Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Metagenomics in Virology

Simon Roux, Department of Energy Joint Genome Institute, Walnut Creek, CA, United States
Jelle Matthijnssens, Rega Institute for Medical Research, KU Leuven, Leuven, Belgium
Bas E Dutilh, Utrecht University, Utrecht, The Netherlands and Radboud University Medical Center, Nijmegen, The Netherlands

Glossary

Metagenomics The study of genetic material (DNA or RNA) extracted from an environmental sample. Recent studies use "shotgun" metagenomics, i.e., the untargeted sequencing of all genomes from all members in the sampled community.

Viromics Viral metagenomics, i.e., shotgun metagenomics applied specifically to the encapsidated fraction of DNA and/or RNA from a sample. The encapsidated fraction is typically obtained through a combination of filtration, precipitation, and DNase/RNase treatment.

Metagenomics Applied to Viruses

Historically, viruses have been primarily explored using laboratory cultivation: new viruses were obtained from clinical or environmental samples through propagation and isolation on cell cultures. This process is, however, biased and challenging to apply at large scales because (i) many viruses depend on host cells that are difficult to maintain as clonal culture in the laboratory, and (ii) even if the cells are available, propagating viruses may require specific conditions distinct from those used to cultivate the cells. These considerations are especially meaningful for viruses with microbial hosts, the vast majority of which remain uncultivated to date.

Metagenomics bypasses this requirement for cultivation and instead relies on the sequencing of viral genomic material extracted directly from a sample (see Box 1 and Fig. 1). Thus far, the history of viral metagenomics has seen two major phases. Initially, entire communities of viruses were assayed by analyzing and comparing short sequencing reads obtained from diverse environments. Because of the fragmented nature of these data, most of these studies had to be conducted at the community scale, and identifying and distinguishing individual viruses in these datasets remained challenging. More recently, bioinformatic advances have enabled the reconstruction of individual viral genome sequences from metagenomes, allowing naturally occurring viruses to be identified and studied at high, genomic resolution. Using a metagenomics approach, entirely new types of viruses can now be discovered, surveyed, and characterized even without cultivation. The unique ability offered by metagenomics to study uncultivated viruses led to the emergence of two parallel and interconnected fields: a clinical one, where metagenomics promises to be a catch-all method for the unbiased surveillance and diagnosis of viral pathogens, and one focused on natural biomes. that aims to describe the diversity of the viral world and understand its ecological and evolutionary drivers and impacts.

Pioneering Viral Metagenomics, One Gene at a Time

When the field of shotgun environmental metagenomics was pioneered in 2002 by the laboratory of Forest Rohwer at San Diego State University, the first datasets consisted of three viral metagenomes (viromes) that, together comprised just under 2,500 short genomic fragments derived from two natural marine viral communities and one human feces sample. While limited in scope and resolution, these and other early viromes provided an unprecedented view of complex viral communities in nature. Both oceanic and human fecal viromes pointed toward the existence of an extensive virus diversity. This diversity of the virosphere was estimated by comparing the sequencing reads within each metagenome, and observing that almost every fragment was unique. Moreover, comparing the short sequencing reads to a reference database of known viral genomes sequences revealed that up to 99% was not similar to any known virus, suggesting that most of the virosphere was yet to be discovered. This uncharted genomic biodiversity became popularly known as "viral dark matter".

In the years that followed, a broader range of environments was progressively surveyed using viromics including freshwater lakes, hot springs, agricultural soils, or human skin, saliva, and gut samples. Improvements in DNA sequencing technologies, especially the advent of the popular pyrosequencing platform, that has since been surpassed and discontinued, increased the scale of these datasets by providing hundreds of thousands of short genomic fragments for each sample. By directly comparing the sequences across these datasets, several studies indicated that virus genes tend to structure by environment rather than by sample location, implying that some of these genes may be globally distributed. In addition, when sampled from the same freshwater and hypersaline ponds across several days, weeks, and months, viral metagenomes revealed that the genetic composition of viral communities was coherent at a broad level, but some individual viral genes experienced rapid changes in relative abundance.

Scaling up From Fragmented Genes to Complete Genomes

While the analyses outlined above were foundational for our current understanding of virus diversity, they were limited by the short length of next-generation sequencing reads which fragmented the view of viral genomes. These limitations were progressively

Box 1 Use of complementary methods to target different types of viruses

A number of approaches have been developed to specifically select and survey the genetic material contained by virus particles in a given sample. Alternatively, viral genomes can also be analyzed from "bulk" metagenomes which include both virus particles and microbial cells. Virus sequences obtained from "bulk" metagenomes will typically reflect viruses infecting their host cell at the time of sampling, either actively replicating or not, while viromes enables a deeper and more focused exploration of the virus diversity in a specific site or sample.

