






Review

Bioinformatic Tools for NGS-Based Metagenomics to Improve the Clinical Diagnosis of Emerging, Re-Emerging and New Viruses

Marta Ibañez-Lligoña ^{1,2,3} , Sergi Colomer-Castell ^{1,2,3}, Alejandra González-Sánchez ⁴, Josep Gregori ¹ , Carolina Campos ^{1,2,3}, Damir Garcia-Cehic ^{1,2} , Cristina Andrés ⁴, Maria Piñana ⁴, Tomàs Pumarola ^{4,5}, Francisco Rodríguez-Frias ^{1,2,6} , Andrés Antón ^{4,5} and Josep Quer ^{1,2,3,*} 

- ¹ Liver Diseases-Viral Hepatitis, Liver Unit, Vall d'Hebron Institut de Recerca (VHIR), Vall d'Hebron Hospital Universitari, Vall d'Hebron Barcelona Hospital Campus, Passeig Vall d'Hebron 119-129, 08035 Barcelona, Spain
 - ² Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), Instituto de Salud Carlos III, Av. Monforte de Lemos, 3-5, 28029 Madrid, Spain
 - ³ Biochemistry and Molecular Biology Department, Universitat Autònoma de Barcelona (UAB), Campus de la UAB, Plaça Cívica, 08193 Bellaterra, Spain
 - ⁴ Microbiology Department, Vall d'Hebron Institut de Recerca (VHIR), Vall d'Hebron Hospital Universitari, Vall d'Hebron Barcelona Hospital Campus, Passeig Vall d'Hebron 119-129, 08035 Barcelona, Spain
 - ⁵ Microbiology Department, Universitat Autònoma de Barcelona (UAB), Campus de la UAB, Plaça Cívica, 08193 Bellaterra, Spain
 - ⁶ Department of Basic Sciences, Universitat Internacional de Catalunya, Sant Cugat del Vallès, 08195 Barcelona, Spain
- * Correspondence: josep.quer@vhir.org



Citation: Ibañez-Lligoña, M.; Colomer-Castell, S.; González-Sánchez, A.; Gregori, J.; Campos, C.; Garcia-Cehic, D.; Andrés, C.; Piñana, M.; Pumarola, T.; Rodríguez-Frias, F.; et al. Bioinformatic Tools for NGS-Based Metagenomics to Improve the Clinical Diagnosis of Emerging, Re-Emerging and New Viruses. *Viruses* **2023**, *15*, 587. <https://doi.org/10.3390/v15020587>

Academic Editors: Leyi Wang and Ganwu Li

Received: 13 January 2023
Revised: 16 February 2023
Accepted: 17 February 2023
Published: 20 February 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: Epidemics and pandemics have occurred since the beginning of time, resulting in millions of deaths. Many such disease outbreaks are caused by viruses. Some viruses, particularly RNA viruses, are characterized by their high genetic variability, and this can affect certain phenotypic features: tropism, antigenicity, and susceptibility to antiviral drugs, vaccines, and the host immune response. The best strategy to face the emergence of new infectious genomes is prompt identification. However, currently available diagnostic tests are often limited for detecting new agents. High-throughput next-generation sequencing technologies based on metagenomics may be the solution to detect new infectious genomes and properly diagnose certain diseases. Metagenomic techniques enable the identification and characterization of disease-causing agents, but they require a large amount of genetic material and involve complex bioinformatic analyses. A wide variety of analytical tools can be used in the quality control and pre-processing of metagenomic data, filtering of untargeted sequences, assembly and quality control of reads, and taxonomic profiling of sequences to identify new viruses and ones that have been sequenced and uploaded to dedicated databases. Although there have been huge advances in the field of metagenomics, there is still a lack of consensus about which of the various approaches should be used for specific data analysis tasks. In this review, we provide some background on the study of viral infections, describe the contribution of metagenomics to this field, and place special emphasis on the bioinformatic tools (with their capabilities and limitations) available for use in metagenomic analyses of viral pathogens.

Keywords: NGS; deep-sequencing; viruses; metagenomics; zoonosis; diagnostic tools

1. Introduction

Viruses, particularly those with an RNA genome, are characterized by high variability, a feature that facilitates easy adaptation to changing environments. When viral genomes isolated from samples of infected patients are sequenced, one sees a mixture of different but closely related genomes that undergoes continuous changes over time, mainly because

the RNA-dependent RNA polymerases essential for replication often lack proofreading mechanisms [1]. This complex mixture of closely related genomes is known as a viral quasispecies, and the continuous changes in quasispecies composition result from competitive selection [2] and cooperation [3] between arising mutants.

The abundance or frequency with which a specific genome is found in the viral quasispecies depends on its fitness (replication efficacy) and other known and unknown viral and host factors [2,3]. As a consequence of the multiple variants produced during replication, the virus may obtain certain advantages in addition to adaptability: reduced sensitivity to antiviral therapy, escape from the immune response and vaccine protection, and the possibility to invade new niches. Hence, because of their inherent characteristics, “new” viruses with pathogenic capability in humans can arise.

Our experience with SARS-CoV-2 has clearly shown that viruses cannot be walled-off and that human travel bans are ineffective for preventing expansion of an infection [4]. A better strategy to face the emergence, re-emergence, or appearance of new infectious genomes would be prompt detection and identification, so that specific tools can be applied to tackle the threat. Immediate implementation of control measures at the start of an infection, when there are only a few cases, would have the greatest success in controlling transmission.

Prompt detection implies an accurate disease diagnosis, but the data show that 50% to 60% of acute infections have an unidentified etiology, and 60% to 80% of meningitis/encephalitis, 50% of acute gastroenteritis, 20% of hemorrhagic fever, and 15% to 25% of acute respiratory infections are incorrectly diagnosed [5–9]. As examples of this situation, in the California encephalitis project including 1570 patients, the cause of encephalitis was not identified in 63% of patients [5], and this figure rose to 80% in a study in France [6]. Diagnostic failure can result in delayed and ineffective treatments, with increases in mortality and excessive health expenditure. Hence, correct identification of the infectious agents associated with human disease is a priority.

Currently available diagnostic tests are limited for detecting new pathologic agents. Identification of a virus during acute infection is of enormous value, but current techniques do not enable us to rule out the presence of other coinfecting agents that may be pathogenic. As a large number of pathogens can cause a syndromic infection, a high throughput method such as next-generation sequencing (NGS) adapted to simultaneously detect any pathogen present could be more advantageous than the use of a large number of individual tests based on current methods [10–12]. Furthermore, NGS [13] and metagenomics could potentially be used in the future in a wide range of diseases, such as cancer or gastrointestinal infections, in which the findings obtained could lead to the identification of new treatment targets [14,15].

The sensitivity to detect a virus in a bodily fluid is determined by three factors: the concentration of the virus in the clinical sample, the amount of total RNA and DNA (which compete with what we intend to detect), and the analytical sequencing depth. Sequencing depth can be resolved using high throughput equipment, but the viral concentration cannot be improved unless large amounts of sample are accessible. To overcome the problem caused by background, enrichment of the sample’s viral content is required. This is achieved by removing ribosomal RNA (rRNA), by DNaseI treatment to significantly reduce the amount of free DNA, by using panels with specific capture probes, or by using specific primers in multiplex amplification. Research still has a long way to go to improve pre-sequencing methods to increase the sensitivity and efficiency of current techniques for enhanced clinical diagnosis of emerging, re-emerging, and new viruses. High-throughput NGS technologies based on metagenomics could be the answer to overcome these limitations.

Next-generation sequencing based on metagenomics is a powerful technique to confront the challenge of genetic identification and characterization of known and unknown viral genomes in a large variety of human cell and tissue samples. Metagenomics can be used to design tests based on PCR, to develop mRNA- and DNA-based vaccines, to design direct-acting compounds that block a pathogen’s specific functions, to study variability and

enable correct classification of an infecting agent, and to identify genetic markers associated with the severity of an infection, antigenicity, and the evolution of new variants [16–21].

