Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Virology

Detecting viral sequences in NGS data

Paul G Cantalupo and James M Pipas

Next generation sequencing (NGS) technologies provide an increasingly important avenue for detecting known viruses, and for discovering novel viruses present in clinical or environmental samples. Several computational pipelines capable of identifying and classifying viral sequences in NGS data have been developed and used to search for viruses in human or animal samples, microbiomes, and in various environments. In this review we summarize the different approaches used to determine viral presence in sequence data. Strategies for avoiding confounding factors such as physical contamination and computational artifacts that lead to false virus identification are discussed. The application of these methodologies to cancer data sets has led to important insights on viruses both as drivers of and biomarkers for specific tumors.

Address
Department of Biological Sciences, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA

Corresponding author: Pipas, James M (pipas@pitt.edu)

Current Opinion in Virology 2019, **39**:41–48

This review comes from a themed issue on **Viruses and cancer**

Edited by **Marta M Gaglia** and **Karl Munger**

For a complete overview see the Issue and the Editorial

Available online 26th August 2019

https://doi.org/10.1016/j.coviro.2019.07.010

Introduction

The advent of cost-effective, high throughput nucleic acid sequencing technologies has enabled new approaches for the detection, identification, and characterization of known and novel viruses. In the past, the ability to detect and discover viruses was limited by the ability to grow the virus in cell culture. The advent of next generation sequencing (NGS) opened an alternative route for virus discovery that bypassed the need for culture. When the DNA or RNA of a specific target organism is sequenced, microorganisms and viruses present in the sample are sequenced as well. Known viruses, that is, viruses whose sequences have been annotated and placed in public sequence databases, can be detected in this data by nucleotide sequence alignment. Novel viruses can also be detected, either by relaxing the nucleotide alignment parameters, or by assembling the sequence reads and aligning open reading frames (ORFs) with protein sequence databases.

Several excellent articles comparing the numerous computational approaches and pipelines available for virus detection from NGS data exist [1••,2,3]. In this brief review focusing on eukaryotic virus discovery, we discuss the power and pitfalls of interpreting data generated by these pipelines, emphasizing the need for intense manual inspection to minimize artifacts and contamination that lead to false positives. Viruses have been established as drivers of several types of human cancer and there is much interest in using computational strategies to uncover additional oncogenic viruses. This article does not deal with the experimental evidence needed to establish a causative role for a virus in cancer. Rather we focus on the methods and data that establish the presence of viruses.

Why look for viruses?

Viruses are extremely diverse agents. Currently the International Committee on Taxonomy of Viruses defines 5560 virus species across 150 families. Viral particles come in a variety of structural forms and sizes, ranging from roughly spherical to linear, to amorphic. Some are comprised of only protein and nucleic acid, some have lipid envelopes, and some partition their genomes among different particles. Furthermore, viral nucleic acids are chemically and topologically diverse. They may consist of DNA or RNA, be single or double stranded, contain covalently linked proteins or chemically modified termini, and exist as linear or circular forms. The entire genome can reside on a single molecule of nucleic acid or can be distributed among separate segments. Finally, there is the computational challenge of identifying diverse members of a biological entity that has no commonly conserved genes such as 16S ribosomal RNA (rRNA) in bacteria. This remarkable diversity makes it very difficult to design a single virion enrichment or nucleic acid purification protocol that captures all viral sequences.

