Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Virology

Emerging technologies in the study of the virome

Sophie E Smith[1,2,*], Wanqi Huang[1,2,*], Kawtar Tiamani[1,2], Magdalena Unterer[1,2], Mohammadali Khan Mirzaei[1,2] and Li Deng[1,2,#,$]

Despite the growing interest in the microbiome in recent years, the study of the virome, the major part of which is made up of bacteriophages, is relatively underdeveloped compared with their bacterial counterparts. This is due in part to the lack of a universally conserved marker such as the 16S rRNA gene. For this reason, the development of metagenomic approaches was a major milestone in the study of the viruses in the microbiome or virome. However, it has become increasingly clear that these wet-lab methods have not yet been able to detect the full range of viruses present, and our understanding of the composition of the virome remains incomplete. In recent years, a range of new technologies has been developed to further our understanding. Direct RNA-Seq technologies bypass the need for cDNA synthesis, thus avoiding biases subjected to this step, which further expands our understanding of RNA viruses. The new generation of amplification methods could solve the low biomass issue relevant to most virome samples while reducing the error rate and biases caused by whole genome amplification. The application of long-read sequencing to virome samples can resolve the shortcomings of short-read sequencing in generating complete viral genomes and avoid the biases introduced by the assembly. Novel experimental methods developed to measure viruses' host range can help overcome the challenges of assigning hosts to many phages, specifically unculturable ones.

Addresses
[1] Institute of Virology, Helmholtz Centre Munich — German Research Centre for Environmental Health, 85764 Neuherberg, Germany
[2] Chair of Microbial Disease Prevention, School of Life Sciences, Technical University of Munich, 85354 Freising, Germany

Corresponding author: Li Deng (li.deng@helmholtz-muenchen.de)
[*] These authors contributed equally to the work.
[#] ORCID: 0000-0003-0225-0663
[$] Research ID: A-7233-2015

Current Opinion in Virology 2022, **54**:101231

This review comes from a themed issue on **The virome in health and disease**

Edited by **Jelle Matthijnssens** and **Evelien Adriaenssens**

For complete overview about the section, refer "The virome in health and disease (2022)"

Available online 25th May 2022

https://doi.org/10.1016/j.coviro.2022.101231

Introduction

Viruses are abundant and widespread biological entities, with approximately 10^{31} virions on Earth [1]. They infect all other biological entities, such as bacteria, archaea, plants, arthropods, mammals, and even other viruses [2,3]. The majority of these viruses are bacteriophages, or phages, which influence multiple aspects of life on Earth by regulating the bacterial abundance, diversity, and metabolism [4–6]. In addition, changes in their community structure in the human body are linked to multiple human diseases [7–9]. Unlike bacteria, viruses lack universal phylogenetic markers for amplicon-based sequencing, making virome assessment more complicated than for bacteria. The development of highly sensitive metagenomic high-throughput sequencing approaches provides an opportunity to investigate the composition of viral communities (the virome) within environmental and clinical samples, allowing the study of the correlation between the virome and disease, specifically improving the detection of unculturable and novel viruses [10]. Despite representing a significant step forward in the field of virome studies, commonly used metagenomic methods often have issues, such as a bias toward dsDNA or circular genomes at the expense of ssDNA and RNA viruses, or a low yield of extracted genomic material requiring amplification, which can introduce further biases [11]. Next-generation short-read sequencing can also have problems resolving repetitive or complex genomic regions. Viruses in the virome identified through metagenomic sequencing cannot easily be linked to any particular host, and identifying which bacteria they infect can also be challenging. Here, we present some of the newest technologies aiming to overcome these problems.

Methods combating bias in metagenomic samples

For metagenomic studies, genomes must first be extracted, before they are prepared for sequencing. Most well-characterized viruses in virome studies are dsDNA viruses, with viruses with ssDNA, RNA, or multipartite genomes [12] remaining poorly represented in metagenomic analyses [13], and this is at least in part due to genome-extraction protocols favoring dsDNA [14–17]. Protocols to specifically extract these genomes have been developed: for ssDNA viruses, alkaline extraction is used to target circular genomes [18], as well as duplex-

Box 1 Computational approaches detecting background contamination in virome data.

In downstream processes, reliable bioinformatic approaches for quality control are required to predict viral genomes and remove potential contaminants *in silico*, for example, aligning sequencing reads to databases of bacterial housekeeping gene *cpn60* chaperonin or microbial rRNA [75]. Recently, M. Zolfo et al. (2019) developed a software, ViromeQC, by integrating 16S/18S, 23S/28S rRNA genes and 31 universal bacterial and archaeal marker genes to quantify the abundance of microbial contaminants from raw metagenomic reads in virome data [76]. Another viral metagenomics analysis platform, METAnnotatorX, can identify viral and nonviral DNA in reads and contigs of sequenced datasets through iterative classification steps using a multidatabase of the National Center of Biotechnology Information (NCBI) (Viral RefSeq, Archaeal RefSeq, Bacterial Refseq, and Whole RefSeq for eukaryotes) or other alternative databases integrated on-demand, for example, GenBank, VirSorter, or IMG/VR databases [77]. Many researchers also used the Virome Decontamination mode of VirSorter to flag potential contamination for removal [78].

specific nucleases that specifically digest double-stranded DNA [19]; and NetoVIR is a protocol designed to use commercial kits to effectively extract RNA genomes in addition to DNA genomes [20]. Metatranscriptomics provides an interesting new way to identify novel ssRNA viruses, which can be extracted and sequenced in the same way as mRNA transcripts. Recent studies have expanded the number of known ssRNA-phage genomes from just 25 [21] to tens of thousands [21–23]. Many of these studies looked at existing RNA-seq data and analyzed it using hidden Markov models to identify viral sequences [21]. There is the potential for newly developed direct RNA-Seq technologies, which sequence RNA without the need for a cDNA step, to further expand what we know about RNA viruses [24], and this technology is already being used for diagnostics of infections caused by RNA viruses [25], and to identify novel pathogenic RNA viruses from the complete RNA extracted from diseased tissue [26]. In addition, virome analyses can be impacted by background contamination due to the small genome size of viruses. Therefore, multiple computational approaches have been developed for background-contamination identification and removal (Box 1).

