in grey. B. Overview of the current standard viromics bioinformatic pipeline and possible analyses performed on metagenome-assembled viral genomes. Steps typically included in a majority of viromics studies are highlighted in black, while other analyses are indicated in grey. The parts of the pipeline requiring tools specifically dedicated to viruses are highlighted in yellow, as opposed to blue highlights for steps leveraging common metagenomic tools.

An emerging viromics analysis workflow for viral ecology and evolution

Over the past decade, a common set of viromics analyses has emerged for the study of viral diversity, ecology, and evolution (Fig. 1b). For viral ecogenomics studies and *de novo* identification of virus genomes, metagenomic reads must first be assembled into contigs, which can also be grouped into predicted genome bins[67–69]. Several tools are available for metagenome

assembly[70], and most viral ecogenomics studies successfully used standard ones such as metaSPAdes[71] and megahit[72], although some optimizations have been proposed specifically for viral genome assemblies[73]. In most microbial metagenomic analyses, assembly is followed by a binning step to group contigs belonging to the same genome and obtain Metagenome-Assembled Genome (MAGs)[67,70]. This step is often omitted for viruses as their smaller genome often leads them to be assembled into a single contig given sufficient sequencing depth[74]. Genome binning for viruses can still be useful to identify cases in which a single virus genome was split into multiple contigs, but must be performed with tools specifically developed for viral genomes to avoid widespread "over-binning" (i.e. bin contamination)[68,69]. One exception is the study of giant viruses (*Nucleocytoviricota*), for which genome binning tools developed for bacteria and archaea seem efficient[75].

Using contigs and/or genome bins obtained after assembly, the main steps of a viromics analysis are the identification, quality control, and clustering of viral sequences to obtain a set of non-redundant viral genomes and genome fragments (Fig. 1b, highlighted in bold). These steps are critical as the choices of tools and cutoffs at this stage will condition the interpretation of most downstream analyses. Current tools for virus sequence identification typically leverage gene content analysis, alignment-free nucleotide composition analysis, or a combination of both[76–79]. Most importantly, all these tools face a sensitivity vs precision trade-off, i.e. conservatively call virus sequences at the risk of missing more novel and divergent viruses, or incorporating these potential new virus sequences at the risk of also including non-viral sequences (Box 2). Combining the output from multiple tools does not necessarily optimise this sensitivity vs precision trade-off, and requires cautious evaluation of the tools and inclusion criteria used[80]. Hence, regardless of the tool(s) selected, researchers should always check the documentation and adjust the minimum confidence score (or equivalent threshold) used to consider a sequence as viral, in accordance with the type of sample analyzed, the planned downstream analyses, and the goal of their study.

Following this first step of viral sequence identification, predicted viral sequences are typically checked for "contamination", i.e. the presence of non-viral regions detected and removed *a posteriori*, and compared to reference databases to predict their "completeness", i.e. how much of the whole genome each contig or bin captures. Automated tools, such as CheckV[81], are available to assist researchers with these quality control steps.

Finally, when integrating virus sequences from multiple samples, predictions from individual metagenomes are most often clustered to form a non-redundant set of viral genomes and genome fragments, using a standard genome-wide ANI (average nucleotide identity) approach[66]. Further selection can also be applied at this stage e.g. requiring a minimum length and/or a minimum predicted completeness, which again must be considered when interpreting results from downstream analyses.

Once established, this set of viral genomes can be analysed to evaluate the functional potential, distribution, taxonomic diversity, and host interactions of the viruses identified. Most of these analyses use tools building from classical metagenomic analysis but specifically adapted for viruses, for instance dedicated functional annotation databases and tools[82,83], specific cutoffs to identify the presence or absence of individual viruses based on read mapping[74], or network-based methods for viral taxonomic classification[84]. These tools and databases are currently at different levels of maturity, with taxonomic classification, host prediction, and automated identification of metabolic genes in viruses (AMGs) in particular being areas of active development and fast-paced methods improvement[85,86].

This emerging viromics workflow offers a unique opportunity to evaluate the predicted distribution, host range, and potential impacts of viruses at ecosystem scales and without any

laboratory cultivation. For several steps, such as viral sequence quality control and clustering, a standard analysis and reporting pipeline has now mostly been adopted[66]. Integrated pipelines performing a combination of all the steps outlined above are also being developed[87,88], providing a user-friendly way for researchers to explore the viral signal in their datasets and perform reproducible and standardized analyses allowing more direct comparisons between studies. The novel and recent nature of these approaches must be recognized however, and caution is warranted when interpreting the results of these viromics analyses. This is especially critical when these results seem to run contrary to established paradigms, for instance the identification of genes and functions never-before-seen in viruses, or a prediction of host range much beyond what has been previously observed[40,89,90]. As much remains to be discovered in the viral world, existing paradigms are likely to shift, yet the results of (semi-)automated viromics workflows must be complemented with follow-up analyses and experiments alongside careful evaluation of the technical and methodological biases and limitations of current viromics approaches to do so[31,37,39,91] [Box 2].

Charting the global virome diversity

Large-scale genome catalogues highlight the global distribution of viruses

Advances in sample processing and sequencing technologies, coupled with improvements in bioinformatic tools including the emerging end-to-end standard viromics pipeline, have led to a significant increase in the number and quality of viral genomes obtained from metagenomes in the last few years. This enabled the establishment of large-scale catalogues of virus genomes, which provide a broad overview of viral diversity across human gut, oceanic, or terrestrial microbiomes[18,19,23,24,92–97]. Other catalogues focus on specific virus groups[38,98–100], or attempt to

Box 2 |Pushing back the frontiers of the virosphere.

In addition to genuine viral genomes from known viral taxa, a number of viral-like elements can be identified in viromics analyses that may, in some cases, represent entirely new types of viruses. Some of these viral-like elements can confuse tools trained to distinguish viral from non-viral sequences and as such can be a source of error, for instance by leading to a gene mistakenly being considered as encoded on a virus genome. Important ones to be aware of when performing viromics analyses include decayed prophages which are remnants of virus genomes integrated in bacterial and archaeal genomes not able to excise and replicate anymore[166], endogenous viral elements (EVEs) which often represent partial fragments of virus genomes in eukaryotic genomes also incapable of replicating and transmitting[167], gene transfer agents (GTAs) which retain genes encoding viral-like particles but do not replicate and encapsidate a viral genome[168], or tailocins which are derived from tailed phage structural proteins but used for microbe-microbe competition[169]. Meanwhile, careful analysis of virus-like sequences identified in viromics analyses can lead to the identification of candidate genomes that may represent entirely new virus and virus-like taxa. Such examples include the polinton-like viruses initially considered as non-viral mobile genetic elements[170], the mirusviruses[171], the "obelisks" RNA elements[172], and possibly the so-called "Borgs" elements initially identified as large plasmids[173]. Hence, viromics can be used to help guide the search for entirely new viruses, such as the elusive RNA viruses of archaea[174], but requires in that case careful and often custom-designed analyses and can not rely exclusively on automated and standard pipelines.

collect distinct virus genomes across the broadest diversity of viral taxa and/or environments possible[17,101]. The latter group includes the IMG/VR v4 database, which currently includes > 5 million virus genomes, and can be used to illustrate emerging patterns of global virus novelty and distribution revealed by viromics.

When considering the entire IMG/VR v4 database, sampling of taxonomic and ecological diversity remains uneven. For instance, metagenome-derived virus genome catalogues remain dominated overall by tailed dsDNA phages (*Caudoviricetes*) and RNA viruses, while other virus groups including atypical dsDNA viruses and ssDNA viruses remain likely undersampled and underrepresented (Fig. 2a). Complete and (near-)complete viral genomes can now be routinely assembled from metagenomes. However, achieving high-quality assemblies is still easier for shorter genomes (e.g., inoviruses compared to caudoviruses), simpler microbiomes (e.g., human gut versus soil samples), and remains biased against rare viruses or those with high population diversity[59,74]. Furthermore, based on accumulation curves, current databases seem far from reaching saturation regardless of the biome, virus taxon, or quality of the virus genome, suggesting that significant additional sampling is required to eventually chart the entire viral diversity on Earth (Fig. 2b).