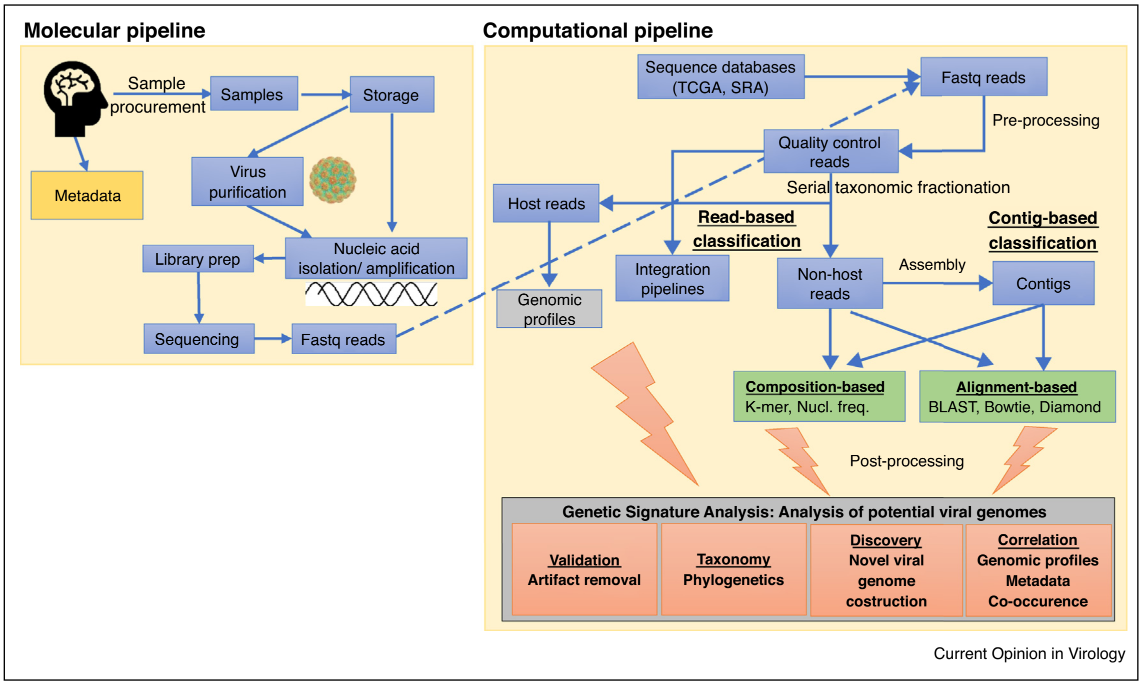
A major reason for identifying viruses is to associate them with disease. In these cases, the samples consist of pathological tissues and control healthy tissues. Once a specific virus has been identified in diseased tissue, a series of additional studies are required to determine if it is the causal agent for the disease, or a passenger, that is, the virus simply grows or persists better in diseased versus healthy tissue, or a contaminant. Other studies seek to define the microbiome, including the virome, of healthy or diseased tissues. Still other studies utilize metagenomic approaches of environmental samples such as water, soil, or air. Finally, some strategies seek to uncover previously undescribed viruses that are very distantly related to those we currently know, or that represent entirely novel types. In fact, a large proportion of metagenomic sequences do not match sequences in public databases (the so-called 'dark matter') [4,5]. Each of these goals presents their own unique challenges in the development and application of virus detection pipelines.

Generating NGS data

All virus detection strategies start with sample collection and nucleic acid isolation. These initial steps are the major determining factors as to the types of viruses that can be detected and the types that will be excluded. The first decision is whether the NGS data will be from DNA sequencing, RNA sequencing, or both. DNA sequence data analysis detects both the virion-associated and replicating genomes of DNA viruses, but most viruses with RNA genomes will be excluded. The exception is retroviruses which convert their RNA genomes into double stranded DNAs using the viral reverse transcriptase. RNA sequence data detect the virion-associated RNA and replicative RNAs and mRNAs of viruses with RNA genomes, as well as mRNAs expressed by DNA viruses. Viruses with DNA genomes that are not being actively transcribed will be excluded. Therefore, the most accurate and informative data are derived from sequencing both DNA and RNA from the same sample. For example, a study of viral sequences in The Cancer Genome Atlas (TCGA) DNA-seq and RNA-seq databases detected Hepatitis C virus (HCV) in liver cancer RNA-seq data, but not DNA-seq (as expected since there is no DNA form of the virus genome), while Hepatitis B virus was detected in both DNA-seq and RNA-seq data [6••].

The first steps in generating NGS data are sample selection, procurement, and storage, coupled with the collection of relevant metadata (Figure 1). Each of these steps are fraught with the issues of potential contamination and/or nucleic acid degradation, both of which can compromise downstream analyses (Table 1). For example, high-quality RNA-seq analysis of cancers requires minimization of the time between resection and storage in liquid nitrogen. Furthermore, samples are usually stored in media, buffers, and glycerol. All these reagents are potentially contaminated with viral sequences that will be detected by downstream pipelines [7•].

Figure 1



Current Opinion in Virology

The workflow of a viral metagenomic experiment.

Viral metagenomics is composed of two major parts: a molecular (left) and computational pipeline (right). The molecular pipeline ultimately results in a set of fastq sequencing reads that become the input for the computational pipeline (dashed arrow). Fastq reads can be obtained from databases such as TCGA and SRA. Fastq reads are pre-processed to remove poor quality sequence information. Then reads are processed in various ways depending on the specific goals of the experiment. During post-processing, the results are scrutinized with Genetic Signature Analysis to remove artifacts and confirm the presence of viral genomes [61••]. Sample metadata are used during post-processing to identify potential correlates of virus presence or expression.

Table 1

Pitfalls and contamination in a viral metagenomics study

Step	Pitfalls and Contamination
Sample collection	Viral contamination as a ‘passenger’
Sample storage	Virus degradation. Microorganism growth (freezer failure)
Virus purification	Loss of viruses that evade purification
Nucleic acid purification	Contamination from laboratory reagents Could exclude specific families of viruses (i.e. DNAseq would exclude RNA viruses) Amplification bias can skew resulting virome profile Loss of low abundance transcripts.
Sequencing	Machine contamination and index switching Sequencing errors in reads may lead to false viral identification or variants.
Preprocessing	Low quality reads missing cutoff may over estimate viral diversity Non-template sequence failed to be removed
Serial Taxonomic Fractionation	Loss of valid viral reads and reduced sensitivity of detection
Assembly	Fragmented assembly or no assembly of low abundance viruses. Requires enough coverage to assemble full genome Potential chimera formation due to high intrapopulation diversity
Alignment-based classification	Slow, resource intensive, more sensitive and novel virus detection Relatively few viral genomes in databases renders classification difficult. Requires maintaining up to date indexes of sequence databases Incorrect database annotations
Composition-based classification	Fast, less sensitive and unlikely to identify divergent sequences Does not allow for species/strain level identification
Software usage in general	Requires understanding of the algorithm used and its assumptions Need to determine similarity thresholds and values for other parameters
Genetic Signature Analysis	The largest time sink in the entire workflow. Requires time consuming manual curation to identify false positives