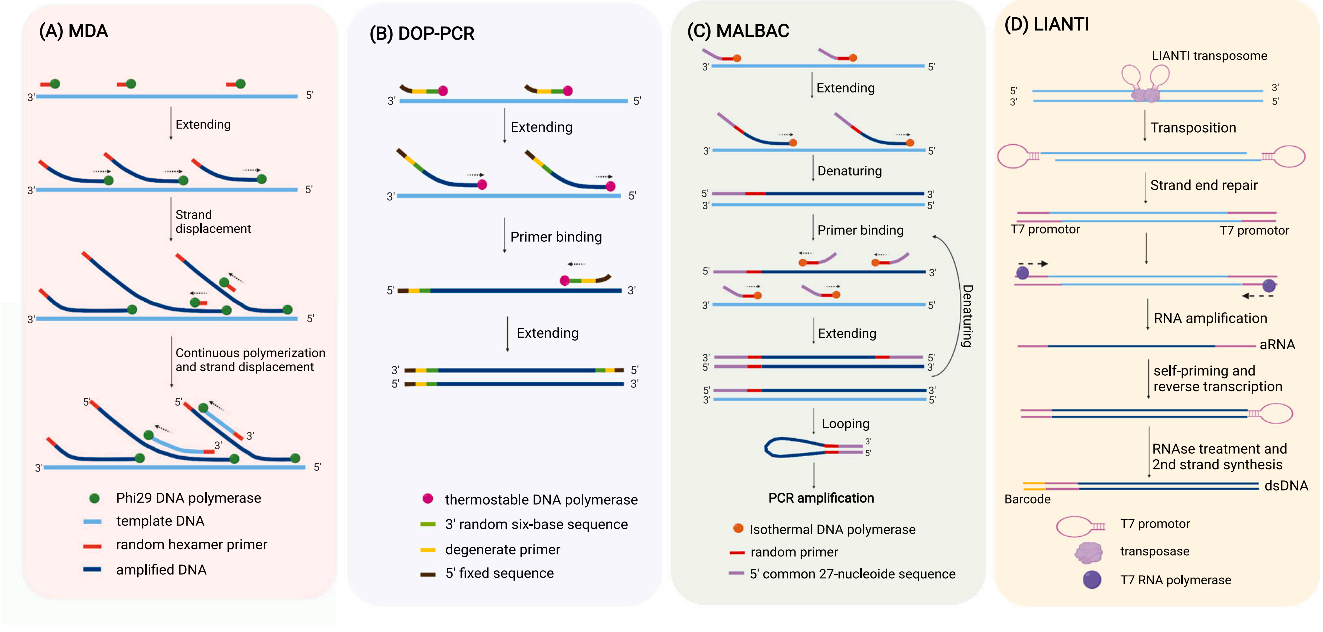
Given that low-biomass samples can yield minimal amounts of extracted genetic material, an additional amplification step before library preparation can be useful in order to circumvent this limitation and generate sufficient input material. This is particularly relevant when discussing the virome as the genome size of viruses is smaller and shorter compared with bacteria [27,28]. The earliest available methods such as whole-genome amplification have been replaced by more modern techniques, which will be discussed here [29], although the underlying mechanism of most amplification methods remains either isothermal amplification or polymerase chain reaction (PCR)-based.

One of the most common amplification methods, which can be used before library preparation, is called multiple-displacement amplification (MDA) [27,30]. Samples are incubated with random phosphorothioate-modified hexamers, dNTPs, buffer, and DNA polymerase (Figure 1) [29]. Instead of using heat to melt the DNA, the polymerase displaces downstream-bound primers while it is copying the template strand. The polymerase's high processivity leads to amplified fragments with an average length of 12 kb [29]. Although MDA can itself show bias, favoring small circular genomes and struggling to amplify templates with a high guanine-cytosine (GC) content [31], its technology has been the basis of several adaptations that aim to reduce this bias. The most widely used of these are degenerate oligonucleotide primer PCR (DOP-PCR), multiple annealing and looping-based amplification cycles (MALBAC), and the very similar PicoPLEX.

In DOP-PCR, degenerate primers bind at low annealing temperatures during the first round and then on a fixed 5′-end with higher temperatures on the second round [28,32] (Figure 1). However, DOP-PCR provides low genome coverage [32]. MALBAC and PicoPLEX improve on the genome coverage and have increased evenness of amplification compared with MDA [32]. These methods involve two steps. Despite increased performance compared with DOP-PCR, the multiple steps are more labor-intensive than the single-step DOP-PCR and are susceptible to contamination with microbial DNA and high error rates [33].

In recent years, a new generation of amplification methods have been developed, several of which improve upon their predecessors by improving the polymerase enzyme. The appropriately named 'improved DOP-PCR' uses a new thermostable DNA polymerase with stronger strand displacement than the original method, as well as an adjusted primer design [34]. This method of amplification produces better-quality amplified DNA compared with both DOP-PCR and PicoPLEX [34]. Another more recently developed method is WGA-X, which builds on MDA, but uses a thermostable mutant of the phi29 polymerase. This method enhances genome recovery, particularly for high GC-content templates when compared with traditional MDA [35]. Linear Amplification via Transposon Insertion (LIANTI) takes a different approach and uses transposons to introduce a T7 promoter, which leads to amplification through transcription, before it is reverse-transcribed into DNA (Figure 1). This means that the amplification is linear, contrasting with the exponential amplification employed by other whole genome amplification (WGA) methods.

Figure 1



Current Opinion in Virology

Comparison of different amplification methods. **(a)** MDA [29], random hexamer primers combine with the denatured template DNA. The commonly Phi29 DNA polymerase initiates primer extension. The newly amplified reverse strands are displaced by polymerase from the template and bound with primers for continuous polymerization. **(b)** DOP-PCR [33], primers contain a 6-bp degenerate random sequence on the 3′-end and a fixed 5′-end, so that it can evenly hybridize to the template at low annealing temperature. Subsequently, thermostable polymerase extends primers at a higher temperature. **(c)** MALBAC [32] primers are designed with a common 27-bp sequence at the 5′-end and an 8-bp random nucleotide sequence at the 3′-end. The specifically designed primers help the amplified strands self-hybridize in DNA loops and the primers only extend along the original template. **(d)** LIANTI: a transposome consists of T7 promoter and transposase is used. Once the transposome randomly binds and integrates into the template, a single-stranded RNA will be amplified from the T7 promoter region along the genomic DNA by T7 RNA polymerase. After reverse transcription, RNAse digestion, and second-strand synthesis, a double-stranded barcoded DNA will be created for the sequencing library [34].

The use of linear amplification reduces error rate and bias, and this method performs better than older methods, in terms of genome coverage, allele dropout, and bias [36].

Although many of these methods have been used to amplify RNA virus genomes after a reverse-transcription step, there are also methods designed and optimized specifically to amplify RNA. Whole-transcriptome amplification is analogous to WGA and works by reverse-transcribing RNA transcripts into cDNA and adding universal priming sites that are used for PCR amplification [37]. Ribo-Single Primer Isothermal Amplification uses DNA–RNA chimeric primers to amplify and covert to cDNA in a single reaction [38], and Sequence-independent single-primer amplification first developed in the early 90s for DNA amplification, has been optimized for RNA [39].