Regardless of the type of sample, viromes are most often generated through a combination of centrifugation, filtration and DNase/RNase treatment, aiming at removing as much of the cellular genomes as possible (Fig. 1). A typical protocol will notably include a filtration through 0.22, 0.45, or 0.8 μm membrane filters to remove bacteria and larger cells. Depending on the initial concentration of virus particles, a concentration step using e.g., iron chloride (FeCl), PolyEthyleneGlycol (PEG), or tangential flow filtration step(s), may be necessary to obtain enough material for sequencing library preparation. Cesium chloride density gradients ultracentrifugation can also be used to further separate viruses from extracellular DNA and large particles in complex samples, although this step can also lead to a substantial loss of viral material. Finally, the virus particles obtained are typically treated with DNase or RNase to remove free DNA and/or RNA. Depending on the type of virus studied, the corresponding protocols for RNA or DNA extraction and sequencing library preparation are then applied, after releasing the genetic material from the virus particle through e.g., a heat shock if necessary.

A critical step in this process is to recover enough material for sequencing. While micrograms of DNA were initially needed, several protocols are now available which only require ~1 ng of DNA. In addition, a DNA/RNA random amplification step, called "whole genome amplification", can also be conducted in order to gather enough input material. This type of approach was initially used in almost every virome study, and revealed important information for example on the unsuspected diversity of ssDNA genome viruses in the environment (see below). However, the whole genome amplification process is inherently biased, and these datasets are not quantitative, i.e., one cannot draw any conclusion about the relative abundance of the viruses identified in these amplified metagenomes. Thus, whole genome amplification methods have now been often replaced by advanced library preparation protocols which require nanogram-scale input but enable quantitative datasets well suited for ecological studies. Alternatively, for cases in which target viruses represent a minor part of the templates, targeted sequence capture approaches have been used, mainly in a clinical framework as they can only be applied to viruses with known genomes but can detect these viruses with a very high sensitivity.

The recovery of virus genomes from bulk metagenomes and from viromes each have their own limitations. For bulk metagenomes, viruses typically represent only a minor fraction of all sequences compared to cellular genomes. This means that the virus genomes obtained this way will tend to be restricted to abundant viruses found in their host cells, while viruses that are not infecting at the time of sampling, viruses with a low frequency of infected hosts, or viruses infecting rare hosts will likely be missed. Viromes provide a deeper description of the virus community, since most of the sequencing data will be obtained from virus genomes. In addition, virus particles will not represent only current infections but a more integrated sampling of all recent successful infections, the timing of which depending on the type of sample and the individual virus decay rate. Yet viromes still suffer from several biases. Notably, the size-based selection of virus particles excludes most of the larger viruses such as the "giant viruses", and viromes also tend to be dominated by viruses with high burst size while under-sampling viruses with low burst size and long infection time. All metagenomes (bulk and viromes) will struggle with very rare viruses, as well as hypervariable viruses which genome will not assemble well. Hence complementary approaches such as targeted capture approach for the former, and long read sequencing for the latter, are being developed (Fig. 1).

Overall, the different methods developed over the last decade to sequence genomes from uncultivated viruses are mostly complementary and can be individually tailored for specific applications. Virus discovery can be achieved through bulk metagenomes or viromes, while viral ecology studies will tend to rely more on viromes as a reflection of virus activity and transport, and metagenomics used as a diagnostic tool in the clinic would be the most likely to use sequence capture. Nevertheless, all these complementary approaches will be needed for achieving a comprehensive picture of viral diversity.

overcome through an increase in sequencing throughput associated with improvements in sequence assembly and analysis tools. The first large-scale assemblies of viral genomes from short metagenomic fragments were published around 2010, and quickly became a standard analysis in the viral metagenomic field so that by the year ~2015, complete or near-complete virus genomes were routinely reported and analyzed in viromics studies.

In only a couple of years, metagenomics has thus transformed the way scientists can identify and study viruses in the environment, as illustrated by the quick rise of virus genomes and genome fragments assembled from metagenomes available in public databases (Fig. 2). In 2010, only 84 viral genomes (fragments) assembled from metagenomes were publicly available, while this number reached 35,000 in 2016, and 775,000 in 2018. Genome sequences of uncultivated viruses are frequently obtained not only from viral metagenomes, i.e., metagenomes from virus-targeted samples, but also from "bulk" metagenomes in which virus particle were not enriched and viral and microbial sequences are mixed. Combined with genome sequences obtained from isolates, these "uncultivated virus genomes" represent the foundation of an extensive mapping of the viral sequence space.