There are two broad strategies for detecting viruses in samples. The first is to isolate and sequence nucleic acid from the entire sample, and then rely on computational approaches to detect viral sequences [8]. This approach can detect a broad array of viral classes but requires very deep sequencing and thus can be quite expensive. The second is to further process the sample in order to enrich for virions. There are several molecular methods designed to obtain virions from tissue or environmental sources with most ending in nuclease treatment to eliminate exogenous nucleic acids while preserving viral genomes that are encapsulated within virions [9•]. At present, no single method enriches for all virion types. Thus, this approach works best if the characteristics of the subject viruses are known so that enrichment procedures can be tailored to virion properties.

The various nucleic acid isolation methods also favor or disfavor certain classes of viruses. In fact, every nucleic acid extraction method will favor some types of viral genomes and exclude others. For example, typical methods for preparing cellular DNA favor large DNA fragments and these methods are inefficient for isolating small, circular DNAs. Whole exon sequencing is even more selective since it disfavors small viral genomes. Similarly, RNA extraction is generally followed by the application of one of two methods for eliminating rRNA. The first is to enrich for polyadenylated RNAs. However, many important types of viruses do not modify the 3′-terminus of viral RNAs with poly(A), and thus this class is excluded from poly(A) selected RNA preparations. The second approach is to deplete RNA preparations of rRNAs using bead-based capture probes [10].

Some virus detection strategies are designed to enrich for viral sequences by including a PCR amplification step. Generally, the PCR primers are designed to detect a specific type or family of viruses. However, recently targeted sequence capture platforms designed to detect a broad array of virus types have been introduced [11,12] as well as microfluidic platforms that purify viral species from a complex mixture [13]. Finally, the nucleic acid isolation and subsequent library preparations for sequencing are major sources of contamination. Each reagent potentially harbors many viruses, either as remnants of vectors used in reagent preparation, or as contaminants inadvertently introduced by investigators [7•,14].

Analyzing NGS data

Since viruses do not contain universally conserved genomic sequences, like 16S rRNA of bacteria, that can be used to establish taxonomy, all metagenomic sequences need to be compared to existing viral sequences for identification. However, viruses are underrepresented in public databases since millions of extant viruses have not been characterized. Sequences derived from these divergent viruses are often difficult to classify because either they align very weakly to a known virus or have no match to any virus in current databases [4,5].

Computational pipelines

Here we aim to provide a general overview of the types of analyses that are used in viral metagenomics and some specific information at each step of a metagenomics pipeline. Processing of NGS data involves a series of steps including 1) preprocessing/quality control, 2) filtering sequences, 3) assembly, 4) taxonomic identification, and 5) validation and analysis (Figure 1). There are many available tools and pipelines that have been developed to perform some or all these steps [15–19]. The diversity of pipelines and bioinformatic methods highlights a lack of consensus in the field on how best to analyze viromes. In addition, several pipelines that detect viral-host integration junctions have been reported [20–22]. Part of the reason for this is that each tool was developed for a specific application whether it be for novel viral discovery, understanding viral ecology or for a diagnostic medical application. It is important to note that virus detection/discovery pipelines do not really discover viruses. They detect sequences in NGS data that are related to known viruses. Several notable reviews have tested and compared many of the computational pipelines [1••,2,3] and comprehensive reviews on viral metagenomics exist [4,23,24,25•,26].

Raw reads, preprocessing, quality control

After performing sequencing, the raw reads must be preprocessed since many reads are of low quality or have low quality ends. FastQC is a popular tool for analyzing the quality of the raw reads [27]. The results obtained inform as to how the raw reads should be processed. For example, since sequencing facilities may or may not remove library adapters from the data, the reads may contain non-template sequences from adapters, barcodes or primers. Cutadapt can be used to search for and trim off non-template sequences that are provided by the user [28]. TagCleaner automatically identifies common sequences at the ends of the reads that may represent non-template sequence [29]. This is especially useful when working with publicly available NGS data. A popular program to remove low quality reads, trim low-quality ends, perform deduplication and remove low entropy reads (nucleotide repeats) is Prinseq [30].