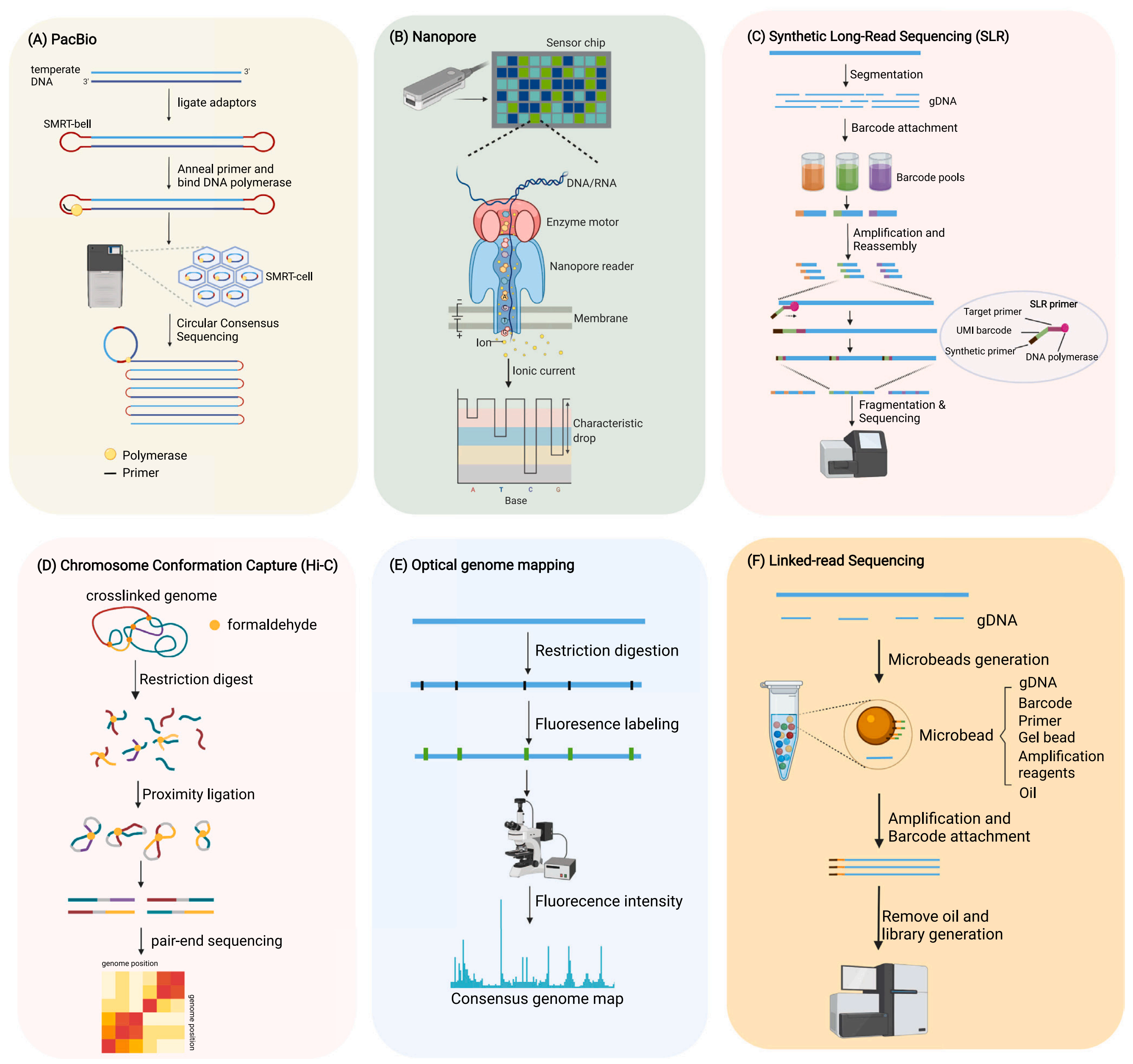
Methods combating *de novo* assembly difficulties

The majority of metagenomic studies are done using high-throughput short-read sequencing technologies, where DNA is digested into small pieces before sequencing, producing reads that are typically around 300-bp long [40]. The development of this method of sequencing, as well as its increasing availability and relatively low cost, was revolutionary for the field of viromics; however, the method has some drawbacks. Short reads must be assembled into a complete genome by looking for overlapping regions, which can make it difficult to resolve highly repetitive or complex regions [41]. For successful assembly, high coverage is required, and this can be a particular problem for the virome, where often only a small amount of genetic material can be recovered [42]. As already discussed, amplification steps may be useful but can bring their own biases. Often, assembled viruses are incomplete, and short fragmented viral contigs are then filtered out and lost during *in silico* size-selection steps [42]. Since short-read sequencing has difficulty resolving repetitive regions such as terminal repeats, it can be difficult to know if viral genomes are complete, and incomplete circular genomes can be mistakenly identified as linear [43].

A number of techniques have been developed to attempt to overcome these shortcomings, including paired-end reads and mate-pair reads, and in the last ten years or so, long-read sequencing has been developed, producing reads that are in excess of 10 kb [40]. There have been two main technologies developed for this purpose. The first is Pacific Biosciences (PacBio) single-molecule real-time sequencing, which uses circular consensus sequencing. A polymerase tethered to the bottom of a small well adds fluorescently labeled nucleotides that can be detected and identified. In this way, it can generate long reads of up to 50 kb [40] (Figure 2). The other main method for long-read sequencing is Nanopore sequencing, which passes DNA molecules through a pore, and measures the ionic current fluctuations — each base generates a different pattern in current fluctuation (Figure 2). Nanopore sequencing can generate reads even longer than PacBio — the record is 4.2 Mb [44], and most average libraries are 10–30 kb [40]. Long-read sequencing with these technologies is able to generate longer contigs than short-read technology, and therefore, more full-length genomes can be assembled [45]. Long-read methods are particularly beneficial for viromics studies as they are able to sequence whole viral genomes in a single read, without requiring any assembly [43]. This also means they are able to successfully sequence terminal repeats and this gives us additional information about phage-packaging strategy [43]. Long reads have been used for metagenomic studies, both alone and alongside short-read sequencing, and have identified novel viruses that were not identified by short-read sequencing alone [46]. However, both methods of long-read sequencing have a reputation for low accuracy [47]. Both methods often introduce indel errors that shift the reading frame and introduce false stop codons [42]. This is a problem for the virome even more than the rest of the metagenome because viral genes are already generally shorter than bacterial genes [42]. Also, the majority of viral genes are of unknown function, which makes it difficult to evaluate the accuracy of gene predictions [42]. These methods also often require a high amount of input DNA [42].