2. What Is Metagenomics?

Metagenomics is a field of NGS that enables identification of microbial communities, and genetic detection, identification, and characterization of disease-causing agents. It has proven to be a key element in genetic characterization of viruses and has led to discoveries that would not have been accomplished using traditional culturing techniques [22]. Current molecular assays target a limited number of pathogens using specific primers or probes, whereas metagenomics can approach all DNA and RNA molecules present in a sample, enabling analysis of the corresponding host genome and its collection of microbes [23]. The capability of metagenomics to detect any genome, including bacteria, viruses, parasites, and fungi in a human sample is of great interest for the diagnosis of infectious diseases. Metagenomic approaches have also been applied to several other research areas: environmental studies (e.g., marine samples, soil, sewage, farm dust) [24–28]; viral infection in Bronze Age human samples, 7000 years old [29,30]; characterization of the human gut microbiome in health, disease, and forensic investigation [31–34]; clinical studies [23,35]; and discovery of new viral pathogens such as SARS-CoV-2 [36,37].

2.1. What Does Metagenomics Involve?

Metagenomic analyses are performed with random primers that contain every possible combination of nucleotides. This results in the presence of all possible hexamers and allows primers to bind to any RNA or DNA molecule in a mixture of genomes. Once amplified, the PCR product can be loaded onto any NGS platform to obtain millions of short sequences (reads) smaller than 600 bases, but usually <300 bases -Illumina (San Diego, CA, USA), ThermoFisher Scientific (Waltham, MA, USA), MGI (Shenzhen, China), Complete Genomics (San José, CA, USA); suppliers- or long reads of around 1 kb or more -PacBio (Menlo Park, CA, USA), Oxford Nanopore Technologies (Oxford, UK); manufacturers-, and analyzed through bioinformatic techniques [38]. This methodology has several desirable advantages. It does not require any prior knowledge of the genomes under study for primer or probe design and is useful for *de novo* sequencing or resequencing. It enables identification of all pathogens present in a wide variety of samples, including cerebrospinal fluid, sputum, serum, plasma, stool, amniotic fluid, and many others [39–43], and can identify the major genomes of viral populations in epidemiological studies, outbreaks, and phylogenetic analyses.

However, the technique also has strong disadvantages. It requires a large quantity of genomic starting material (e.g., a high viral load) and the analysis is quite complex. The capability to amplify any DNA or RNA genome at random can lead to an underrepresentation or even a loss of minority genomes (low sensitivity), as DNA from the host genome and commensal microorganisms are also amplified. Another important factor to take into account is the risk of contamination during sample collection and the analytical process, which can complicate interpretation of the results. RNA contamination or cross-contamination can be managed by including several controls and quality check points [44]. For example, negative controls consisting of RNase-free water could be added to detect undesirable contamination or human error [44,45]. An additional approach is to include controls in the data analysis to identify contaminants [44].

Briefly, the laboratory protocols for viral metagenomics include sample collection, virus enrichment, DNA/RNA extraction, and library preparation for sequencing. Sample collection and processing varies considerably and depends on the type of biological fluid to analyze [46]. After obtaining the free virions in suspension, several viral enrichment techniques can be applied, as will be reviewed later. Nucleic acid extraction is mainly based on the use of silica spin columns, including lysis buffers containing chaotropic salts and detergents, and optional use of DNA or RNA carriers [47]. Finally, library preparation involves amplification enrichment after fragmentation using kits such as the

Illumina TruSeq RNA Library Prep or Nextera XT DNA [44]. The proper choice among the various available options in each step is key to enhancing the quality of the results for each specific aim.

Of note, recent advances in sequencing technology have enabled long-read sequencing (LRS). With LRS, reads longer than 10 kbp can be generated [48], facilitating the process of mapping against a reference genome and identifying different species in complex samples. Furthermore, the evolution of single-molecule real-time (SMRT) technology through circular consensus sequencing (CCS) has overcome the problem that LRS typically has with base accuracy per read [49]. Long-read sequencing can be performed using various platforms, such as PacBio and Oxford Nanopore Technologies.

2.2. Alternatives to Metagenomics

NGS for metagenomics, whole-genome sequencing (WGS), and targeted deep-sequencing are currently the best tools available for genetic identification and characterization of viruses. With the use of these techniques we can correctly classify a virus, determine its variability, identify viral genetic markers associated with virulence, and consider antigenicity and susceptibility to antivirals based on pre-existing knowledge, when available [16,17,19].

2.3. Enrichment of Viruses in Metagenomic Samples

One of the main issues in virome sequencing is related to the viral titer present in the sample. Usually, viruses account for only a small percentage of the total of genomes present, with most of them belonging to other types of organisms. Depending on sample type, several methods have been used to concentrate the virions present [50]. Viruses in water can be concentrated based on their surface charge, using flocculation-precipitation methods or negative/positive membrane filters [51,52]. Size-based techniques can be used for viral concentration in both water and cell culture supernatants [53,54]. Selection of cell-free DNA (cfDNA) in plasma based on size can enrich viral DNA, as cfDNA size varies according to the species [55]. Finally, the ViroCap project has created a sequence capture panel for viral RNA from 34 families of viruses, notably increasing viral coverage [56].

3. Available Bioinformatic Tools

A number of tools designed for various purposes are now available for use in metagenomics [57,58]. Our focus in this review is to examine analytical tools that facilitate identification of emerging, re-emerging, and new viruses in samples having an animal origin.

There are five main steps in metagenomic assemblies to identify viruses: quality control (QC); quality trimming; read assembly and assembly QC (which are optional); and taxonomic classification of assemblies with two different aims—the identification of known viruses that have already been sequenced and the identification of viruses that have not been sequenced or are unknown. Metagenome binning is an additional step that can be performed before taxonomic profiling. The aim of binning is to cluster assembled sequences according to their origin.

3.1. Bioinformatic Tools for Data QC

The first step in metagenomics would be to perform sequence QC, as it is essential to remove technical errors from the analysis. The main objective of this step is to pre-process the data to eliminate undesirable adapter sequences, excessively short reads, low-quality reads or nucleotides, and others that may be present. Several programs can be used in this step, depending on the data being analyzed (Table 1).

Table 1. Bioinformatic programs for data quality control in short-read and long-read sequencing. All websites/GitHub Repository links were accessed on 23 December 2022.

Program	Type of Data	Website/GitHub Repository
FastQC	Short reads	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
MultiQC	Short reads	https://multiqc.info
LongQC	Long reads	https://github.com/yfukasawa/LongQC
MinionQC	Long reads	https://github.com/roblanf/minion_qc

For short-read sequencing data, QC can be performed with FastQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>, accessed on 23 December 2022), which checks the quality of the data and generates a report summarizing its metrics. The same type of report is available with other QC programs, such as MultiQC [59], which has the same functionality as FastQC with one main difference, it can be applied to several fastQ files simultaneously and generate one main report for all the files provided.

For the preprocessing of long-read sequencing data, longQC [60] or MinionQC [61] can be used to determine sequence quality. These have been applied to data obtained with Nanopore's MinION and other long-read sequencers.

3.2. Bioinformatic Tools for Data Pre-Processing

After having identified technical errors within the metagenomic data, such as excessively short reads, low-quality reads, and adapters, the next step is to eliminate them to avoid false positives and negatives. In addition, removal of undesirable sequences will decrease computational time and cost in the following steps. Several tools can be used for the removal of unwanted sequences or nucleotides, depending on the pre-processing aim and what has been identified during the QC step.

3.2.1. Tools for Quality Trimming

In metagenomics and in most RNA sequencing methods, the first step would be to trim unwanted elements identified through the data QC programs (Table 1). The main program for this step used in short-read sequencing is Trimmomatic, a bioinformatic tool designed to remove low-quality reads and adapters [62]. Another possibility is cutadapt, which identifies and removes adapters and other sequence types [63] (Table 2).

Table 2. Bioinformatic programs for data trimming in short-read and long-read sequencing. All websites/GitHub Repository links were accessed on 23 December 2022.

Program	Type of Data	Website/GitHub Repository
Trimmomatic	Short reads	http://www.usadellab.org/cms/?page=trimmomatic
Fastp	Short reads	https://github.com/OpenGene/fastp
Cutadapt	Short reads	https://cutadapt.readthedocs.io/en/stable/
SOAPnuke	Short reads	https://github.com/BGI-flexlab/SOAPnuke
NanoPack	Long reads	https://github.com/wdecoster/nanopack
SequelTools	Long reads	https://github.com/ISUgenomics/SequelTools

Long-read sequencing fastq files can be trimmed with other programs (Table 2). For example, NanoPack [64] can be used to process data from long-read sequencing and to visualize QC results. SequelTools [65], which has the same functionality as Nanopack, is another option.