Serial taxonomic fractionation (STF)

Next, each sequence read is associated with a specific organism by sequence alignment. Viral sequences are often at very low abundance, since in most instances, the vast majority of reads are derived from the host species from which the virus was obtained. In order to reduce computational time, many pipelines first identify host reads and remove them from further analysis, a process that has been called host subtraction or filtering. This greatly reduces the number of reads that need to be matched to viral databases. However, in many applications, including tumor virology, analysis of the host reads provide valuable information for downstream analyses such as associating the presence of a virus with specific gene expression signatures or mutation profiles [6••,31••]. Thus, we prefer the term ‘serial taxonomic fractionation’ by analogy with column chromatography in biochemistry. STF allows the successive separation of reads into taxonomic groups such as human, fungal, or bacterial. The taxonomy level depends on the selected databases to be used. STF provides for 1) determining the genomic profiles of host reads (i.e. transcript levels or mutations), 2) improving the speed of downstream pipeline steps, and 3) improving the accuracy of viral taxonomic assignments.

For example, when analyzing human cancer samples, the set of quality reads will be dominated by human sequences. The number of human reads will be dependent on the type of sample preparation. Whole genome DNA sequencing and total RNA sequencing generate a very high number of reads, while enriched virion preparations generate a relatively low number of reads. Several studies have aggressively filtered human reads [6••,32,33]. One reason is that there are known gaps in the human reference genome [34] and performing more filtering steps improves the chances to identify a read as human. Additional filtering steps can be introduced depending on the goals of the experiment and the samples being studied. For example, sequences can be filtered against vector (VectorDB NCBI), rRNA, and bacterial sequences. However, even modest filtering may remove non-human sequences, thereby decreasing sensitivity of virus detection. These reads can be recovered by aligning them to non-human databases such as viruses and bacteria.

Assembly

Assembly of complete viral genomes from a complex mixture of sequences from many organisms is challenging. Assembly programs are written to assemble sequences from one organism and assume even coverage. They flag regions of uneven higher coverage as repeats instead of considering the possibility that more coverage might represent a highly abundant virus [35]. Intrapopulation variation between similar species can induce the generation of chimeric contigs. Because of these issues, assembly of viral genomes may result in artificial chimeras or several contigs that come from a single genome. Given these complexities, it is remarkable that complete viral genome assemblies from a metagenome can be achieved.

Assembly can be performed in several different ways. Reference-based assembly requires the use of a reference genome for read alignment. *De novo* assembly attempts to determine the connections between each read to all others. Finally, a combination of these approaches increases the accuracy of assembly [15,36].

De novo assembly of all reads can generate incorrect assemblies, for instance when two similar organisms are present in the sample. This can be mitigated by only assembling reads that belong to a particular taxonomic rank. Assemblers use different algorithms such as the overlap-layout-consensus (OLC) or a de Bruijn graph (DBG) method [37,38]. It has been reported that integrating both types of assemblers improves the accuracy and length of the contigs [39•]. Several studies discussing how choice of assemblers affects virome characterization have been published [35,37,40].

Most assembled contigs are fragments of complete virus genomes. Increasing the accuracy and the length of genome assemblies is crucial for two primary reasons: 1) increases the chance of obtaining a significant alignment when performing taxonomic identification (see below) and 2) increases the efficiency of manual curation and analysis of the organism. Alignment-free similarity methods exist to help close the gaps in genome assemblies. K-mer profiling of non-overlapping sequences from the same genome has been shown to have a similar k-mer frequency signature and could be used to distinguish between different viruses [41,42] as well as coverage profiling [43].

Taxonomic identification

The goal of viral metagenomics is to determine the taxonomic makeup of a sample or collection of samples. This requires matching each read or contig to known virus sequences. There are several ways this is performed: 1) Alignment-based algorithms that perform nucleotide or protein alignments like BLAST, Bowtie or fast protein search tools like Diamond or RapSearch [44–47]. Most of these tools are relatively slow but are highly sensitive to detect distant sequence relationships; 2) Composition-based algorithms that do not rely on similarity matching such as exact k-mer matching and nucleotide frequencies [48,49]. These approaches are rapid but suffer from identifying distant relationships; and 3) Probabilistic methods that use hidden Markov models of protein domains such as HMMER to identify remote homology [50].

After obtaining results from the above analyses, a decision is made to classify each sequence by setting cutoffs on search scores such as E-value [25•]. Then, the sequence is classified by simply selecting the top search hit. More thorough classification can be performed by utilizing all the significant search results. For example, the lowest common ancestor (LCA) taxon can be determined using MEGAN [51] or Bayesian mixture models can be used [52–54]. Additionally, phylogenetics should be used if the goal is to determine if a virus genome is part of known species, genera or family [55].

Determining viral presence

When does NGS data detect a virus, and when is viral presence biologically significant?

Detecting a virus in both DNA-seq and RNA-seq data from the same sample and with the same polymorphisms would provide the most confidence that a virus is detected in a sample. This is only possible for viruses with a DNA genome or DNA intermediate genomes. With sequencing data, one read alignment to a virus may reveal the presence of a virus; therefore, setting a cutoff for the number of read alignment for viruses is problematic. This is because viral infections may be localized to a specific region of a tissue, and/or only a subset of the susceptible cells may be infected at a given time. Thus, there are many cases where there is a low number of virions or infected cells present. In these cases, even a single viral read may indicate infection. In a clinical setting this would demand follow-up studies such as PCR.