While still incomplete, these large-scale genome catalogues can still provide useful information about the ecological processes governing the distribution and population diversity of viruses. For instance, multiple studies have now observed that many viruses seem to be globally distributed geographically, but restricted to specific ecosystems or biomes, likely reflecting the distribution of their host(s)[19,24,89]. Similarly, in IMG/VR v4, many vOTUs are detected in samples from similar ecosystems but collected thousands of kilometers away from each other (Fig. 2c). This illustrates how viral communities diversity and structure can now be studied at scale through the comparison of metagenome-derived virus genomes. In that context, temporal sampling at

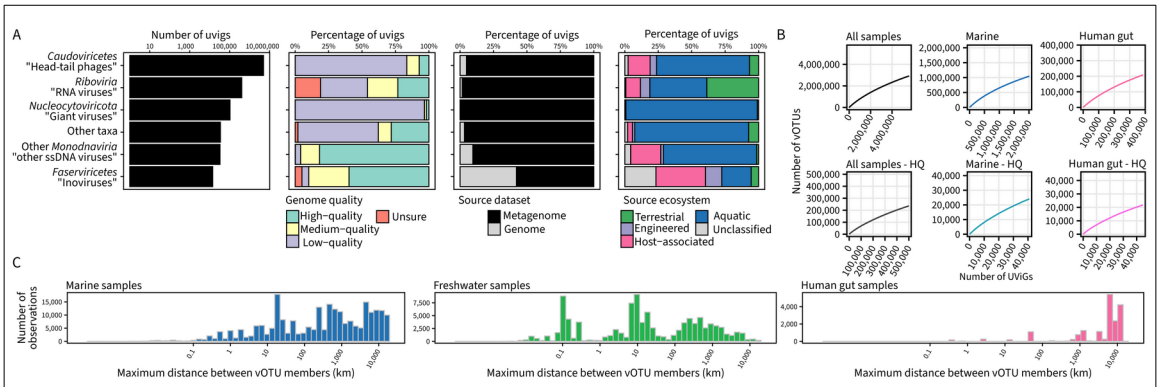

Figure 2: **Global diversity and distribution of viruses currently captured from metagenomes.** A. Number of sequences assigned to different viral taxa in the IMG/VR v4 database17 (left panel). The three subsequent panels show the distribution of genome quality, source dataset, and ecosystem of origin within each taxonomic group. B. Accumulation curve of vOTUs (y-axis) as a function of the number of viral genomes considered (x-axis), in the IMG/VR v4. From left to right, the panels show accumulation curves in all, marine, and human gut samples. The top row shows accumulation curves including all sequences, while the bottom row panels only include genomes predicted as high-quality (completeness ≥90%). C. Distribution of the maximum geographic distance within members of a given vOTU observed in marine (left), freshwater (middle) and human gut (right) samples. For each vOTU, only metagenome-derived sequences were considered, and pairwise geographic distance was evaluated based on latitude and longitude information provided for each sample. For all panels, only high-confidence IMG/VR v4 sequences were considered.

individual sites can be used to reveal short- and long-term variations in viral community composition[42,102–105], while viromics studies across ecological gradients can provide insights into the influence of viruses on microbiome-driven biogeochemical processes[29,106–108].

The vast functional potential encoded in viral genomes

Analyzing viral genomes assembled from metagenomes already confirmed that viruses encode an extensive and mostly uncharacterized genetic diversity. This was already expected from early metagenomics studies[14] and has been confirmed since with larger datasets[95,109–111]. In the current IMG/VR v4 database for instance, 77% of all genes predicted have no detectable similarity to any protein conserved domain without noticeable variation across ecosystems (Fig. 3a). This novel functional diversity is in part linked to a large number of relatively short and conserved ORFs, most predicted to display novel protein folds[109,111]. These novel viral protein families likely include many involved in host cell manipulation, anti-defense mechanisms, and host lysis, as such proteins are expected to be highly variable due to virus-host arms race dynamics[112]. Functional assignment approaches not relying on sequence similarity to known references but involving protein structure prediction and comparison[113,114] or genomic and protein language models[115,116] will likely be required to predict putative functions for these genes.

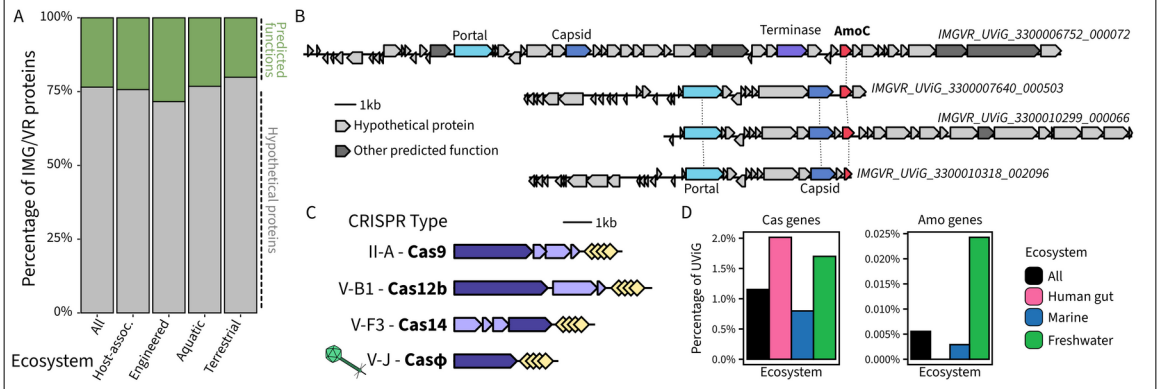

Figure 3: **Illustration of the vast functional potential encoded by uncultivated viruses.** A. Percentage of predicted proteins with an assigned product (green) or indicated as "hypothetical proteins" (grey) in high-confidence viral sequences from the IMG/VR v4 database. This percentage is provided for all sequences (left), and for specific ecosystems (right). B. Genomic neighborhoods of predicted archaeal viruses encoding the AmoC AMG, adapted from31. C. Genomic neighborhood of selected Cas loci, including the phage-encoded Casphi (bottom), adapted from34. D. Percentage of high-confidence virus sequences from IMG/VR v4 including a predicted protein functionally annotated as a Cas gene (left) or AmoC (right). This percentage is provided for all sequences, and for specific ecosystems.

Among functionally annotated genes, next to the ones directly involved in viral genome replication and capsid formation, viral-encoded genes directly or indirectly involved in host interaction and/or host cell takeover have been the focus of extensive research as they may represent important mechanisms by which viruses reprogram host cell metabolism as well as unique platforms for biotechnological applications (Fig. 3b and 3c). Viral-encoded genes involved in reprogramming and/or complementing host metabolism are typically designated as Auxiliary Metabolic Genes (AMGs)[25] (Fig. 3b). While initially identified and characterised in cyanophage isolates[117–119], a broader diversity of potential AMGs have since been predicted in metagenome-assembled genomes[27,120,121]. For instance, AMGs with the potential to degrade some pesticide

precursors and pollutants, as well as AMGs involved in sulfur compound cycling, have been identified from metagenomes and potentially associated with specific ecological niches[28,122–125]. Importantly however, true AMGs are, almost by definition, rare across viral genomes (Fig. 3d), and thus prone to high false-discovery rates. Given the large scale of current viromics analyses and the expected rarity of AMGs, even a very low error rate can lead to a substantial number of false positive AMG predictions, and these computational predictions should always be interpreted and presented with caution[86].

Meanwhile, viruses can also encode enzymes of interest for biotechnological applications. For instance, new divergent and hypercompact CRISPR-Cas systems were identified in uncultivated phage genomes and validated *in vitro* as efficient genome editors[34,126] (Fig. 3c). More generally, the general compactness of virus-encoded genes compared to their host homologs was previously noted, and probably results from the stronger limitation on genome length experienced by viruses[127]. Similarly, uncultivated virus genomes may be the source of novel DNA modification enzymes, such as methylases[128,129]. Finally, viral genes are also mined for potential therapeutic applications, in particular lysins, which may represent potential new classes of antibiotics[130,131]. Leveraging the extensive diversity of uncultivated viruses, the natural host specificity and adaptation of viruses, and recent developments in high-throughput synthetic biology, represents a promising path forward for the development of new applications and therapeutics based on phage genes[132].