Despite recent advances improving the accuracy of both methods [40], they remain too inaccurate for some applications. There are also bioinformatics methods that can correct errors and polish reads, however, this can be time-consuming and intensive work [48]. Long-read sequencing also remains more expensive than short reads [40]. Therefore, methods that combine the accuracy and familiarity of the cheaper short-read sequencing technology with the benefits of long reads have been developed. These methods must find a way to maintain long-range data within the scope of short reads, so that synthetic long reads (SLR) can be built, while the accuracy, low cost, and other benefits of short-read sequencing can be maintained. These methods can be divided into SLR, and linked-read sequencing. In both cases, long DNA fragments are spatially separated before being digested and barcoded. Then, library preparation proceeds as normal for short-read sequencing. This additional barcoding step allows for the original long fragment to be reassembled before additional assembly [49]. The difference between the two approaches lies in the barcode coverage — SLR sequencing allows the entire fragment to be fully reassembled, whereas in linked-read

Figure 2



Current Opinion in Virology

New sequencing technologies. **(a)** PacBio sequencing [36]. Adaptors are used to circularize the template strands. Primers bind to the adaptor and the DNA polymerase binds to the primer. The polymerase attaches one fluorescent base at a time, the order of the fluorescent signals gives the sequence. **(b)** Nanopore sequencing [36]. The DNA molecule is passed through a pore, causing fluctuations in ionic current. Each base causes a different-sized drop in current, and the sequence can be identified from the changes in current. **(c)** Synthetic long-read sequencing [40]. Long transcripts are physically separated, digested into small segments, and barcoded. They are then sequenced using short-read technology such as Illumina, before the barcodes are used to reassemble the original long reads. **(d)** Chromosome-conformation capture [41]. Formaldehyde is added to cross-link genome, before being digested into small fragments. Cross-linked pieces are ligated to each other to form chimeras, which can then be sequenced. **(e)** Optical genome mapping [42]. The genome is digested by enzymes at fixed motifs, as well as fluorescently labeled. A microscope can then be used to visualize the locations where the enzyme has bound, and a consensus genome map can then be created. **(f)** Linked-read sequencing [40,43,44]. Similarly to synthetic long-read sequencing, long DNA fragments are physically separated, digested, and barcoded, in many cases using barcodes bound to a microbead. The fragments can then be sequenced using short-read technology and assembled, without the need for a long-read assembly step.

sequencing, fewer reads are barcoded in such a way that the fragment cannot be fully reconstructed, however, long-range data are retained and can be used bioinformatically for assembly [49] (Figure 2). Different methods of the initial separation of DNA fragments have been developed. Some methods, such as the GemCode platform used by company 10x Genomics, separate fragments into compartments using microfluidics. The fragments are separated into individual droplets, or gems, where they are digested and barcoded. They are then amplified before being released from the droplet, where they can undergo library preparation and short-read sequencing [50]. This platform only requires 1 ng of input DNA, however, it does require use of an expensive machine that is able to perform the microfluidics [49]. More recently developed methods, such as CPT-Seq [51], stLFR [52], and TELL-Seq [49], have focused on developing ways of partitioning the molecules without the need for expensive microfluidics, using equipment that is commonly found in most laboratories. In these methods, microbeads are each coated with a single barcode, removing the need to partition altogether. A transposon transfers the barcodes from the beads to the DNA molecules at the same time as fragmentation, meaning that the reaction can happen in a single PCR tube [49]. These methods reduce the amount of intensive work required, do not require any equipment beyond standard consumables, and use readily available sequencing technology such as Illumina sequencing machines [45,47,48].

An alternative method of maintaining long-range info in short-read sequencing is through proximity ligation, where DNA, which is physically near to each other, is linked in some way before sequencing. In the context of metagenomics, it is assumed that linked DNA came from the same cell and is therefore from the same organism, and this information can be used to correctly assemble the metagenome. It can also be used to give information about the interactions between viruses and their bacterial hosts in the microbiome, a topic that is discussed further later in this review. The most commonly used method of proximity ligation in metagenomic studies is chromosome-conformation capture (3C) technology, where chromatin is cross-linked using formaldehyde, then digested, and sequenced. Ligated sequences must have been physically close to each other before digestion, giving information on the genomic sequence and also the 3D organization of the genome [53] (Figure 2). A 3C approach for metagenomics, called meta3C, was developed as long ago as 2014 [54], and use of this method has allowed metagenomics to be performed on complex communities containing closely related strains [55–58]. However, proximity ligation can require a high amount of input DNA and can lead to issues with assembly, including false inversion or scaffold misplacement [59].

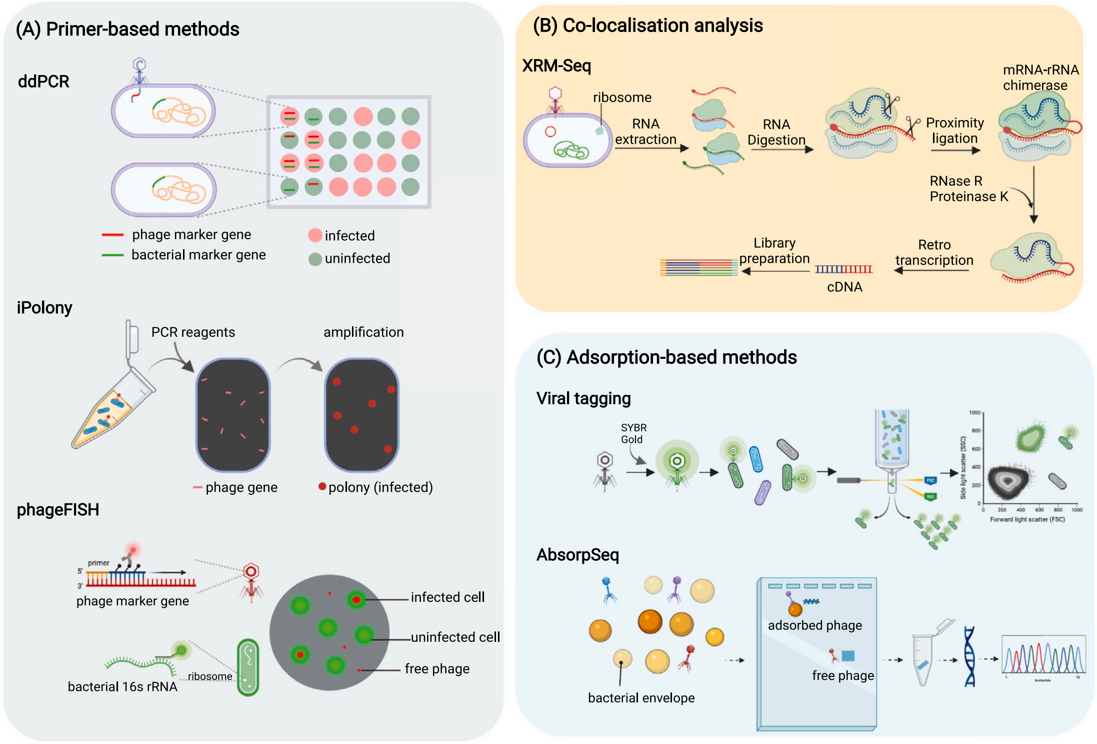
One alternative method to sequencing is genome mapping [56,57]. Historically, this was done by using restriction enzymes to digest an unknown piece of DNA to create a unique ‘fingerprint’ of restriction sites [61] (Figure 2). More modern methods use enzymes to incorporate fluorescence that can be mapped using a light microscope. The major benefit of using fluorescent optical maps over restriction mapping is that they maintain the order of the restriction sites, which restriction mapping does not [62]. Optical mapping can produce sequence reads significantly longer than even long-read sequencing, and the introduction of microfluidics has made it higher throughput than it once was [59]. However, species identification requires that the map be matched to an expected map generated from a known sequence, so on its own, it is not suitable for identifying novel species [60]. It can, however, be used as a scaffold for assembly using short-read sequences [59].