Identifying the presence of biologically relevant viruses requires the elimination of contaminants and computational artifacts. Then the remaining viruses are assumed to be residents of the sample. Several studies have identified ubiquitous reagent or laboratory contaminants responsible for false positive identifications [33,56–58]. Detection of a laboratory virus can be determined by comparing the single nucleotide variants in the lab strain to the one detected in the sequencing run. Other common viral contaminants include PhiX174 (used as a control for NGS) and Simian Virus 40, whose many genomic elements exist in a wide array of plasmids [6••].

Another source of contamination occurs on the sequencing machine itself. There are two types that have been described: carry-between (within a run) and carry-over (across separate runs) [59,60]. Carry-between can occur across samples during the run by index switching where free indexes are incorrectly used to prime library fragments [60]. Machine contamination of HeLa cell nucleic acid was identified as the most likely culprit for the misidentification of HPV18 in several human cancers [61••]. A potential mitigation strategy would be to amplify template with a primer [25•,62]. Then, any raw reads that do not contain the primer sequence are contaminating reads. Also, alternating barcodes relative to previous sequencing runs could be used but this requires knowledge of the history of the sequencing machine runs.

Finally, computational artifacts can also cause false-positive identifications. Sequencing records in databases such as Genbank sometimes inaccurately annotate the taxonomy of a sequence. This manifests as bacteriophage sequences annotated as bacteria or bacteriophage sequences annotated as a eukaryotic virus [62]. The detection of a granulovirus in human cancer was attributed to alignment of bacterial rRNA reads to the virus

[63]. Another study reported that the RefSeq records for two bunyaviruses, Shamonda and Simbu virus, have stretches of 40–70 bases of perfect or near perfect identity to human rRNA. The false identification was due to the Bowtie aligner not finding a seed length long enough to align the human rRNA reads that contained several inaccurate base-calls at low quality positions [6••].

NGS databases: the gift that keeps on giving

All viral metagenomic studies reveal a vast sea of sequences that do not align to any organism in any of the current databases. It is highly likely that a large portion of this, so called dark matter, represents uncharacterized viruses. As more annotated genomes are added to databases, these 'hidden' viruses will be revealed. This makes analysis of publicly available NGS data more attractive. For example, the Short Read Archive (SRA) database at NCBI, represents a golden opportunity to discover novel viruses. Often the original analysis of NGS data misses the presence of novel viruses either because closely related viruses are not in the database, the specific computational tools miss the relevant viral sequences, or the focus of the study is not aimed at virus detection. As NGS databases rapidly expand, older NGS data may reveal alignments to sequences not present when the sequencing was done. For example, a novel polyomavirus was discovered in a metagenome sample from a Pomona leaf-nosed bat and several novel RNA viruses were discovered in a spider metagenome [64,65].

Viruses and cancer

Surveys of human tumor DNA-seq and RNA-seq data have confirmed the known associations of specific viruses with some cancers. As expected, virtually all cervical carcinomas have been found to harbor HPV, generally integrated into the genomes of tumor cells [66]. Similarly, HPV is found in a subset of head and neck cancers [6••,67,68]. Approximately 20% of stomach cancers contain EBV and HBV or HCV have been detected in a subset of liver cancers [6••,69–71]. Some surprising associations have also been detected such as integrated HPV and Human polyomavirus 1 (BKV) in a small number of bladder cancers [6••,72]. There is strong evidence supporting a driver role for each of these viruses in cancer. However, these surveys also detected passenger viruses in both normal and tumor tissue. It is intriguing to speculate on the properties of some tumors that attract these viruses, for example, an immunosuppressed microenvironment, or a metabolic state that favors virus replication. These are questions ripe for future investigations.

Two forces rapidly converging favor the discovery of additional cancer-associated viruses. One is the explosion in sequencing of human tumor and normal tissues. Much of this is driven by intense studies on the basic molecular biology of tumors, such as those driven by the TCGA. Increasingly, the movement toward personalized cancer therapies demands a significant increase in sequencing efforts. These factors mean that data to be mined for viruses is increasing rapidly. The second force is derived from the increased number of viral metagenomic studies. As more and more viral species, including novel new types of genome organizations, are discovered and annotated, the ability to uncover distant relationships between viruses increases. Thus, as viral sequences are recruited from metagenomic dark matter and placed in annotated databases, the sensitivity of virus search and discovery platforms expands. This is a field ripe for discovery.

Declaration of interest

None.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. We thank Drs. Coyne Drummond, Ping An and Travis Mavrich for their critical comments on the manuscript.

References and recommended reading

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

• of special interest
•• of outstanding interest

1. •• Nooij S, Schmitz D, Vennema H, Kroneman A, Koopmans MPG: **Overview of virus metagenomic classification methods and their biological applications**. *Front Microbiol* 2018, **9**:749

An excellent critical appraisal of 49 published computational workflows for viral metagenomics and provides decision trees for which workflow to choose for specific applications.