Future perspectives for viromics studies

Modeling and inferring virus-host dynamics and interactions

Despite all the new discoveries enabled by viromics approaches, a key challenge remains in connecting uncultivated viruses, known only from their genome content and ecological distribution, to their potential host(s), which significantly limits our understanding of these viruses. While host prediction is, intuitively, a critical part of any viromics pipeline, and has been a topic of extensive work and tool development over the last decade, connecting uncultivated viruses to their host *in silico* remains challenging. The most accurate methods currently available compare new uncultivated virus genomes to databases of known virus marker genes associated with a given host range[133,134], or to genomes of candidate hosts to identify signals of host adaptation and virus-host arms race[85,135]. Current *in silico* host prediction tools now typically integrate these different information, often through a machine learning approach, to provide the most reliable prediction possible[136,137] (Fig. 4a). Unfortunately, *in silico* host predictions remain limited by the host and virus reference databases currently available, and the complexity of the signals observed. Pragmatically, this means that for some environments (e.g. soil) and virus groups (e.g. RNA viruses), the majority of uncultivated viruses remain without any predicted host[19,136]. Even when available, host predictions are currently limited to a host taxon, e.g. prediction of the host family or genus, and better datasets and methods are needed to eventually link viruses to individual host species and/or strains[138,139].

Beyond sequence alignment-based methods, other approaches have been explored focusing on the co-localization of viruses and hosts. This type of analysis has been attempted in different ways, either *in silico* with e.g. abundance correlation between viruses and potential hosts[135,140], or *in vitro* through methods like proximity ligation that inform on pieces of DNA physically located next to each other in a sample[41,141–144]. While promising, these analyses can be complicated to interpret, since direct co-occurrence or co-localization does not necessarily translate into virus-host

interactions. For instance, depending on the type and timing of infection cycle, virus and host abundances will not necessarily be directly correlated[145]. Similarly, some virus genomes may be present in non-host microbial cells, confounding co-localization signals[40] (Fig. 4b and c). Pragmatically, this means that these approaches may be very efficient in some cases, e.g. inactive prophages residing in host genomes, but challenged in other situations, e.g. lytic virus infecting a microbe itself engaged in multiple microbe-microbe interactions (Fig. 4b). Moving forward, significantly improving the rate, resolution, and reliability of host predictions will likely require transitioning from "black box" tools to prediction frameworks explicitly modelling the different steps of viral infection (host cell entry, host cell takeover, genome replication, virion production, and host cell lysis), as well as the complex virus-host ecological networks and dynamics[146]. Recent AI developments aided by the fast-paced growth of virus, host, and virus-host interactions databases represent such an opportunity to more quantitatively evaluate the possibility and probability of infection of a given host strain by a given virus, and eventually better consider viruses in large-scale microbiome and ecosystem models[147,148].

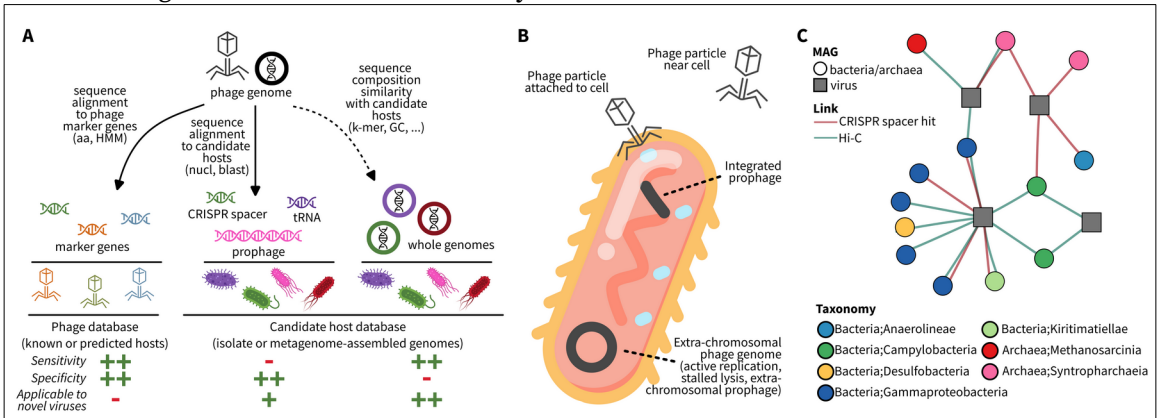

Figure 4: **Host prediction methods and their interpretation in viromics.** A. Overview of the three main types of in silico host prediction approaches typically used in viromics pipelines. From left to right, sequence similarity to known viral marker genes, sequence similarity to specific sequences from candidate host genomes (e.g. CRISPR spacer, tRNA, or proviruses), and overall sequence composition similarity between virus and candidate host genome. B. Schematic representation of the different situations in which a virus genome would be in close vicinity or within a microbial cell, and thus captured through host prediction methods such as Hi-C cross-linking and CRISPR spacer matching. C. Example of a virus-microbe interaction network derived from Hi-C and CRISPR analysis of a deep-sea microbial mat metagenomic dataset, showing some viruses seemingly interacting with microbes from different phyla and domains, adapted from[40].

Assessing the activity of uncultivated viruses

In addition to linking viruses and hosts, another critical aspect of viral ecology studies is the evaluation of viral activity, i.e. the identification and quantification of active viral replication through time and space in a microbiome. Considering the extensive viral diversity recovered by metagenomics, understanding which viruses are actively replicating or dormant across different ecological conditions will be key to properly estimate their potential impacts on microbiome processes. Metatranscriptomics, or shotgun RNA sequencing, currently provides the most straightforward and accessible window into viral activity. Matching metatranscriptome reads to metagenome-assembled viral genomes enables the detection of actively transcribed viral genes, which itself reflects an active infection since the transcription of viral genes should only happen in

a host cell (Fig. 5). RNA-Seq studies on cultivated virus-host model systems already led to the identification of different transcriptional activity patterns that may be used as signatures of different infection rates and stages, for instance lysogenic vs lytic infection in phages[149–152]. These patterns can now be searched in metatranscriptomes to help distinguish between different types of infection cycles occurring in a microbiome[102,153,154].

