Large - scale investigation of zoonotic viruses in the era of high - throughput sequencing

AUTHOR(S):

Kawasaki, Junna; Tomonaga, Keizo; Horie, Masayuki

CITATION:

Kawasaki, Junna ...[et al]. Large - scale investigation of zoonotic viruses in the era of high - throughput sequencing. Microbiology and Immunology 2023, 67(1): 1-13

ISSUE DATE:

2023-01

URL:

http://hdl.handle.net/2433/284162

Received: 18 June 2022 | Revised: 28 September 2022 | Accepted: 16 October 2022

DOI: 10.1111/1348-0421.13033

REVIEW

Microbiology and Immunology

Large-scale investigation of zoonotic viruses in the era of high-throughput sequencing

Junna Kawasaki[1,2,3] | Keizo Tomonaga[1,2,4] | Masayuki Horie[5,6]

[1]Laboratory of RNA Viruses, Department of Virus Research, Institute for Frontier Life and Medical Sciences, Kyoto University, Kyoto, Japan

[2]Laboratory of RNA Viruses, Department of Mammalian Regulatory Network, Graduate School of Biostudies, Kyoto University, Kyoto, Japan

[3]Faculty of Science and Engineering, Waseda University, Tokyo, Japan

[4]Department of Molecular Virology, Graduate School of Medicine, Kyoto University, Kyoto, Japan

[5]Division of Veterinary Sciences, Graduate School of Life and Environmental Sciences, Osaka Prefecture University, Osaka, Japan

[6]Osaka International Research Center for Infectious Diseases, Osaka Prefecture University, Osaka, Japan

Correspondence

Junna Kawasaki, Laboratory of RNA Viruses, Department of Virus Research, Institute for Frontier Life and Medical Sciences, Kyoto University, Kyoto 606-8507, Japan.
Email: jrt13mpmuk@gmail.com

Masayuki Horie, Osaka International Research Center for Infectious Diseases, Osaka Prefecture University, Osaka 598-8531, Japan.
Email: mhorie@omu.ac.jp

Funding information

Ministry of Education, Culture, Sports, Science and Technology, Grant/Award Numbers: JP16H06429, JP16H06430, JP16K21723, JP17H05821, JP19H04833; Japan Society for the Promotion of Science, Grant/Award Numbers: JP18K19443, JP19J22241, JP19K22530, JP20H05682, JP21H01199, JP22J00010; Waseda Research Institute for Science and Engineering, Grant-in-Aid for Young Scientists (Early Bird)

Abstract

Zoonotic diseases considerably impact public health and socioeconomics. RNA viruses reportedly caused approximately 94% of zoonotic diseases documented from 1990 to 2010, emphasizing the importance of investigating RNA viruses in animals. Furthermore, it has been estimated that hundreds of thousands of animal viruses capable of infecting humans are yet to be discovered, warning against the inadequacy of our understanding of viral diversity. High-throughput sequencing (HTS) has enabled the identification of viral infections with relatively little bias. Viral searches using both symptomatic and asymptomatic animal samples by HTS have revealed hidden viral infections. This review introduces the history of viral searches using HTS, current analytical limitations, and future potentials. We primarily summarize recent research on large-scale investigations on viral infections reusing HTS data from public databases. Furthermore, considering the accumulation of uncultivated viruses, we discuss current studies and challenges for connecting viral sequences to their phenotypes using various approaches: performing data analysis, developing predictive modeling, or implementing high-throughput platforms of virological experiments. We believe that this article provides a future direction in large-scale investigations of potential zoonotic viruses using the HTS technology.

KEYWORDS

data reusability, high-throughput sequencing, RNA virus, virome analysis, zoonosis

INTRODUCTION

Zoonotic viruses have repeatedly threatened the human population. Previous studies have reported that RNA viruses caused approximately 94% of zoonotic diseases documented between 1990 and 2010.[1] Therefore, RNA viral investigations in animals have been emphasized in preparing for future viral zoonoses, particularly in searching for viruses capable of infecting humans.

High-throughput sequencing (HTS) is a comprehensive method for determining genetic sequences in a sample, enabling us to search for viral sequences with little bias. The HTS

Abbreviations: COVID-19, coronavirus disease 2019; GTEx, Genotype-Tissue Expression; GVP, Global Virome Project; HTS, high-throughput sequencing; ICTV, International Committee on Taxonomy of Viruses; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; SRA, sequence read archive; STAT, Sequence Taxonomic Analysis Tool; TSA, transcriptome shotgun assembly.

technology has been involved with three main accomplishments in the field of virology. First, it has been shown that our interests are greatly biased toward pathogenic viruses, and an enormous number of viruses in ecology may have been overlooked.[2] In fact, the number of virus species classified by the International Committee on Taxonomy of Viruses (ICTV) has increased exponentially according to studies using HTS (https://talk.ictvonline.org/taxonomy/p/taxonomy_releases). Second, HTS has helped reduce the time required to identify causative agents of emerging infectious diseases and has led to the rediscovery of viruses responsible for unexplained diseases. Third, the accumulation of viral sequences identified using HTS has promoted data-driven research, providing insights into viral epidemiology and evolution.

This review summarizes recent research on viral sequence identification using HTS and the ongoing efforts to understand viral phenotypes using different approaches: performing sequencing data analysis, developing predictive modeling, and implementing high-throughput platforms for virological experiments. Here, we primarily focused on RNA viral research for the following reasons: (i) the majority of zoonotic diseases are caused by RNA virus infections,[1] and (ii) recent studies using HTS have rapidly revealed RNA viral diversity.[2]

First, we introduced a series of studies on viral identification using HTS, focusing on the sample types. HTS has been primarily used to determine the causative viruses of infectious disease using samples collected from animals showing evident symptoms. Conversely, HTS analyses using various samples from individuals, including apparently healthy ones, have identified hidden viral infections and broadened our understanding of viral diversity.[3–5] Furthermore, recent studies reported that large-scale investigations of viral infections reusing HTS data from public databases can be a powerful approach for identifying both pathogenic and nonpathogenic viral infections.[6–9]

Next, we argue the current status and challenges to connect the "sequence data" to the "phenotypic data" of viruses. It is impractical to investigate the infectivity and pathogenicity of all viruses experimentally; therefore, several approaches are needed to understand viral phenotypes: sequencing data analysis or prediction modeling development. In particular, prediction for viral host range or infectivity would help evaluate the zoonotic potential of viruses or prioritize certain viruses for experimental validation. Furthermore, developing a novel platform for high-throughput experiments is important, such as screening virus infectivity and sorting virus-infected cells. These different approaches to obtaining viral phenotypic data will help to prepare us for future viral zoonoses.

Why do we need to investigate viruses in animals?

Many infectious diseases are caused by viral transmission from animals to humans, including influenza viruses, coronaviruses, ebolaviruses, and poxviruses. Specifically, the animal-to-human transmission of influenza viruses has repeatedly occurred.[10,11] One of the most remarkable cases was the 2009 H1N1 pandemic. It was estimated that this virus led to approximately 201,200 deaths from respiratory illness and 83,300 from cardiovascular disease during the first year of the pandemic.[12] Sequence analysis showed that this pandemic was caused by a virus generated by the reassortment of human H3N2, swine H1N1, and avian influenza viral genes. The H1N1pdm09 virus was initially transmitted from pigs to humans and spread via human-to-human transmissions.[13–15]

As another case, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infected over 608 million people and caused more than 6.5 million deaths by September 2022 (https://covid19.who.int/). SARS-CoV-2 or related viruses have been identified in various animals, such as bats, pangolins, white-tailed deer in the wild, farmed minks, cats and dogs as companion animals, nonhuman primates, and large cats in zoos, and it has been reported that SARS-CoV-2 was transmitted between humans and animals.[16,17] Based on these cases, it is necessary to identify viruses capable of causing future zoonoses by investigating the human–animal interface.