3.2.2. Tools for Filtering Untargeted Reads

The second filtering step is to eliminate reads of no interest, which can be derived from various sources. When targeting viral reads, we would have to remove reads belonging to the host genome and contaminants. This is a key step to decrease the execution time

and computational expense in taxonomic classification of dataset sequences. In addition, it reduces false positives and can prevent assembly of chimeric virus–host sequences [66]. Several strategies are commonly used to remove these sequences from the data.

For example, read mappers can be applied to remove all sequences mapping a selected reference genome, which could belong to the host genome or possible contaminants. For short reads, mappers such as BWA [67], bowtie2 [68], and BMAP [69] are available (Table 3). Other tools such as FastQ-Screen (https://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/, accessed on 23 December 2022) can identify sequences belonging to specific genomes. These programs determine the proportion of host genome or other specific contaminant sequences against target reads.

Table 3. Bioinformatic mappers for filtering out untargeted or contaminant sequences in short-read and long-read sequencing data. All websites/GitHub Repository links were accessed on 23 December 2022.

Program	Type of Data	Website/GitHub Repository
BWA	Short reads, long reads	https://bio-bwa.sourceforge.net
Bowtie2	Short reads	https://github.com/BenLangmead/bowtie2
BMAP	Short reads, long reads	https://jgi.doe.gov/data-and-tools/software-tools/bbtools/bb-tools-user-guide/bbmap-guide/
Minimap2	Long reads	https://github.com/lh3/minimap2

The filtering of long reads can be performed by some of the above-mentioned read mappers, such as BWA [70] and BMAP [69], or more specific ones, such as minimap2, which was particularly designed for long-read sequencing data [71].

Alternative approaches are available to reduce the number of unwanted sequences. Certain bioinformatic tools have been programmed to identify specific sequences belonging to specific taxa. In this case, once the reads have been trimmed, the sequences are passed through a filtering program that only selects reads with certain features. An example is VirusHunter (<https://bio.tools/virushunter>, accessed on 23 December 2022), used to identify viral sequences in NGS data. RINS is another option, as it has been designed to identify non-human sequences [72] and can filter out reads from a human source.

In some situations, other RNA sequence types such as ribosomal (rRNA), mitochondrial (mtRNA), or messenger (mRNA) types from untargeted taxa may have to be removed from metagenomic data [66]. RiboDetector (<https://github.com/hzi-bifo/RiboDetector>, accessed on 23 December 2022) can be used for this purpose, as it is designed to identify rRNA, which can thus be filtered out to improve the following analyses.

Another approach involves taxonomic profiling of the reads before assembly. With this strategy, sequences other than those belonging to viruses can be filtered out, and viral sequences retained for further analysis. This option can be carried out with taxonomic classifiers, such as kraken2 [73] and kaiju [74], which will be further explained later.

3.3. Bioinformatic Tools for Assembly

3.3.1. Tools for Short-Read Assembly

To perform taxonomic assignment and identify the viruses present, we must first restore the metagenomes. This implies the generation of contigs, sets of sequences that have been overlapped to provide a longer, continuous sequence.

The main type of assembly used in metagenomics is called *de novo* genome assembly. This can be performed by overlap layout consensus, based on overlapping the read ends, or with various algorithms, such as de Bruijn graphs, which split the sequences into smaller fragments, called k-mers, thereby reducing the time and computational effort needed for analysis [75]. In addition, several programs are available to perform *de novo* assembly, for example, MEGAHIT [76], a bioinformatic assembler tool optimized for metagenomes, or metaSPADES [77] and IDBA-UD [78], which are also optimized for metagenomes (Table 4).

Table 4. Bioinformatic tools for metagenome assembly for short-read, long-read, and hybrid assemblies. All websites/GitHub Repository links were accessed on 23 December 2022.

Program	Read Length	Algorithm	Website/GitHub Repository
MEGAHIT	Short reads	De Bruijn graph	https://github.com/voutcn/megahit
metaSPADES	Short reads	De Bruijn graph	https://github.com/ablab/spades
IDBA-UD	Short reads	De Bruijn graph	https://i.cs.hku.hk/~alse/hkubrg/projects/idba_ud/
MetaVelvet	Short reads	De Bruijn graph	http://metavelvet.dna.bio.keio.ac.jp
Omega2	Short reads	Overlap layout consensus	https://github.com/qiumingyao/omega2
metaFlye	Long reads	Overlap layout consensus	https://github.com/fenderglass/Flye
Canu	Long reads	Overlap layout consensus	https://github.com/marbl/canu
NECAT	Long reads (Nanopore)	String graph	https://github.com/xiaochuanle/NECAT
HybridSPADES	Hybrid	De Bruijn graph	https://github.com/ablab/spades
OPERA-MS	Hybrid	De Bruijn graph	https://github.com/CSB5/OPERA-MS
HASLR	Hybrid	De Bruijn graph	https://github.com/vpc-ccg/haslr
Wegan	Hybrid	Synthetic scaffolding graph	https://github.com/adigenova/wengan

Nonetheless, the assembly of sequences is quite complex. In metagenomic analyses, where there is a huge volume of data and unequal representation of the microbial community, genomes which have a small presence are usually underrepresented [79]. Furthermore, de novo assembly can generate errors, as a single sample can contain sequences from very similar organisms.

Another strategy, reference-based assembly, can also be used in metagenomics. However, appropriate reference genomes may not be available in all cases, and this approach does not allow the identification of new viruses, or viruses that have not been sequenced previously.

3.3.2. Tools for Long-Read Assembly

Various programs have been designed to assemble long reads. These are specific for this kind of technology and take into account the higher error rate associated with these reads. Some examples are metaFlye [80], Canu [81], and NECAT [82]. These tools can be used with data from various techniques, from Nanopore sequencing to PacBio, even in high-fidelity reads (Table 4).

3.3.3. Tools for Hybrid Assembly

As was mentioned above, short reads are known to have very low error rates, but they cannot be used in some cases, such as in the assembly of highly variable regions. Long reads can provide continuous coverage of very long regions, but they have high error rates [55]. The optimal situation would be to combine both these features: coverage of long regions with the reliability of short reads.

Accordingly, various programs have been developed to perform hybrid assembly, specifically in metagenomic studies. OPERA-MS [83], implemented with the de Bruijn graph algorithm, works by first assembling short reads to produce contigs, and then mapping both long and short reads to the contigs. Once this is completed, long reads are then used to connect the contigs. Finally, the contigs are clustered according to their genomic distance and difference in coverage [83]. HybridSPADES [84] is another program that can perform hybrid assembly. It is based on the de Bruijn graph algorithm, which first constructs an assembly graph using short reads, and then maps long reads to this first assembly, closing the gaps and resolving repeats using long reads [84]. HASLR [85] and Wengan [86] are additional tools that perform assembly in a manner similar to HybridSPADES (Table 4).

3.4. Bioinformatic Tools for Quality Control of Metagenome Assembly

Once the metagenome has been assembled, the quality of the assembly should be determined. The tools for this purpose can be classified into two main categories: those that require reference genomes for QC such as MetaQUAST [87], which uses references

to calculate statistics for the assembly; and those that do not need references, such as DeepMAsED [88], which uses machine learning to identify misassemblies, or REAPR [89], which computes basic statistics using paired-end reads mapped to the assembly (Table 5). In general, it can be difficult to work with references in metagenomic studies, as reference genomes are often unavailable or of very poor quality.

Table 5. Bioinformatic tools for quality control of metagenome assembly. All websites/GitHub Repository links were accessed on 23 December 2022.

Program	Type of Data	Website/GitHub Repository
MetaQuast	Reference-based	http://bioinf.spbau.ru/metaquast
DeepMAsED	Non-reference	https://github.com/leylabmpi/DeepMAsED
REAPR	Non-reference	https://www.sanger.ac.uk/tool/reapr/
CheckM	Non-reference	https://github.com/Ecogenomics/CheckM
BUSCO	Non-reference	https://busco.ezlab.org
VALET	Non-reference	https://www.cccb.umd.edu/software/valet

BUSCO [90] is another assembly evaluator, in this case working with program-based reference datasets to assess completeness of the metagenome [90]. CheckM [91] can also be used for assembly evaluation. It provides estimates of the completeness of the genome, such as GC content. Finally, VALET (<https://github.com/marbl/VALET>, accessed on 23 December 2022) can be applied to detect misassemblies in metagenomic data, as it can bin contigs by coverage and avoid false positives and false negatives due to uneven coverage depth [57,92]

Quality control statistics for genome assembly may not be appropriate for assessment of metagenome assemblies. For example, the N50 measure describes the quality of the assembly based on the minimum contig length, which represents half the genome. Nonetheless, this is not representative of metagenomic data, which contain a wide range of genomes having different sizes. Thus, there is a need to develop QC statistics for metagenome assembly. As an alternative to the programs mentioned, reads could be mapped to contigs to determine the quality of the assembly. This would enable a more general assembly check.