Viral Metagenomics in the Clinic

Over the past few decades, a number of molecular techniques, such as (q)PCR or ELISA, have been developed and used to detect pathogenic viruses in clinical samples. However, these techniques can only detect previously known viruses, and often require

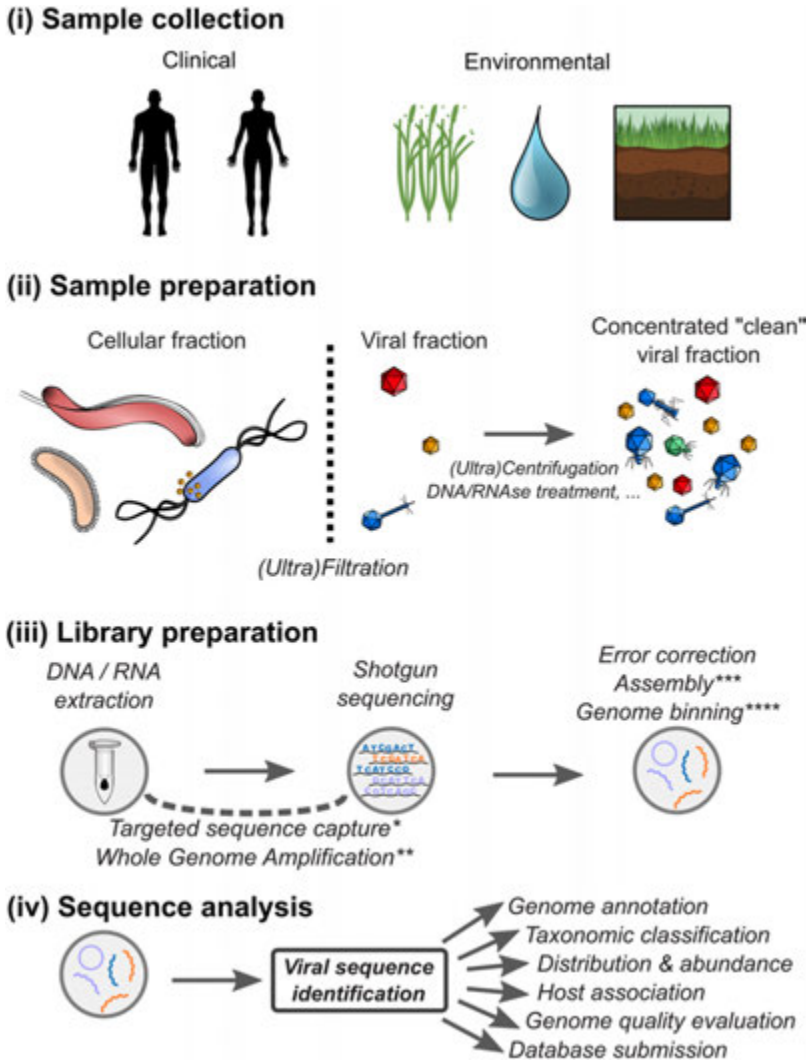

Fig. 1 Overview of the viral metagenomics workflow. The overall process used to generate and analyze viral metagenomes can be divided into four major steps: (i) Collection of environmental and/or clinical sample, (ii) sample preparation, (iii) library preparation, and (iv) sequence analysis. The sample preparation step can target either the cellular fraction (left) or the viral fraction (right) in which case viral particles are often further concentrated and purified to remove free nucleic acids. *Targeted sequence capture can be applied to the extracted DNA/RNA to enrich for a specific virus. **While whole genome amplification was initially used routinely for viral metagenomes, it has now been supplanted by methods enabling the preparation of more quantitative libraries from low input ($\sim$ 1 ng), hence whole genome amplification is now primarily used in single-cell or single-virus-particle experiments. ***The genome assembly can be bypassed if using long-read sequencing technologies, although these long-read datasets require a more careful error correction. ****Genome binning, i.e., the identification of multiple contigs assembled from a metagenome and corresponding to the same genome, is typically only used for large genomes (e.g., $>$500 kb), and individual contigs are directly analyzed instead for most viruses.

specific assays for each pathogen. Metagenomics instead offers the possibility to detect known and novel viruses without prior knowledge from a single analysis, and is thus well suited and already applied to study emerging and/or rare viruses, as well as cases which remain negative using the available diagnostic tests (see below). A number of challenges remain however for viral metagenomics to become a standard clinical procedure. First, given the current cost associated with sample processing and sequencing, metagenomes are still more expensive and slower than ELISA's or qPCR assays. Second, there is no generally validated bioinformatics pipelines than can perform a rapid, sensitive, and specific analysis of the obtained data on a bench top computer. Finally, physicians will have to be trained and guided to deal with the obtained breadth of data. Specifically, it is becoming clear that each individual is chronically infected with a dozen or more eukaryotic viruses (many of which have not been associated with any disease, e.g., anelloviruses), and that many known viral pathogens can also cause asymptomatic infections. Therefore, a physician might get a list of viruses (and other potentially pathogenic or unknown organisms), and it will be a challenge to identify the actual cause of a particular disease. Nevertheless, the price of sample preparation and high throughput sequencing has declined enormously in the last decade including with the development of smaller and faster machines, while automatic virome