Methods combating lack of host information in viromics data

One piece of information that metagenomics struggles to give us is the host of any phage that is identified in the virome. Despite the fact that phages make up the vast majority of the virome, the interactions between phages and their hosts remain poorly characterized. There are various bioinformatics methods that attempt to assign a host to a phage genome; however, they all have their flaws, for example, some phages can be matched to spacers in the host CRISPR–cas system. However, this cannot work for bacteria that do not have a CRISPR system [63,64]. In recent years, various experimental methods have been developed to solve this problem. These can be divided into three types: those that use PCR to identify phage-marker genes within host bacteria, those that identify viruses and hosts that are spatially colocalized, and finally those that aim to identify viruses that have adsorbed to their host.

The first, and oldest, set of methods are those that use primers to amplify phage-marker genes and link them in a variety of ways to their host (Figure 3). In all of these methods, individual bacterial cells must first be separated: in microfluidic chambers, as in microfluidic PCR and droplet-digital PCR (ddPCR) [65]; spatially separated within a polyacrylamide gel as with the ipolony method [66]; or isolated in an emulsion droplet as with emulsion-paired isolation-concentration PCR (epicPCR) [67]. The next step is to amplify the target gene with PCR. With ddPCR, two PCR reactions take place in each chamber — one targeting phage genes, and one targeting host genes. If both reactions amplify their target gene, it can be concluded that the phage in that chamber infects the host in that chamber [68]. With the ipolony method, primers targeted at viral genes are embedded in the gel along with the phage-infected

Figure 3



Current Opinion in Virology

Methods of identifying phage hosts. **(a)** Primer-based method. In digital-droplet PCR [60,61] (ddPCR), bacterial cells are separated within microfluidic chambers before PCR is performed targeting both a phage-marker gene and a bacterial-marker gene. If both marker genes are amplified, the chamber contains both a phage and its bacterial host. In iPolony [62], phage-infected bacteria are embedded in a polyacrylamide gel and primers targeting a viral marker gene are used to create 'polonies' where a bacteria is infected by a phage. In phageFISH [63], fluorescent probes targeting phage and bacterial-marker genes visualize the colocalization of phage and bacterial DNA. **(b)** Colocalization-based methods. In XRM-Seq [64], ribosomes are cross-linked with their mRNA transcripts and ligated to form rRNA–mRNA chimeras that are then sequenced. The presence of viral mRNA alongside host rRNA in these chimeras identifies phage-host pairs**.** **(c)** Adsorption-based methods. In VT [65,66], fluorescently stained phages are added to a target bacterial population. Fluorescent phage-host pairs can then be separated using FACS and sequenced to identify them. In adsorp-seq [59], free phages move freely through an agarose gel, leaving just phages bound to their host in the wells. These can then be sequenced and identified.

bacteria. If an amplification sphere, or PCR colony (polony) is seen, it identifies that a phage-infected bacteria is present [66]. epicPCR uses a fusion PCR reaction to join and amplify viral and host-marker genes within phage-infected cells. These fused amplicons can then be sequenced to identify both the host and the virus [67]. Another method that uses phage-specific probes in a slightly different way is phageFISH. This uses the concept of fluorescence *in situ* hybridization (FISH) to visualize the colocalization of phage and host DNA. Here, probes are labeled with molecules that bind to and activate fluorescently labeled antibodies. In addition, to degenerate primers for a phage-specific gene, primers for bacterial rRNA genes are used to visualize host cells — this allows the covisualization of both the phage and the host DNA inside the infected cell, suggesting that the phage is infecting the bacteria [69]. The degenerate primers used for these methods are designed based on metagenomic data to amplify as wide a range of phages as possible. However, the design and optimization of these primers can be time-consuming and laborious [67], and are limited because they only target known genes. There are no universally conserved sequences in viruses, and it is inevitable that there will be bias toward known sequences [70].

Other methods have been developed to identify the colocalization of phage DNA with host DNA without the use of primers (Figure 3). One approach to do this is through Hi-C. As previously discussed, this involves ligating DNA together, which is physically close together. If phage and host DNA are found ligated together, it must be assumed that the phage DNA was inside the host cell and therefore is able to infect that host. Phage genomes are assigned to host genomes based on physical proximity [71]. Another similar method based on colocalization is XRM-Seq. In this method, ribosomes are cross-linked as in Hi-C, however, at this stage, the ribosomes and total RNA are extracted and digested with an enzyme that only targets RNA that has been cross-linked. Then, digested RNA is circularized and noncircular RNA is degraded, leaving only cross-linked transcripts. The RNA is then converted to cDNA and sequenced. Host/virus chimeras once again suggest that the virus was inside the host cell and therefore must be able to infect that host [72].

A different approach to identifying phage hosts is through identifying and sequencing phage-host pairs, where the phage has adsorbed to the host in preparation for infection (Figure 3). One method that does this is viral tagging (VT). In VT, phages are stained with a fluorescent DNA strain and then mixed with either a target host population [73], or a mixed population containing a mixture of different possible hosts [74]. The fluorescent phages bind to their hosts, and then fluorescence-activated cell sorting (FACS) is used to separate the bacteria that have a phage attached, and therefore have a fluorescent signal, from those that do not. Single-cell sorting means that individual phage-host pairs can then be sequenced, and both can be identified from their sequence [73]. Another more recently developed method is adsorp-Seq. This takes advantage of the different ways that bacteria with phages bound move differently through an agarose gel compared with free phages. Phages are mixed with hosts and then run on a gel. Phages that have not bound to the host will run freely away from the well, whereas phage-host pairs will remain in the well, where they can be extracted and sequenced [64].

Conclusion

The development of high-throughput sequencing technologies was revolutionary for viromics research; however, it has still left some gaps in our knowledge and understanding of viral communities. In recent years, a range of technologies has been developed, which aim to overcome the problems associated with short-read sequencing and fill in these gaps. Many of these technologies are brand new, and their full potential in the field of viromics may be yet to be seen.