An essential step to identify viruses in metagenomic analyses is to perform taxonomic profiling. There are two main methods to achieve this task: the first is to classify reads according to taxonomies, and the second is to establish taxonomic groups by contigs. Both methods have advantages and disadvantages. In taxonomic profiling with contigs (i.e., with assembled reads), classification is conducted with longer sequences. However, there is a risk that some contigs might be chimeric. Taxonomic profiling with reads is less statistically significant; the sequences are shorter, although a larger number of sequences are analyzed [93]. This approach could provide a more diverse result, but the computational expense would be higher.

3.5. Tools for Identification of Known Viruses

Taxonomic profilers use reference databases to compare reads/contigs with given sequences. Some profilers have been implemented using k-mers, such as kraken2 [73], bracken [94], CLARK [95], and Centrifuge [96] (Table 6). There are also protein-based programs, which first translate sequences to enable a comparison with reference protein databases. For example, kaiju can input reads or contigs, which are translated to proteins and then queued to the system to find matches. DIAMOND [97] and MM-seqs2 [98] are additional protein-based programs. Tools such as MetaPhlan4 [99], IG-Gsearch [100], and GOTCHA [101] involve the use of gene markers to align sequences to gene marker-related databases.

Table 6. Bioinformatic taxonomic profilers, including algorithm, use of custom databases, and website or GitHub repository. All websites/GitHub Repository links were accessed on 23 December 2022.

Program	Type of Classifier	Allows Custom Databases	Website/GitHub Repository
Kraken2	k-mers	Yes	https://ccb.jhu.edu/software/kraken2/
Kraken-HLL	k-mers	Yes	https://github.com/Krischan/krakenhll
KrakenUniq	k-mers	Yes	https://github.com/fbreitwieser/krakenuniq
Centrifuge	k-mers	Yes	https://github.com/DaehwanKimLab/centrifuge
Ganon	k-mers	Yes	https://github.com/pirovc/ganon
Bracken	k-mers	Yes	https://github.com/jenniferlu717/Bracken
MetaCache	k-mers	Yes	https://github.com/muellan/metacache
CLARK	k-mers	Yes	http://clark.cs.ucr.edu
VirusTaxo	k-mers	No	https://omics-lab.com/virustaxo
Metavir2	k-mers	No	https://github.com/jhayer/metavir
k-SLAM	k-mers	Yes	https://github.com/aindj/k-SLAM
Taxonomer	k-mers	No	http://taxonomer.com
LMAT	k-mers	No	https://computing.llnl.gov/projects/livermore-metagenomics-analysis-toolkit
Sourmash	k-mers	No	https://sourmash.readthedocs.io/en/latest/
metaOthello	k-mers	Yes	https://github.com/xa6xa6/metaOthello
ProPhyle	k-mers	No	https://prophyle.github.io
TaxMaps	k-mers	Yes	https://github.com/nygenome/taxmaps#sge
Kaiju	Protein-coding	Yes	https://kaiju.binf.ku.dk
DIAMOND	Protein-coding	Yes	https://github.com/bbuchfink/diamond
MMseqs2	Protein-coding	Yes	https://github.com/soedinglab/MMseqs2
IGGsearch	Marker gene	No	https://github.com/snayfach/IGGsearch
MetaPhlan3	Marker gene	No	https://huttenhower.sph.harvard.edu/metaphlan
GOTTCHA	Marker gene	No	https://lanl-bioinformatics.github.io/GOTTCHA/
DeepVirFinder	CNN	No	https://github.com/jessieren/DeepVirFinder
BLAST	Alignment-based	Yes	https://blast.ncbi.nlm.nih.gov/Blast.cgi
DUDes	DUD	No	https://github.com/pirovc/dudes
MCP	Alignment-based	Yes	https://microba.com/microbiome-research/

Other programs are based on algorithms, such as BLAST or DUDes [102], which execute a new algorithm using the DUD (Deepest Uncommon Descent) strategy [103]. Some bioinformatic tools have been specifically programmed to study the virome, including VirusTaxo [104], Metavir2 [105], and DeepVirFinder [106], whose main algorithm is based on convolutional neural networks (CNN).

On the other hand, some tools, such as MetaPhlan4 [99] and MCP (Microbiota Community Profiler), contain sequences from unidentified metagenomics-assembled genomes. This allows the identification of viruses that are not available in corresponding databases. However, MCP can only be used to identify bacterial, archaeal, eukaryotic, and viral sequences in microbiota studies [107].

It is important to keep in mind that each taxonomic profiler performs differently, and that a variety of algorithms and reference databases are used. This multiplicity can lead to differing results according to the program applied, and a wide range of time and computational expenses. K-mer-based taxonomic profilers seem to be the most computationally efficient, although they have high memory demands. Marker-based classifiers have lower memory requirements, but they can only classify reads or contigs from specific regions. Alignment-based bioinformatic programs are more computationally expensive than the others [107].

Tools for Identification of Novel Viruses

Programs to identify viral sequences without the need for any reference are now available. VirSorter [108] and VirFinder [106] are two such bioinformatic tools. VirFinder [106] is a k-mer-based R package that can identify viral contigs with good predictive accuracy, whereas VirSorter can identify novel viral sequences in a diverse microbial dataset [105].

Specific taxonomic profilers, such as MCP [107], that include unidentified metagenome-assembled genomes (MAGs) retrieved from microbiota studies, enable the detection of sequences that are found in other metagenomic datasets, but have not been uploaded to the main databases. However, not many profilers can do this. It would be of great value to

upload unidentified MAGs to provide a consortium of unrecognized sequences and enable a more efficient and effortless analysis.

Despite these advances, better information and tools are needed to enable recognition of emergent and new viruses whose sequences are not present in reference databases. Robust approaches and programs should be developed to detect new viral sequences in metagenomic analyses. As an example, once the metagenome has been assembled, a possible strategy could be to classify sequences according to their proportion of tetranucleotides. This would allow similar sequences to be clustered together, possibly indicating a similar origin. However, this method would need to be reinforced by analysis of the sequence characteristics, such as ORFs and protein-coding sequences. Several available bioinformatic programs have been designed to do this, such as Prodigal [109], which can find protein-coding sequences within contigs.

3.6. Bioinformatic Tools for Contig Binning

An optional step, contig binning, can be performed before taxonomic profiling. The main goal of contig binning is to cluster contigs according to species, as these sequences usually do not cover the whole genome [57]. Several tools using different core algorithms for short-read sequencing can be used for this purpose (Table 7). One example is CONCOCT [110], a program that allows the clustering of metagenomic contigs according to nucleotide composition and coverage data. Another is GraphBin [111], which uses the assembly's connectivity information to cluster contigs.

Table 7. Bioinformatic tools to perform metagenomic binning in short-read and long-read datasets or after assembly. All websites/GitHub Repository links were accessed on 23 December 2022.

Program	Read Length	Website/GitHub Repository
MetaBAT	Short reads	https://bitbucket.org/berkeleylab/metabat/src/master/
GroopM	Short reads	https://github.com/centre-for-microbiome-research/GroopM
MaxBin	Short reads	https://sourceforge.net/projects/maxbin/
CONCOCT	Short reads	https://github.com/BinPro/CONCOCT
MyCC	Short reads	https://sourceforge.net/projects/sb2nhri/files/MyCC/
SolidBin	Short reads	https://github.com/sufforest/SolidBin
BMC3C	Short reads	http://mlda.swu.edu.cn/codes.php?name=BMC3C
COCACOLA	Short reads	https://github.com/younglululu/COCACOLA
GraphBin	Short reads	https://github.com/metagentools/GraphBin
METAMVGL	Short reads	https://github.com/ZhangZhenmiao/METAMVGL
VAMB	Short reads	https://github.com/RasmussenLab/vamb
LRBinner	Long reads	https://github.com/anuradhawick/LRBinner
MEGAN-LR	Long reads	http://software-ab.cs.uni-tuebingen.de/download/megan6/megan-lr/
BusyBee Web	Long reads	https://ccb-microbe.cs.uni-saarland.de/busybee

Metagenomic binning is not restricted to contigs. It can also be conducted with reads, specifically, long reads, and with the use of MEGAN-LR [112], BusyBee [113], or LRBinner [114] (Table 6).