Initiative for zoonotic viral research: PREDICT and Global Virome Project

Zoonotic diseases substantially impact public health and socioeconomics. Several international projects have been established to prepare for subsequent viral zoonoses. The PREDICT project (https://p2.predict.global/) has investigated viruses at the human–animal interface since 2009 and has worked to construct platforms for virus infection surveillance, creating a transdisciplinary collaborative team and developing analysis pipelines or databases.[18–20] From 2009 to 2020, by investigating more than 164,000 animal and human samples, this project succeeded in detecting over 1100 viruses, including filoviruses and coronaviruses that have repeatedly caused infectious diseases in humans (https://ohi.vetmed.ucdavis.edu/programs-projects/predict-project).

The Global Virome Project (GVP; https://www.globalviromeproject.org/) was launched in 2018 to further elucidate animal viral diversity because the GVP team estimated that 631,000–827,000 unknown viruses capable of infecting humans are present in mammals and birds.[1] This project planned to conduct virus searches using more samples than those in the PREDICT project to identify unknown viruses. However, the feasibility and cost-effectiveness of these projects have been discussed[21–23]; thus, we now face the need to reconsider the significance and prospect of virus searches.

HTS for virus identification

Classical virological experiments have been conducted to identify viral infections; for example, viral cultivation by

specimen inoculation into a cell culture system, morphological analysis of virus particles and infected cells, or serological tests.[24] Although these experiments can provide detailed information on viral phenotypes, they are time-consuming and labor-intensive. Furthermore, as virus-specific detection tools, such as antibodies or PCR primers, are used in most classical experiments, it is often difficult to comprehensively detect viruses other than the target. Thus, these classical viral detection methods have limitations in terms of labor and comprehensiveness.

By contrast, the HTS technology has enabled the rapid identification of viral sequences with relatively low bias. There are two major types of sequencing methods using the HTS technology for identifying viral sequences: untargeted and targeted sequencing. These methods differ in their preparation of sequence libraries. It should be noted that the sensitivity and range of the viral detection depend on the sequencing library preparation. In the first method, the untargeted sequencing, sequence libraries are constructed by nucleic acids extracted from every component, including viruses and host organisms in a given sample. Thus, this method effectively detects various viral sequences in a sample. However, viral sequences are present in very small amounts compared with host sequences, leading to low viral detection sensitivity. To improve virus detection sensitivity without narrowing the detection range, processes to remove nonviral sequences are needed; for example, size filtration for collecting viral particles from a sample or nuclease treatment to degrade nucleic acids that are not protected by viral particles. In the second method, the target enrichment sequencing, sequence libraries are prepared to consist mainly of specific viral sequences. For example, PCR amplification or sequence capturing using tagged probes are conducted to enrich target viral sequences after nucleic acid extraction. The targeted sequencing method shows high viral detection sensitivity and can reconstruct high-quality viral genomes by sequence assembly (see Box 1 for details of sequence assembly analysis). However, this method cannot detect sequences other than the target viral sequences, which is unsuitable for searching a wide range of viruses.

After read sequences are determined by HTS, two methods are primarily used for identifying viral sequences in HTS data: the read-based method and the assembly-based method (details in Box 1).

Virus identification using the HTS technology

Here, we summarize three major accomplishments involving identifying viral sequences using the HTS technology and investigating accumulated sequences. First, the viruses that we have recognized to date are only a small fraction of the entire virosphere diversity.[2,24] The number of virus species in the ICTV has increased exponentially, and such rapid elucidation of viral diversity has been associated with studies using HTS, especially virus metagenomic or metatranscriptomic analysis.[2] Notably, a recent study discovered a new DNA virus family, Redondoviridae, even in humans, the most intensively investigated species for viral infections.[25] These findings emphasize that our understanding of viral diversity remains incomplete.

Second, the HTS technology has enabled rapid and comprehensive investigations into the causes of viral infectious diseases. Classical virological experiments, such as virus isolation and culture, often required months or years to identify candidate causative viruses; however, the HTS technology can identify them with nearly complete genome sequences in a matter of days or weeks. For example, the outbreak of Nipah virus infection was recognized in September 1998, and the genetic sequence of the causative virus was not reported until 2000.[26] By contrast, in the SARS-CoV-2 pandemic, the epidemic of pneumonia cases was reported in December 2019, and the viral genomic sequences determined using HTS were globally shared in January 2020 (https://www.who.int/news/item/27-04-2020-who-timeline—covid-19). These cases indicate that HTS has greatly reduced the time needed to identify the causative agents of infectious diseases, even considering advances that have occurred in the global scientific community over the past 20 years. Furthermore, HTS has enabled the identification of pathogenic viruses associated with previously unexplained diseases.[27–30] For example, investigation of samples from patients who died of encephalitis revealed infection with Borna disease virus 1, which causes encephalitis in horses and sheep.[31–33] These cases are examples of re-examining a disease of unknown cause using HTS, which led to identifying a pathogenic virus and providing valuable clues for future investigation and countermeasures to viral infectious diseases. Therefore, a viral search in clinical samples using HTS is a powerful approach to elucidate the associations between diseases and viral infections.

Third, the accumulation of viral sequences has accelerated data-driven research.[34,35] In particular, sequences of SARS-CoV-2 or influenza viruses have been intensively collected and registered in databases, such as GISAID (https://www.gisaid.org/) or International Nucleotide Sequence Database (https://www.insdc.org/). These virus sequences are used for molecular epidemiological analysis to define variants of concern in the coronavirus disease 2019 (COVID-19) pandemic (https://www.gisaid.org/hcov19-variants/), or to develop vaccines for influenza viruses (https://www.gisaid.org/references/human-influenza-vaccine-composition/). As another example, GLUE has been reported as a surveillance platform for viral sequences accumulated in public databases, enabling the tracking of viral transmission routes or identifying drug resistance–associated mutations.[36,37] Such analysis platforms using accumulated viral sequence data contribute to predicting the spatiotemporal viral spread or evaluating which intervention strategies effectively control viral infections.

Expanding our understanding of viral diversity by correcting sampling bias

As mentioned in the "Virus identification using the HTS technology" section, almost all studies have focused

BOX 1 Methods for detecting viral infections in HTS data

There are two major methods to search for viral sequences in HTS data: read-based and assembly-based methods (Figure 1). Herein, we outline the methods and summarize their advantages and limitations. Furthermore, we will list recent advances against these methods' limitations in the "Challenges of virus searches using public HTS-related data" section.

In the read-based method, viral infections are investigated by mapping HTS reads to the reference viral genomes. In the assembly-based method, contig sequences are reconstructed using HTS reads. Thereafter, viral sequences are searched by comparison between the contigs and reference viral genomes. Sequence assembly is often performed using HTS reads that are unmapped to host genomic sequences for the following reasons: (i) saving computational costs and operation time, or (ii) avoiding mis-assembly (e.g., generating chimeric sequences with endogenous viral elements in the host genome[38,39]). The quality of viral genomic sequences reconstructed by sequence assembly can be validated by mapping HTS reads to the assembled sequence and checking the uniformity of the read coverage throughout the sequence. The read- and assembly-based methods are not exclusive, and most studies use both methods depending on their advantages and limitations.[27,40–42]