Author contributions

All authors listed have contributed significantly to this work and approved it for publication.

Conflict of interest statement

There are no interests to declare.

Acknowledgements

This work was funded by the German Research Foundation (DFG Emmy Noether program, Proj. no. 273124240, SFB 1371, Proj. no. 395357507), and the European Research Council Starting Grant (ERC StG 803077) awarded to L.D.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest

1. Mushegian AR: **Are there 10 31 virus particles on earth, or more, or fewer?** *J Bacteriol* 2020, **202**:e00052-20.

2. Zhang YZ, Shi M, Holmes EC: **Using metagenomics to characterize an expanding virosphere**. *Cell* 2018, **172**:1168-1172.

3. Temmam S, Monteil-Bouchard S, Robert C, Pascalis H, Michelle C, Jardot P, Charrel R, Raoult D, Desnues C: **Host-associated metagenomics: a guide to generating infectious RNA viromes**. *PLoS One* 2015, **10**:e0139810.

4. Coutinho FH, Cabello-Yeves PJ, Gonzalez-Serrano R, Rosselli R, López-Pérez M, Zemskaya TI, Zakharenko AS, Ivanov VG, Rodriguez-Valera F: **New viral biogeochemical roles revealed through metagenomic analysis of Lake Baikal**. *Microbiome* 2020, **8**:1-15.

5. Weitz JS, Wilhelm SW: **Ocean viruses and their effects on microbial communities and biogeochemical cycles**. *F1000 Biol Rep* 2012, **4**:17.

6. Federici S, Nobs SP, Elinav E: **Phages and their potential to modulate the microbiome and immunity**. *Cell Mol Immunol* 2021, **18**:889-904.

7. Khan Mirzaei M, Deng L: **New technologies for developing phage-based tools to manipulate the human microbiome**. *Trends Microbiol* 2021, **30**:131-142, https://doi.org/10.1016/J.TIM.2021.04.007

8. Mohammadali Khan Mirzaei A, Anik Ashfaq Khan M, Ghosh P, Deng L: **Bacteriophages isolated from stunted children can regulate gut bacterial communities in an age-specific manner**. *Cell Host Microbe* 2020, **27**:199-212.e5, https://doi.org/10.1016/j.chom.2020.01.004

9. Ma T, Ru J, Xue J, Schulz S, Mirzaei MK, Janssen KP, Quante M, Deng L: **Differences in gut virome related to Barrett esophagus and esophageal adenocarcinoma**. *Microorganisms* 2021, **9**:1701.

10. Plyusnin I, Kant R, Jaä¨skelaïnen AJ, Sironen T, Holm L, Vapalahti O, Smura T: **Novel NGS pipeline for virus discovery from a wide spectrum of hosts and sample types**. *Virus Evol* 2020, **6**:veaa091.

11. Conceição-Neto N, Yinda KC, van Ranst M, Matthijnssens J: **NetoVIR: modular approach to customize sample preparation procedures for viral metagenomics**. Methods in Molecular Biology. Humana Press Inc; 2018:85-95.

12. Sicard A, Michalakis Y, Gutiérrez S, Blanc S: **The strange lifestyle of multipartite viruses**. *PLoS Pathog* 2016, **12**:e1005819.

13. Steward GF, Culley AI, Mueller JA, Wood-Charlson EM, Belcaid M, Poisson G: **Are we missing half of the viruses in the ocean?** *ISME J* 2013, **7**:672-679.

14. Wolf YI, Silas S, Wang Y, Wu S, Bocek M, Kazlauskas D, Krupovic M, Fire A, Dolja VV, Koonin EV: **Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome**. *Nat Microbiol* 2020, **5**:1262-1270.

15. Carding SR, Davis N, Hoyles L: **The human intestinal virome in health and disease**. *Aliment Pharmacol Ther* 2017, **46**:800-815.

16. Mushegian A, Shipunov A, Elena SF: **Changes in the composition of the RNA virome mark evolutionary transitions in green plants**. *BMC Biol* 2016, **14**:1-14.

17. Wolf YI, Kazlauskas D, Iranzo J, Lucía-Sanz A, Kuhn JH, Krupovic M, Dolja VV, Koonin EV: **Origins and evolution of the global RNA virome**. *mBio* 2018, **9**:e02329-18.

18. Bimboim HC, Doly J: **A rapid alkaline extraction procedure for screening recombinant plasmid DNA**. *Nucleic Acids Res* 1979, **7**:1513.

19. Anisimova VE, Barsova EV, Bogdanova EA, Lukyanov SA, Shcheglov AS: **Thermolabile duplex-specific nuclease**. *Biotechnol Lett* 2009, **31**:251-257.

20. Conceição Neto N, Zeller M, Lefrère H, de Bruyn P, Beller L, Deboutte W, Yinda CK, Lavigne R, Maes P, van Ranst M, *et al*.: **NetoVIR: a reproducible protocol for virome analysis**. *Protocol Exch* 2016, **1838**:85-95, https://doi.org/10.1038/protex.2016.029

21. Callanan J, Stockdale SR, Shkoporov A, Draper LA, Ross RP, Hill C: **Expansion of known ssRNA phage genomes: from tens to over a thousand**. *Sci Adv* 2020, **6**:eaay5981.

22. Starr EP, Nuccio EE, Pett-Ridge J, Banfield JF, Firestone MK: **Metatranscriptomic reconstruction reveals RNA viruses with the potential to shape carbon cycling in soil**. *Proc Natl Acad Sci USA* 2019, **116**:25900-25908.

23. Wu H, Pang R, Cheng T, Xue L, Zeng H, Lei T, Chen M, Wu S, Ding Y, Zhang J, *et al*.: **Abundant and diverse RNA viruses in insects revealed by RNA-Seq analysis: ecological and evolutionary implications**. *mSystems* 2020, **5**:e00039-20.

24. • Wongsurawat T, Jenjaroenpun P, Taylor MK, Lee J, Tolardo AL, Parvathareddy J, Kandel S, Wadley TD, Kaewnapan B, Athipanyasilp N, *et al*.: **Rapid sequencing of multiple RNA viruses in their native form**. *Front Microbiol* 2019, **10**:68-72.

Direct RNA seq is used for detection and identification of RNA viral pathogens.

25. Leigh DM, Schefer C, Cornejo C: **Determining the suitability of MinION's direct RNA and DNA amplicon sequencing for viral subtype identification**. *Viruses* 2020, **12**:801.

26. Wang Z, Neupane A, Feng J, Pedersen C, Marzano SYL: **Direct metatranscriptomic survey of the sunflower microbiome and virome**. *Viruses* 2021, **13**:1867.

27. Parras-Moltó M, López-Bueno A: **Methods for enrichment and sequencing of oral viral assemblages: saliva, oral mucosa, and dental plaque viromes**. *Methods Mol Biol* 2018, **1838**:143-161.