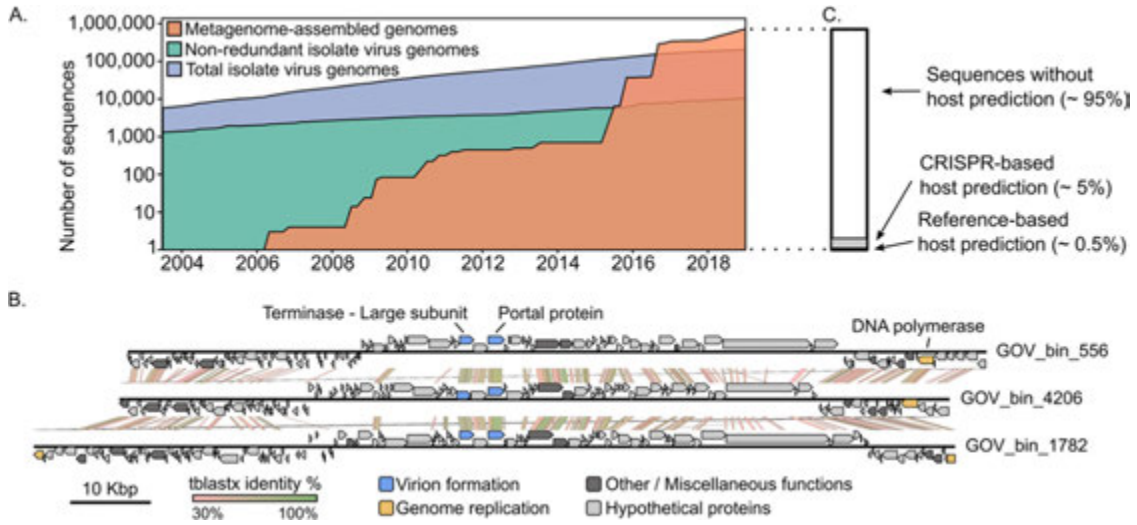

Fig. 2 Size of virus genome databases over time, host linkage information, and examples of uncultivated virus genomes. A. The total number of genomes from isolates was based on queries to the NCBI nucleotide database portal, while the number of uncultivated virus genomes was estimated by compiling data from the literature and from the IMG/VR database. The number of sequences is displayed on a $\log_{10}$ scale. B. Comparison of 3 complete viral genomes assembled from viral metagenomes sampled from the Indian, Pacific, and Atlantic oceans, through the *Tara* Oceans expedition. These sequences were identified and analyzed as part of the "Global Ocean Virome" dataset (GOV). Predicted genes are colored by functional annotation. C. Overview of the host predictions available for uncultivated virus genomes in the IMG/VR database. Host prediction was based on signals including sequence similarity with isolate viruses, prophages, and CRISPR spacers derived from known bacterial and archaeal genomes.

analysis pipelines are being actively developed, so that metagenomics will likely be available in the near future as a routine test allowing physicians to get a viral diagnosis from a biological sample in a matter of minutes to hours in their home office or on the clinic bedside.

Metagenomic Discovery of New Viral Pathogens

Currently, metagenomics is most often used in a diagnostic context when both conventional and enhanced molecular testing fail to identify a causative agent in a sample. These cases can represent a significant fraction of patients for diseases such as acute diarrhea, for which an etiological agent is identified in only ~60% of cases. In this framework, metagenomic analysis can lead to the discovery of unexpected or novel viruses that are associated with a specific set of symptoms.

First, metagenomics can successfully identify known viruses in unexpected sample types. These studies include the detection of enterovirus D68 in clinical samples (rectal, throat, and oral swabs as well as blood samples) in cases of acute flaccid paralysis, the detection of herpes simplex virus 1 (HSV-1) in cerebrospinal fluid samples of a patient with encephalitis, and the detection of mumps vaccine virus from the brain biopsy of a patient with chronic encephalitis. In addition, new human pathogens only distantly related to known viruses have also been discovered with metagenomics. These include the Bas-Congo virus, a rhabdovirus that was associated with a 2009 hemorrhagic fever outbreak in the African Congo, as well as novel rhinovirus, bocavirus, arenavirus, and parechoviruses. Finally, entirely novel types of potentially pathogenic viruses have been described through metagenomics, including previously unknown cycloviruses, cosaviruses, and klasseviruses.

Diagnostics through viral metagenomics has also been applied to non-human animals as well as plants, and similarly revealed new potential viral pathogens in organisms showing unexplained symptoms. Multiple new virus types including novel parvoviruses, polyomaviruses, sapoviruses, and picornaviruses were for example identified in livestock samples (porcine and bovine), while a large diversity of persistent RNA viruses were newly identified across several groups of plants. However, it is important to note that the detection of a (novel) virus in a sample from a patient with an illness of unknown etiology does not prove causation, even in cases of a demonstrated significant association between the presence of the virus sequence and the observed symptoms. Hence, metagenomics will often be the first step of a longer process involving attempts to propagate the virus in culture, or monitoring healthy individuals exposed to the suspected pathogen (see below "Future of viral metagenomics: major challenges and upcoming innovations").