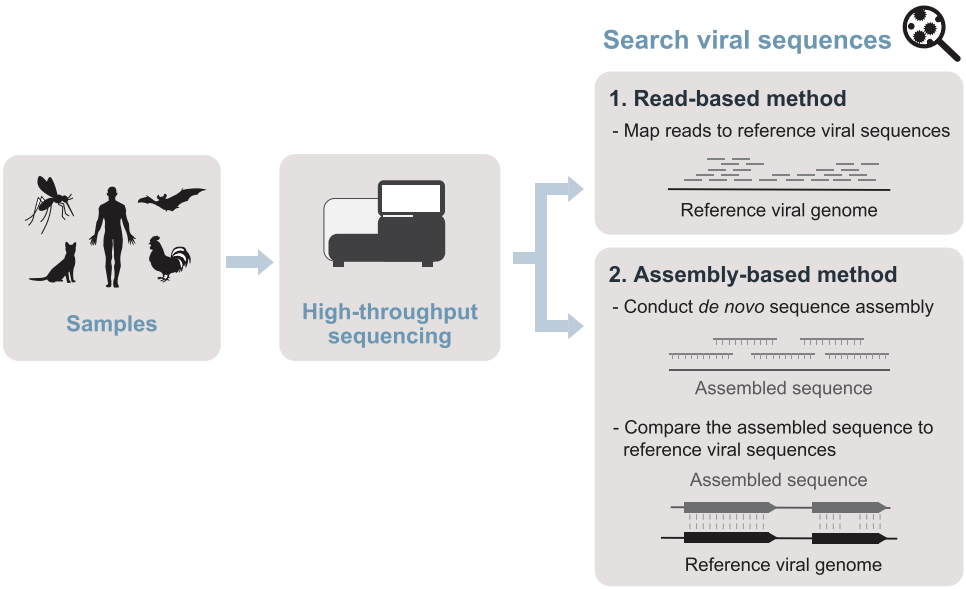

FIGURE 1 Methods for viral sequence search in HTS data. The genetic information in a sample can be determined by HTS and used to identify viral sequences. There are two major methods for searching viral sequences in HTS data: the read-based method and the assembly-based method (details in Box 1).

(1) Read-based method

Advantages

- The computational costs, including operation time, are lower than those of the assembly-based method.
- The infection can be detected with high sensitivity when investigating known viruses.

Limitations

- It is often challenging to detect viruses distantly related to reference viral genomes by sequence similarity search, especially using short-read HTS data,[43] as many mapping tools are designed to not allow extensive mismatches. To solve this issue, sequence comparison methods between HTS reads and reference viral genomes have been developed, such as using the translated sequences (details in the "Current research studies reusing public HTS data for viral searches: read-based method" section).[6,44] This is because the protein sequences or their structures are more conserved than nucleotide sequences.[43,45,46]
- The read-based method cannot be used to obtain viral genome sequences, essential for downstream experiments to characterize viral phenotypes.

- The risk of false-positive detection of viral infection due to contamination from experimental resources or environments should be noted if a small number of virus-derived reads is detected.[47]

(2) Assembly-based method

Advantages

- Sequence assembly can reconstruct partial or whole viral sequences, although short-read HTS length ranges between tens and hundreds of nucleotides. A similarity search using assembled sequences as queries can be used to detect viruses highly divergent from reference sequences. Sequence similarity search tools, such as BLAST, measure the alignment quality by similarity and matched length between two sequences; the sensitivity of similarity detection can be increased using a longer query sequence.[43]
- Reconstructed viral genomic sequences can be used for downstream experimental analysis, such as a reverse genetics system for generating infectious viruses and characterizing their phenotypes (see also the limitation relevant to this point in the next paragraph).

Limitations

- This method consists of several analysis steps (extract HTS reads unmapped to the host genomic sequence, *de novo* sequence assembly, and/or sequence comparison with known viral sequences) and requires relatively higher computational costs.
- Depending on the amount of virus-derived reads and their diversity, it is difficult to determine the full-length genomic sequences.[48]
- Sequence assembly constructs a consensus sequence that can originate from several quasispecies variants. We should note that authentic viral phenotypes may not be observed in experiments using only consensus sequences. Such variant information can be checked by analyses that map HTS reads to the assembled sequence. Furthermore, single-virus genomics has also been developed to address this issue.[49]
- The mapping analysis using viral contig sequence and HTS data, in which the virus was detected, is needed to quantitatively evaluate the viral infection level.

primarily on "pathogenic viruses." However, virus identification in various samples, including those from apparently healthy animals, is valuable for surveying potential zoonotic pathogens. It has been shown that many zoonotic viruses are not pathogenic in their natural hosts but cause severe diseases in humans.[24,50,51] Such cases indicate that viral pathogenesis in animals cannot necessarily be used as an indicator of whether viruses possess zoonotic potential, and surveys focusing on animal samples with evident symptoms may miss identifying viruses capable of causing zoonoses. Thus, enriching the virome catalog will enable the prompt identification and characterization of the causative agents of emerging infectious diseases in humans or animals in the future.

In addition, the expansion of sequence databases by identifying a wide variety of viruses is expected to improve viral detection sensitivity. As virus searches in HTS data depend on sequence similarity to known viruses (Box 1), the expansion of search space by the sequence accumulation of phylogenetically diverse viruses can lead to discovering unknown virus sequences.[43,52] Recent studies have successively identified phylogenetically distinct viruses from mammalian or avian viruses in animals that have rarely been used for virus searches, such as fish, amphibians, reptiles, and invertebrates.[3–5] Taken together, virus searches in various samples can contribute to discovering potentially zoonotic viruses hidden in natural hosts and improving viral detection rate in bioinformatics analyses.

Reusing publicly available HTS data for large-scale investigation of viral infections

Virus searches using various animals, including asymptomatic ones, can identify hidden viral infections. However, it is expected that the viral detection rate in such investigations, without focusing on diseases, is relatively low. Thus, although large-scale investigations are needed to compensate for a low viral detection rate, the cost and labor involved can be problematic.

Recent studies have attempted to reuse publicly available HTS data, which accumulate in databases, to solve this issue. For example, 62 Peta bases of HTS data were publicly available in the NCBI sequence read archive (SRA) database

on March 2022 (https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi). HTS data originated from various animal species, containing asymptomatic virus–infected samples. Therefore, reusing public HTS data allows us to conduct large-scale investigations of viral infections while saving the costs for random sampling and moderating the effects of sampling bias.

Current research studies reusing public HTS data for viral searches: read-based method

Here, we present three studies that investigated viral infections using public HTS data by the read-based method. First, one study[6] screened for viral infections in 5.7 million HTS data using the read-based method. Notably, to improve virus detection sensitivity of the read-based method, the authors used a sequence comparison method that calculates the similarities between translated HTS read sequences and known viral proteins (see also Box 1). Furthermore, new coronavirus lineages were discovered in nonmammalian aquatic vertebrate samples by performing *de novo* sequence assembly. The authors disclosed viral infection profiles in publicly available HTS data on an open-access database (https://serratus.io/), which is a helpful platform for monitoring the spread of viral infections or discovering new viruses. Indeed, new bornaviruses were reported by a study using this database.[53]

In the second study,[8] the authors investigated HTS data from healthy human tissue samples available on the Genotype-Tissue Expression (GTEx) Project (https://gtexportal.org/home/). The authors revealed different patterns in the immune response depending on the infected viruses by association analysis between viral abundance and host transcriptome. Furthermore, in this study, the specimens provided by GTEx biobank were used for pathological examination to confirm disease symptoms associated with viral infections. One of the advantages of using data sets from integrated biobanks, consisting of HTS data and their original samples, is that further experimental validations for viral infectivity or pathogenesis can be performed. By contrast, original samples of public HTS data are not usually available, and it is often challenging to further analyze detected viruses.