28. Regnault B, Bigot T, Ma L, Pérot P, Temmam S, Eloit M: **Deep impact of random amplification and library construction methods on viral metagenomics results**. *Viruses* 2021, **13**:253.

29. Long N, Qiao Y, Xu Z, Tu J, Lu Z: **Recent advances and application in whole-genome multiple displacement amplification**. *Quant Biol* 2020, **8**:279-294.

30. Deleye L, Tilleman L, van der Plaetsen AS, Cornelis S, Deforce D, van Nieuwerburgh F: **Performance of four modern whole genome amplification methods for copy number variant detection in single cells**. *Sci Rep* 2017, **7**:1-9.

31. Sabina J, Leamon JH: **Bias in whole genome amplification: causes and considerations**. *Methods Mol Biol* 2015, **1347**:15-41.

32. Huang L, Ma F, Chapman A, Lu S, Xie XS: **Single-cell whole-genome amplification and sequencing: methodology and applications**. *Annu Rev Genom Hum Genet* 2015, **16**:79-102.

33. Gawad C, Koh W, Quake SR: **Single-cell genome sequencing: current state of the science**. *Nat Rev Genet* 2016, **17**:175-188.

34. Blagodatskikh KA, Kramarov VM, Barsova EV, Garkovenko A v, Shcherbo DS, Shelenkov AA, Ustinova VV, Tokarenko MR, Baker SC, Kramarova TV, *et al*.: **Improved DOP-PCR (iDOP-PCR): A robust and simple WGA method for efficient amplification of low copy number genomic DNA**. *PLoS One* 2017, **12**:e0184507.

35. Stepanauskas R, Fergusson EA, Brown J, Poulton NJ, Tupper B, Labonté JM, Becraft ED, Brown JM, Pachiadaki MG, Povilaitis T, *et al*.: **Improved genome recovery and integrated cell-size analyses of individual uncultured microbial cells and viral particles**. *Nat Commun* 2017, **8**:1-10.

36. Chen C, Xing D, Tan L, Li H, Zhou G, Huang L, Xie XS: **Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI)**. *Science* 2017, **356**:189-194.

37. Tomlins SA, Mehra R, Rhodes DR, Shah RB, Rubin MA, Bruening E, Makarov V, Chinnaiyan AM: **Whole transcriptome amplification for gene expression profiling and development of molecular archives**. *Neoplasia* 2006, **8**:153-162.

38. • Zhang Y, Gao J, Huang Y, Wang J: **Recent developments in single-cell RNA-Seq of microorganisms**. *Biophys J* 2018, **115**:173-180.

Used both long-read and short-read sequencing in combination to assemble the marine virome.

39. Chrzastek K, Lee DH, Smith D, Sharma P, Suarez DL, Pantin-Jackwood M, Kapczynski DR: **Use of Sequence-Independent, Single-Primer-Amplification (SISPA) for rapid detection, identification, and characterization of avian RNA viruses**. *Virology* 2017, **509**:159-166.

40. • Logsdon GA, Vollger MR, Eichler EE: **Long-read human genome sequencing and its applications**. *Nat Rev Genet* 2020, **21**:597-614.

Newly developed linked-read sequencing method which can be performed in a single tube without the need for any complicated equipment.

41. Kuleshov V, Jiang C, Zhou W, Jahanbani F, Batzoglou S, Snyder M: **Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome**. *Nat Biotechnol* 2016, **34**:64-69.

42. Warwick-Dugdale J, Solonenko N, Moore K, Chittick L, Gregory AC, Allen MJ, Sullivan MB, Temperton B: **Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands**. *PeerJ* 2019, **7**:e6800.

43. Beaulaurier J, Luo E, Eppley JM, Uyl P, den, Dai X, Burger A, Turner DJ, Pendelton M, Juul S, Harrington E, *et al*.: **Assembly-free single-molecule sequencing recovers complete virus genomes from natural microbial communities**. *Genome Res* 2020, **30**:437-446.

44. Jain M: **From kilobases to "whales": a Short History of Ultra-long Reads and High-throughput Genome Sequencing**; 2021.

45. Bickhart DM, Watson M, Koren S, Panke-Buisse K, Cersosimo LM, Press MO, van Tassell CP, van Kessel JAS, Haley BJ, Kim SW, *et al*.: **Assignment of virus and antimicrobial resistance genes to microbial hosts in a complex microbial community by combined long-read assembly and proximity ligation**. *Genome Biol* 2019, **20**:153.

46. Cao J, Zhang Y, Dai M, Xu J, Chen L, Zhang F, Zhao N, Wang J: **Profiling of human gut virome with oxford nanopore technology**. *Med Microecol* 2020, **4**:100012.

47. Volden R, Palmer T, Byrne A, Cole C, Schmitz RJ, Green RE, Vollmers C: **Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA**. *Proc Natl Acad Sci USA* 2018, **115**:9726-9731.

48. Chen Y, Nie F, Xie SQ, Zheng YF, Dai Q, Bray T, Wang YX, Xing JF, Huang ZJ, Wang DP, *et al*.: **Efficient assembly of nanopore reads via highly accurate and intact error correction**. *Nat Commun* 2021, **12**:1-10.

49. Chen Z, Pham L, Wu TC, Mo G, Xia Y, Chan PL, Porter D, Phan T, Che H, Tran H, *et al*.: **Ultralow-input single-tube linked-read library method enables short-read second-generation sequencing systems to routinely generate highly accurate and economical long-range sequencing information**. *Genome Res* 2020, **30**:898-909.

50. Zheng GXY, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, Kyriazopoulou-Panagiotopoulou S, Masquelier DA, Merrill L, Terry JM, *et al*.: **Haplotyping germline and cancer genomes using high-throughput linked-read sequencing**. *Nat Biotechnol* 2016, **34**:303.

51. Zhang F, Christiansen L, Thomas J, Pokholok D, Jackson R, Morrell N, Zhao Y, Wiley M, Welch E, Jaeger E, *et al*.: **Haplotype phasing of whole human genomes using bead-based barcode partitioning in a single tube**. *Nat Biotechnol* 2017, **35**:852-857.

52. Wang O, Chin R, Cheng X, Wu M, Mao Q, Tang J, Sun Y, Anderson E, Lam HK, Chen D, Zhou Y, Wang L, Fan F, Zou Y, Xie Y, Zhang RY, Drmanac S, Nguyen D, Xu C, Villarosa C, Gablenz S, Barua N, Nguyen S, Tian W, Liu JS, Wang J, Liu X, Qi X, Chen A, Wang H, Dong Y, Zhang W, Alexeev A, Yang H, Wang J, Kristiansen K, Xu X, Drmanac R, Peters BA: **Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing,**

haplotyping, and de novo assembly. *Genome Res* 2019, **29**:798-808, https://doi.org/10.1101/gr.245126.118

53. Belton J-M, Mccord RP, Gibcus J, Naumova N, Zhan Y, Dekker J: **Hi-C: a comprehensive technique to capture the conformation of genomes**. *Methods* 2012, **58**:268-276.