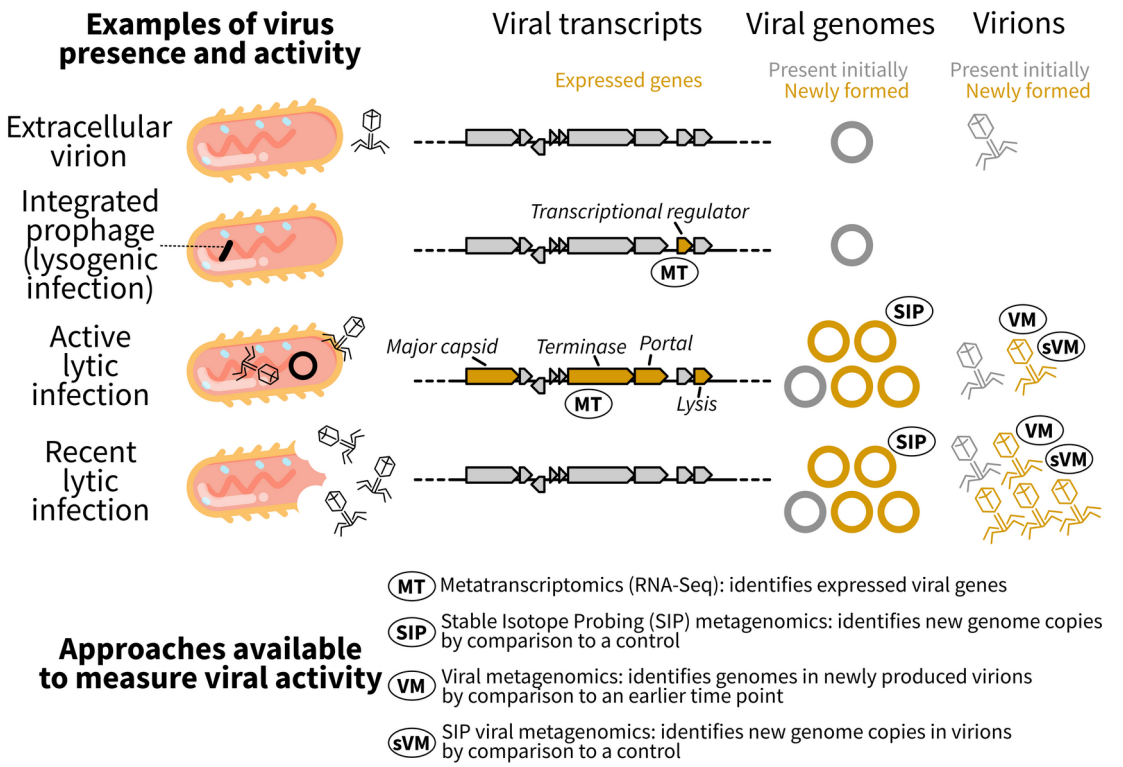

Figure 5: **Complementary approaches for in situ viral activity measurement.** The main schematic shows a comparison between different types and stages of infections (rows), and their typical respective signals in terms of gene expression, viral genome replication, and new virion production (columns). The products (transcripts, genomes, or virions) newly formed are highlighted in yellow, while the ones present in the control sample (e.g. previous sampling point in a time series) are highlighted in grey. Methods are indicated with acronyms and indicated next to the type of activity they can detect. MT: metatranscriptomics, VM: viral metagenomics, SIP: Stable Isotope Probing.

Although the application of metatranscriptomics to measure the activity of uncultivated viruses is promising, its conceptual and analytical frameworks are still under development. Specifically, new workflows need to be established and validated to integrate metagenomic and metatranscriptomic data to robustly quantify the different types of transcriptional activity for individual viruses. These workflows must also account for challenges such as limited sequencing depth, variations in viral abundance, differential gene expression levels, and the asynchronicity of infections within a microbiome (Fig. 5). This latter challenge, leading to a blurred signal when different infection stages co-occur in the same sample, may be addressed by single-cell RNA sequencing, which can enable the identification of viral activity separately for individual cells[155–159]. Finally, metatranscriptomics is mostly applicable to evaluate the transcriptional activity of DNA viruses, while transcripts and genomes can be impossible to distinguish for RNA viruses. In some cases, the detection of replication intermediaries (for ssRNA viruses) or differential

coverage between coding and non-coding strand (for dsRNA viruses) may still provide insights into the activity of RNA viruses, but here again analysis frameworks and pipelines remain to be fully developed[102].

Other approaches have been used to evaluate uncultivated virus activity and can complement metatranscriptomics. For instance, comparing the relative abundance of virus genomes in viral-fraction and cellular-fraction metagenomes across a time series can reflect virion production and thus recent active infections[103,160]. Similarly, Stable Isotope Probing (SIP) metagenomics can be used to identify actively replicating genomes, which can reflect ongoing viral replication in environmental samples[161–163]. As these different methods rely on different signals (gene transcription, genome replication, viral capsid formation), they will reflect virus activity on different spatial and temporal scales (Fig. 5). A key next step will be to understand and characterise the specific constraints and limitations of each method, and determine how to best combine them to explore different aspects of viral activity in microbiomes.

Conclusions

Viromics is now the primary method used to explore and characterise global viral diversity. It already unveiled fundamental aspects of viral ecology, from the global distribution of viruses and their vast genetic diversity to the potential association of specific viruses and genes with key processes shaping global metabolism and human health. This broad characterization of viruses across ecosystems will likely continue and even accelerate in the next few years as more datasets are being generated, including through large-scale standardised efforts aiming at characterising the viral diversity of specific ecosystems such as the human microbiome.

As a global map of viruses on Earth emerges, and the "virus discovery" phase of viromics progressively slows down, we expect the field to expand towards the detailed characterization of specific viruses and viral genes identified as especially important, e.g. showing high prevalence or connected to a host or process of interest. These viromics-guided detailed characterization of individual viruses and virus genes will require strong interdisciplinary teams able to bridge from large-scale microbiome data analysis to high-throughput *in vitro* molecular characterization approaches, including novel synthetic biology and biological engineering methods. Given the incredible diversity of viruses and viral genes on Earth, improved computational pipelines, most likely leveraging recent and upcoming developments in the AI field, will also be required to efficiently identify and rank potential targets for follow-up experiments. In that context, some of the most pressing challenges in the current viromics pipeline include (i) functionally characterize the vast diversity of novel genes encoded by viruses, (ii) linking uncultivated viruses to their host, or even better their host range, and (iii) modelling viral activity and virus-host dynamics over time from 'omics data. With improved methods for *in vitro* characterization of novel viruses and virus genes alongside a better understanding of microbiome diversity and dynamics, viromics will play a critical role in the precise evaluation of virus roles across ecosystems. It will also be instrumental in identifying viruses and genes with potential biotechnological applications, such as phage therapy and broader microbiome manipulation.

References

1. Koonin, E. V., Kuhn, J. H., Dolja, V. V. & Krupovic, M. Megataxonomy and global ecology of the virosphere. *The ISME Journal* **18**, wrad042 (2024).
2. Koonin, E. V. *et al.* Global Organization and Proposed Megataxonomy of the Virus World. *Microbiol Mol Biol Rev* **84**, e00061-19 (2020).
3. Suttle, C. A. Viruses: unlocking the greatest biodiversity on Earth. *Genome* **56**, 542–544 (2013).
4. Bratbak, G., Egge, J. K. & Heldal, M. Viral mortality of the marine alga Emiliania huxleyi (Haptophyceae) and termination of algal blooms. *Marine Ecology Progress Series* **93**, 39–48 (1993).
5. Coy, S. R., Gann, E. R., Pound, H. L., Short, S. M. & Wilhelm, S. W. Viruses of Eukaryotic Algae: Diversity, Methods for Detection, and Future Directions. *Viruses* **10**, 487 (2018).
6. Lennon, J. T. & Martiny, J. B. H. Rapid evolution buffers ecosystem impacts of viruses in a microbial food web. *Ecology Letters* **11**, 1178–1188 (2008).
7. Albright, M. B. N. *et al.* Experimental evidence for the impact of soil viruses on carbon cycling during surface plant litter decomposition. *ISME Communications* **2**, 24 (2022).
8. Weitz, J. S. & Wilhelm, S. W. Ocean viruses and their effects on microbial communities and biogeochemical cycles. *F1000 biology reports* **4**, 17 (2012).
9. Runa, V., Wenk, J., Bengtsson, S., Jones, B. V. & Lanham, A. B. Bacteriophages in Biological Wastewater Treatment Systems: Occurrence, Characterization, and Function. *Front. Microbiol.* **12**, (2021).
10. Carreira, C. *et al.* Integrating viruses into soil food web biogeochemistry. *Nat Microbiol* **9**, 1918–1928 (2024).
11. Roux, S. A Viral Ecogenomics Framework To Uncover the Secrets of Nature's "Microbe Whisperers". *mSystems* **4**, 1–5 (2019).
12. Vela, J. D. & Al-Faliti, M. Emerging investigator series: the role of phage lifestyle in wastewater microbial community structures and functions: insights into diverse microbial environments. *Environmental Science: Water Research & Technology* **9**, 1982–1991 (2023).
13. Pires, D. P., Cleto, S., Sillankorva, S., Azeredo, J. & Lu, T. K. Genetically Engineered Phages : a Review of Advances over the Last Decade. *Microbiology and Moleculat Biology Reviews* **80**, 523–543 (2016).
14. Edwards, R. A. & Rohwer, F. Viral metagenomics. *Nature Reviews Microbiology* **3**, 504–510 (2005).
15. Breitbart, M. *et al.* Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 14250–5 (2002).
16. Breitbart, M. *et al.* Metagenomic Analyses of an Uncultured Viral Community from Human Feces. *Journal of bacteriology* **185**, 6220–6223 (2003).
17. Camargo, A. P. *et al.* IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Research* **51**, D733–D743 (2023).
18. Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D. & Lawley, T. D. Massive expansion of human gut bacteriophage diversity. *Cell* **184**, 1098-1109.e9 (2021).
19. Ma, B. *et al.* Biogeographic patterns and drivers of soil viromes. *Nat Ecol Evol* **8**, 717–728 (2024).
20. Dutilh, B. E. *et al.* A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat Commun* **5**, 4498 (2014).
21. Yutin, N. *et al.* Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nature Microbiology* **3**, 38–46 (2018).
22. Edwards, R. A. *et al.* Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. *Nat Microbiol* **4**, 1727–1736 (2019).
23. Roux, S. *et al.* Ecogenomics and potential biogeochemical impacts of uncultivated globally abundant ocean viruses. *Nature* **537**, 689–93 (2016).
24. Horst, A. M. ter *et al.* Minnesota peat viromes reveal terrestrial and aquatic niche partitioning for local and global viral populations. *Microbiome* **9**, 1–18 (2021).
25. Breitbart, M., Bonnain, C., Malki, K. & Sawaya, N. A. Phage puppet masters of the marine microbial