The third study, the *k*-mer–based Sequence Taxonomic Analysis Tool (STAT), which taxonomically classifies HTS reads into the viral family or genus, can be used to screen viral sequences from HTS data.[54] The taxonomic composition of read sequences in public HTS data is available on the NCBI SRA database, which helps to determine the number of virus-derived reads in HTS data before the investigation. Furthermore, checking the taxonomic composition of HTS reads using this tool is useful for validating the accuracy of virus–host relationships or the quality of assembled sequences. For example, (i) if a virus is detected in HTS data containing sequence reads from multiple animals, it would be difficult to define the true host of the virus; (ii) we should suspect contamination from experimental or environmental resources when viral sequences are detected in animals unreported as the host thus far; or (iii) sequence assembly may construct chimeric viral sequences if genetically related viruses are in the same HTS data. Indeed, this study reported that SARS-CoV-2 sequences were detected in bacterial HTS data, suggesting contamination during the pandemic. Therefore, this tool would enable the rapid screening of virus-infected samples and serve as a new indicator for quality control of HTS data.

Current research reusing public HTS data for viral searches: assembly-based method

Next, we present two types of research using the assembly-based method. Our previous study performed *de novo* sequence assembly using over 46,000 public RNA-seq data from mammals and birds, which detected approximately 900 RNA viral infections.[7] Interestingly, this study indicated that viral infections were detected approximately three times more frequently in birds than in mammals. This analytic approach and the subsequent results may be useful in determining which animals should be prioritized for further investigation of viral infection (see also the "Challenges of virus searches using public HTS-related data" section). Furthermore, this study identified novel RNA viruses genetically similar to human pathogenic viruses. Using HTS metadata, we also investigated characteristics of the newly identified viruses, such as geographic distribution, tissue tropism, and pathogenesis. These results support that reusing HTS data can identify unknown viruses and that HTS metadata analysis can help obtain viral phenotype data.

Chang et al.[9] investigated RNA virus infections using the assembled sequences registered on the NCBI transcriptome shotgun assembly (TSA) database (https://www.ncbi.nlm.nih.gov/genbank/tsa/). This study identified 1833 RNA virus genomes in TSA data of 711 arthropod species. Remarkably, 882 RNA virus groups represented less than 75% amino acid identity to known viruses and were expected to correspond to novel virus species or genera. Combined with the previous section, these five studies demonstrate the effectiveness of reusing public HTS-related data to identify known and unknown viral infections.

Challenges of virus searches using public HTS-related data

Virus searches using publicly available HTS or TSA data enable large-scale virus surveillance, but several challenges remain.[55] Here, we discuss five issues regarding investigations of viral infections using public HTS-related data: (i) detecting viral infections, (ii) connecting virus–host relationships, (iii) developing effective analytical platforms, (iv) availability of HTS metadata, and (v) evaluating the cost-effectiveness of the virus search in a data-driven manner.

First, as most public HTS data were not obtained by virus-targeted sequencing methods, there may be several issues: (i) detection bias depending on viral genomic types, and (ii) low abundance of viral sequence reads. Regarding the first issue, it has been pointed out that positive-sense single-stranded RNA viruses tended to be frequently detected in public RNA-seq data because the poly-A–enriched sequencing method is primarily used, and many positive-sense single-stranded RNA viruses possess a poly-A tract at the 3′ end of their genome.[56] Regarding the second issue, it is often challenging to obtain full-length viral genomes due to the low abundance of viral sequence reads in public HTS data (Box 1). Several approaches would be useful to obtain longer viral sequences: (i) developing more efficient sequencing assembly methods,[48,57] (ii) performing coassembly analysis using combined HTS data from samples considered to be infected with the same viruses,[58,59] or (iii) using long-read HTS data.[60]

Second, it is difficult to accurately link virus–host relationships, even if viral sequences are identified in HTS data.[55] This issue is common for viral identification; however, more careful validations are needed in research reusing public HTS data, because it is challenging to control contamination effects in sampling or sequencing steps. Importantly, previous studies reported that many viral sequences detected in HTS data may originate from environmental or experimental resources.[47,54,61–65] As described in the previous sections, data analysis can help connect the virus–host relationships. For example, the spread of viral infections in the host population can be confirmed if the same viral infections are detected in other individuals.[7] In addition, it would be useful to determine the quality of the HTS data based on the taxonomic composition of sequence reads. If there are sequence reads from multiple animals in HTS data, it is difficult to determine which animal is the true host.[54] Developing predictive modeling based on viral sequence features may also help impute their host information (details in Box 2). Furthermore, a recent study reported a high-throughput experimental approach, XRM-seq, which can accurately connect virus–host relationships by detecting cross-linked sequences consisting of viral messenger RNA and host ribosomal RNA.[66]

The third is developing and maintaining an analytical platform to continue reusing HTS data for viral infection screening. This issue is related to the trade-off between computational costs and virus detection sensitivity. The read-based method is superior to the assembly-based method of computational costs, including operation time. However, the sensitivity of virus detection is higher with the assembly-based method (see also Box 1). One solution is to develop a read-based method for increasing viral detection rate, such as calculating similarities between translated HTS read sequences and known viral proteins.[6,44] Another solution, using assembled sequence data published in the TSA database, can save computational resources for sequence assembly.[9] However, the assembled sequence data that originated from a limited number of HTS data are currently available in the TSA database. We believe that further sharing of assembled sequences can promote viral searches in public data. In addition, organizing viral sequences accumulated in public databases thus far will contribute to constructing effective analytical platforms. Although numerous viral genomic sequences are registered in public databases, careful validation of the quality of viral sequences is needed when reusing such public data. This is because sequence quality can significantly impact subsequent analyses, such as phylogenetic tree construction. Therefore, careful curation of public viral sequences is useful for accurately understanding viral diversity. As a representative, the NCBI RefSeq database provides curated viral genomic sequences. However, such curated sequences are limited in number; for example, 58,754 viral genomic sequences are available in the NCBI RefSeq database, in contrast to 10,166,748 virus sequences registered in the NCBI GenBank database in September 2022. In the future, automating verification for sequence quality by checking genome completeness or gene annotations and constructing their databases would increase the number of available curated viral sequences (also see the “Data analysis: gene annotation and taxonomic classification” section in Box 2).

Fourth, it is necessary to improve the availability of HTS metadata to investigate viral phenotypes. HTS metadata contains information on the sample and experiment, for example, original animals and tissues, health status, sampled locations, library preparation method, and/or sequencing strategy. As mentioned in the “Current research reusing public HTS data for viral searches: assembly-based method” section, the spread of viral infections can be investigated by detecting virus-derived sequence reads in public HTS data if their metadata contains information on sampling areas.[7] Thus, sharing HTS metadata can accelerate data analysis to investigate characteristics of viral infections.[67] However, it has been pointed out that there may be little benefit in sharing data details for the HTS data generators, resulting in the low availability of metadata. Indeed, some metadata contain limited information, such as the animal species from which the HTS data were derived. In response, some researchers have proposed constructing a system that evaluates the contribution of sample collectors or data generators by assigning DOIs (digital object identifiers) to the data.[23]

Finally, we believe that evaluating the virus detection rate in large-scale investigations using public HTS data can provide useful information for reconsidering the cost-effectiveness of virus searches. For example, our previous study examined the virus detection rate for each host animal.[7] Such association analysis can be conducted by focusing on different aspects, for example, the ecological characteristics or habitats of host animals. It should be noted that there may be several issues pointed out above, such as a virus detection bias due to the sequencing methods or difficulty in connecting virus–host relationships. Nevertheless, these evaluations using public HTS data can

BOX 2 Connecting sequence data with phenotypic data of viruses

Many virus sequences identified using HTS often lack phenotypic information, such as taxonomic classification or the host. Multiple approaches can be used to obtain such viral phenotypes: data analysis, development of prediction modeling, and implementation of high-throughput platforms for virological experiments. This section discusses the current status and challenges of each approach (Figure 2).