Epidemiological Surveillance and Environmental Monitoring

In parallel to the diagnosis application, metagenomics is also very well suited for environmental surveillance. Species representing important reservoirs of viruses with high zoonotic pandemic potential such as mosquitoes, rodents, and bats have been specifically targeted in this context. A recent study investigating the virome of more than 200 invertebrate species (a fraction of known invertebrate species), identified more than 1,400 novel RNA viruses, exemplifying that the diversity of unknown eukaryotic viruses

in the environment is enormous and only poorly characterized. Since the majority of human pandemics have a zoonotic origin, one hope is that such metagenomic surveillance will allow a faster identification of novel pandemic viruses during outbreaks, as well as identify their natural reservoirs. This knowledge is crucial for an appropriate and fast response from a medical and global health perspective. As an example, in the last two decades zoonotic coronaviruses were able to jump from bats to humans and pigs. Both the SARS (Severe Acute Respiratory Syndrome virus) and MERS (Middle East Respiratory Syndrome) viruses caused large-scale disease outbreaks in humans, whereas SADS (Swine Acute Diarrhea Syndrome) caused an epidemic in the swine industry. Ongoing efforts to characterize the virome of such reservoir animals will facilitate the implementation of control measure to prevent epidemics or enforce appreciate actions to stop ongoing epidemics. In an ideal situation, obtained environmental virome data in combination with biochemical experiments could help with the early identification of candidate viruses with the potential to transfer to a human host. For instance, a combination of metagenomics and DNA synthesis-based experiments revealed that a novel coronavirus (WIV1-CoV) initially detected in bat samples could be prime for transfer and emergence into human hosts.

Metagenomic analysis can also be leveraged in response to viral outbreaks, for example to rapidly determine viral subtypes in a novel infection source. This has been applied to cases of influenza infections as well as for a novel wild type Ebola virus outbreak, for which metagenomic approaches could correctly identify the causative agent, even in cases where traditional methods were unsuccessful because the wild type virus was too distantly related to known Ebola viruses. A correct and rapid identification of these viruses could enable the application of the correct therapeutics and guide preventive efforts against potential epidemics.

Characterizing the Global Viral Diversity

While viruses of humans, animals, and plants may have direct clinical or economic relevance, the vast majority of the (estimated) 10^{31} virus particles on Earth infect micro-organisms, including bacteria, archaea, protists, fungi, and other environmental microbes. Initial studies of environmental viral diversity focused on human feces, coastal and open ocean, freshwater lakes, as well as hypersaline and hot geothermal ponds, because protocols for efficient separation of virus particles from microbial cells were first developed for aquatic samples. Importantly though, recent innovations and technology improvements now enable application of viromics to more complex samples such as soil, groundwater, or ice cores, helping to expand our view of global viral diversity both in the human microbiome and in the environment.

Identifying Globally Dominant Bacteriophages

A striking example of a viromics discovery is that of a highly abundant bacteriophage, named "crAssphage", that was assembled from a set of human fecal viromes in 2014. The crAssphage genome was identified by combining information from 12 individual viromes, which yielded a high-confidence 97 kb sequence with matching 5′ and 3′ ends, suggesting that it represented a complete circular genome. This crAssphage genome was mostly unrelated to any isolated phage genome known at the time: from the 80 predicted proteins, less than half (39) were even remotely similar to known proteins or domains, and only 25 had a predicted function, such as "phage structural protein" or "DNA helicase". While clearly novel, crAssphage was also found to be uniquely abundant and ubiquitous: its genome was detected across 940 metagenomes, primarily from human feces, at average levels that were six times higher than all other known phages combined. By applying several independent computational host-prediction approaches, a bacterial host (*Bacteroides*) was predicted. Thus, in this instance, metagenomics revealed what remains to date the most abundant and widespread phage associated with the human gut microbiome, which had until then evaded detection through classical approaches like laboratory cultivation and PCR.

Assembling genomes of uncultivated viruses can not only identify some of the most abundant and widespread viruses in an ecosystem, but these sequences also represent foundational data for targeted follow-up experiments aimed at further characterizing these novel viruses. In the case of crAssphage, two major studies leveraged this initial genome sequence to better understand the diversity and host of these phages. First, predicted proteins from the original crAssphage genome were used as "bait" to identify related phages in a broad range of metagenomes. This revealed an extensive and diverse group of "crAss-like" phages predicted to represent a new family within the *Caudovirales* order, that may be related to *Podoviridae*. Genome comparisons within this new family also enabled the identification of conserved structural and replication gene modules. Meanwhile, another study was able to isolate a member of the crAssphage-like family through broth enrichment on *Bacteroides intestinalis* strains isolated from human gut samples, confirming the computational predictions from bioinformatic analyses that these phages were likely infecting *Bacteroidetes* hosts and had a *Podoviridae*-like morphology.