2. Tangherlini M, Dell'Anno A, Zeigler Allen L, Riccioni G, Corinaldesi C: **Assessing viral taxonomic composition in benthic marine ecosystems: reliability and efficiency of different bioinformatic tools for viral metagenomic analyses**. *Sci Rep* 2016, **6**:28428.

3. Rose R, Constantinides B, Tapinos A, Robertson DL, Prosperi M: **Challenges in the analysis of viral metagenomes**. *Virus Evol* 2016, **2**:vew022.

4. Mokili JL, Rohwer F, Dutilh BE: **Metagenomics and future perspectives in virus discovery**. *Curr Opin Virol* 2012, **2**:63-77.

5. Reyes A, Semenkovich NP, Whiteson K, Rohwer F, Gordon JI: **Going viral: next-generation sequencing applied to phage populations in the human gut**. *Nat Rev Microbiol* 2012, **10**:607-617.

6. •• Cantalupo PG, Katz JP, Pipas JM: **Viral sequences in human cancer**. *Virology* 2018, **513**:208-216

This paper is a comprehensive survey of viral sequences in multiple human cancers. It proposes that HPV drives tumorigenesis in a small number of bladder cancers and highlights many contamination issues in viral metagenomics.

7. • Asplund M, Kjartansdóttir KR, Mollerup S, Vinner L, Fridholm H, Herrera JAR, Friis-Nielsen J, Hansen TA, Jensen RH, Nielsen IB *et al.*: **Contaminating viral sequences in high-throughput sequencing viromics: a linkage study of 700 sequencing libraries**. *Clini Microbiol Infect* 2019

This paper critically examines laboratory components to identify viral contaminating sequences.

8. Shi M, Zhang YZ, Holmes EC: **Meta-transcriptomics and the evolutionary biology of RNA viruses**. *Virus Res* 2017, **243**:83-90.

9. • Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F: **Laboratory procedures to generate viral metagenomes**. *Nat Protoc* 2009, **4**:470-483

A comprehensive review of methods for purifying different types of viruses and isolating nucleic acid from viruses.

10. Schuierer S, Carbone W, Knehr J, Petitjean V, Fernandez A, Sultan M, Roma G: **A comprehensive assessment of RNA-seq protocols for degraded and low-quantity samples**. *BMC Genomics* 2017, **18**:442.

11. Briese T, Kapoor A, Mishra N, Jain K, Kumar A, Jabado OJ, Lipkin WI: **Virome capture sequencing enables sensitive viral diagnosis and comprehensive virome analysis**. *MBio* 2015, **6**.

12. Wylie TN, Wylie KM, Herter BN, Storch GA: **Enhanced virome sequencing through solution-based capture enrichment**. *Genome Res* 2015, **25**:1910-1920.

13. Han HS, Cantalupo PG, Rotem A, Cockrell SK, Carbonnaux M, Pipas JM, Weitz DA: **Whole-genome sequencing of a single viral species from a highly heterogeneous sample**. *Angew Chem* 2015, **54**:13985-13988.

14. Breitwieser FP, Pertea M, Zimin A, Salzberg SL: **Human contamination in bacterial genomes has created thousands of spurious proteins**. *Genome Res* 2019, **29**:954-960.

15. Ajami NJ, Wong MC, Ross MC, Lloyd RE, Petrosino JF: **Maximal viral information recovery from sequence data using VirMAP**. *Nat Commun* 2018, **9**:3205.

16. Zhao G, Wu G, Lim ES, Droit L, Krishnamurthy S, Barouch DH, Virgin HW, Wang D: **VirusSeeker, a computational pipeline for virus discovery and virome composition analysis**. *Virology* 2017, **503**:21-30.

17. Naccache SN, Federman S, Veeraraghavan N, Zaharia M, Lee D, Samayoa E, Bouquet J, Greninger AL, Luk KC, Enge B *et al.*: **A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples**. *Genome Res* 2014, **24**:1180-1192.

18. Flygare S, Simmon K, Miller C, Qiao Y, Kennedy B, Di Sera T, Graf EH, Tardif KD, Kapusta A, Rynearson S *et al.*: **Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling**. *Genome Biol* 2016, **17**:111.

19. Kostic AD, Ojesina AI, Pedamallu CS, Jung J, Verhaak RG, Getz G, Meyerson M: **PathSeq: software to identify or discover microbes by deep sequencing of human tissue**. *Nat Biotechnol* 2011, **29**:393-396.

20. Katz JP, Pipas JM: **SummonChimera infers integrated viral genomes with nucleotide precision from NGS data**. *BMC Bioinformatics* 2014, **15**:348.

21. Chen X, Kost J, Sulovari A, Wong N, Liang WS, Cao J, Li D: **A virome-wide clonal integration analysis platform for discovering cancer viral etiology**. *Genome Res* 2019, **29**:819-830.

22. Wang Q, Jia P, Zhao Z: **VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data**. *PLoS One* 2013, **8**:e64465.

23. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N: **Shotgun metagenomics, from sampling to analysis**. *Nat Biotechnol* 2017, **35**:833-844.