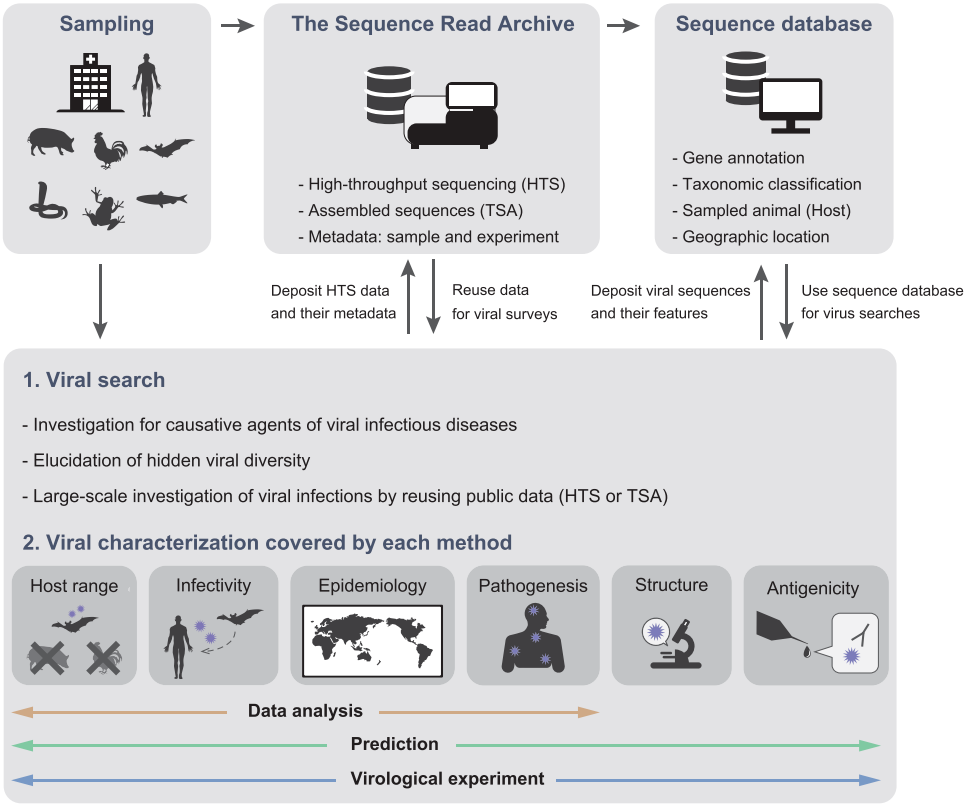

FIGURE 2 Virus identification using HTS and the method used to obtain viral phenotypic data. Virus infections have been investigated using various samples: humans or nonhumans, and clinical or asymptomatic. Virus searches using HTS have enabled the rapid identification of causative agents of viral infectious diseases and the elucidation of hidden viral diversity. Recent studies have performed large-scale investigations of viral infections reusing public data, such as HTS or TSA. As most viral sequences identified in HTS data often lack biological characteristics, several approaches are needed to obtain viral phenotypic information: performing data analysis, developing predictive modeling, or implementing high-throughput platforms of virological experiments (details in Box 2).

Data analysis: gene annotation and taxonomic classification

Gene annotation and taxonomic classification are essential for understanding viral diversity and evolution. However, manually performing gene annotation and virus classification is unrealistic, considering the rapid increase in viral sequences. Thus, an automated system of gene annotation and viral classification based only on viral sequences should be established.[68,69] Such a system will reduce the burden on researchers and ensure the reproducibility of the analysis by eliminating human errors, which can accelerate our understanding of virosphere diversity.

Gene annotation and taxonomic classification are closely related, and automated protocols are expected. The standard procedure for virus classification is (i) to identify protein-coding regions in the viral genome, (ii) to collect conserved protein sequences at the order or family levels, and (iii) to perform phylogenetic analysis using the collected sequences. Several tools have already been developed for each step of this protocol, and their combination will enable the automation of viral gene annotation and taxonomic classification. However, there are still challenges, such as errors in gene prediction due to low conservation among viral proteins (details were described in Simmonds and colleagues[68,69]). In addition, it is necessary to consider protein isoforms generated by RNA splicing or RNA editing in gene annotation[60,70] or recombination between viruses in phylogenetic analysis.[71,72]

Currently, several tools can be used for animal viral gene annotation[73–76] or taxonomic classification.[77–79] VADR is also available for the validation of viral gene annotation, such as norovirus, dengue virus, or SARS-CoV-2.[73]

GRAVity is an analytic pipeline for family-level viral classification using only sequence information, which has been reported to provide viral classification consistently with the ICTV taxonomy rules.[78] Regarding both gene annotation and taxonomic classification, such tools will further develop to increase the number of applicable viruses.

Prediction: host range and infectivity

Viral host range and infectivity are critical phenotypes used to determine whether further experiments are needed to investigate the zoonotic potentials. As pointed out in the "Challenges of virus searches using public HTS-related data" section, identifying viral sequences cannot surely support their infections. Furthermore, another review article reported that the host information of approximately 40% among the viral nucleotide sequences infecting nonhuman hosts is not registered in the GenBank database,[55] reflecting the difficulty in connecting virus–host relationships.

We believe that predictive modeling can compensate for information on viral host range or infectivity.[18,19,23,80–85] These predictions would help design downstream experiments and prioritize viruses for further characterization. In addition, interpretable models for prediction may enable us to understand the mechanisms underlying viral infectivity. However, these prediction results should be carefully considered because there may be potential bias in virological data, such as focusing on pathogenic viruses in humans and livestock. Furthermore, the necessity of validating prediction results by virological experiments has also been also pointed out.[22,23,86]

Current prediction models for host range and infectivity have been developed by incorporating various types of information.[18,19,23,80–85] For example, some studies used phylogenetic relationships or sequence features of viruses,[80–82] whereas others used environmental, host, and virus factors with risk assessment by experts.[18,19] Notably, a recent study ranked viruses according to the risk of infecting humans using a machine learning model with only viral sequence signatures.[81] Interestingly, the authors suggested that this sequence-based approach could predict human-infecting viruses, even if they belong to different virus families. Thus, this prediction approach is expected to extract features that control viral host range and infectivity, providing clues to clarify mechanisms of viral adaptation to humans.

Prediction: viral protein structure and antigenicity

Here, we introduce recent studies regarding prediction protein structure and antigenicity. These phenotypes are important, especially in medical research, such as vaccine and antiviral development. Furthermore, as a different aspect from medical applications, predicting viral protein structures can elucidate viral diversity and evolution. For example, protein structure information may help identify unknown viruses previously classified as "dark matter" because protein structure is highly conserved among diverse viruses.[43,46,52] In addition, comparisons of viral protein structures may enable the deep tracking of viral evolutionary history.[87,88] Thus, the prediction of viral protein structures will improve antiviral strategy and deepen our understanding of viral diversity.

Alphafold2 is expected to innovate for predicting viral protein structures.[89] Recent studies reported the prediction of viral protein structures by AlphaFold2, which showed high concordance to experimentally validated ones.[90,91] However, it has been pointed out that several challenges remain, such as the prediction of conformational changes or post-translational modifications.[92] As another example, changes in viral antigenicity and the mutation effects on viral antigenicity have been predicted.[93,94] A recent study that applied a model for natural language processing to viral sequence information reported the prediction of escape patterns of influenza virus and SARS-CoV-2 from the host immune system.[95] Interestingly, this study also indicated complex rules for balancing between viral fitness (e.g., replication efficacy or binding affinity to viral entry receptor) and escape patterns, which are analogous to "grammar" and "meaning" in natural language. These insights may lead to an understanding of the evolutionary principles of viruses.

Virological experiments: infectivity screening at the cellular level

The first step in virological experiments is the selection of the cells for viral infection. An automated system with processes, such as inoculating specimens onto cell plates, transfecting artificially synthesized viral genomes, or obtaining morphological and sequence data of infected cells, will reduce the time and labor required for virus isolation.[96] Furthermore, plates seeded with cells derived from different animals or tissues would help perform high-throughput screening for viral zoonotic potential or candidates of reservoir animals.