In 2016, a comprehensive effort to chart viral diversity across the global oceans yielded a similar observation. This study detected more than 15,000 viral genome fragments, and grouped them into clusters of closely related viruses, approximately consistent with genera in the viral taxonomy. Two out of the four most highly abundant and ubiquitous clusters were entirely novel and had not been described before, while the other two were similar to known bacteriophages. With viral metagenomics being applied to a larger set of samples and environments, and with bioinformatic analyses including genome assembly and interpretation constantly improving, novel groups of dominant and widespread viruses may thus be progressively revealed across many environments.

Unveiling New Uncultivated Giant Viruses

Another group of viruses whose known diversity has been vastly expanded through metagenomics are the so-called "giant viruses", dsDNA viruses with a uniquely large virion ($\sim$0.5–1 μm) and genome (often $>$1 Mb), blurring the boundaries between "simple" viruses and "complex" cellular life. Following the isolation and characterization of the first giant virus in 2004 ("*Acanthamoeba polyphaga* mimivirus"), around 50 other members of this group have been isolated, the vast majority by using an *Acanthamoeba* host. However, metagenome analyses suggest that the true diversity of giant viruses vastly exceeds the number of isolates.

As early as 2013, an analysis of 17 metagenomes revealed that giant viruses could be found in the ocean at concentrations of $\sim 10^4$ genomes/ml. These initial studies were based on the detection of marker genes, since the technologies available at the time did not enable the assembly of complex and large genomes like those of giant viruses. More recently, four complete or near-complete giant virus genomes could be assembled from metagenomes of a wastewater treatment plant. This revealed a new subgroup of giant viruses named Klosneuviruses which comprised some genomes with the largest set of translation system components found at the time in any virus, including aminoacyl transfer RNA synthetases with specificity for all 20 amino acids. Undoubtedly, as our collective ability to assemble large genomes from metagenomes increases, the giant virus diversity will keep expanding.

Revealing the Extraordinary Diversity of ssDNA and RNA Viruses

While most sequencing technologies are designed for dsDNA templates (see **Box 1**), our knowledge of single-stranded DNA (ssDNA) and RNA viruses has also been transformed by metagenomics. In both cases, specific sample processing steps are required to access these genomes, however their relatively short length (usually $<$20 kb) means that complete genomes are routinely assembled from total community shotgun metagenomes that target all the nucleic acids in an environment. As for dsDNA viruses, metagenomics revealed that these ssDNA and RNA viruses were much more diverse and broadly distributed than previously inferred from isolation and cultivation approaches.

Enrichment for circular ssDNA viruses can be achieved through phi29-based whole genome amplification, which is known to over-amplify small circular ssDNA templates. Pragmatically, this translates into viral metagenomes that are dominated by ssDNA viruses with circular genomes, which helped shed a new light on the diversity of two major groups: bacteriophages from the *Microviridae* family, and eukaryotic viruses from the CRESS DNA (Circular REp-encoding ssDNA) supergroup. The latter saw the more striking expansion: until 2009, these viruses were known exclusively in plants and vertebrates, specifically pigs and birds, yet in less than a decade, CRESS DNA viruses were detected in metagenomes sampled from primates, arthropods, and unicellular protists, as well as diverse aquatic, terrestrial, and man-made ecosystems. Hence, while the exact host range and impact of these viruses remain to be fully characterized, metagenomics already revealed that ssDNA viruses are ubiquitous and can be found associated with all types of cellular hosts.

For RNA viruses, several additional sample processing steps have to be performed to preferentially sequence viral RNA, typically including reverse transcription and random amplification. The most comprehensive study of RNA virus diversity to date included samples from $>$ 220 invertebrate species across 9 phyla, and led to the discovery of nearly 1,500 novel viruses across the 13 major clades of RNA viruses. In addition, the assembly of complete genomes provided new insights on the recombination patterns of these viruses, highlighting a remarkable propensity of RNA viruses to exchange or acquire genes horizontally, both with other viruses and with their host. RNA viruses were also detected in a much broader host range than currently known from isolates, although these host associations now have to be confirmed through laboratory experiments since virus detection in metagenomes does not equate active infection.

Leveraging Time Series to Track Virus Populations Dynamics

Improvements in metagenomics protocols post $\sim$2015 enabled the analysis of dozens of samples in parallel. In the field of viral metagenomics, this increased capacity has been leveraged specifically to analyze viral signal along time series and thus investigate virus-host dynamics in nature. Such datasets have notably been obtained from freshwater lakes, for which recurrent sampling across months or years can be done, and which usually harbor a high concentration of viruses. These first explorations of environmental viral diversity across months and seasons indicated that viruses display a large range of relative abundance patterns, from "ephemeral" ones with a single peak in abundance to "constitutive" ones detected in virtually all samples. Some of these patterns were seasonal and possibly linked to similar patterns of abundance for their microbial hosts, while other viruses displayed drastic changes from one year to the next. For instance, although longitudinal virome studies of the human gut are scarce, available data suggests a rather stable population of gut viruses (almost exclusively phages) in adults over time, whereas the infant gut virome is much more variable and may be dominated by eukaryotic viruses at particular time point coinciding with an acute enteric infection.