4. Conclusions

Numerous recent advances have been achieved in the field of metagenomics. This technique can aid in the discovery of new viruses, prediction of outbreaks, and diagnosis of certain diseases for clinical purposes, among others. The rapid evolution of long-read sequencing platforms can benefit metagenomic analyses by producing more reliable results. In metagenomics in virology, appropriate sample processing is important to detect viromes without underrepresentation.

Despite this progress, further developments are needed. For example, consensus guidelines would be of great value to promote proper data analysis. As is seen in this review, numerous programs using various approaches are available to analyze this type of data, and several pipelines have been developed to enable faster and easier analysis.

However, these are based on very different processes and there is still a lack of consensus regarding their performance for various tasks. Another key undertaking is to maintain the related databases updated, as these are essential for identification of taxa within the data.

Funding: This study was partially supported by Pla Estratègic de Recerca i Innovació en Salut (PERIS)—Direcció General de Recerca i Innovació en Salut (DGRIS), Catalan Health Ministry, Generalitat de Catalunya; the Spanish Network for Research in Infectious Diseases (REIPI RD16/0016/0003) from the European Regional Development Fund (ERDF); Centro para el Desarrollo Tecnológico Industrial (CDTI) from the Spanish Ministry of Economy and Business, grant number IDI-20200297; grants PI19/00301 and PI22/00258 from Instituto de Salud Carlos III, cofinanced by the European Regional Development Fund (ERDF); and Gilead’s biomedical research project GLD21/00006. S.C-C is a recipient of a predoctoral fellowship, FPU, from Ministerio de Universidades (FPU21/04150).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We thank Celine Cavallo for English language support.

Conflicts of Interest: No competing interests to declare.

References

1. Perlman, S.; Netland, J. Coronaviruses post-SARS: Update on replication and pathogenesis. *Nat. Rev. Microbiol.* **2009**, *7*, 439–450. [[CrossRef](#)] [[PubMed](#)]
2. Domingo, E. Mutation Rates and Rapid Evolution of RNA Viruses. *Evol. Biol. Viruses* **1994**, 161–184.
3. Vignuzzi, M.; Stone, J.K.; Arnold, J.J.; Cameron, C.E.; Andino, R. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature* **2005**, *439*, 344–348. [[CrossRef](#)]
4. Fischer, C.; Maponga, T.G.; Yadouleton, A.; Abílio, N.; Aboce, E.; Adewumi, P.; Afonso, P.; Akorli, J.; Andriamandimby, S.F.; Anga, L.; et al. Gradual emergence followed by exponential spread of the SARS-CoV-2 Omicron variant in Africa. *Science* **2022**, *378*, eadd8737. [[CrossRef](#)] [[PubMed](#)]
5. Glaser, C.A.; Honarmand, S.; Anderson, L.J.; Schnurr, D.P.; Forghani, B.; Cossen, C.K.; Schuster, F.L.; Christie, L.J.; Tureen, J.H. Beyond Viruses: Clinical Profiles and Etiologies Associated with Encephalitis. *Clin. Infect. Dis.* **2006**, *43*, 1565–1577. [[CrossRef](#)]
6. Mailles, A.; Stahl, J.P. Infectious Encephalitis in France in 2007: A National Prospective Study. *Clin. Infect. Dis.* **2009**, *49*, 1838–1847. [[CrossRef](#)]
7. Studahl, M.; Lindquist, L.; Eriksson, B.-M.; Günther, G.; Bengner, M.; Franzen-Röhl, E.; Fohlman, J.; Bergström, T.; Aurelius, E. Acute Viral Infections of the Central Nervous System in Immunocompetent Adults: Diagnosis and Management. *Drugs* **2013**, *73*, 131–158. [[CrossRef](#)]
8. Fernandez-Cassi, X.; Martínez-Puchol, S.; Silva-Sales, M.; Cornejo, T.; Bartolome, R.; Bofill-Mas, S.; Girones, R. Unveiling Viruses Associated with Gastroenteritis Using a Metagenomics Approach. *Viruses* **2020**, *12*, 1432. [[CrossRef](#)]
9. Racska, L.D.; Kraft, C.S.; Olinger, G.; Hensley, L. Viral Hemorrhagic Fever Diagnostics. *Clin. Infect. Dis.* **2016**, *62*, 214–219. [[CrossRef](#)]
10. Kennedy, P.G.E. Viral encephalitis. *J. Neurol.* **2005**, *252*, 268–272. [[CrossRef](#)]
11. Edridge, A.W.D.; Deijs, M.; van Zeggeren, I.E.; Kinsella, C.M.; Jebbink, M.F.; Bakker, M.; van de Beek, D.; Brouwer, M.C.; van der Hoek, L. Viral Metagenomics on Cerebrospinal Fluid. *Genes* **2019**, *10*, 332. [[CrossRef](#)]
12. Brown, J.R.; Bharucha, T.; Breuer, J. Encephalitis diagnosis using metagenomics: Application of next generation sequencing for undiagnosed cases. *J. Infect.* **2018**, *76*, 225–240. [[CrossRef](#)]
13. Tafazoli, A.; Guchelaar, H.-J.; Miltik, W.; Kretowski, A.J.; Swen, J.J. Applying Next-Generation Sequencing Platforms for Pharmacogenomic Testing in Clinical Practice. *Front. Pharmacol.* **2021**, *12*, 2025. [[CrossRef](#)]
14. Banerjee, J.; Mishra, N.; Dhas, Y. Metagenomics: A new horizon in cancer research. *Meta Gene* **2015**, *5*, 84–89. [[CrossRef](#)]
15. Ungaro, F.; Massimino, L.; Furfaro, F.; Rimoldi, V.; Peyrin-Biroulet, L.; D’Alessio, S.; Danese, S. Metagenomic analysis of intestinal mucosa revealed a specific eukaryotic gut virome signature in early-diagnosed inflammatory bowel disease. *Gut Microbes* **2018**, *10*, 149–158. [[CrossRef](#)]
16. Hwang, J.Y.; Ahn, S.J.; Kwon, M.; Seo, J.S.; Hwang, S.D.; Jee, B.Y. Whole-genome next-generation sequencing and phylogenetic characterization of viral haemorrhagic septicaemia virus in Korea. *J. Fish Dis.* **2020**, *43*, 599–607. [[CrossRef](#)]
17. Santiago-Rodriguez, T.M.; Hollister, E.B. Human Virome and Disease: High-Throughput Sequencing for Virus Discovery, Identification of Phage-Bacteria Dysbiosis and Development of Therapeutic Approaches with Emphasis on the Human Gut. *Viruses* **2019**, *11*, 656. [[CrossRef](#)]