Virological experiments: sorting virus-infected cells

Developing methods to separate virus-infected cells from noninfected cells would enable efficient virus isolation and culture. Cell-sorting techniques have been developed in several research fields. The advantages of these methods are that the phenotypic information can be obtained without destroying the cells, and sorted cells can be used for

downstream experiments. For example, "Ghost cytometry" reportedly sorts targeted cells in microseconds by machine learning using fluorescence signals or fluorescence label–free waveforms obtained from cells as input.[97,98] As another example, previous studies using prokaryotes reported that it is possible to discriminate gene expression patterns or species using Raman spectra signature.[99–102] Interestingly, a study using virus particles as samples has reported Raman spectra signatures specific for virus species or strains.[103] The application of such techniques, detecting morphological or biomolecular signatures, may enable the efficient sorting of virus-infected cells.

Virological experiments: investigation of viral infection at the antibody level

The recognition of viral antigens by the host immune system induces antibody production, and the antibody responses can be maintained over years or decades. In other words, antibodies can be considered as "footprints of viral infections" and provide information regarding past or present viral infections. Thus, antibody detection can be useful for investigating their host range or geographic distribution. A previous study has reported a high-throughput serological testing platform based on a bacteriophage display method, VirScan, which can investigate antibodies targeting more than 1000 viral strains.[104] One of the advantages of this method is that it is less invasive because the viral infection history can be determined using 1 μL of blood. Using animal samples for the serological investigations may identify viral reservoir hosts or trace viral infection spreads.

provide (i) an indicator of how much and what type of samples should be surveyed and (ii) an overview of the cost-effectiveness of virus searches. Such information can help adjust future viral surveillance plans in a data-driven manner.

CONCLUSION

The HTS technology has enabled the rapid and comprehensive identification of viral sequences. However, our understanding of viral diversity is still incomplete, and it is essential to continue further viral infection surveys. In this article, we first summarized recent studies using public HTS-related data for surveillance of viral infections, showing that public HTS-related data can be a powerful resource to elucidate RNA viral diversity. Furthermore, we addressed several challenges, such as detecting or characterizing viruses in public HTS-related data, to investigating viral infections more efficiently. Thus, this article offers a perspective on the reusability of public HTS data for the large-scale investigations of RNA viral infections. Considering that HTS data from various animal samples have been accumulated in databases, continuing viral surveys reusing publicly available HTS data can deepen our understanding of the global virome.

Second, as most viruses identified by HTS analysis lack biological characteristics, this article summarized several approaches for obtaining viral phenotypic data: performing data analysis and compensating by predictive modeling. Viral host range and infectivity are critical phenotypes for zoonotic potential; however, it is often difficult to define true host species even if viral sequences were identified in HTS data. We believe that analyzing HTS metadata or developing predictive models can provide supportive information for viral hosts, which can help plan downstream experiments and prioritize which viruses should be experimentally validated. The development of high-throughput experimental platforms is also needed to obtain large-scale data sets of viral phenotypes and to validate genotype–phenotype connections. We believe that this article provides a direction for constructing a feedback system between viral sequence and phenotypic data, which will help us to prepare for future zoonotic viruses.

AUTHOR CONTRIBUTIONS

Junna Kawasaki and Masayuki Horie prepared the figures and wrote the initial draft of the manuscript. All authors revised and approved the final manuscript.

ACKNOWLEDGMENTS

We are grateful for the insightful discussions during a research workshop (October 22–24, 2021) organized by two study groups: Bioinfowakate (https://www.bioinfowakate.org/) and Virology "Wakate" Network (https://youngvirologistnw.weebly.com/about.html). We thank Editage for editing and reviewing this manuscript for the English language accuracy. This study was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI (Grants-in-Aid for Scientific Research) JP22J00010 (J. K.), JP19J22241 (J. K.), JP19K22530 (K. T.), JP20H05682 (K. T.), JP18K19443 (M. H.), and JP21H01199 (M. H.); by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) KAKENHI JP16H06429 (K. T.), JP16K21723 (K. T.), JP16H06430 (K. T.), JP17H05821 (M. H.), and JP19H04833 (M. H.); by the Waseda Research Institute for Science and Engineering, Grant-in-Aid for Young Scientists (Early Bird) (J. K.); by the JSPS Core-to-Core Program (K. T.); and by the Joint Usage/Research Center Program on inFront, Kyoto University (K. T.).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Data sharing not applicable—no new data generated, or the article describes entirely theoretical research.

ORCID

Junna Kawasaki https://orcid.org/0000-0002-6609-5300
Keizo Tomonaga http://orcid.org/0000-0003-0405-7103
Masayuki Horie http://orcid.org/0000-0003-4682-7698