Time series metagenomes are especially interesting to discover and predict virus-host associations, and to analyze dynamics of known virus-host pairs. The former approach already provided host prediction for several giant viruses that are so far known exclusively from metagenome assemblies, and suggested that these may be linked to uncultivated protist hosts. The latter raised the intriguing possibility of complex and diverse virus-host relationships occurring in nature: while the expected patterns would be a strong correlation between virus and host abundances with possibly a short lag in the virus signal in a typical predator-prey

fashion, these large-scale metagenome time series instead suggested that some of the viruses could peak prior to a peak in abundance of their host, while other virus-host pairs showed no similarity in relative abundance at all. These conflicting results likely reflect the complex interactions at play between viruses and microbes in nature, including variable host ranges, from viruses infecting a unique host strain to others infecting multiple host species sometimes across different genera, as well as the spectrum of infection dynamics from fast-acting lytic viruses to slower, temperate, and even chronic ones, and the development of resistance to the virus among the host population. Despite these numerous challenges in their analysis, time series metagenomes are poised to become a key approach to complement laboratory experiments and untangle the intricate relationships between viruses and their hosts.

Future of Viral Metagenomics: Major Challenges and Upcoming Innovations

Metagenomics has quickly become a major tool for exploring viral diversity, yet several challenges need to be addressed in order to fully leverage the potential of these methods. First, metagenomes built from limited input material are still difficult to reliably obtain and interpret, and do not yet provide a comprehensive and quantitative view of the viral community present in the sample. This includes environments with low biomass such as some human tissues, hydrothermal vents, ice cores, or ancient samples, but also samples with a thick substrate or matrix to which cells and virus particles tend to adhere such as human lung mucus or coral samples. Improvement in the recovery of cells and virions from this type of substrates and in the generation of quantitative libraries from sub-nanogram input will help better survey these viral communities.

The second major challenge lies in the absence of direct host information for genomes assembled from metagenomes. In a clinical context, this means that one of Koch's postulates, which requires that the candidate etiological agent be isolated from a diseased organism and grown in pure culture, cannot be fulfilled. Already, several smacoviruses which had been detected in human samples metagenomes and suspected to represent new human viral pathogens have been found to likely infect prokaryotic cells from the human microbiome instead. In a similar way, evidence is emerging that picobirnaviruses, which are believed to be eukaryotic viruses, might actually infect bacterial cells. These examples should thus serve as a cautionary tale when trying to detect entirely new viral pathogens from mixed samples containing both human and microbial cells. A modified Koch's postulate for the metagenomic era has been proposed in which potential new pathogens must first be present and more abundant in the diseased subject compared to matched control. Then, experiments using either a sample from a disease subject or an artificial virus obtained through DNA synthesis and expression in cell cultures must be performed to demonstrate that this agent induces disease in another healthy subject. While not trivial, these additional experiments based on metagenomic results could still lead to the identification of viral pathogens much more quickly than classic culture techniques.

In an ecological context, associating uncultivated viruses to their host is also critical to understand their impact on microbial communities and to meaningfully integrate viruses into ecosystem models. Because viral ecology studies typically include hundreds to thousands of viruses of interest, these host associations are typically derived from in silico approaches based on various types of genome sequence comparison. While methods for in vitro confirmation of these metagenome-derived virus-host pairs are currently being developed, they will need to improve both in terms of scale and resolution to provide meaningful host association for the vast diversity of uncultivated viruses.

Among the expected technological improvements, two stand out as likely to benefit the field of viral metagenomics in the near future. First, long-read sequencing technologies are progressively amenable to the sequencing of environmental viral communities. Pragmatically, this means that instead of having to assemble virus genomes from short reads, a process which can yield potentially erroneous and/or incomplete genome sequences, a complete viral genome could be sequenced as a single read. Once broadly available, these long-reads metagenomes will not only bypass assembly issues but also provide valuable information about virus genome evolution by enabling whole-genome phasing of polymorphisms. Meanwhile, in an epidemiological context, long-read sequencing technologies associated with miniaturized devices, streamlined sample preparation, and live scanning of the sequencing results offers unique possibility for real-time surveillance or diagnostics. This is especially the case for the MinION sequencer based on Nanopore sequencing technology, allowing the identification of viral pathogens from a patient sample in less than 6 h, compared to more than 20 h for other sequencing technologies. The computational framework to analyze and share these types of data in a timely, safe, and meaningful way remains to be built, however it is likely that metagenomics through portable genome sequencers will become a major component of the epidemiological toolkit in the near future.