24. Thomas T, Gilbert J, Meyer F: **Metagenomics - a guide from sampling to data analysis**. *Microb Inform Exp* 2012, **2**:3.

25. Delwart EL: **Viral metagenomics**. *Rev Med Virol* 2007, **17**:115-131
• A comprehensive review, and one of the first, on viral metagenomics.

26. Rosario K, Breitbart M: **Exploring the viral world through metagenomics**. *Curr Opin Virol* 2011, **1**:289-297.

27. Andrews S: *FastQC A Quality Control tool for High Throughput Sequence Data*. 2014 http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

28. Martin M: **Cutadapt removes adapter sequences from high-throughput sequencing reads**. *EMBnet.journal* 2011, **17**:10-12.

29. Schmieder R, Lim YW, Rohwer F, Edwards R: **TagCleaner: identification and removal of tag sequences from genomic and metagenomic datasets**. *BMC Bioinformatics* 2010, **11**:341.

30. Schmieder R, Edwards R: **Quality control and preprocessing of metagenomic datasets**. *Bioinformatics* 2011, **27**:863-864.

31. Johnson ME, Cantalupo PG, Pipas JM: **Identification of head and neck cancer subtypes based on human papillomavirus presence and E2F-regulated gene expression**. *mSphere* 2018, **3**
•• This paper defined head and neck cancer subtypes by associating the presence of a virus with specific gene expression signatures or mutation profiles.

32. Dimon MT, Wood HM, Rabbitts PH, Arron ST: **IMSA: integrated metagenomic sequence analysis for identification of exogenous reads in a host genomic background**. *PLoS One* 2013, **8**:e64546.

33. Yozwiak NL, Skewes-Cox P, Stenglein MD, Balmaseda A, Harris E, DeRisi JL: **Virus identification in unknown tropical febrile illness cases using deep sequencing**. *PLoS Negl Trop Dis* 2012, **6**:e1485.

34. Li R, Li Y, Zheng H, Luo R, Zhu H, Li Q, Qian W, Ren Y, Tian G, Li J *et al.*: **Building the sequence map of the human pan-genome**. *Nat Biotechnol* 2010, **28**:57-63.

35. Smits SL, Bodewes R, Ruiz-Gonzalez A, Baumgartner W, Koopmans MP, Osterhaus AD, Schurch AC: **Assembly of viral genomes from metagenomes**. *Front Microbiol* 2014, **5**:714.

36. Lin YY, Hsieh CH, Chen JH, Lu X, Kao JH, Chen PJ, Chen DS, Wang HY: **De novo assembly of highly polymorphic metagenomic data using in situ generated reference sequences and a novel BLAST-based assembly pipeline**. *BMC Bioinformatics* 2017, **18**:223.

37. White DJ, Wang J, Hall RJ: **Assessing the impact of assemblers on virus detection in a de novo metagenomic analysis pipeline**. *J Comput Biol* 2017, **24**:874-881.

38. Miller JR, Koren S, Sutton G: **Assembly algorithms for next-generation sequencing data**. *Genomics* 2010, **95**:315-327.

39. Deng X, Naccache SN, Ng T, Federman S, Li L, Chiu CY, Delwart EL: **An ensemble strategy that significantly improves de novo assembly of microbial genomes from metagenomic next-generation sequencing data**. *Nucleic Acids Res* 2015, **43**:e46
• This paper thoroughly compares 13 metagenomic assemblers using simulated reads as well as real viral metagenomes.

40. Sutton TDS, Clooney AG, Ryan FJ, Ross RP, Hill C: **Choice of assembly software has a critical impact on virome characterisation**. *Microbiome* 2019, **7**:12.

41. Trifonov V, Rabadan R: **Frequency analysis techniques for identification of viral genetic data**. *MBio* 2010, **1**.

42. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM: **Mash: fast genome and metagenome distance estimation using MinHash**. *Genome Biol* 2016, **17**:132.

43. Smits SL, Bodewes R, Ruiz-Gonzalez A, Baumgartner W, Koopmans MP, Osterhaus AD, Schurch AC: **Recovering full-length viral genomes from metagenomes**. *Front Microbiol* 2015, **6**:1069.

44. Zhao Y, Tang H, Ye Y: **RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data**. *Bioinformatics* 2012, **28**:125-126.

45. Buchfink B, Xie C, Huson DH: **Fast and sensitive protein alignment using DIAMOND**. *Nat Methods* 2015, **12**:59-60.

46. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: **BLAST+: architecture and applications**. *BMC Bioinformatics* 2009, **10**:421.

47. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nat Methods* 2012, **9**:357-359.

48. Wood DE, Salzberg SL: **Kraken: ultrafast metagenomic sequence classification using exact alignments**. *Genome Biol* 2014, **15**:R46.

49. Zielezinski A, Vinga S, Almeida J, Karlowski WM: **Alignment-free sequence comparison: benefits, applications, and tools**. *Genome Biol* 2017, **18**:186.

50. Eddy SR: **A new generation of homology search tools based on probabilistic inference**. *Genome Inform* 2009, **23**:205-211.