realm. *Nat Microbiol* **3**, 754–766 (2018).
26. Sieradzki, E. T., Ignacio-Espinoza, J. C., Needham, D. M., Fichot, E. B. & Fuhrman, J. A. Dynamic marine viral infections and major contribution to photosynthetic processes shown by spatiotemporal picoplankton metatranscriptomes. *Nature Communications* **10**, (2019).
27. Sharon, I. *et al.* Photosystem I gene cassettes are present in marine virus genomes. *Nature* **461**, 258–262 (2009).
28. Kieft, K. *et al.* Ecology of inorganic sulfur auxiliary metabolism in widespread bacteriophages. *Nature Communications* **12**, 1–16 (2021).
29. Emerson, J. B. *et al.* Host-linked soil viral ecology along a permafrost thaw gradient. *Nature Microbiology* **3**, 870–880 (2018).
30. Chen, L.-X. *et al.* Large freshwater phages with the potential to augment aerobic methane oxidation. *Nat Microbiol* **5**, 1504–1515 (2020).
31. Ahlgren, N. A., Fuchsman, C. A., Rocap, G. & Fuhrman, J. A. Discovery of several novel, widespread, and ecologically distinct marine Thaumarchaeota viruses that encode amoC nitrification genes. *The ISME Journal* **13**, 618–631 (2019).
32. Braga, L. P. P. *et al.* Viruses direct carbon cycling in lake sediments under global change. *Proceedings of the National Academy of Sciences* **119**, e2202261119 (2022).
33. Schwartz, D. A. *et al.* Human-Gut Phages Harbor Sporulation Genes. *mBio* **14**, e00182-23 (2023).
34. Pausch, P. *et al.* Crispr-casphi from huge phages is a hypercompact genome editor. *Science* **369**, 333–337 (2020).
35. Palmer, M. *et al.* Diversity and Distribution of a Novel Genus of Hyperthermophilic Aquificae Viruses Encoding a Proof-Reading Family-A DNA Polymerase. *Frontiers in Microbiology* **11**, 1–18 (2020).
36. Daly, R. A. *et al.* Viruses control dominant bacteria colonizing the terrestrial deep biosphere after hydraulic fracturing. *Nature Microbiology* **4**, 352–361 (2019).
37. Medvedeva, S. *et al.* Three families of Asgard archaeal viruses identified in metagenome-assembled genomes. *Nat Microbiol* **7**, 962–973 (2022).
38. Medvedeva, S., Borrel, G., Krupovic, M. & Gribaldo, S. A compendium of viruses from methanogenic archaea reveals their diversity and adaptations to the gut environment. *Nat Microbiol* **8**, 2170–2182 (2023).
39. Rambo, I. M., Langwig, M. V., Leão, P., De Anda, V. & Baker, B. J. Genomes of six viruses that infect Asgard archaea from deep-sea sediments. *Nat Microbiol* **7**, 953–961 (2022).
40. Hwang, Y., Roux, S., Coclet, C., Krause, S. J. E. & Girguis, P. R. Viruses interact with hosts that span distantly related microbial domains in dense hydrothermal mats. *Nat Microbiol* **8**, 946–957 (2023).
41. Marbouty, M., Thierry, A., Millot, G. A. & Koszul, R. MetaHiC phage-bacteria infection network reveals active cycling phages of the healthy human gut. *eLife* **10**, 1–51 (2021).
42. Arkhipova, K. *et al.* Temporal dynamics of uncultured viruses: a new dimension in viral diversity. *The ISME Journal* **12**, 199–211 (2017).
43. Knowles, B. *et al.* Lytic to temperate switching of viral communities. *Nature* **531**, 533–537 (2016).
44. Ignacio-Espinoza, J. C., Ahlgren, N. A. & Fuhrman, J. A. Long-term stability and Red Queen-like strain dynamics in marine viruses. *Nat Microbiol* **5**, 265–271 (2020).
45. Garmaeva, S. *et al.* Transmission and dynamics of mother-infant gut viruses during pregnancy and early life. *Nat Commun* **15**, 1945 (2024).
46. Zhang, F. *et al.* Longitudinal dynamics of gut bacteriome, mycobiome and virome after fecal microbiota transplantation in graft-versus-host disease. *Nat Commun* **12**, 65 (2021).
47. Lam, S. *et al.* Roles of the gut virome and mycobiome in faecal microbiota transplantation. *The Lancet Gastroenterology & Hepatology* **7**, 472–484 (2022).
48. Conceição-Neto, N., Yinda, K. C., Van Ranst, M. & Matthijnssens, J. NetoVIR: Modular Approach to Customize Sample Preparation Procedures for Viral Metagenomics. in *The Human Virome: Methods and Protocols* (eds. Moya, A. & Pérez Brocal, V.) 85–95 (Springer, New York, NY, 2018). doi:10.1007/978-1-4939-8682-8_7.
49. Kleiner, M., Hooper, L. V. & Duerkop, B. A. Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes. *BMC Genomics* **16**, 7 (2015).

50. Soria-Villalba, A. *et al.* Comparison of Experimental Methodologies Based on Bulk-Metagenome and Virus-like Particle Enrichment: Pros and Cons for Representativeness and Reproducibility in the Study of the Fecal Human Virome. *Microorganisms* **12**, 162 (2024).
51. Hayes, S., Mahony, J., Nauta, A. & van Sinderen, D. Metagenomic Approaches to Assess Bacteriophages in Various Environmental Niches. *Viruses* **9**, 127 (2017).
52. Trubl, G. *et al.* Optimization of viral resuspension methods for carbon-rich soils along a permafrost thaw gradient. *PeerJ* **4**, e1999 (2016).
53. Santos-Medellin, C. *et al.* Viromes outperform total metagenomes in revealing the spatiotemporal patterns of agricultural soil viral communities. *ISME Journal* **15**, 1956–1970 (2021).
54. Kosmopoulos, J. C., Klier, K. M., Langwig, M. V., Tran, P. Q. & Anantharaman, K. Viromes vs. mixed community metagenomes: choice of method dictates interpretation of viral community ecology. *Microbiome* **12**, 195 (2024).
55. Lücking, D., Mercier, C., Alarcón-Schumacher, T. & Erdmann, S. Extracellular vesicles are the main contributor to the non-viral protected extracellular sequence space. *ISME Communications* **3**, 112 (2023).
56. Shkoporov, A. N. *et al.* The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. *Cell Host & Microbe* **26**, 527-541.e5 (2019).
57. Labonté, J. M. *et al.* Single-cell genomics-based analysis of virus–host interactions in marine surface bacterioplankton. *ISME J* **9**, 2386–2399 (2015).
58. Jarett, J. K. *et al.* Insights into the dynamics between viruses and their hosts in a hot spring microbial mat. *ISME Journal* **14**, 2527–2541 (2020).
59. Martinez-Hernandez, F. *et al.* Single-virus genomics reveals hidden cosmopolitan and abundant viruses. *Nature communications* **8**, 15892 (2017).
60. Lund, M. C. *et al.* Diverse microviruses circulating in invertebrates within a lake ecosystem. *Journal of General Virology* **105**, 002049 (2024).
61. Lopez, J. K. M. *et al.* Genomes of Bacteriophages Belonging to the Orders Caudovirales and Petitvirales Identified in Fecal Samples from Pacific Flying Fox (Pteropus tonganus) from the Kingdom of Tonga. *Microbiology Resource Announcements* **11**, e00038-22 (2022).
62. Marine, R. *et al.* Caught in the middle with multiple displacement amplification: the myth of pooling for avoiding multiple displacement amplification bias in a metagenome. *Microbiome* **2**, 3 (2014).
63. Kim, K.-H. & Bae, J.-W. Amplification Methods Bias Metagenomic Libraries of Uncultured Single-Stranded and Double-Stranded DNA Viruses. *Appl Environ Microbiol* **77**, 7663–7668 (2011).
64. Hillary, L. S., Adriaenssens, E. M., Jones, D. L. & McDonald, J. E. RNA-viromics reveals diverse communities of soil RNA viruses with the potential to affect grassland ecosystems across multiple trophic levels. *ISME COMMUN.* **2**, 34 (2022).
65. Potapov, S. *et al.* RNA-Seq Virus Fraction in Lake Baikal and Treated Wastewaters. *International Journal of Molecular Sciences* **24**, 12049 (2023).
66. Roux, S. *et al.* Minimum information about an uncultivated virus genome (MIUVIG). *Nature Biotechnology* **37**, 29–37 (2019).
67. Pinto, Y. & Bhatt, A. S. Sequencing-based analysis of microbiomes. *Nat Rev Genet* **25**, 829–845 (2024).
68. Johansen, J. *et al.* Genome binning of viral entities from bulk metagenomics data. *Nat Commun* **13**, 965 (2022).
69. Kieft, K., Adams, A., Salamzade, R., Kalan, L. & Anantharaman, K. vRhyme enables binning of viral genomes from metagenomes. *Nucleic Acids Research* **50**, e83 (2022).
70. Meyer, F. *et al.* Critical Assessment of Metagenome Interpretation: the second round of challenges. *Nat Methods* **19**, 429–440 (2022).
71. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
72. Li, D. *et al.* MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**, 3–11 (2016).
73. Antipov, D., Raiko, M., Lapidus, A. & Pevzner, P. A. Metaviral SPAdes: assembly of viruses from metagenomic data. *Bioinformatics* **36**, 4126–4129 (2020).