18. Yll, M.; Cortese, M.F.; Murillo, M.G.; Orriols, G.; Gregori, J.; Casillas, R.; González, C.; Sopena, S.; Godoy, C.; Vila, M.; et al. Conservation and variability of hepatitis B core at different chronic hepatitis stages. *World J. Gastroenterol.* **2020**, *26*, 2584–2598. [[CrossRef](#)]
19. Chen, Q.; Perales, C.; Soria, M.E.; García-Cehic, D.; Gregori, J.; Rodríguez-Frías, F.; Buti, M.; Crespo, J.; Calleja, J.L.; Tabernero, D.; et al. Deep-sequencing reveals broad subtype-specific HCV resistance mutations associated with treatment failure. *Antivir. Res.* **2020**, *174*, 104694. [[CrossRef](#)]
20. Greaney, A.J.; Starr, T.N.; Gilchuk, P.; Zost, S.J.; Binshtein, E.; Loes, A.N.; Hilton, S.K.; Huddleston, J.; Eguia, R.; Crawford, K.H.; et al. Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition. *Cell Host Microbe* **2021**, *29*, 44–57.e9. [[CrossRef](#)]
21. Svraka, S.; Rosario, K.; Duizer, E.; Van Der Avoort, H.; Breitbart, M.; Koopmans, M. Metagenomic sequencing for virus identification in a public-health setting. *J. Gen. Virol.* **2010**, *91*, 2846–2856. [[CrossRef](#)] [[PubMed](#)]
22. Bashir, Y.; Singh, S.P.; Konwar, B.K. Metagenomics: An Application Based Perspective. *Chin. J. Biol.* **2014**, *2014*, 146030. [[CrossRef](#)]
23. Chiu, C.Y.; Miller, S.A. Clinical metagenomics. *Nat. Rev. Genet.* **2019**, *20*, 341. [[CrossRef](#)] [[PubMed](#)]
24. Zayed, A.A.; Wainaina, J.M.; Dominguez-Huerta, G.; Pelletier, E.; Guo, J.; Mohssen, M.; Tian, F.; Pratama, A.A.; Bolduc, B.; Zablocki, O.; et al. Cryptic and abundant marine viruses at the evolutionary origins of Earth’s RNA virome. *Science* **2022**, *376*, 156–162. [[CrossRef](#)] [[PubMed](#)]
25. Ufarté, L.; Laville, É.; Duquesne, S.; Potocki-Veronese, G. Metagenomics for the discovery of pollutant degrading enzymes. *Biotechnol. Adv.* **2015**, *33*, 1845–1854. [[CrossRef](#)] [[PubMed](#)]
26. Hendriksen, R.S.; Munk, P.; Njage, P.; Van Bunnik, B.; McNally, L.; Lukjancenko, O.; Röder, T.; Nieuwenhuijse, D.; Pedersen, S.K.; Kjeldgaard, J.; et al. Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. *Nat. Commun.* **2019**, *10*, 1124. [[CrossRef](#)]
27. Kwok, K.T.T.; de Rooij, M.M.T.; Messink, A.B.; Wouters, I.M.; Smit, L.A.M.; Cotten, M.; Heederik, D.J.J.; Koopmans, M.P.G.; Phan, M.V.T. Establishing farm dust as a useful viral metagenomic surveillance matrix. *Sci. Rep.* **2022**, *12*, 16308. [[CrossRef](#)]
28. Kristensen, D.M.; Mushegian, A.R.; Dolja, V.V.; Koonin, E.V. New dimensions of the virus world discovered through metagenomics. *Trends Microbiol.* **2010**, *18*, 11–19. [[CrossRef](#)]
29. Allentoft, M.E.; Sikora, M.; Sjögren, K.-G.; Rasmussen, S.; Rasmussen, M.; Stenderup, J.; Damgaard, P.B.; Schroeder, H.; Ahlström, T.; Vinner, L.; et al. Population genomics of Bronze Age Eurasia. *Nature* **2015**, *522*, 167–172. [[CrossRef](#)]
30. Mühlemann, B.; Jones, T.C.; Damgaard, P.D.B.; Allentoft, M.E.; Shevnina, I.; Logvin, A.; Usmanova, E.; Panyushkina, I.P.; Boldgiv, B.; Bazartseren, T.; et al. Ancient hepatitis B viruses from the Bronze Age to the Medieval period. *Nature* **2018**, *557*, 418–423. [[CrossRef](#)]
31. Wang, W.-L.; Xu, S.-Y.; Ren, Z.-G.; Tao, L.; Jiang, J.-W.; Zheng, S.-S. Application of metagenomics in the human gut microbiome. *World J. Gastroenterol.* **2015**, *21*, 803. [[CrossRef](#)]
32. Lloyd-Price, J.; Abu-Ali, G.; Huttenhower, C. The healthy human microbiome. *Genome Med.* **2016**, *8*, 51. [[CrossRef](#)]
33. Pascal, V.; Pozuelo, M.; Borruel, N.; Casellas, F.; Campos, D.; Santiago, A.; Martinez, X.; Varela, E.; Sarrabayrouse, G.; Machiels, K.; et al. A microbial signature for Crohn’s disease. *Gut* **2017**, *66*, 813–822. [[CrossRef](#)]
34. Hampton-Marcell, J.T.; Lopez, J.V.; Gilbert, J.A. The human microbiome: An emerging tool in forensics. *Microb. Biotechnol.* **2017**, *10*, 228. [[CrossRef](#)]
35. Wilson, M.R.; O’Donovan, B.D.; Gelfand, J.M.; Sample, H.A.; Chow, F.C.; Betjemann, J.P.; Shah, M.P.; Richie, M.B.; Gorman, M.P.; Hajj-Ali, R.A.; et al. Chronic Meningitis Investigated via Metagenomic Next-Generation Sequencing. *JAMA Neurol.* **2018**, *75*, 947–955. [[CrossRef](#)]
36. Zhou, P.; Yang, X.-L.; Wang, X.-G.; Hu, B.; Zhang, L.; Zhang, W.; Si, H.-R.; Zhu, Y.; Li, B.; Huang, C.-L.; et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **2020**, *579*, 270–273. [[CrossRef](#)]
37. Wu, F.; Zhao, S.; Yu, B.; Chen, Y.-M.; Wang, W.; Song, Z.-G.; Hu, Y.; Tao, Z.-W.; Tian, J.-H.; Pei, Y.-Y.; et al. A new coronavirus associated with human respiratory disease in China. *Nature* **2020**, *579*, 265–269. [[CrossRef](#)]
38. Quer, J.; Colomer-Castell, S.; Campos, C.; Andrés, C.; Piñana, M.; Cortese, M.F.; González-Sánchez, A.; Garcia-Cehic, D.; Ibáñez, M.; Pumarola, T.; et al. Next-Generation Sequencing for Confronting Virus Pandemics. *Viruses* **2022**, *14*, 600. [[CrossRef](#)]
39. Lim, Y.W.; Evangelista, J.S.; Schmieder, R.; Bailey, B.; Haynes, M.; Furlan, M.; Maughan, H.; Edwards, R.; Rohwer, F.; Conrad, D. Clinical Insights from Metagenomic Analysis of Sputum Samples from Patients with Cystic Fibrosis. *J. Clin. Microbiol.* **2014**, *52*, 425. [[CrossRef](#)]
40. Lee, J.-H.; Choi, J.-P.; Yang, J.; Won, H.-K.; Park, C.S.; Song, W.-J.; Kwon, H.-S.; Kim, T.-B.; Kim, Y.-K.; Park, H.-S.; et al. Metagenome analysis using serum extracellular vesicles identified distinct microbiota in asthmatics. *Sci. Rep.* **2020**, *10*, 15125. [[CrossRef](#)]
41. Molton, J.S.; Lee, I.R.; Bertrand, D.; Ding, Y.; Kalimuddin, S.; Lye, D.C.; Nagarajan, N.; Gan, Y.-H.; Archuleta, S. Stool metagenome analysis of patients with Klebsiella pneumoniae liver abscess and their domestic partners. *Int. J. Infect. Dis.* **2021**, *107*, 1–4. [[CrossRef](#)]
42. Wang, H.; Yang, G.X.; Hu, Y.; Lam, P.; Sangha, K.; Siciliano, D.; Swenerton, A.; Miller, R.; Tilley, P.; Von Dadelszen, P.; et al. Comprehensive human amniotic fluid metagenomics supports the sterile womb hypothesis. *Sci. Rep.* **2022**, *12*, 6875. [[CrossRef](#)] [[PubMed](#)]