51. Huson DH, Auch AF, Qi J, Schuster SC: **MEGAN analysis of metagenomic data**. *Genome Res* 2007, **17**:377-386.

52. Hong C, Manimaran S, Shen Y, Perez-Rogers JF, Byrd AL, Castro-Nallar E, Crandall KA, Johnson WE: **PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples**. *Microbiome* 2014, **2**:33.

53. Prabhakaran S, Rey M, Zagordi O, Beerenwinkel N, Roth V: **HIV haplotype inference using a propagating dirichlet process mixture model**. *IEEE/ACM Trans Comput Biol Bioinform* 2014, **11**:182-191.

54. Morfopoulou S, Plagnol V: **Bayesian mixture analysis for metagenomic community profiling**. *Bioinformatics* 2015, **31**:2930-2938.

55. Roossinck MJ, Martin DP, Roumagnac P: **Plant virus metagenomics: advances in virus discovery**. *Phytopathology* 2015, **105**:716-727.

56. Naccache SN, Greninger AL, Lee D, Coffey LL, Phan T, Rein-Weston A, Aronsohn A, Hackett J Jr, Delwart EL, Chiu CY: **The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns**. *J Virol* 2013, **87**:11966-11977.

57. Grard G, Fair JN, Lee D, Slikas E, Steffen I, Muyembe JJ, Sittler T, Veeraraghavan N, Ruby JG, Wang C *et al.*: **A novel rhabdovirus associated with acute hemorrhagic fever in central Africa**. *PLoS Pathog* 2012, **8**:e1002924.

58. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW: **Reagent and laboratory contamination can critically impact sequence-based microbiome analyses**. *BMC Biol* 2014, **12**:87.

59. Nelson MC, Morrison HG, Benjamino J, Grim SL, Graf J: **Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys**. *PLoS One* 2014, **9**:e94249.

60. Sinha R, Stanley G, Gulati GS, Ezran C, Travaglini KJ, Wei E, Chan CKF, Nabhan AN, Su T, Morganti RM *et al.*: **Index switching causes "spreading-of-signal" among multiplexed samples in Illumina HiSeq 4000 DNA sequencing**. *bioRxiv* 2017:125724.

61. •• Cantalupo PG, Katz JP, Pipas JM: **HeLa nucleic acid contamination in the cancer genome atlas leads to the misidentification of human papillomavirus 18**. *J Virol* 2015, **89**:4051-4057

An important paper that details how contamination, such as HeLa cells, can lead to the misidentification of viral presence in human tumors.

62. Cantalupo PG, Calgua B, Zhao G, Hundesa A, Wier AD, Katz JP, Grabe M, Hendrix RW, Girones R, Wang D, Pipas JM: **Raw sewage harbors diverse viral populations**. *MBio* 2011, **2**.

63. Tang KW, Alaei-Mahabadi B, Samuelsson T, Lindh M, Larsson E: **The landscape of viral expression and host gene fusion and adaptation in human cancer**. *Nat Commun* 2013, **4**:2513.

64. Cantalupo PG, Buck CB, Pipas JM: **Complete genome sequence of a polyomavirus recovered from a pomona leaf-nosed bat (*Hipposideros pomona*) metagenome data set**. *Genome Announc* 2017, **5**.

65. Debat HJ: **An RNA virome associated to the golden orb-weaver spider *Nephila clavipes***. *Front Microbiol* 2017, **8**:2097.

66. Cancer-Genome-Atlas-Research-Network: **Integrated genomic and molecular characterization of cervical cancer**. *Nature* 2017, **543**:378-384.

67. Parfenov M, Pedamallu CS, Gehlenborg N, Freeman SS, Danilova L, Bristow CA, Lee S, Hadjipanayis AG, Ivanova EV, Wilkerson MD *et al.*: **Characterization of HPV and host genome interactions in primary head and neck cancers**. *Proc Natl Acad Sci U S A* 2014, **111**:15544-15549.

68. Cancer-Genome-Atlas-Research-Network: **Comprehensive genomic characterization of head and neck squamous cell carcinomas**. *Nature* 2015, **517**:576-582.

69. Cancer-Genome-Atlas-Research-Network: **Comprehensive molecular characterization of gastric adenocarcinoma**. *Nature* 2014, **513**:202-209.

70. Strong MJ, Xu G, Coco J, Baribault C, Vinay DS, Lacey MR, Strong AL, Lehman TA, Seddon MB, Lin Z *et al.*: **Differences in gastric carcinoma microenvironment stratify according to EBV infection intensity: implications for possible immune adjuvant therapy**. *PLoS Pathog* 2013, **9**:e1003341.

71. The-Cancer-Genome-Atlas-Research-Network: **Comprehensive and integrative genomic characterization of hepatocellular carcinoma**. *Cell* 2017, **169**:1327-1341.e23.

72. Cancer-Genome-Atlas-Research-Network: **Comprehensive molecular characterization of urothelial bladder carcinoma**. *Nature* 2014, **507**:315-322.