74. Roux, S., Emerson, J. B., Eloe-Fadrosh, E. A. & Sullivan, M. B. Benchmarking viromics: An in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ* **5**, e3817 (2017).
75. Schulz, F. *et al.* Advantages and Limits of Metagenomic Assembly and Binning of a Giant Virus. *mSystems* **5**, 10.1128/msystems.00048-20 (2020).
76. Camargo, A. P. *et al.* Identification of mobile genetic elements with geNomad. *Nat Biotechnol* 1–10 (2023) doi:10.1038/s41587-023-01953-y.
77. Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **8**, 90 (2020).
78. Guo, J. *et al.* VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 37 (2021).
79. Ren, J. *et al.* Identifying viruses from metagenomic data using deep learning. *Quant Biol* **8**, 64–77 (2020).
80. Hegarty, B. *et al.* Benchmarking informatics approaches for virus discovery: caution is needed when combining in silico identification methods. *mSystems* **9**, e01105-23 (2024).
81. Nayfach, S. *et al.* CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nature Biotechnology* (2020) doi:10.1038/s41587-020-00774-7.
82. Terzian, P. *et al.* PHROG: families of prokaryotic virus proteins clustered using remote homology. *NAR Genomics and Bioinformatics* **3**, lqab067 (2021).
83. Shaffer, M. *et al.* DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Research* **48**, 8883–8900 (2020).
84. Jang, H. B. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nature Biotechnology* **37**, 632–639 (2019).
85. Coclet, C. & Roux, S. Global overview and major challenges of host prediction methods for uncultivated phages. *Current Opinion in Virology* **49**, 117–126 (2021).
86. Pratama, A. A. *et al.* Expanding standards in viromics: in silico evaluation of dsDNA viral genome identification, classification, and auxiliary metabolic gene curation. *PeerJ* **9**, e11447 (2021).
87. Zhou, Z., Martin, C., Kosmopoulos, J. C. & Anantharaman, K. ViWrap: A modular pipeline to identify, bin, classify, and predict viral–host relationships for viruses from metagenomes. *iMeta* **2**, e118 (2023).
88. Coclet, C., Camargo, A. P. & Roux, S. MVP: a modular viromics pipeline to identify, filter, cluster, annotate, and bin viruses from metagenomes. *mSystems* **9**, e00888-24 (2024).
89. Páez-Espino, D. *et al.* Uncovering Earth's virome. *Nature* **536**, 425–430 (2016).
90. Enault, F. *et al.* Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. *The ISME journal* 053025 (2016) doi:10.1038/ismej.2016.90.
91. Anantharaman, K. *et al.* Sulfur Oxidation Genes in Diverse Deep-Sea Viruses. *Science* **344**, 757–760 (2014).
92. Nayfach, S. *et al.* Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nature Microbiology* **6**, 960–970 (2021).
93. Gregory, A. C. *et al.* The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut. *Cell Host & Microbe* **28**, 724-740.e8 (2020).
94. An, L. *et al.* Global diversity and ecological functions of viruses inhabiting oil reservoirs. *Nat Commun* **15**, 6789 (2024).
95. Brum, J. R. *et al.* Ocean plankton. Patterns and ecological drivers of ocean viral communities. *Science* **348**, 1261498 (2015).
96. Gregory, A. C. *et al.* Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell* **177**, 1109-1123.e14 (2019).
97. Graham, E. B. *et al.* A global atlas of soil viruses reveals unexplored biodiversity and potential biogeochemical impacts. *Nat Microbiol* **9**, 1873–1883 (2024).
98. Neri, U. *et al.* Expansion of the global RNA virome reveals diverse clades of bacteriophages. *Cell* **185**, 4023-4037.e18 (2022).

99. Zayed, A. A. et al. Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. *Science* **376**, 156–162 (2022).
100. Sakaguchi, S. *et al.* NeoRdRp: A Comprehensive Dataset for Identifying RNA-dependent RNA Polymerases of Various RNA Viruses from Metatranscriptomic Data. *Microbes and Environments* **37**, (2022).
101. Wang, R. H. *et al.* PhageScope: a well-annotated bacteriophage database with automatic analyses and visualizations. *Nucleic Acids Research* **52**, D756–D761 (2024).
102. Coclet, C. *et al.* Virus diversity and activity is driven by snowmelt and host dynamics in a high-altitude watershed soil ecosystem. *Microbiome* **11**, 237 (2023).
103. Santos-Medellín, C., Blazewicz, S. J., Pett-Ridge, J., Firestone, M. K. & Emerson, J. B. Viral but not bacterial community successional patterns reflect extreme turnover shortly after rewetting dry soils. *Nat Ecol Evol* **7**, 1809–1822 (2023).
104. Bolaños, L. M., Michelsen, M. & Temperton, B. Metagenomic time series reveals a Western English Channel viral community dominated by members with strong seasonal signals. *The ISME Journal* **18**, wrae216 (2024).
105. Sun, C. L. *et al.* Virus ecology and 7-year temporal dynamics across a permafrost thaw gradient. *Environmental Microbiology* **26**, e16665 (2024).
106. Muscatt, G., Cook, R., Millard, A., Bending, G. D. & Jameson, E. Viral metagenomics reveals diverse virus-host interactions throughout the soil depth profile. *mBio* **14**, e02246-23 (2023).
107. Coutinho, F. H., Rosselli, R. & Rodríguez-Valera, F. Trends of Microdiversity Reveal Depth-Dependent Evolutionary Strategies of Viruses in the Mediterranean. *mSystems* **4**, 10.1128/msystems.00554-19 (2019).
108. Gao, S.-M. *et al.* Depth-related variability in viral communities in highly stratified sulfidic mine tailings. *Microbiome* **8**, 89 (2020).
109. Pavlopoulos, G. A. *et al.* Unraveling the functional dark matter through global metagenomics. *Nature* **622**, 594–602 (2023).
110. Zayed, A. A. *et al.* efam: an e xpanded, metaproteome-supported HMM profile database of viral protein fam ilies. *Bioinformatics* 2–9 (2021) doi:10.1093/bioinformatics/btab451.
111. Fremin, B. J. *et al.* Thousands of small, novel genes predicted in global phage genomes. *Cell Reports* **39**, 110984 (2022).
112. Chevallereau, A., Pons, B. J., van Houte, S. & Westra, E. R. Interactions between bacterial and phage communities in natural environments. *Nat Rev Microbiol* **20**, 49–62 (2022).
113. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* **42**, 243–246 (2024).
114. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
115. Shao, B. & Yan, J. A long-context language model for deciphering and generating bacteriophage genomes. *Nat Commun* **15**, 9392 (2024).
116. Hwang, Y., Cornman, A. L., Kellogg, E. H., Ovchinnikov, S. & Girguis, P. R. Genomic language model predicts protein co-regulation and function. *Nat Commun* **15**, 2880 (2024).
117. Sullivan, M. B. *et al.* Prevalence and Evolution of Core Photosystem II Genes in Marine Cyanobacterial Viruses and Their Hosts. *PLoS biology* **4**, e234 (2006).
118. Lindell, D. *et al.* Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 11013–11018 (2004).
119. Mann, N. H., Cook, A., Millard, A., Bailey, S. & Clokie, M. Bacterial photosynthesis genes in a virus. *Nature* **424**, 741–741 (2003).
120. Puxty, R. J. & Millard, A. D. Functional ecology of bacteriophages in the environment. *Current Opinion in Microbiology* **71**, 102245 (2023).
121. Brown, T. L., Charity, O. J. & Adriaenssens, E. M. Ecological and functional roles of bacteriophages in contrasting environments: marine, terrestrial and human gut. *Current Opinion in Microbiology* **70**, 102229 (2022).
122. Zheng, X. *et al.* Organochlorine contamination enriches virus-encoded metabolism and pesticide