REFERENCES

1. Carroll D, Daszak P, Wolfe ND, et al. The Global Virome Project. Science. 2018;359(6378):872–4.
2. Greninger AL. A decade of RNA virus metagenomics is (not) enough. Virus Res. 2018;244:218–29.
3. Shi M, Lin X-D, Tian J-H, et al. Redefining the invertebrate RNA virosphere. Nature. 2016;540(7634):539–43.
4. Shi M, Lin X-D, Chen X, et al. The evolutionary history of vertebrate RNA viruses. Nature. 2018;556(7700):197–202.
5. Geoghegan JL, Di Giallonardo F, Cousins K, Shi M, Williamson JE, Holmes EC. Hidden diversity and evolution of viruses in market fish. Virus Evol. 2018;4(2):vey031.
6. Edgar RC, Taylor J, Lin V, et al. Petabase-scale sequence alignment catalyses viral discovery. Nature. 2022;602(7895):142–7.
7. Kawasaki J, Kojima S, Tomonaga K, Horie M, Moscona A. Hidden viral sequences in public sequencing data and warning for future emerging diseases. mBio. 2021;12(4):e01638–21.
8. Kumata R, Ito J, Takahashi K, Suzuki T, Sato K. A tissue level atlas of the healthy human virome. BMC Biol. 2020;18(1):55.
9. Chang T, Hirai J, Hunt BPV, Suttle CA. Arthropods and the evolution of RNA viruses. bioRxiv. 2021; https://doi.org/10.1101/2021.05.30.446314
10. Wille M, Holmes EC. The ecology and evolution of influenza viruses. Cold Spring Harbor Perspect Med. 2020;10(7):a038489.
11. Long JS, Mistry B, Haslam SM, Barclay WS. Host and viral determinants of influenza A virus species specificity. Nat Rev Microbiol. 2019;17(2):67–81.
12. Dawood FS, Iuliano AD, Reed C, et al. Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study. Lancet Infect Dis. 2012;12(9):687–95.
13. Trifonov V, Khiabanian H, Rabadan R. Geographic dependence, surveillance, and origins of the 2009 influenza A (H1N1) virus. N Engl J Med. 2009;361(2):115–9.
14. Garten RJ, Davis CT, Russell CA, et al. Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. Science. 2009;325(5937):197–201.
15. Smith GJ, Vijaykrishna D, Bahl J, et al. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature. 2009;459(7250):1122–5.
16. Meekins DA, Gaudreault NN, Richt JA. Natural and experimental SARS-CoV-2 infection in domestic and wild animals. Viruses. 2021;13(10):1993.
17. Sharun K, Dhama K, Pawde AM, et al. SARS-CoV-2 in animals: potential for unknown reservoir hosts and public health implications. Vet Q. 2021;41(1):181–201.
18. Olival KJ, Hosseini PR, Zambrana-Torrelio C, Ross N, Bogich TL, Daszak P. Host and viral traits predict zoonotic spillover from mammals. Nature. 2017;546(7660):646–50.
19. Grange ZL, Goldstein T, Johnson CK, et al. Ranking the risk of animal-to-human spillover for newly discovered viruses. Proc Nat Acad Sci. 2021;118(15):e2002324118.
20. Joly D, Johnson CK, Goldstein T, et al. The first phase of PREDICT: surveillance for emerging infectious zoonotic diseases of wildlife origin (2009-2014). Int J Infect Dis. 2016;53:31–2.
21. Jonas O, Seifman R. Do we need a Global Virome Project? Lancet Glob Health. 2019;7(10):e1314–e6.
22. Holmes EC, Rambaut A, Andersen KG. Pandemics: spend on surveillance, not prediction. Nature. 2018;558(7709):180–2.
23. Carlson CJ, Farrell MJ, Grange Z, et al. The future of zoonotic risk prediction. Philos Trans R Soc, B. 2021;376(1837):20200358.
24. Zhang Y-Z, Chen Y-M, Wang W, Qin X-C, Holmes EC. Expanding the RNA virosphere by unbiased metagenomics. Annu Rev Virol. 2019;6(1):119–39.
25. Abbas AA, Taylor LJ, Dothard MI, et al. Redondoviridae, a family of small, circular DNA viruses of the human oro-respiratory tract associated with periodontitis and critical illness. Cell Host Microbe. 2019;25(5):719–29.
26. Chua KB, Bellini WJ, Rota PA, et al. nipah virus: a recently emergent deadly paramyxovirus. Science. 2000;288(5470):1432–5.
27. Chiu CY, Miller SA. Clinical metagenomics. Nat Rev Genet. 2019;20(6):341–55.
28. Chiu CY. Viral pathogen discovery. Curr Opin Microbiol. 2013;16(4):468–78.
29. Niller HH, Angstwurm K, Rubbenstroth D, et al. Zoonotic spillover infections with Borna disease virus 1 leading to fatal human encephalitis, 1999–2019: an epidemiological investigation. Lancet Infect Dis. 2020;20(4):467–77.
30. Tappe D, Pörtner K, Frank C, et al. Investigation of fatal human Borna disease virus 1 encephalitis outside the previously known area for human cases, Brandenburg, Germany – a case report. BMC Infect Dis. 2021;21(1):787.
31. Rubbenstroth D, Schlottau K, Schwemmle M, Rissland J, Beer M. Human bornavirus research: back on track! PLoS Pathog. 2019;15(8):e1007873.
32. Korn K, Coras R, Bobinger T, et al. Fatal encephalitis associated with borna disease virus 1. N Engl J Med. 2018;379(14):1375–7.
33. Schlottau K, Forth L, Angstwurm K, et al. Fatal encephalitic borna disease virus 1 in solid-organ transplant recipients. N Engl J Med. 2018;379(14):1377–9.
34. Grad YH, Lipsitch M. Epidemiologic data and pathogen genome sequences: a powerful synergy for public health. Genome Biol. 2014;15(11):538.
35. Hill V, Ruis C, Bajaj S, Pybus OG, Kraemer MUG. Progress and challenges in virus genomic epidemiology. Trends Parasitol. 2021;37(12):1038–49.
36. Singer JB, Thomson EC, McLauchlan J, Hughes J, Gifford RJ. GLUE: a flexible software system for virus sequence data. BMC Bioinformatics. 2018;19(1):532.
37. Campbell K, Gifford RJ, Singer J, et al. Making genomic surveillance deliver: a lineage classification and nomenclature system to inform rabies elimination. bioRxiv. 2021. https://doi.org/10.1101/2021.10.13.464180
38. Katzourakis A, Gifford RJ. Endogenous viral elements in animal genomes. PLoS Genet. 2010;6(11):e1001191.
39. Aiewsakun P, Katzourakis A. Endogenous viruses: connecting recent and ancient viral evolution. Virology. 2015;479-480:26–37.
40. Conesa A, Madrigal P, Tarazona S, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17(1):181.
41. Lapidus AL, Korobeynikov AI. Metagenomic data assembly – the way of decoding unknown microorganisms. Front Microbiol. 2021;12(653):613791.
42. Peri S, Roberts S, Kreko IR, et al. Read mapping and transcript assembly: a scalable and high-throughput workflow for the processing and analysis of ribonucleic acid sequencing data. Front Genet. 2020;10:1361.
43. Krishnamurthy SR, Wang D. Origins and challenges of viral dark matter. Virus Res. 2017;239:136–42.
44. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nature Methods. 2015;12(1):59–60.