Complementarily, the throughput of sample preparation protocols and short-read sequencing approaches is likely to keep increasing at a fast pace. Concretely, these technological improvements will translate into a lower cost per sample, and an increased ability to process hundreds of samples in parallel in a timely fashion, in particular through laboratory robotics automation. For the detection of viral pathogens as well as the exploration of viral diversity and virus-host interactions in nature, this increased throughput will provide the opportunity to generate e.g., high-resolution time-series, possibly including paired cellular and viral size fractions with multiple replicates per sample, enabling more robust and sensitive data analyses.

Eventually, a fully developed virus metagenomics toolkit will enable the accurate identification of viruses in natural, clinical, and biotechnological samples for monitoring and diagnostics purposes. Moreover, as bioinformatics analysis tools advance, the reconstruction of full viral genome sequences will allow predictions to be made for the most important viruses in different environments, leading to the reconstruction of environmental virus-host networks and, when combined with other 'omics' approaches, the comprehensive evaluation of viral activity across an entire ecosystem. Collectively, these studies should lead to a

deeper understanding of viral impacts on ecological, evolutionary, and metabolic processes as well as information on potentially new viral pathogens and putative molecular virus-host interactions which could then be further characterized through targeted laboratory experiments. Hence viral metagenomics will remain a central and fundamental way to interrogate the viral world in many research fields.

Further Reading

Bibby, K., 2013. Metagenomic identification of viral pathogens. Trends in Biotechnology 31, 275–279.
Breitbart, M., Bonnain, C., Malki, K., Sawaya, N.A., 2018. Phage puppet masters of the marine microbial realm. Nature Reviews Microbiology 3, 754–766.
Conceição-Neto, N., *et al.*, 2015. Modular approach to customise sample preparation procedures for viral metagenomics: A reproducible protocol for virome analysis. Scientific Reports 12 (5), 16532.
Dutilh, B.E., Reyes, A., Hall, R.J., Whiteson, K.L., 2017. Virus discovery by metagenomics: The (im)possibilities. Frontiers in Microbiology 8, 5–8.
Gardy, J., Loman, N.J., Rambaut, A., 2015. Real-time digital pathogen surveillance – The time is now. Genome Biology 16, 15–17.
Greninger, A.L., 2018. A decade of RNA virus metagenomics is (not) enough. Virus Resarch 244, 218–229.
Hall, R.J., Draper, J.L., Nielsen, F.G.G., Dutilh, B.E., 2015. Beyond research: A primer for considerations on using viral metagenomics in the field and clinic. Frontiers in Microbiology 6, 224.
Mokili, J.L., Rohwer, F., Dutilh, B.E., 2012. Metagenomics and future perspectives in virus discovery. Current Opinion in Virology 2, 63–77.
Racaniello, V., 2016. Moving beyond metagenomics to find the next pandemic virus. Proceedings of the National Academy of Sciences of the United States of America 113, 2812–2814.
Rosario, K., Duffy, S., Breitbart, M., 2012. A field guide to eukaryotic circular single-stranded DNA viruses: Insights gained from metagenomics. Archives of Virology 157, 1851–1871.
Roux, S., *et al.*, 2019. Minimum Information about an Uncultivated Virus Genome (MIUViG). Nature Biotechnology 37, 29–37.
Roux, S., Brum, J.R., 2019. A viral reckoning: Viruses emerge as essential manipulators of global ecosystems. Environmental Microbiology Reports 11, 1–6.
Shkoporov, A.N., Hill, C., 2019. Bacteriophages of the human gut: The 'known unknown' of the microbiome. Cell Host Microbe 25, 195–209.
Sullivan, M.B., 2014. The phage metagenomic revolution. In: Rohwer, F., Youle, M., Maughan, H., Hisakawa, N. (Eds.), Life in Our Phage World. San Diego, CA: Wholon, pp. 2–55. (p. 2-55-70).
Williamson, K.E., Fuhrmann, J.J., Wommack, K.E., Radosevich, M., 2017. Viruses in soil ecosystems: An unknown quantity within an unexplored territory. Annual Review of Virology 4, 201–219.
Zhang, Y.-Z., Shi, M., Holmes, E.C., 2018. Using metagenomics to characterize an expanding virosphere. Cell 172, 1168–1172.

Relevant Websites

https://img.jgi.doe.gov/cgi-bin/vr/main.cgi
IMG/VR – Collection of viral genomes assembled from metagenomes.
http://ivirus.us
iVirus – Analysis of viromes.
http://metavir-meb.univ-bpclermont.fr/
MetaVir – Analysis of viromes.
https://www.protocols.io/groups/verve-net
VERVE Net – Viral ecology collaboration network.
http://viromes.org
VIROME – Analysis of viromes.