43. Manso, C.; Bibby, D.; Mohamed, H.; Brown, D.W.G.; Zuckerman, M.; Mbisa, J.L. Enhanced Detection of DNA Viruses in the Cerebrospinal Fluid of Encephalitis Patients Using Metagenomic Next-Generation Sequencing. *Front. Microbiol.* **2020**, *11*, 1879. [[CrossRef](#)] [[PubMed](#)]
44. Fitzpatrick, A.H.; Rupnik, A.; O’Shea, H.; Crispie, F.; Keaveney, S.; Cotter, P. High Throughput Sequencing for the Detection and Characterization of RNA Viruses. *Front. Microbiol.* **2021**, *12*, 190. [[CrossRef](#)] [[PubMed](#)]
45. Ogunbayo, A.E.; Sabiu, S.; Nyaga, M.M. Evaluation of extraction and enrichment methods for recovery of respiratory RNA viruses in a metagenomics approach. *J. Virol. Methods* **2023**, *314*, 114677. [[CrossRef](#)] [[PubMed](#)]
46. Wylezich, C.; Papa, A.; Beer, M.; Höper, D. A Versatile Sample Processing Workflow for Metagenomic Pathogen Detection. *Sci. Rep.* **2018**, *8*, 13108. [[CrossRef](#)] [[PubMed](#)]
47. Klenner, J.; Kohl, C.; Dabrowski, P.W.; Nitsche, A. Comparing Viral Metagenomic Extraction Methods. *Curr. Issues Mol. Biol.* **2017**, *24*, 59–70. [[CrossRef](#)]
48. Athanasopoulou, K.; Boti, M.A.; Adamopoulos, P.G.; Skourou, P.C.; Scorilas, A. Third-Generation Sequencing: The Spearhead towards the Radical Transformation of Modern Genomics. *Life* **2022**, *12*, 30. [[CrossRef](#)]
49. Zaragoza-Solas, A.; Haro-Moreno, J.M.; Rodriguez-Valera, F.; López-Pérez, M. Long-Read Metagenomics Improves the Recovery of Viral Diversity from Complex Natural Marine Samples. *Msystems* **2022**, *7*, e00192-22. [[CrossRef](#)]
50. Forés, E.; Rusiñol, M.; Itarte, M.; Martínez-Puchol, S.; Calvo, M.; Bofill-Mas, S. Evaluation of a virus concentration method based on ultrafiltration and wet foam elution for studying viruses from large-volume water samples. *Sci. Total. Environ.* **2022**, *829*, 154431. [[CrossRef](#)]
51. Katayama, H.; Shimasaki, A.; Ohgaki, S. Development of a Virus Concentration Method and Its Application to Detection of Enterovirus and Norwalk Virus from Coastal Seawater. *Appl. Environ. Microbiol.* **2002**, *68*, 1033–1039. [[CrossRef](#)]
52. Calgua, B.; Rodriguez-Manzano, J.; Hundesa, A.; Suñen, E.; Calvo, M.; Bofill-Mas, S.; Girones, R. New methods for the concentration of viruses from urban sewage using quantitative PCR. *J. Virol. Methods* **2013**, *187*, 215–221. [[CrossRef](#)]
53. Le Bideau, M.; Robresco, L.; Baudoin, J.-P.; La Scola, B. Concentration of SARS-CoV-2-Infected Cell Culture Supernatants for Detection of Virus-like Particles by Scanning Electron Microscopy. *Viruses* **2022**, *14*, 2388. [[CrossRef](#)]
54. Ichim, C.V.; Wells, R. Generation of high-titer viral preparations by concentration using successive rounds of ultracentrifugation. *J. Transl. Med.* **2011**, *9*, 137. [[CrossRef](#)]
55. Phung, Q.; Lin, M.J.; Xie, H.; Greninger, A.L. Fragment Size-Based Enrichment of Viral Sequences in Plasma Cell-Free DNA. *J. Mol. Diagn.* **2022**, *24*, 476–484. [[CrossRef](#)]
56. Wylie, T.N.; Wylie, K.M.; Herter, B.N.; Storch, G.A. Enhanced virome sequencing using targeted sequence capture. *Genome Res.* **2015**, *25*, 1910–1920. [[CrossRef](#)]
57. Yang, C.; Chowdhury, D.; Zhang, Z.; Cheung, W.K.; Lu, A.; Bian, Z.; Zhang, L. A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 6301–6314. [[CrossRef](#)]
58. Luthra, A.; Mastrogiacomo, B.; Smith, S.A.; Chakravarty, D.; Schultz, N.; Sanchez-Vega, F. Computational methods and translational applications for targeted next-generation sequencing platforms. *Genes Chromosom. Cancer* **2022**, *61*, 322–331. [[CrossRef](#)]
59. Ewels, P.; Magnusson, M.; Lundin, S.; Käller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **2016**, *32*, 3047–3048. [[CrossRef](#)]
60. Fukasawa, Y.; Ermini, L.; Wang, H.; Carty, K.; Cheung, M.-S. LongQC: A Quality Control Tool for Third Generation Sequencing Long Read Data. *G3 Genes Genomes Genet.* **2020**, *10*, 1193–1196. [[CrossRef](#)]
61. Lanfear, R.; Schalamun, M.; Kainer, D.; Wang, W.; Schwessinger, B. MinIONQC: Fast and simple quality control for MinION sequencing data. *Bioinformatics* **2018**, *35*, 523–525. [[CrossRef](#)] [[PubMed](#)]
62. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [[CrossRef](#)] [[PubMed](#)]
63. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **2011**, *17*, 10–12. [[CrossRef](#)]
64. De Coster, W.; D’Hert, S.; Schultz, D.T.; Cruts, M.; Van Broeckhoven, C. NanoPack: Visualizing and processing long-read sequencing data. *Bioinformatics* **2018**, *34*, 2666–2669. [[CrossRef](#)]
65. Hufnagel, D.E.; Hufford, M.B.; Seetharam, A.S. SequelTools: A suite of tools for working with PacBio Sequel raw sequence data. *BMC Bioinform.* **2020**, *21*, 2666. [[CrossRef](#)]
66. Nooij, S.; Schmitz, D.; Vennema, H.; Kroneman, A.; Koopmans, M. Overview of Virus Metagenomic Classification Methods and Their Biological Applications. *Front. Microbiol.* **2018**, *9*, 749. [[CrossRef](#)]
67. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [[CrossRef](#)]
68. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [[CrossRef](#)]
69. Bushnell, B. *BBMap: A Fast, Accurate, Splice-Aware Aligner*; Lawrence Berkeley National Lab. (LBNL): Berkeley, CA, USA, 2014.
70. Li, H.; Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **2010**, *26*, 589–595. [[CrossRef](#)]
71. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **2018**, *34*, 3094–3100. [[CrossRef](#)]