45. Chen J, Guo M, Wang X, Liu B. A comprehensive review and comparison of different computational methods for protein remote homology detection. Brief. Bioinform. 2016;19(2):231–44.
46. Illergård K, Ardell DH, Elofsson A. Structure is three to ten times more conserved than sequence - a study of structural response in protein cores. Proteins: Struct, Funct, Bioinf. 2009;77(3):499–508.
47. Jurasz H, Pawłowski T, Perlejewski K. Contamination issue in viral metagenomics: problems, solutions, and clinical perspectives. Front Microbiol. 2021;12:745076.
48. Sutton TDS, Clooney AG, Ryan FJ, Ross RP, Hill C. Choice of assembly software has a critical impact on virome characterisation. Microbiome. 2019;7(1):12.
49. Martínez Martínez J, Martinez-Hernandez F, Martinez-Garcia M. Single-virus genomics and beyond. Nat Rev Microbiol. 2020;18(12):705–16.
50. Bean AG, Baker ML, Stewart CR, et al. Studying immunity to zoonotic diseases in the natural host – keeping it real. Nat Rev Immunol. 2013;13(12):851–61.
51. Seal S, Dharmarajan G, Khan I. Evolution of pathogen tolerance and emerging infections: a missing experimental paradigm. eLife. 2021;10:e68874.
52. Wang D. 5 challenges in understanding the role of the virome in health and disease. PLoS Pathog. 2020;16(3):e1008318.
53. Pfaff F, Rubbenstroth D. Two novel bornaviruses identified in colubrid and viperid snakes. Arch Virol. 2021;166(9):2611–4.
54. Katz KS, Shutov O, Lapoint R, Kimelman M, Brister JR, O'Sullivan C. STAT: a fast, scalable, MinHash-based k-mer tool to assess Sequence Read Archive next-generation sequence submissions. Genome Biol. 2021;22(1):270.
55. Cobbin JCA, Charon J, Harvey E, Holmes EC, Mahar JE. Current challenges to virus discovery by meta-transcriptomics. Curr Opin Virol. 2021;51:48–55.
56. Altmäe S, Molina NM, Sola-Leyva A. Omission of non-poly(A) viral transcripts from the tissue level atlas of the healthy human virome. BMC Biol. 2020;18(1):179.
57. Deng Z, Delwart E. ContigExtender: a new approach to improving de novo sequence assembly for viral metagenomics data. BMC Bioinformatics. 2021;22(1):119.
58. Iwamoto M, Shibata Y, Kawasaki J, et al. Identification of novel avian and mammalian deltaviruses provides new insights into deltavirus evolution. Virus Evol. 2021;7(1):veab003.
59. Iyer MK, Niknafs YS, Malik R, et al. The landscape of long noncoding RNAs in the human transcriptome. Nature Genet. 2015;47(3):199–208.
60. Boldogkői Z, Moldován N, Balázs Z, Snyder M, Tombácz D. Long-read sequencing – a powerful tool in viral transcriptome research. Trends Microbiol. 2019;27(7):578–92.
61. Asplund M, Kjartansdóttir KR, Mollerup S, et al. Contaminating viral sequences in high-throughput sequencing viromics: a linkage study of 700 sequencing libraries. Clin Microbiol Infect. 2019;25(10):1277–85.
62. Lusk RW. Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data. PLoS One. 2014;9(10):e110808.
63. Porter AF, Cobbin J, Li C-X, Eden J-S, Holmes EC. Metagenomic identification of viral sequences in laboratory reagents. Viruses. 2021;13(11):2122.
64. Feehan BJ, Penin AA, Mukhin AN, et al. Novel mammalian orthorubulavirus 5 discovered as accidental cell culture contaminant. Viruses. 2019;11(9):777.
65. Ngoi CN, Siqueira J, Li L, et al. Corrigendum: the plasma virome of febrile adult Kenyans shows frequent parvovirus B19 infections and a novel arbovirus (Kadipiro virus). J Gen Virol. 2017;98(3):517.
66. Ignacio-Espinoza JC, Laperriere SM, Yeh Y-C, et al. Ribosome-linked mRNA-rRNA chimeras reveal active novel virus host associations. bioRxiv. 2020. https://doi.org/10.1101/2020.10.30.332502
67. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data. 2016;3(1):160018.
68. Simmonds P, Adams MJ, Benkő M, et al. Virus taxonomy in the age of metagenomics. Nat Rev Microbiol. 2017;15(3):161–8.
69. Roux S, Adriaenssens EM, Dutilh BE, et al. Minimum information about an uncultivated virus genome (MIUViG). Nature Biotechnol. 2019;37(1):29–37.
70. Cross ST, Michalski D, Miller MR, Wilusz J. RNA regulatory processes in RNA virus biology. Wiley Interdiscip Rev RNA. 2019;10(5):e1536.
71. Martin DP, Varsani A, Roumagnac P, et al. RDP5: a computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets. Virus Evol. 2020;7(1):veaa087.
72. Lam HM, Ratmann O, Boni MF. Improved algorithmic complexity for the 3SEQ recombination detection algorithm. Mol Biol Evol. 2017;35(1):247–51.
73. Schäffer AA, Hatcher EL, Yankie L, et al. VADR: validation and annotation of virus sequence submissions to GenBank. BMC Bioinformatics. 2020;21(1):211.
74. Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Tatusova T. FLAN: a web server for influenza virus genome annotation. Nucleic Acids Res. 2007;35(Web Server):W280–W4.
75. Wang S, Sundaram JP, Stockwell TB. VIGOR extended to annotate genomes for additional 12 different viruses. Nucleic Acids Res. 2012;40(W1):W186–W92.
76. Shean RC, Makhsous N, Stoddard GD, Lin MJ, Greninger AL. VAPiD: a lightweight cross-platform viral annotation pipeline and identification tool to facilitate virus genome submissions to NCBI GenBank. BMC Bioinformatics. 2019;20(1):48.
77. Gorbalenya AE, Lauber C. Bioinformatics of virus taxonomy: foundations and tools for developing sequence-based hierarchical classification. Curr Opin Virol. 2022;52:48–56.
78. Aiewsakun P, Simmonds P. The genomic underpinnings of eukaryotic virus taxonomy: creating a sequence-based framework for family-level virus classification. Microbiome. 2018;6(1):38.
79. Bao Y, Chetvernin V, Tatusova T. Improvements to pairwise sequence comparison (PASC): a genome-based web tool for virus classification. Arch Virol. 2014;159(12):3293–304.
80. Babayan SA, Orton RJ, Streicker DG. Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes. Science. 2018;362(6414):577–80.
81. Mollentze N, Babayan SA, Streicker DG. Identifying and prioritizing potential human-infecting viruses from their genome sequences. PLoS Biol. 2021;19(9):e3001390.
82. Young F, Rogers S, Robertson DL. Predicting host taxonomic information from viral genomes: a comparison of feature representations. PLoS Comput Biol. 2020;16(5):e1007894.
83. Mollentze N, Streicker DG. Viral zoonotic risk is homogenous among taxonomic orders of mammalian and avian reservoir hosts. Proc Nat Acad Sci. 2020;117(17):9423–30.
84. Pepin KM, Lass S, Pulliam JRC, Read AF, Lloyd-Smith JO. Identifying genetic markers of adaptation for surveillance of viral host jumps. Nat Rev Microbiol. 2010;8(11):802–13.
85. Albery GF, Becker DJ, Brierley L, et al. The science of the host–virus network. Nat Microbiol. 2021;6(12):1483–92.
86. Wille M, Geoghegan JL, Holmes EC. How accurately can we assess zoonotic risk. PLoS Biol. 2021;19(4):e3001135.
87. Krupovic M, Dolja VV, Koonin EV. Origin of viruses: primordial replicators recruiting capsids from hosts. Nat Rev Microbiol. 2019;17(7):449–58.
88. Krupovic M, Dolja VV, Koonin EV. The LUCA and its complex virome. Nat Rev Microbiol. 2020;18(11):661–70.
89. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9.

90. Robertson AJ, Courtney JM, Shen Y, Ying J, Bax A. Concordance of X-ray and AlphaFold2 models of SARS-CoV-2 main protease with residual dipolar couplings measured in solution. J Am Chem Soc. 2021;143(46):19306–10.
91. Jumper J, Tunyasuvunakool K, Kohli P, Hassabis D, Team tA. Computational predictions of protein structures associated with COVID-19 Version 3. 2022. [cited 2020 Aug]. https://deepmind.com/research/open-source/computational-predictions-of-protein-structures-associated-with-COVID-19
92. Higgins MK. Can we alphafold our way out of the next pandemic? J Mol Biol. 2021;433(20):167093.
93. Gouma S, Anderson EM, Hensley SE. Challenges of making effective influenza vaccines. Annu Rev Virol. 2020;7(1):495–512.
94. Cai X, Li JJ, Liu T, Brian O, Li J. Infectious disease mRNA vaccines and a review on epitope prediction for vaccine design. Briefings Funct Genomics. 2021;20(5):289–303.
95. Hie B, Zhong ED, Berger B, Bryson B. Learning the language of viral evolution and escape. Science. 2021;371(6526):284–8.
96. Gao A, Murphy RR, Chen W, et al. Progress in robotics for combating infectious diseases. Sci Robot. 2021;6(52):eabf1462.
97. Ota S, Horisaki R, Kawamura Y, et al. Ghost cytometry. Science. 2018;360(6394):1246–51.
98. Ugawa M, Kawamura Y, Toda K, et al. In silico-labeled ghost cytometry. eLife. 2021;10:e67660.
99. Germond A, Ichimura T, Horinouchi T, Fujita H, Furusawa C, Watanabe TM. Raman spectral signature reflects transcriptomic features of antibiotic resistance in *Escherichia coli*. Commun Biol. 2018;1(1):85.
100. Ho C-S, Jean N, Hogan CA, et al. Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning. Nat Commun. 2019;10(1):4927.
101. Kobayashi-Kirschvink KJ, Nakaoka H, Oda A, et al. linear regression links transcriptomic data and cellular Raman spectra. Cell Systems. 2018;7(1):104–17.
102. Kanno N, Kato S, Ohkuma M, Matsui M, Iwasaki W, Shigeto S. Machine learning-assisted single-cell Raman fingerprinting for in situ and nondestructive classification of prokaryotes. iScience. 2021;24(9):102975.
103. Yeh Y-T, Gulino K, Zhang Y, et al. A rapid and label-free platform for virus capture and identification from clinical samples. Proc Nat Acad Sci. 2020;117(2):895–901.
104. Xu GJ, Kula T, Xu Q, et al. Comprehensive serological profiling of human populations using a synthetic human virome. Science. 2015;348(6239):aaa0698.

How to cite this article: Kawasaki J, Tomonaga K, Horie M. Large-scale investigation of zoonotic viruses in the era of high-throughput sequencing. Microbiol Immunol. 2023;67:1–13.
https://doi.org/10.1111/1348-0421.13033