degradation associated auxiliary genes in soil microbiomes. *ISME J* **16**, 1397–1408 (2022).
123. Johansen, J. *et al.* Centenarians have a diverse gut virome with the potential to modulate metabolism and promote healthy lifespan. *Nat Microbiol* **8**, 1064–1078 (2023).
124. Gao, R. *et al.* Ecological drivers and potential functions of viral communities in flooded arsenic-contaminated paddy soils. *Science of The Total Environment* **872**, 162289 (2023).
125. Kieft, K. *et al.* Virus-associated organosulfur metabolism in human and environmental systems. *Cell Reports* **36**, 109471 (2021).
126. Al-Shayeb, B. *et al.* Diverse virus-encoded CRISPR-Cas systems include streamlined genome editors. *Cell* **185**, 4574-4586.e16 (2022).
127. Thompson, L. R. *et al.* Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proceedings of the National Academy of Sciences of the United States of America* **108**, E757-64 (2011).
128. Hiraoka, S. *et al.* Diverse DNA modification in marine prokaryotic and viral communities. *Nucleic Acids Research* **50**, 1531–1550 (2022).
129. Seong, H. J., Roux, S., Hwang, C. Y. & Sul, W. J. Marine DNA methylation patterns are associated with microbial community composition and inform virus-host dynamics. *Microbiome* **10**, 157 (2022).
130. Fu, Y. *et al.* DeepMineLys: Deep mining of phage lysins from human microbiome. *Cell Reports* **43**, 114583 (2024).
131. Pottie, I., Vázquez Fernández, R., Van de Wiele, T. & Briers, Y. Phage lysins for intestinal microbiome modulation: current challenges and enabling techniques. *Gut Microbes* **16**, 2387144 (2024).
132. Fischetti, V. Development of Phage Lysins as Novel Therapeutics: A Historical Perspective. *Viruses* **10**, 310 (2018).
133. Coutinho, F. H. *et al.* RaFAH: Host prediction for viruses of Bacteria and Archaea based on protein content. *Patterns* **2**, (2021).
134. Amgarten, D., Iha, B. K. V., Piroupo, C. M., Silva, A. M. da & Setubal, J. C. vHULK, A new tool for bacteriophage host prediction based on annotated genomic features and deep neural networks. *bioRxiv* 0–15 (2020) doi:10.1101/2020.12.06.413476.
135. Edwards, R. A., McNair, K., Faust, K., Raes, J. & Dutilh, B. E. Computational approaches to predict bacteriophage–host relationships. *FEMS Microbiology Reviews* **40**, 258–272 (2016).
136. Roux, S. *et al.* iPHoP: An integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria. *PLoS Biol* **21**, e3002083 (2023).
137. Wang, W. *et al.* A network-based integrated framework for predicting virus–prokaryote interactions. *NAR Genomics and Bioinformatics* **2**, lqaa044 (2020).
138. Boeckaerts, D. *et al.* Prediction of Klebsiella phage-host specificity at the strain level. *Nat Commun* **15**, 4355 (2024).
139. Gaborieau, B. *et al.* Prediction of strain level phage–host interactions across the Escherichia genus using only genomic information. *Nat Microbiol* **9**, 2847–2861 (2024).
140. Meng, L. *et al.* Quantitative Assessment of Nucleocytoplasmic Large DNA Virus and Host Interactions Predicted by Co-occurrence Analyses. *mSphere* **6**, 10.1128/msphere.01298-20 (2021).
141. Wu, R. *et al.* Hi-C metagenome sequencing reveals soil phage–host interactions. *Nat Commun* **14**, 7666 (2023).
142. Uritskiy, G. *et al.* Accurate viral genome reconstruction and host assignment with proximity-ligation sequencing. 2021.06.14.448389 Preprint at https://doi.org/10.1101/2021.06.14.448389 (2021).
143. Marbouty, M., Baudry, L., Cournac, A. & Koszul, R. Scaffolding bacterial genomes and probing host-virus interactions in gut microbiome by proximity ligation (chromosome capture) assay. *Science Advances* **3**, e1602105 (2017).
144. Du, Y., Fuhrman, J. A. & Sun, F. ViralCC retrieves complete viral genomes and virus-host pairs from metagenomic Hi-C data. *Nat Commun* **14**, 502 (2023).
145. Coenen, A. R. & Weitz, J. S. Limitations of Correlation-Based Inference in Complex Virus-Microbe Communities. *mSystems* **3**, e00084-18 (2018).
146. Bastien, G. E. *et al.* Virus-host interactions predictor (VHIP): Machine learning approach to resolve