72. Bhaduri, A.; Qu, K.; Lee, C.S.; Ungewickell, A.; Khavari, P.A. Rapid identification of non-human sequences in high-throughput sequencing datasets. *Bioinformatics* **2012**, *28*, 1174–1175. [[CrossRef](#)]
73. Wood, D.E.; Lu, J.; Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **2019**, *20*, 257. [[CrossRef](#)]
74. Menzel, P.; Ng, K.L.; Krogh, A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.* **2016**, *7*, 11257. [[CrossRef](#)]
75. Khan, A.R.; Pervez, M.T.; Babar, M.E.; Naveed, N.; Shoaib, M. A Comprehensive Study of De Novo Genome Assemblers: Current Challenges and Future Prospective. *Evol. Bioinform.* **2018**, *14*, 1176934318758650. [[CrossRef](#)]
76. Li, D.; Liu, C.-M.; Luo, R.; Sadakane, K.; Lam, T.-W. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **2015**, *31*, 1674–1676. [[CrossRef](#)]
77. Nurk, S.; Meleshko, D.; Korobeynikov, A.; Pevzner, P.A. metaSPAdes: A new versatile metagenomic assembler. *Genome Res.* **2017**, *27*, 824–834. [[CrossRef](#)]
78. Peng, Y.; Leung, H.C.M.; Yiu, S.M.; Chin, F.Y.L. IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **2012**, *28*, 1420–1428. [[CrossRef](#)]
79. Lapidus, A.L.; Korobeynikov, A.I. Metagenomic Data Assembly—The Way of Decoding Unknown Microorganisms. *Front. Microbiol.* **2021**, *12*, 653. [[CrossRef](#)]
80. Kolmogorov, M.; Bickhart, D.M.; Behsaz, B.; Gurevich, A.; Rayko, M.; Shin, S.B.; Kuhn, K.; Yuan, J.; Pevnikov, E.; Smith, T.P.L.; et al. metaFlye: Scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **2020**, *17*, 1103–1110. [[CrossRef](#)]
81. Koren, S.; Walenz, B.P.; Berlin, K.; Miller, J.R.; Bergman, N.H.; Phillippy, A.M. Canu: Scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **2017**, *27*, 722–736. [[CrossRef](#)]
82. Chen, Y.; Nie, F.; Xie, S.-Q.; Zheng, Y.-F.; Dai, Q.; Bray, T.; Wang, Y.-X.; Xing, J.-F.; Huang, Z.-J.; Wang, D.-P.; et al. Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nat. Commun.* **2021**, *12*, 60. [[CrossRef](#)] [[PubMed](#)]
83. Bertrand, D.; Shaw, J.; Kalathiyappan, M.; Ng, A.H.Q.; Kumar, M.S.; Li, C.; Dvornicic, M.; Soldo, J.P.; Koh, J.Y.; Tong, C.; et al. Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nat. Biotechnol.* **2019**, *37*, 937–944. [[CrossRef](#)] [[PubMed](#)]
84. Antipov, D.; Korobeynikov, A.; McLean, J.S.; Pevzner, P.A. hybridSPAdes: An algorithm for hybrid assembly of short and long reads. *Bioinformatics* **2016**, *32*, 1009–1015. [[CrossRef](#)] [[PubMed](#)]
85. Haghshenas, E.; Asghari, H.; Stoye, J.; Chauve, C.; Hach, F. HASLR: Fast Hybrid Assembly of Long Reads. *Iscience* **2020**, *23*, 101389. [[CrossRef](#)]
86. Di Genova, A.; Buena-Atienza, E.; Ossowski, S.; Sagot, M.-F. Efficient hybrid de novo assembly of human genomes with WENGAN. *Nat. Biotechnol.* **2020**, *39*, 422–430. [[CrossRef](#)]
87. Mikheenko, A.; Saveliev, V.; Gurevich, A. MetaQUAST: Evaluation of metagenome assemblies. *Bioinformatics* **2016**, *32*, 1088–1090. [[CrossRef](#)]
88. Mineeva, O.; Rojas-Carulla, M.; Ley, R.; Schölkopf, B.; Youngblut, N.D. DeepMAS-ED: Evaluating the quality of metagenomic assemblies. *Bioinformatics* **2020**, *36*, 3011–3017. [[CrossRef](#)]
89. Hunt, M.; Kikuchi, T.; Sanders, M.; Newbold, C.; Berriman, M.; Otto, T.D. REAPR: A universal tool for genome assembly evaluation. *Genome Biol.* **2013**, *14*, R47. [[CrossRef](#)]
90. Manni, M.; Berkeley, M.R.; Seppey, M.; Zdobnov, E.M. BUSCO: Assessing Genomic Data Quality and Beyond. *Curr. Protoc.* **2021**, *1*, e323. [[CrossRef](#)]
91. Parks, D.H.; Imelfort, M.; Skennerton, C.T.; Hugenholtz, P.; Tyson, G.W. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **2015**, *25*, 1043–1055. [[CrossRef](#)]
92. Olson, N.D.; Treangen, T.J.; Hill, C.M.; Cepeda-Espinoza, V.; Ghurye, J.; Koren, S.; Pop, M. Metagenomic assembly through the lens of validation: Recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Briefings Bioinform.* **2019**, *20*, 1140–1150. [[CrossRef](#)]
93. Rodríguez-Brazzarola, P.; Pérez-Wohlfeil, E.; Díaz-Del-Pino, S.; Holthausen, R.; Trelles, O. Analyzing the Differences between Reads and Contigs When Performing a Taxonomic Assignment Comparison in Metagenomics. In Proceedings of the Bioinformatics and Biomedical Engineering: 6th International Work-Conference, IWBBIO 2018, Granada, Spain, 25–27 April 2018; Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 450–460. [[CrossRef](#)]
94. Lu, J.; Breitwieser, F.P.; Thielen, P.; Salzberg, S.L. Bracken: Estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **2017**, *3*, e104. [[CrossRef](#)]
95. Ounit, R.; Wanamaker, S.; Close, T.J.; Lonardi, S. CLARK: Fast and accurate classification of metagenomic and genomic sequences using discriminative *k*-mers. *BMC Genom.* **2015**, *16*, 236. [[CrossRef](#)]
96. Kim, D.; Song, L.; Breitwieser, F.P.; Salzberg, S.L. Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Res.* **2016**, *26*, 1721–1729. [[CrossRef](#)]
97. Buchfink, B.; Reuter, K.; Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **2021**, *18*, 366–368. [[CrossRef](#)]
98. Mirdita, M.; Steinegger, M.; Breitwieser, F.; Söding, J.; Karín, E.L. Fast and sensitive taxonomic assignment to metagenomic contigs. *Bioinformatics* **2021**, *37*, 3029–3031. [[CrossRef](#)]

99. Blanco-Miguez, A.; Beghini, F.; Cumbo, F.; McIver, L.J.; Thompson, K.N.; Zolfo, M.; Manghi, P.; Dubois, L.; Huang, K.D.; Thomas, A.M.; et al. Extending and Improving Metagenomic Taxonomic Profiling with Uncharacterized Species with MetaPhlAn 4. *bioRxiv* **2022**. [[CrossRef](#)]
100. Nayfach, S.; Shi, Z.J.; Seshadri, R.; Pollard, K.S.; Kyrpides, N.C. New insights from uncultivated genomes of the global human gut microbiome. *Nature* **2019**, *568*, 505–510. [[CrossRef](#)]
101. Freitas, T.A.K.; Li, P.-E.; Scholz, M.B.; Chain, P.S.G. Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. *Nucleic Acids Res.* **2015**, *43*, e69. [[CrossRef](#)]
102. Piro, V.C.; Lindner, M.S.; Renard, B.Y. DUDes: A top-down taxonomic profiler for metagenomics. *Bioinformatics* **2016**, *32*, 2272–2280. [[CrossRef](#)]
103. Zhang, Z.; Schwartz, S.; Wagner, L.; Miller, W. A Greedy Algorithm for Aligning DNA Sequences. *J. Comput. Biol.* **2000**, *7*, 203–214. [[CrossRef](#)] [[PubMed](#)]
104. Raju, R.S.; Al Nahid, A.; Dev, P.C.; Islam, R. VirusTaxo: Taxonomic classification of viruses from the genome sequence using k-mer enrichment. *Genomics* **2022**, *114*, 110414. [[CrossRef](#)] [[PubMed](#)]
105. Roux, S.; Tournayre, J.; Mahul, A.; Debroas, D.; Enault, F. Metavir 2: New tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinform.* **2014**, *15*, 76. [[CrossRef](#)]
106. Ren, J.; Ahlgren, N.A.; Lu, Y.Y.; Fuhrman, J.A.; Sun, F. VirFinder: A novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **2017**, *5*, 69. [[CrossRef](#)]
107. Parks, D.H.; Rigato, F.; Vera-Wolf, P.; Krause, L.; Hugenholtz, P.; Tyson, G.W.; Wood, D.L.A. Evaluation of the Microba Community Profiler for Taxonomic Profiling of Metagenomic Datasets From the Human Gut Microbiome. *Front. Microbiol.* **2021**, *12*, 731. [[CrossRef](#)] [[PubMed](#)]
108. Roux, S.; Enault, F.; Hurwitz, B.L.; Sullivan, M.B. VirSorter: Mining viral signal from microbial genomic data. *PeerJ* **2015**, *3*, e985. [[CrossRef](#)]
109. Hyatt, D.; Chen, G.-L.; Locascio, P.F.; Land, M.L.; Larimer, F.W.; Hauser, L.J. Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **2010**, *11*, 119. [[CrossRef](#)]
110. Alneberg, J.; Bjarnason, B.S.; de Bruijn, I.; Schirmer, M.; Quick, J.; Ijaz, U.Z.; Lahti, L.; Loman, N.; Andersson, A.; Quince, C. Binning metagenomic contigs by coverage and composition. *Nat. Methods* **2014**, *11*, 1144–1146. [[CrossRef](#)]
111. Mallawaarachchi, V.; Wickramarachchi, A.; Lin, Y. GraphBin: Refined binning of metagenomic contigs using assembly graphs. *Bioinformatics* **2020**, *36*, 3307–3313. [[CrossRef](#)]
112. Huson, D.H.; Albrecht, B.; Bağcı, C.; Bessarab, I.; Górska, A.; Jolic, D.; Williams, R.B.H. MEGAN-LR: New algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. *Biol. Direct* **2018**, *13*, 6. [[CrossRef](#)]
113. Laczny, C.C.; Kiefer, C.; Galata, V.; Fehlmann, T.; Backes, C.; Keller, A. BusyBee Web: Metagenomic data analysis by bootstrapped supervised binning and annotation. *Nucleic Acids Res.* **2017**, *45*, W171–W179. [[CrossRef](#)]
114. Wickramarachchi, A.; Lin, Y. Binning long reads in metagenomics datasets using composition and coverage information. *Algorithms Mol. Biol.* **2022**, *17*, 14. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.