54. Marbouty M, Cournac A, Flot JF, Marie-Nelly H, Mozziconacci J, Koszul R: **Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms**. *Elife* 2014, **3**:e03318.

55. Burton JN, Liachko I, Dunham MJ, Shendure J: **Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps**. *G3: Genes Genomes Genet* 2014, **4**:1339-1346.

56. Beitel CW, Froenicke L, Lang JM, Korf IF, Michelmore RW, Eisen JA, Darling AE: **Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products**. *PeerJ* 2014, **2**:e415.

57. Foutel-Rodier T, Thierry A, Koszul R, Marbouty M: **Generation of a metagenomics proximity ligation 3C library of a mammalian gut microbiota**. Methods in Enzymology. Academic Press; 2018:183-195.

58. Bickhart DM, Kolmogorov M, Tseng E, Portik DM, Korobeynikov A, Tolstoganov I, Uritskiy G, Liachko I, Sullivan ST, Shin SB, *et al.*: **Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities**. *Nat Biotechnol* 2022, **40**:711-719, https://doi.org/10.1038/s41587-021-01130-z

59. • Yuan Y, Chung CYL, Chan TF: **Advances in optical mapping for genomic research**. *Comput Struct Biotechnol J* 2020, **18**:2051-2062.

Adsorp-Seq: a novel method identifying phage-host pairs using agarose gels.

60. Bouwens A, Deen J, Vitale R, D'huys L, Goyvaerts V, Descloux A, Borrenberghs D, Grussmayer K, Lukes T, Camacho R, *et al.*: **Identifying microbial species by single-molecule DNA optical mapping and resampling statistics**. *NAR Genom Bioinform* 2020, **2**:lqz007.

61. Abid HZ, Young E, McCaffrey J, Raseley K, Varapula D, Wang HY, Piazza D, Mell J, Xiao M: **Customized optical mapping by CRISPR–Cas9 mediated DNA labeling with multiple sgRNAs**. *Nucleic Acids Res* 2021, **49**:e8.

62. • Bogas D, Nyberg L, Pacheco R, Azevedo NF, Beech JP, Gomila M, Lalucat J, Manaia CM, Nunes OC, Tegenfeldt JO, *et al.*: **Applications of optical DNA mapping in microbiology**. *Biotechniques* 2017, **62**:255-267.

Use of the polony or PCR colony method to link phages to their hosts.

63. Somerville V, Lutz S, Schmid M, Frei D, Moser A, Irmler S, Frey JE, Ahrens CH: **Long-read based de novo assembly of low-complexity metagenome samples results in finished genomes and reveals insights into strain diversity and an active phage system**. *BMC Microbiol* 2019, **19**:143.

64. • de Jonge PA, von Meijenfeldt FAB, Costa AR, Nobrega FL, Brouns SJJ, Dutilh BE: **Adsorption sequencing as a rapid method to link environmental bacteriophages to hosts**. *iScience* 2020, **23**:101439.

A new method linking phages to hosts based on the physical proximity of viral mRNA transcripts and host rRNA.

65. Tadmor AD, Ottesen EA, Leadbetter JR, Phillips R: **Probing individual environmental bacteria for viruses by using microfluidic digital PCR**. *Science* 2011, **333**:58.

66. Mruwat N, Carlson MCG, Goldin S, Ribalet F, Kirzner S, Hulata Y, Beckett SJ, Shitrit D, Weitz JS, Armbrust EV, *et al.*: **A single-cell polony method reveals low levels of infected Prochlorococcus in oligotrophic waters despite high cyanophage abundances**. *ISME J* 2020, **15**:41-54.

67. • Sakowski EG, Arora-Williams K, Tian F, Zayed AA, Zablocki O, Sullivan MB, Preheim SP: **Interaction dynamics and virus–host range for estuarine actinophages captured by epicPCR**. *Nat Microbiol* 2021, **6**:630-642.

PCR-based method for linking phages to hosts by fusing host and viral genes within an emulsion droplet.

68. Morella NM, Yang SC, Hernandez CA, Koskella B: **Rapid quantification of bacteriophages and their bacterial hosts in vitro and in vivo using droplet digital PCR**. *J Virol Methods* 2018, **259**:18-24.

69. Allers E, Moraru C, Duhaime MB, Beneze E, Solonenko N, Barrero-Canosa J, Amann R, Sullivan MB: **Single-cell and population level viral infection dynamics revealed by phageFISH, a method to visualize intracellular and free viruses**. *Environ Microbiol* 2013, **15**:2306-2318.

70. Kumar A, Murthy S, Kapoor A: **Evolution of selective-sequencing approaches for virus discovery and virome analysis**. *Virus Res* 2017, **239**:172-179, https://doi.org/10.1016/j.virusres.2017.06.005

71. 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**. *Sci Adv* 2017, **3**:e1602105.

72. Cesar Ignacio-Espinoza J, Laperriere SM, Yeh Y-C, Weissman J, Hou S, Long M, Fuhrman JA: **Ribosome-linked mRNA-rRNA chimeras reveal active novel virus host associations**. *bioRxiv* 2020, **1**:332502, https://doi.org/10.1101/2020.10.30.332502

73. Deng L, Ignacio-Espinoza JC, Gregory AC, Poulos BT, Weitz JS, Hugenholtz P, Sullivan MB: **Viral tagging reveals discrete populations in Synechococcus viral genome sequence space**. *Nature* 2014, **513**:242-245.

74. Džunková M, D'Auria G, Moya A: **Direct sequencing of human gut virome fractions obtained by flow cytometry**. *Front Microbiol* 2015, **6**:955.

75. Breitbart M: **Marine viruses: truth or dare**. *Ann Rev Mar Sci* 2012, **4**:425-448.

76. Zolfo M, Pinto F, Asnicar F, Manghi P, Tett A, Bushman FD, Segata N: **Detecting contamination in viromes using ViromeQC**. *Nat Biotechnol* 2019, **37**:1408-1412.

77. Milani C, Casey E, Lugli GA, Moore R, Kaczorowska J, Feehily C, Mangifesta M, Mancabelli L, Duranti S, Turroni F, *et al.*: **Tracing mother-infant transmission of bacteriophages by means of a novel analytical tool for shotgun metagenomic datasets: METAnnotatorX**. *Microbiome* 2018, **6**:1-16.

78. Roux S, Enault F, Hurwitz BL, Sullivan MB: **VirSorter: mining viral signal from microbial genomic data**. *PeerJ* 2015, **2015**:e985.