microbial virus-host interaction networks. *PLOS Computational Biology* **20**, e1011649 (2024).
147. Kumar, M., Ji, B., Zengler, K. & Nielsen, J. Modelling approaches for studying the microbiome. *Nat Microbiol* **4**, 1253–1267 (2019).
148. Sokol, N. W. *et al.* Life and death in the soil microbiome: how ecological processes influence biogeochemistry. *Nat Rev Microbiol* **20**, 415–430 (2022).
149. Howard-Varona, C. *et al.* Regulation of infection efficiency in a globally abundant marine Bacteriodetes virus. *The ISME Journal* **00**, 1–12 (2016).
150. Lindell, D. *et al.* Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. *Nature* **449**, 83–86 (2007).
151. Owen, S. V. *et al.* A window into lysogeny: revealing temperate phage biology with transcriptomics. *Microbial Genomics* **6**, (2020).
152. Blasdel, B. G., Chevallereau, A., Monot, M., Lavigne, R. & Debarbieux, L. Comparative transcriptomics analyses reveal the conservation of an ancestral infectious strategy in two bacteriophage genera. *ISME J* **11**, 1988–1996 (2017).
153. Stough, J. M. A. *et al.* Molecular prediction of lytic vs lysogenic states for Microcystis phage : Metatranscriptomic evidence of lysogeny during large bloom events. *PLoS ONE* 1–17 (2017).
154. Merges, D. *et al.* Metatranscriptomics reveals contrasting effects of elevation on the activity of bacteria and bacterial viruses in soil. *Molecular Ecology* **32**, 6552–6563 (2023).
155. Kuchina, A. *et al.* Microbial single-cell RNA sequencing by split-pool barcoding. *Science* **371**, eaba5257 (2021).
156. Shen, Y. *et al.* High-throughput single-microbe RNA sequencing reveals adaptive state heterogeneity and host-phage activity associations in human gut microbiome. *Protein & Cell* pwae027 (2024) doi:10.1093/procel/pwae027.
157. Putzeys, L. *et al.* Exploring the transcriptional landscape of phage–host interactions using novel high-throughput approaches. *Current Opinion in Microbiology* **77**, 102419 (2024).
158. Hevroni, G., Vincent, F., Ku, C., Sheyn, U. & Vardi, A. Daily turnover of active giant virus infection during algal blooms revealed by single-cell transcriptomics. *Science Advances* **9**, eadf7971 (2023).
159. Fromm, A. *et al.* Single-cell RNA-seq of the rare virosphere reveals the native hosts of giant viruses in the marine environment. *Nat Microbiol* **9**, 1619–1629 (2024).
160. Van Goethem, M. W., Swenson, T. L., Trubl, G., Roux, S. & Northen, T. R. Characteristics of wetting-induced bacteriophage blooms in biological soil crust. *mBio* **10**, 1–15 (2019).
161. Barnett, S. E. & Buckley, D. H. Metagenomic stable isotope probing reveals bacteriophage participation in soil carbon cycling. *Environmental Microbiology* **25**, 1785–1795 (2023).
162. Trubl, G. *et al.* Active virus-host interactions at sub-freezing temperatures in Arctic peat soil. *Microbiome* **9**, 208 (2021).
163. Ngo, V. Q. H. *et al.* Establishing Host–Virus Link Through Host Metabolism: Viral DNA SIP Validation Using T4 Bacteriophage and E. coli. *Curr Microbiol* **81**, 266 (2024).
164. Charon, J. *et al.* Consensus statement from the first RdRp Summit: advancing RNA virus discovery at scale across communities. *Front. Virol.* **4**, (2024).
165. Simmonds, P. *et al.* Virus taxonomy in the age of metagenomics. *Nat Rev Microbiol* **15**, 161–168 (2017).
166. Casjens, S. Prophages and bacterial genomics: what have we learned so far?: Prophage genomics. *Molecular Microbiology* **49**, 277–300 (2003).
167. Holmes, E. C. The Evolution of Endogenous Viral Elements. *Cell Host & Microbe* **10**, 368–377 (2011).
168. Lang, A. S., Westbye, A. B. & Beatty, J. T. The Distribution, Evolution, and Roles of Gene Transfer Agents in Prokaryotic Genetic Exchange. *Annual Review of Virology* **4**, 87–104 (2017).
169. Scholl, D. Phage Tail–Like Bacteriocins. *Annual Review of Virology* **4**, 453–467 (2017).
170. Krupovic, M., Bamford, D. H. & Koonin, E. V. Conservation of major and minor jelly-roll capsid proteins in Polinton (Maverick) transposons suggests that they are bona fide viruses. *Biology Direct* **9**, 6 (2014).
171. Gaïa, M. *et al.* Mirusviruses link herpesviruses to giant viruses. *Nature* **616**, 783–789 (2023).
172. Zheludev, I. N. *et al.* Viroid-like colonists of human microbiomes. *Cell* **0**, (2024).

173. Banfield, J. *et al.* Convergent evolution of viral-like Borg archaeal extrachromosomal elements and giant eukaryotic viruses. 2024.11.05.622173 Preprint at https://doi.org/10.1101/2024.11.05.622173 (2024).
174. Bolduc, B. *et al.* Identification of Novel Positive-Strand RNA Viruses by Metagenomic Analysis of Archaea-Dominated Yellowstone Hot Springs. *J Virol* **86**, 5562–5573 (2012).

Highlighted references

2. Koonin, E. V. et al. Global Organization and Proposed Megataxonomy of the Virus World. Microbiol Mol Biol Rev 84, e00061-19 (2020).

This review proposes a global genome-based viral classification framework that integrates both isolate and metagenome-assembled viral genomes.

15. Breitbart, M. et al. Genomic analysis of uncultured marine viral communities. Proceedings of the National Academy of Sciences of the United States of America 99, 14250–5 (2002).

This is one of the first viral metagenomic analyses from an environmental sample, highlighting the high number of novel genes encoded by viruses.

17. Camargo, A. P. et al. IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. Nucleic Acids Research 51, D733–D743 (2023).

IMG/VR is a large database integrating metagenome-assembled viral genomes from a broad range of ecosystems.

20. Dutilh, B. E. et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nat Commun 5, 4498 (2014).

This first description of the CrAssphage genome, which highlighted the potential of viromics for the discovery of novel highly-abundant viruses.

25. Breitbart, M., Bonnain, C., Malki, K. & Sawaya, N. A. Phage puppet masters of the marine microbial realm. Nat Microbiol 3, 754–766 (2018).

This review presents an overview of the auxiliary metabolic genes (AMGs) discovered at the time in marine phages, and highlights the different cellular processes possibly impacted.

34. Pausch, P. et al. Crispr-casphi from huge phages is a hypercompact genome editor. Science 369, 333–337 (2020).

This study highlights the unique features relevant for biotechnological applications of a phage-encoded Cas gene initially discovered via viromics.

40. Hwang, Y., Roux, S., Coclet, C., Krause, S. J. E. & Girguis, P. R. Viruses interact with hosts that span distantly related microbial domains in dense hydrothermal mats. Nat Microbiol 8, 946–957 (2023).

This study of deep-sea metagenomes reports a potentially broad range of interactions for some viruses, and highlights several potential mechanisms for such interactions to be observed.

54. Kosmopoulos, J. C., Klier, K. M., Langwig, M. V., Tran, P. Q. & Anantharaman, K. Viromes vs. mixed community metagenomes: choice of method dictates interpretation of viral community ecology. Microbiome 12, 195 (2024).

This benchmarking study highlights the differences between and complementarity of different metagenomics approaches for viromics studies.

59. Martinez-Hernandez, F. et al. Single-virus genomics reveals hidden cosmopolitan and abundant viruses. Nature communications 8, 15892 (2017).

This application of single-virus genomics to oceanic samples uncovers a widespread virus population difficult to assemble from metagenomes, highlighting a potential blind spot for some viromics analyses.

66. Roux, S. et al. Minimum information about an uncultivated virus genome (MIUVIG). Nature Biotechnology 37, 29–37 (2019).

This consensus paper outlines key approaches for recovery and analysis of uncultivated virus genomes, and the critical metadata to report when submitting these genomes to public databases.

91. Anantharaman, K. et al. Sulfur Oxidation Genes in Diverse Deep-Sea Viruses. Science 344, 757–760 (2014).

This report and analysis of sulfur oxidation AMGs in metagenome-assembled virus genomes significantly expanded the list of metabolism potentially redirected during viral infections.

103. Santos-Medellín, C., Blazewicz, S. J., Pett-Ridge, J., Firestone, M. K. & Emerson, J. B. Viral but not bacterial community successional patterns reflect extreme turnover shortly after rewetting dry soils. Nat Ecol Evol 7, 1809–1822 (2023).

This time-series viromics analysis integrating bulk soil and viral fractions samples highlights distinct successional patterns between microbial and viral communities.

109. Pavlopoulos, G. A. et al. Unraveling the functional dark matter through global metagenomics. Nature 622, 594–602 (2023).

This global re-analysis of public metagenomes reveals and describes a large number of novel protein families, many seemingly encoded by viruses.

119. Mann, N. H., Cook, A., Millard, A., Bailey, S. & Clokie, M. Bacterial photosynthesis genes in a virus. Nature 424, 741–741 (2003).

This first description of photosynthesis genes encoded by phages, reported here from isolate cyanophages, spurred the search for and subsequent discoveries of a large number of AMGs.

155. Kuchina, A. et al. Microbial single-cell RNA sequencing by split-pool barcoding. Science 371, eaba5257 (2021).

This study describes the development and application of a single-cell RNA sequencing method for prokaryotes, that provides a unique opportunity for detailed characterization of virus-host interactions.

164. Charon, J. et al. Consensus statement from the first RdRp Summit: advancing RNA virus discovery at scale across communities. Front. Virol. 4, (2024).

This consensus statement reports on the discussions led at the first RdRP summit, which gathered experts from different fields around the topic of sequence-based RNA virus discovery.

165. Simmonds, P. et al. Virus taxonomy in the age of metagenomics. Nat Rev Microbiol 15, 161–168 (2017).

This perspective represents a key step and statement by the ICTV towards the integration of metagenome-derived virus genomes in the formal virus taxonomy.

Acknowledgments

The work conducted by the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231.

Competing interests statement

The authors declare no competing interests.